Studies in Autonomic,
Data-driven and Industrial Computing

Neha Sharma
Mandar Bhatavdekar *Editors*

World of Business with Data and Analytics

Springer

Studies in Autonomic, Data-driven and Industrial Computing

The book series Studies in Autonomic, Data-driven and Industrial Computing (SADIC) aims at bringing together valuable and novel scientific contributions that address new theories and their real world applications related to autonomic, data-driven, and industrial computing. The area of research covered in the series includes theory and applications of parallel computing, cyber trust and security, grid computing, optical computing, distributed sensor networks, bioinformatics, fuzzy computing and uncertainty quantification, neurocomputing and deep learning, smart grids, data-driven power engineering, smart home informatics, machine learning, mobile computing, internet of things, privacy preserving computation, big data analytics, cloud computing, blockchain and edge computing, data-driven green computing, symbolic computing, swarm intelligence and evolutionary computing, intelligent systems for industry 4.0, as well as other pertinent methods for autonomic, data-driven, and industrial computing.

The series will publish monographs, edited volumes, textbooks and proceedings of important conferences, symposia and meetings in the field of autonomic, data-driven and industrial computing.

Neha Sharma · Mandar Bhatavdekar
Editors

World of Business with Data and Analytics

Springer

Editors
Neha Sharma
Tata Consultancy Services Ltd.
Pune, Maharashtra, India

Mandar Bhatavdekar
Tata Consultancy Services Ltd.
Pune, Maharashtra, India

ISSN 2730-6437 ISSN 2730-6445 (electronic)
Studies in Autonomic, Data-driven and Industrial Computing
ISBN 978-981-19-5688-1 ISBN 978-981-19-5689-8 (eBook)
https://doi.org/10.1007/978-981-19-5689-8

Article note

The book was inadvertently published with unnecessary/incorrect preface. The Preface have been updated.

Preface

Data and Analytics are widely used in the business world to derive meaningful insights from the huge pool of information that gets generated continuously. It is all about harnessing the power of data for business growth. This book covers research work with the breadth of ventures, variety of challenges and finest of techniques used by subject matter experts from the corporate world. It's content highlights the real-life business problems that are relevant to any industry and technology environment.

The book helps us become a contributor to and accelerator of Artificial Intelligence, Data Science and Analytics, deploy a structured life-cycle approach to data-related issues, apply appropriate analytical tools- & techniques to analyze data and deliver differentiated solutions. It also brings out the story telling element in a compelling fashion using data and analytics. This aids the readers to achieve quantitative and qualitative outcomes and augments various business actions in different domains like Energy, Manufacturing, Healthcare, BFSI, Security, etc.

Section 1: Supply Chain Management

This book starts by presenting an unexpected business problem in the form of changing situation of demand and supply that arose due to the pandemic. Existing forecasting, demand planning, supply planning techniques and approaches are no longer relevant to solve the purpose. The first chapter aims to explore a new approach to deal with distorted enterprise data for successful demand planning and how to move forward with historical data for further improvement in analytics.

Section 2: Energy

The second section highlights a couple of challenges in the energy sector and proposes analytical solutions for the same. The first problem discussed is of rupture and leakage from the pipelines which are responsible for transporting close to 90% of crude oil, natural gas, and natural gas liquids. Data and analytics-driven methodologies are suggested to predict anomalies in advance and support timely maintenance of those specific cross sections of the pipelines and thereby prevent leakages and incidents. The second challenge addressed is of aging infrastructure, growing power demand and the rising community of consumers who have evolved to producing their own electricity thereby getting converted to prosumers. The solution helps organizations with grid optimization and forecasting of overall energy requirements and

pricing. The third problem highlights the global challenges of electricity trading business in developed countries like the USA, Europe, Australia, the UK, etc. The proposed solution includes data and AI to empower the decisions of utility majors on trade-offs between profits, revenue, and risk in view of the supply chain, changing generation mix, consumption and load patterns, regulatory compliances, innovation and incentivization.

Section 3: Manufacturing/Healthcare/Finance

The third domain that is covered is that of manufacturing to address the problem of surface damage classification and segmentation to automate the manual inspection of images and suggests a two-phase analytical approach as a solution to ensure the high quality and standard of the product. The fourth domain is healthcare, where the problem of extracting the features of electrocardiogram (ECG) signals and then classifying them in order to automate cardiac disorders and provide timely treatment to patients, is addressed. The solution proposed is a combined approach for automatic detection of cardiac arrhythmias: discrete wavelet transformation method to extract the features of ECG signals and probabilistic neural network model for signal classification. The fifth domain is of finance where the study has been conducted on how mutual fund advisors can help in planning of the unmet needs of their clients and suggest them different investment options with the use of data and analytics.

Section 4: Security

The next section deliberates on an important aspect of any business i.e. 'security'. The first problem discussed is the complexity of usage of customer's personal data due to heightened regulations . The proposed solution is named as 'iMask' which is an AI-based framework/tool that has a great potential to identify such Personally Identifiable Information from a variety of multiple images automatically using computer vision technology to either encrypt or redact them from the original image. The second problem is related to common vulnerabilities and exposure that can be easily induced in open-source software and exploited by hackers, leading to malicious attacks. Artificial intelligence solution provides us resources to tackle these complex problems. The third solution proposed is a new secure network model which combines both signature-based detection and data mining techniques which act as an efficient detection and response system against intrusion, it is a well-established and evolved system.

Section 5: Advanced Topics

The last section discusses few advanced topics that are relevant to data and analytics in the business world. The first chapter discusses the methodologies for various optimization applications aided by Machine Learning using cloud usage data. The second chapter presents unsupervised learning from machine learning combined with the power of vector algebra to mine deeper insights from the texts. These insights could act as labels for building new supervised models if needed and these predicted insights could be converted into actionable business outcomes through visualizations. The final chapter explains the significance of blending two emerging technologies in AI/ML – Explainable AI (XAI) and Machine Learning Operations (ML Ops) and demonstrates a focused use case that derives value from leveraging XAI to enhance ML Ops.

Fascinated by real-life application of data and advance analytics? this is the go-to book which strongly promotes the different dimensions of data analytics through interesting use cases. It advocates the application of data analytics in the coming decades and encourages the readers to embrace a data attitude and be resilient to this inevitable change. This book pioneers the use of data and analytics applications in various industrial segments which will help gain intelligent, actionable insights from the same.

Pune, India

Neha Sharma
Mandar Bhatavdekar

Acknowledgements

It is exciting to publish the first issue of the A&I Kosh book *World of Business with Data and Analytics*! This book that highlights business challenges of organizations and provides intelligent solutions, which has been a dream for the Analytics and Insights team. We are highly appreciative of Mr. Aninda Bose, Senior Publishing Editor from Springer Nature, for his reassurance and steadfast guidance in this unique endeavour from corporate sector.

We, the editors of the book, take this opportunity to express our heartfelt gratitude to the authors for high-quality contributions which deal with state-of-the-art topics in the areas of data science and artificial intelligence. Without their support, this book would not have become a reality. We extend our deepest gratitude and appreciation to our affiliations—'Tata Consultancy Services', 'Analytics and Insights Unit' and 'Strategy and Marketing Sub-Unit'. We would also like to thank Mr. Dinanath Kholkar, Vice President and Global Head, Partner Ecosystems and Alliances, Tata Consultancy Services, for his constant motivation to take on futuristic and unconventional initiatives.

A special mention of Ms. Dimple Pal and Ms. Smita Taitwale is a must for their unstinted and focused support towards the making of this book. Our sincere appreciation goes to the entire Leadership and fraternity of TCS for their moral support and encouragement to work hard toward this initiative of knowledge democratization.

Neha Sharma
Mandar Bhatavdekar

Contents

About the Editors

Neha Sharma is a data science crusader who advocates its application for achieving sustainable goals, solving societal, governmental and business problems as well as promotes the use of open data. She has more than 22 years of experience and presently working with Tata Consultancy Services and is a Founder Secretary, Society for Data Science. Prior to this she has worked as Director of premier Institute of Pune, that run post-graduation courses like MCA and MBA. She is an alumnus of a premier College of Engineering and Technology, Bhubaneshwar and completed her Ph.D. from prestigious Indian Institute of Technology, Dhanbad. She is a Senior IEEE member, Secretary—IEEE Pune Section and ACM Distinguished Speaker. She is an astute academician and has organized several national and international conferences and published several research papers. She is the recipient of "Best Ph.D. Thesis Award" and "Best Paper Presenter at International Conference Award" at National Level. She is a well-known figure among the IT circle, and well sought over for her sound knowledge and professional skills. Neha Sharma has been instrumental in integrating teaching with the current needs of the Industry and steering students towards their bright future.

Mandar Bhatavdekar is working with Tata Consultancy Services as a Global Head, Data & Analytics Practice, Enterprise Growth Group. He has vast experience of three decades, a Computer Engineering Graduate from VJTI, Mumbai. His focus areas include driving business and operational strategy, elevating the TCS brand in data, analytics and AI world, by bringing in thought leadership, improving our analyst positioning and customer mindshare. He comes with a rich multi-disciplinary experience within TCS from services to product development, sales to delivery, across diverse industries. He is a certified enterprise architect and has earlier led many crucial transformational engagements. On the personal front, he loves music, drama and poetry. He is a friendly face for most people who have been associated with him.

Chapter 1
Dynamic Demand Planning for Distorted Historical Data Due to Pandemic

Anuj Prakash, Gajanan Bejgamwar, and Saurabh Varshney

Abstract COVID-19 has caused a series of cascading crises in demand planning and the overall supply chain. Two side impacts of COVOID-19 can be seen for most companies, change in the order volume and order mix of demand, and supply chain disruption caused due to uncertainties. Historical sales and forecasts are no longer relevant for demand planning. And uncertainty around what people are looking for? What they are ordering? How orders are placed has impacted the supply side of the global supply chain? Organizations and industries cannot afford to be taken by surprise where they have just started to use data for better decision-making, and COVID-19 has induced lots of uncertainties in data itself. If the business needs to pull through this situation, they need to adopt new changing realities of demand and supply. Existing forecasting, demand planning, supply planning techniques, and approaches are no longer relevant to solve the purpose. Successful demand planning must consider a new approach to understand data and ways to predict the future in the best possible way. This chapter aims to explore a new approach to deal with distorted enterprise data for successful demand planning and move forward with historical data for further analytics improvement.

Keywords Demand · Demand Planning · Forecasting · Pandemic · Distorted Data

A. Prakash (✉)
Tata Consultancy Services, Think Campus, Electronic Gity-11, Bangalore, India
e-mail: anuj.prakash@tcs.com

G. Bejgamwar
Tata Consultancy Services, Sahyadri Park, Hinjewadi Phase 3, Pune-411057, India

S. Varshney
Infinity IT Park, Tata Consultancy Services, Malad East, Mumbai-400097, India

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_1

1 Introduction

The operational side of a successful business is driven by Sales and Operations planning (S&OP), which is driven by accurate demand forecasting. Almost all supply chain forecasting systems use some form of statistical forecasting model driven by historical data. Many forecasting and demand planning processes are abandoned or considered failures due to COVID-19 and related data issues. Since the outbreak of COVID-19, one of the intricate issues is to predict the market demand and supply planning, and it becomes much more complicated while considering different locations of business.

Demand forecasts are the primary driver of demand and supply planning, and they should be provided in near real-time to upstream manufacturers for production, logistics providers for distribution planning, marketing (what to promote and when), and executive management (revenue and financial planning). A demand plan, as a result of demand planning, describes how the company will make sure that the company can produce and/or supply the product at the needed time at an acceptable cost. Furthermore, the demand planning function should ensure needs not only to make sure that there are enough goods to sell, but also to make sure, based on sales cycles, inventory turns, and other measurements, that demand planners are able to replenish inventories as needed [1]. Most companies depend on a forecasting approach for better demand planning. They wish to improve demand planning by understanding future demand which is a critical element in planning future supply and implement demand planning with improved forecasting. The forecast is termed as statistically based initial estimate of future demand. A demand plan is an estimate of future demand derived by a consensus-driven review and approval of the forecast. In effect, a forecast is subject to several planning and verification processes to generate the demand plan [2]. But COVID-19 and its impact raised new challenges in the overall demand planning process due to operational disturbances and changes in demand patterns which is the major input for any demand forecasting.

Forecasting models developed before and during the pre-covid period failed to see a massive demand shift. These models are trained using historical data and due to the COVID-19 pandemic data itself has changed tremendously. Most demand planning approaches and forecasting models, which were developed before the pandemic, were of little use during and post-pandemic. Now, it is demanding different sustainable approaches for the long term as consumer behavior is continuously changing during and post COVID-19 pandemic.

Changing consumer behavior is creating immense uncertainties and disruption in demand and supply. Companies have seen sudden demand rise, drop, or sometimes both, and the resulting confusion makes it hard for demand planners to forecast future requirements. The pandemic is expected to increase levels of noise in the data in near future as well. When a highly disruptive event like a pandemic occurs, historical data might be insufficient or even irrelevant, for future planning. The sudden disruption in demand caused by the pandemic has severely affected time series-based forecasting models. Traditional demand planning approaches may no longer be a viable

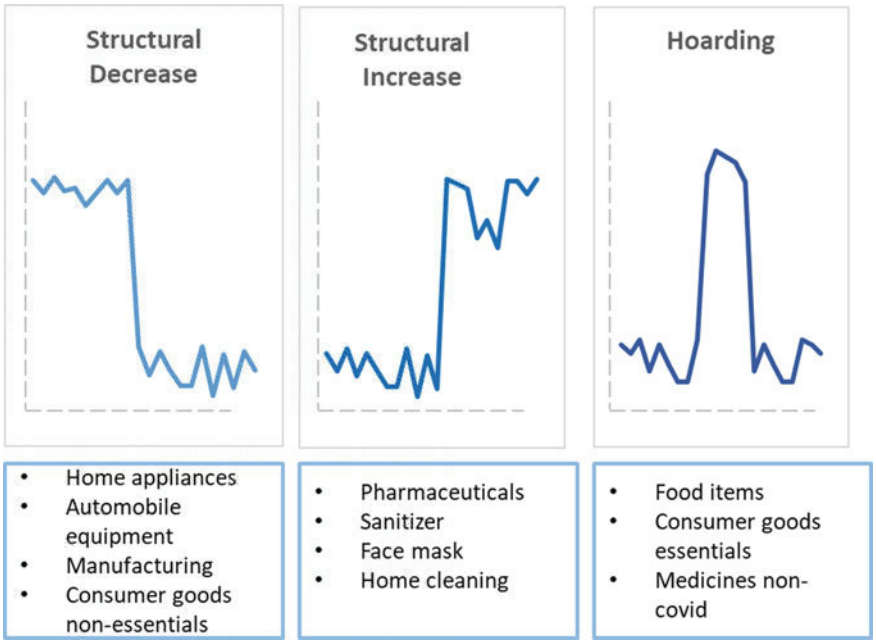


Fig. 1 Change in demand pattern

prediction option as these forecasting methods rely on the assumption that the past is indicative of the future. Those forecasting models do not consider the occurrence of random events, such as the COVID-19 pandemic. The sudden disruption in demand caused by the pandemic has severely affected forecasting models. The change in demand patterns is structural decrease, structural increase, and hoarding with their examples has been shown in Fig. 1. COVID-19 pandemic triggered unprecedented demand shocks both up and down and amplified volatility across many industries. This abrupt and unexpected change in consumer behavior will induce acute and widespread biases in traditional demand forecast models which are heavily dependent on historical data.

The impact of this disruption could be enormous in the mid to long term because it makes it more challenging to interpret, to pattern, and to predict consumer behavior. In the current circumstance, this is much clear that businesses should not rely only on internal data (before and during COVID-19) to predict the future. The change in consumer demand pattern caused by COVID-19 has created an unforeseen problem in demand planning majorly data deficiency and quality. The data before the crisis cannot effectively predict new trends. Structural changes in data lead to unforeseen supply chain vulnerabilities and disruptions. This leads to costing millions of dollars to business and the whole supply chain will collapse due to insufficient planning.

AI-driven analytics and a wide attire of critical business decisions now have a quality issue as data is distorted. The old rule book is lacking now, and this is the

right time to transform. Therefore, there is a need to rethink about a new approach for forecasting. Data science can help to develop the sustainable and dynamic approach for forecasting.

In this chapter, the new approach of demand planning with distorted data has been introduced. The demand planning was an important task for a planner, but earlier it was affected by other internal/external factors like other products, business strategy, weather, market condition, etc. During and post-pandemic (COVID-19), the demand was distorted very much, which makes it very difficult to forecast the demand with higher accuracy. There were different phases of demand like decrease, stable, and recovery phase. To identify each phase and forecast accordingly become a challenge for the practitioners. Therefore, they are seeking for such a solution, which can give demand forecast with higher accuracy and is so adaptive that it can change with the demand pattern. The chapter describes the relevant studies of demand planning during the pandemic and introduces a new and innovative solution of demand planning which is dynamic demand planning using supply data and CRM data, which is the most suitable solution with distorted data. The action plan for the practitioners along with the benefits realized by business is also an integrated part of the new approach.

The rest of the chapter is organized as follows: Sect. 2 summarizes the relevant studies in past literature and Sect. 3 describes the innovative approach of demand planning along with execution plan; whereas Sect. 4 explains the benefits realized by various businesses. Finally, Sect. 5 presents the conclusions along with the future perspective of the research.

2 Literature Review

In this section, various eminent studies on uncertainty in the supply chain due to COVID-19, demand prediction, and planning are discussed that have been conducted before and during the pandemic.

Demand Planning

Demand planning is a set of operations for exploitation of demand forecasting in supply chain planning to manage supply and demand, it helps decision-makers about the size of production, stocking inventory planning, and resource planning. Demand planning is an essential part of any organization to drive profit.

All decisions in the whole supply chain should be based on already fixed (accepted) customer orders and planned sales or forecasts, the latter are determined in the Demand Planning process. Therefore, the performance of each supply chain entity depends on the quality of the demand plan [3]. Demand planning therefore includes such activities as demand forecasting, inventory management, capacity planning, production planning and scheduling, and materials requirements planning as outlined by Bolten [4]. As highlighted by Vladimira and Michal, a demand plan is an essential business tool for most organizations [5]. It is also a necessary source of information

Fig. 2 Evolution of demand planning

for production and operations management but also for other areas like marketing, finance, or human resources. Demand planning not only includes forecasting it is a necessary part of business planning, decision-making, and strategic and tactical planning using forecasting. Integration of demand planning with the supply chain brings effective decision making, reduced lead time, and reduce cost by using effectively managing production and inventory. Demand planning has evolved over a period of time from mere forecasting to demand collaboration with supply chain partners, its evolution clearly outlined by Crum and Palmatier [6] as shown in Fig. 2.

Demand planning includes such activities as demand forecasting, inventory management, capacity planning, production planning and scheduling, and materials requirements planning [4]. Other authors see no differences between demand planning and demand forecasting [7–9]. Nari and Diane developed a Seasonal Autoregressive Integrated Moving Average with external variables (SARIMAX) model which tries to account for all the effects due to the demand influencing factors, to forecast the daily sales of perishable foods in a retail store [10]. Demand planning is a dynamic process and the current dynamic market environment forces business to react faster to quickly varying customer's demands. COVID-19 has impacted all aspects of the business from demand planning to execution and needs to study thoroughly to reduce its impact on business profitability by incorporating different impact mitigation methodologies, for doing so we need to understand the overall impact of COVID-19 on business and their operations.

Impact of COVID-19

Impact of COVID-19 on business as outlined by Sube et al. as due to strict lockdown, the manufacturing and logistics activities have been suspended, and it has affected the demand and supply of various products as a result of restrictions imposed on shopkeepers and retailers [11]. The reaction of people, governments, companies, consumers, and media has all created a simultaneous demand and supply shock. The holistic impact of COVID-19 as outlined [12] has changed the whole world. How we run businesses, deal with customers, employees, community, and other all sorts of stakeholders in organizations and society became a challenge in times of this health crisis which has extreme ramifications for sustainable business as well as society. Businesses have been forced to find creative and innovative ways of addressing some of the great challenges facing our world resulting from the COVID-19 pandemic.

The COVID-19 has not taken only lives but severely paralyzed business operations, supply chain, and logistic operations of every industry that includes the automotive sector, tourism industry, aviation industry, oil industry, construction industry, telecom sector, food industry, and healthcare industry. The COVID-19 challenge as highlighted by Arnesh includes operations, safety, supply chain, training, emergency responses, awareness, incident management, recreating business models, digitalization, and other unanticipated impacts such as consumer behavior [13]. Weersink et al. also identify capacity constraints due to social distancing in the workplace leading to operational challenges [14]. The impact of COVID-19 is omnipresent in every type of unit of business from all sectors; in this study we are concentrating its impact on.

- Supply chain
- Demand planning.

Impact of COVID-19 on Supply Chain

COVID-19 has impacted the global economy from all sectors, and the supply chain is another channel through which COVID-19 accelerates its impact on business finance. Because of the forced shutdown, businesses are experiencing enormous uncertainty in every aspect of their operations such as demand planning and supply chain planning. We have enough evidence from a different market, region, and industry that the global supply chain has been disrupted due to the crisis. In each and every business all the operations are connected through the supply chain. Haren and Simchi-Levi noticed a high impact of the COVID-19 outbreak on supply chain and manufacturing operations and predicts the consequences of the global supply chain during the second quarter of 2020 [15]. The worldwide supply chain includes distribution, packaging, as well as sourcing of raw materials and due to COVID-19 and related lockdowns are disrupting transportation [16]. Due to the exceptional lockdown measures imposed by the government, as a consequence of the emerging coronavirus disease, COVID-19, production and consumption systems have undergone significant changes which is impact as highlighted by [17]. A similar view also mentioned by Hobbs as Lockdown gives rise to a shortage of labor force and logistics disruptions eventually resulted in supply-side shocks to the food supply chain [18]. Moreover, it brings a sudden surge in the demand-side of food supply chains due to the panic buying and hoarding behavior of the people.

Impact of COVID-19 on Demand Planning

According to Tony Gray, with the onset of the COVID-19 pandemic and the associated infection containment efforts, consumer behavior changed significantly [19]. Also Jagdish Shet pointed out that the COVID-19 pandemic and the lockdown and social distancing mandates have disrupted the consumer habits of buying as well as shopping [20]. COVID-19 across countries drives changes in immediate actual needs (healthcare and food) and in consumer behavior (for example panic buying and overstocking at home). Such changes put an enormous strain on the respective

supply chains [21], which later affects business performance. Supply chain disruptions have been known to cause significant challenges and can affect organization performance [22].

Unusual demand shifts caused by the COVID-19 pandemic have created a highly unpredictable environment. Many analysts labeled the COVID-19 pandemic as a 'Black Swan' event and treated this as an outlier event, but this is a misunderstanding about what a 'Black Swan' is. An event is to be classified as a Black Swan if "nothing in the past can convincingly point to its possibility". If we break this down further, we have experienced pandemics frequently, with the latest only a decade ago, in 2009 (Swine Flu). Over the course of human history there have always been so-called Black Swan events. By combining statistical data with information on events, the Black Swan project provides a tool to aid domain experts in identifying important events and their impact [23] have also suggested an approach to extract the events from historical data, as given in Fig. 3. Event identification is halfway approach, it is also important how to predict and fulfilling the demand.

Requirement of a new approach highlighted by John Cranfield as moving beyond a static, certain approach to the structure of preferences will help us understand observed behavior (e.g., stockpiling) and possibly understanding future behaviors in a COVID-19 world. For a successful supply chain management, the planner must consider new approaches to understanding production demands and how to best meet

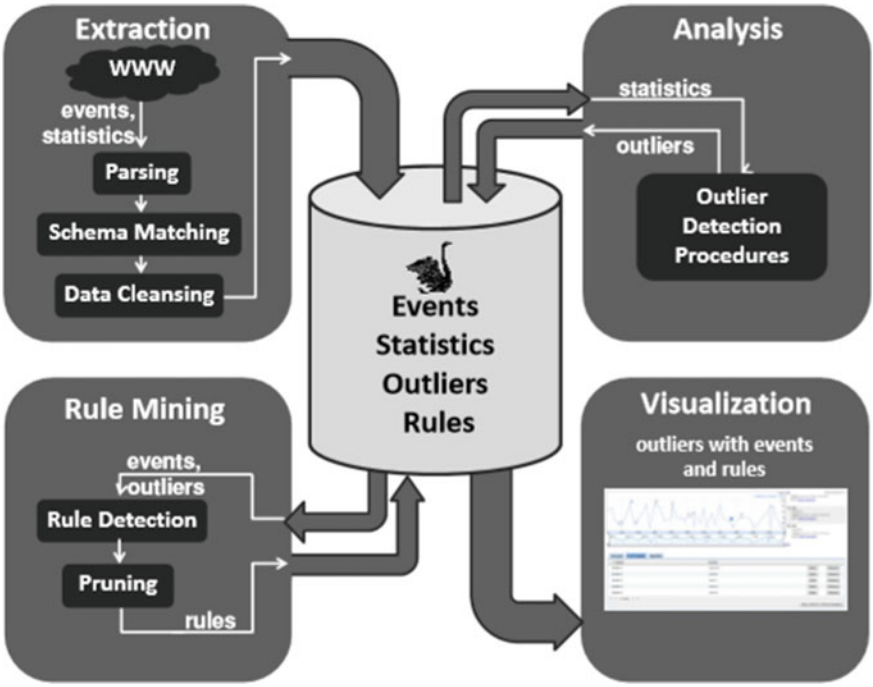


Fig. 3 System architecture

them in a reliable, consistent, and sustainable manner [24]. The new approaches have been highlighted by Dave Nelson and he suggested that customer relationship management (CRM) tools and processes and the sales and operations planning (S&OP) processes are the best way to deal with the situation. This allows demand planners and supplies chain executives within an organization to conduct a more effective demand planning process. CRM data help to improve the demand planning process for the organization which has experienced a major disruption in demand and related planning process. While S&OP practices are often used for validating statistical forecasts generated by planners. CRM data provide the insights about outside-in (consumer behavior) view while S&OP gives insights about the inside-out (operations' response to market demand) view [25]. Advanced S&OP and CRM will help to improve overall demand planning in the current scenario. The other aspect of demand planning is understanding demand drivers. This has been clearly outlined by Vincent et al. as understanding demand and its drivers in the pre-pandemic era were about understanding the underlying demographic characteristics of the customers and short-run price and promotion decisions of the supplier [26].

The impact of COVID-19, on the supply chain from the perspective of demand planning and supply planning must be properly studied in order to propose strategies from the lessons learned. The next step to improve demand planning is the selection of the right data and methodology which is used for forecasting. The new dynamic demand planning methodology must include a deep analysis of input and output generated during this pandemic. Traditional demand factors still determine demand to some extent, but it is more important to introduce new drivers as individuals make decisions that change rapidly with the changing environment. Dynamic demand planning defines the next significant competitive advantage in demand planning and supply chain efficiency.

To summarize this, all consumption and consumer behavior are anchored to time and location, and it is expected that most habits will return to normal but the impact it left will be forever. It requires a new approach and methodology to deal with this distorted data for demand planning in the coming days.

3 Methodology

In the current situation, the solution for demand forecasting using time series techniques is not effective as the historical data is distorted and the forecast accuracy is very low. Therefore, we required such a solution that can work with distorted historical demand data and improve the forecast accuracy. In this chapter, two different methodologies have been discussed according to business type: (i) For commodities or Business-to-Business (B2B) environment and (ii) Business-to-Customer (B2C Environment.

In commodities or B2B environment, the major characteristics are stable market conditions with fluctuating demand, few customers, large-scale orders, derived demand, and less fluctuating prices. This is applicable for Energy, Resource, and

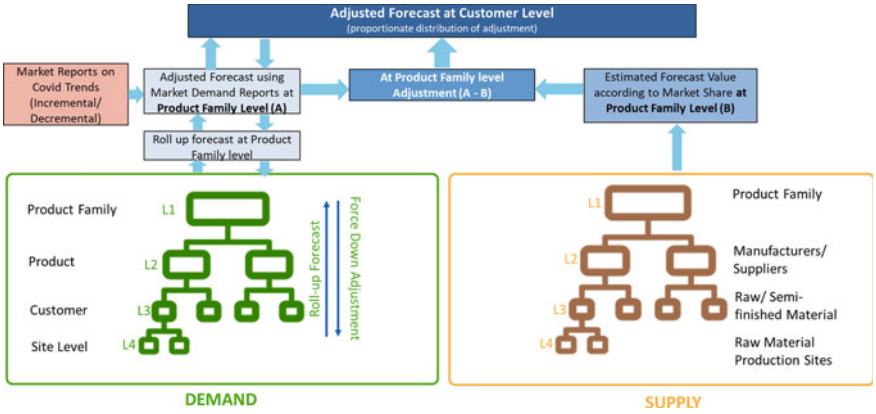


Fig. 4 Demand forecasting for commodities and B2B environment

Utilities, (chemical industry, petrochemical industry, pharma industry, metal industry, etc.) Manufacturing industry (vendors for OEMs, heavy equipment industry, Automobile, etc.) Hi-tech industry, etc. In case of commodities, it is important to consider the supply of raw materials and the market share of the company. For demand side forecasting, the end consumer demand trend should be considered, and information can be extracted from market reports. The adjustment with supply-side forecasting will improve the overall forecasting accuracy. The final adjusted forecast will be distributed up to the lowest level of the hierarchy. The approach has been depicted in Fig. 4.

The figure has shown the hierarchy on both demand and supply side. On the demand side, the demand signals are collected at the site level, which is showing the end consumption of material. For the materials having regular time series demand signals, the forecast is done using traditional time series forecasting techniques like ARIMA, S/ARIMAX, moving average, etc. If there is any cause-effect relationship, the demand is forecasted using regression analysis. For lumpy demand, Croston's method is the most suitable forecasting method. The site level forecast is rolled up at the customer level and then the product level. After that, the forecast will be rolled up at the product family level and product family level forecast will be adjusted with market trends due to pandemic (e.g. the market size of alcohol-based sanitizers will be increased by 45% in next quarter). On the supply side, the forecasting is done at the production site level of raw materials, which is based on market reports or syndicated research. According to a fair share, it is divided into raw/semi-finished material, which will be used for final product manufacturing. The supply is further split according to manufacturer or supplier according to market share. Now, both forecasted demand and supply are obtained at the product family level, which can be compared and adjusted. The forecasted demand is already adjusted using market trends, pandemic trends, and market reports. The adjusted forecasted demand is readjusted using supply forecast value. Finally, the two-level adjusted demand forecast

is further distributed based on fair share i.e., if the consumer/site is having a bigger share, the adjustment will be more. Therefore, the obtained forecast of demand will be more accurate and dynamically adjusted according to changes in demand patterns.

In B2C environment, the market conditions are very dynamic, driven by sentiments, impacted by local incidents. It includes mostly FMCG/beverages products while commodity-based products are very few. This is applicable for retail industry (FMCG, beauty & personal, home care, staple, etc.), CPG industry (beverages, e-commerce, electronics etc.), Travel, Tourism, and Hospitality industry, etc. As the B2C market is driven by market sentiments, so it is important to consider CRM data to understand the buying trend of net-worth customers. The approach has been shown in Fig. 5. The forecasted demand is adjusted by CRM data and later with pandemic data drivers.

In this approach, the demand is forecasted based on the historical demand data. The historical data is distorted; therefore, the CRM data is used to make it more accurate. From the historical demand data, the base demand is calculated based on various indexes (daily, weekly, holiday, etc.). Further, the incremental daily demand is calculated by subtracting the base demand from the daily demand. Finally, the incremental daily demand is forecasted using various traditional time series forecasting techniques. Simultaneously, the CRM data is analyzed, and customers are classified into various clusters according to purchasing and loyalty like High valued, medium valued, and low valued. According to the buying behavior of high valued customers, the weightages are provided to various items like higher weightage to most purchased items. The weightages (high weightage indicates high concentrated purchasing) are assigned to SKUs according to the buying behavior. The forecasted demand is adjusted with weightage of most purchased items. The forecasted demand will be updated using SKU weightages. On the other hand, the SKUs are clustered based on their demand patterns (different phases of demand during and post-pandemic, as discussed in the previous section) which helps to identify demand drivers. Later, based on drivers, the demand is calculated using the regression method. Therefore, the future demand is predicted for each of cluster and SKU. Finally, the CRM-adjusted demand is readjusted based on predicted demand. The finally adjusted demand will be more accurate and there will be less impact of distorted data.

The forecasted demand generated using historical data and the CRM adjustment process will be fine-tuned using the forecast generated from the regression model. As a composite and dynamic analytics approach, many different threads must be woven together to improve the resilience and forecast accuracy during and post-pandemic era. The final adjusted forecast will be more accurate. To implement the dynamic demand forecasting, the data analytics steps should be followed as given in Fig. 6.

First, the data is collected from various data sources like historical demand data, CRM data, pandemic data, external factors, and market research data. The data is harmonized first and prepared for further analysis. First, the missing data is completed using mean/median/interpolation/extrapolation. After that, the outlier data is treated using various techniques like normalization, exclusion, etc. Finally, the negative values are treated using exclusion or with business intervention. The trend and pattern

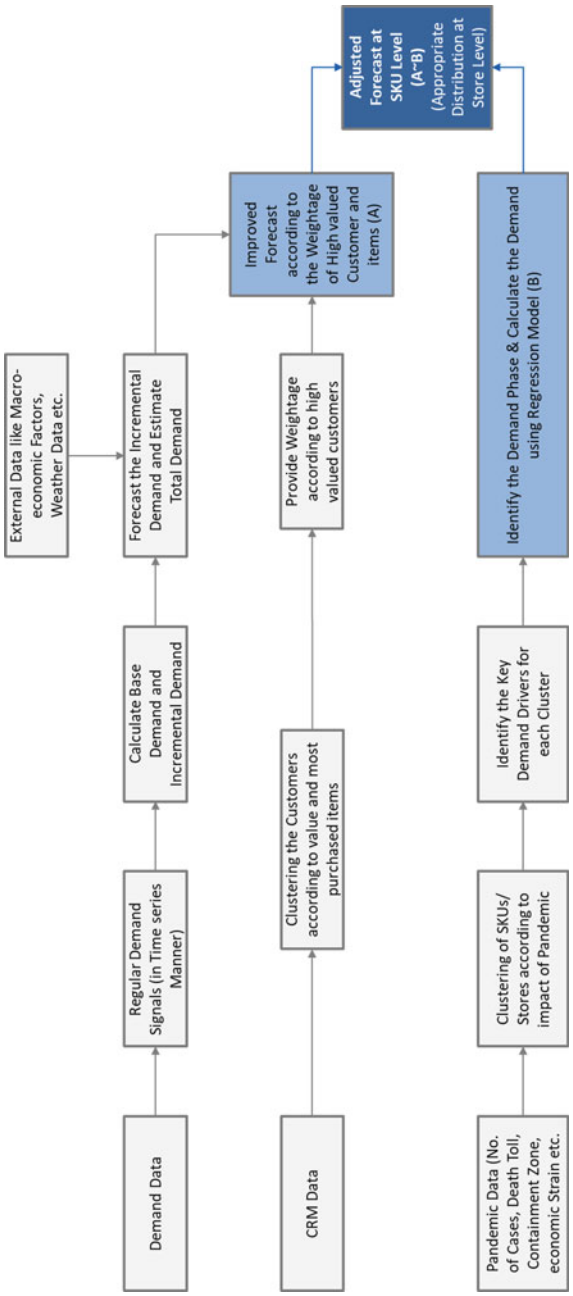


Fig. 5 Dynamic demand forecasting in B2C environment

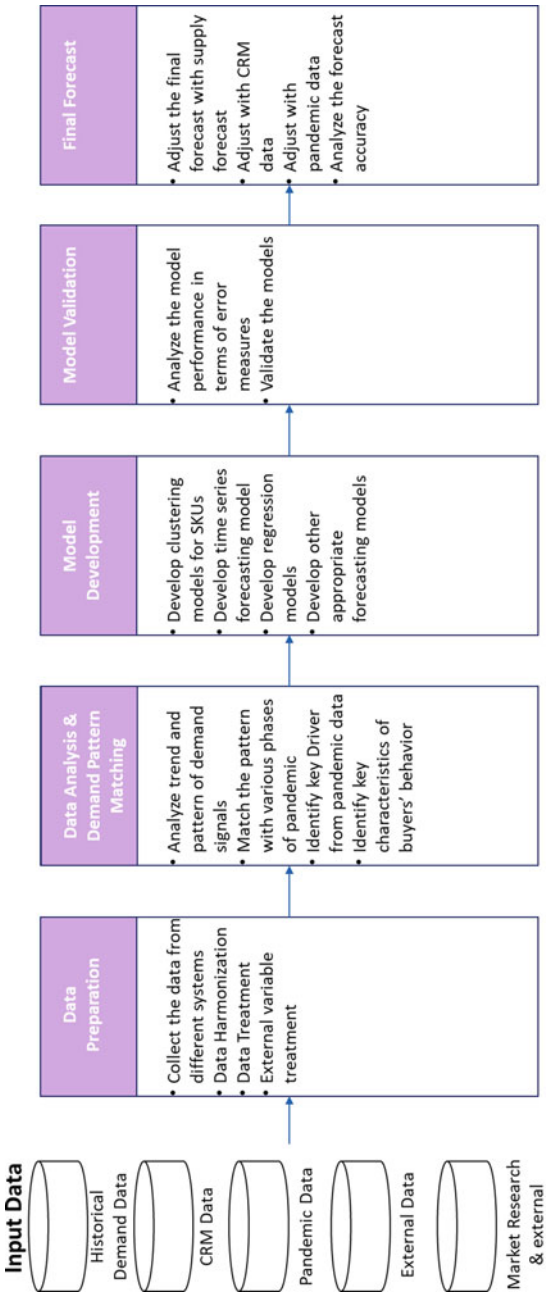


Fig. 6 Data analytics approach for dynamic demand forecasting

of demand data should be analyzed and matched with the stage of the pandemic (structural decrease, structural increase, and hoarding). Simultaneously, the supply data (for commodities), CRM data, and pandemic data (for B2C) should be analyzed and key characteristics of buyers & key drivers from pandemic data (no. of cases, death rate, etc.) should be identified. Later, all the forecasting models for forecasting (time series, causal, regression, ML, etc.) are developed. The accuracy of the forecasting model is calculated using MAPE and model parameters are fine-tuned based on the MAPE value. Finally, the forecast will be adjusted with supply/CRM and pandemic data. This approach is enough robust to capture market dynamics and successfully deals with distorted data.

In the next section, the possible benefits of dynamic demand planning to the organization are discussed. Both quantitative and qualitative benefits are discussed.

4 Results

In the pandemic era, the demand signals are highly distorted, and forecasting based on such distorted data is having low accuracy. Consequently, other planning problems will be generated like improper demand planning, under/over-stocking of inventory, lower demand satisfaction, lower case fill rate, etc. The proposed approach, which can capture all the dynamism of the market and provide the accurate forecast, can provide the benefits in various ways like:

- Improved Forecast Accuracy by
  - Proper categorization of SKUs according to their characteristics and demand trend
  - Capture all the dynamics of the market in pandemic
  - Selection of appropriate techniques of forecasting according to demand signals
  - Quantitative value of adjusted forecast using CRM and pandemic.
- Better Demand Planning
  - The demand planning will be improved with higher forecast accuracy by capturing market dynamics using different parameters
  - SKU are classified according to demand patterns.
- Better Replenishment Planning
  - The outside view (Supply, CRM, pandemic) will help for better replenishment planning
  - Improved forecasting helps to manage inventory in a better way even if demand is sporadic which captures market dynamics.
- Better Production Planning
  - With accurate forecasted demand and having visibility on replenishment plan, the production planning will be better without any unmet demand.

- Maximum Demand Satisfaction
  - With improved demand planning and production planning, the market demand can be satisfied in a better manner.

The proposed approach provides a lot of other qualitative benefits like higher customer satisfaction, improved supply chain resilience, higher NPS score, and improved service level, and those will be resulted in higher market share. Therefore, the organizations can be benefitted in both quantitative and qualitative manner. They can penetrate deep into the market and capture demand signals in effective manner, whereas they can also improve customer satisfaction scores, which will help to improve their brand value. The dynamic demand planning approach can help to improve other supply chain planning like inventory, capacity, production, distribution, etc. The proposed approach can be used by any size of industry like big, medium, and small industries. It will give the most benefit to those industries which are successfully capturing the data (CRM, supply, etc.) and having harmonized data as data lake.

The action points, limitations, and benefits of the proposed approach from the practitioners' point of view have been discussed in the next section.

5 Conclusion

In the past we have observed that life-threatening diseases, natural calamities, or political turmoil are some of the factors that cause supply chain disruptions large-scale uncertainties in demand planning. But COVID-19 is exceptional to some extent as it has suspended almost all operations of the business. Since the situation is rapidly evolving during and post pandemic era, it is highly recommended to regularly retrain and re-run the impact analysis, at a higher frequency. As a supply chain manager, it is important to develop the proper planning strategies to overcome pandemic impact. To develop such a planning strategy, it is essential that they must capture market dynamics and leverages all the new practices as the proposed approach this includes: Capture/collect more pandemic/CRM related data for better prediction, identify future demand drivers, identify pandemic drivers, change forecasting models according to demand behavior, regularly capture and understand supply chain dynamics, Increase communication with customers, restructure and redesign demand planning process, plan product portfolio differently, etc.

The market situations are changing rapidly, and decision making changing continuously due to it. Therefore, only data analytics is the key to make wise decisions. If the appropriate data (e.g. CRM data) is not collected, it should be started immediately. From an analysis perspective, it is good to understand the impact of various other factors (qualitative and quantitative) and those factors should be included in the analysis. The present approach includes all the possible variables and suggests two-level adjustment for the forecast according to demand behavior and other external data. The approach is also different according to the industry type. The managers

can improve the overall resilience of the supply chain by proper demand planning using this approach.

COVID-19 has brought significant changes in consumer behavior and disruption in demand planning. It will impose new consumption habits in near future as well which need to be captured appropriately using new emerging technologies and incorporate them in the demand planning process.

References

1. Heines S (2008) The product manager's desk reference. McGraw-Hill Professional, New York, NY
2. Gattorna J, Ogulin R, Reynolds MW (2003) Gower handbook of supply chain management. Gower Publishing, Ltd.
3. Stadtler H, Stadtler H, Kilger C, Kilger C, Meyr H, Meyr H (2015) Supply chain management and advanced planning: concepts, models, software, and case studies. Springer
4. Bolton J (1998) Effective demand management: are you limiting the performance of your own supply chain. Gower Publishing Limited, Aldershot, England
5. Vlckova V, Patak M (2011) Barriers of demand planning implementation. Econ Manag 1(16):1000–1005
6. Crum C, Palmatier GE (2003) Demand management best practices: process, principles, and collaboration. J. Ross Publishing
7. Sheldon DH (2006) World class master scheduling: best practices and lean six sigma continuous improvement. J. Ross Publishing
8. Voudouris C, Owusu G, Dorne R, Lesaint D (2007) Service chain management: technology innovation for the service business. Springer Science & Business Media
9. Kilger C, Wagner M (2010) Demand planning. In: Stadtler H, Kilger C (ed) Supply chain management and advanced planning. Springer, Berlin, Germany, pp 133–160
10. Arunraj NS, Ahrens D, Fernandes M (2016) Application of SARIMAX model to forecast daily sales in food retail industry. Int J Oper Res Inf Syst 7(2):1–21
11. Singh S, Kumar R, Panchal R, Tiwari MK (2020) Impact of COVID-19 on logistics systems and disruptions in food supply chain. Int J Prod Res 1–16
12. Nurunnabi M, Alhawal HM, Hoque Z (2020) Impact of COVID-19: how CEOs respond to SMEs recovery planning in Saudi Arabia. White Paper 3
13. Telukdarie A, Munsamy M, Mohlala P (2020) Analysis of the Impact of COVID-19 on the food and beverages manufacturing sector. Sustainability 12(22):9331
14. Weersink A, von Massow M, McDougall B (2020) Economic thoughts on the potential implications of COVID-19 on the Canadian dairy and poultry sectors. Can J Agric Econ/Rev Can D'Agroeconomie 68(2):195–200
15. Haren P, Simchi-Levi D (2020) How coronavirus could impact the global supply chain by mid-march. Harward Business Review
16. https: //www.globaltrademag.com/food-sector-faces-multipronged-consequences-of-covid-19-outbreak/2020
17. Aldaco R, Hoehn D, Laso J, Margallo M, Ruiz-Salmón J, Cristobal J, Kahhat R, Villanueva-Rey P, Bala A, Batlle-Bayer L, Fullana-I-Palmer P (2020) Food waste management during the COVID-19 outbreak: a holistic climate, economic and nutritional approach. Sci Total Environ 742:140524
18. Hobbs JE (2020) Food supply chains during the COVID-19 pandemic. Can J Agric Econ/Rev Can D'Agroeconomie 68(2):171–176
19. Gray T (2020) Retail demand planning for pandemic disruptions and the new normal. Tata Consultancy Services Limited, White Paper

20. Sheth J (2020) Impact of Covid-19 on consumer behavior: will the old habits return or die? J Bus Res 117:280–283
21. Nikolopoulos K, Punia S, Schäfers A, Tsinopoulos C, Vasilakis C (2021) Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. Eur J Oper Res 290(1):99–115
22. Hendricks KB, Singhal VR (2003) The effect of supply chain glitches on shareholder wealth. J Oper Manag 21(5):501–522
23. Lorey J, Naumann F, Forchhammer B, Mascher A, Retzlaff P, ZamaniFarahani A, Discher S, Faehnrich C, Lemme S, Papenbrock T, Peschel RC (2011) Black swan: augmenting statistics with event data. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp 2517–2520
24. Cranfield JA (2020) Framing consumer food demand responses in a viral pandemic. Can J Agric Econ/Rev Can D'Agroeconomie 68(2):151–156
25. Nelson D (2021) The new demand planning: leveraging CRM for post-pandemic supply chain success. River Rock Advisors
26. Nijs VR, Dekimpe MG, Steenkamps JBE, Hanssens DM (2018) The category-demand effects of price promotions. In: Long-term impact of marketing: a compendium, pp 187–233

Chapter 2
Cognitive Models to Predict Pipeline Leaks and Ruptures

Sugato Samanta, Kiran Kumar Ganapathi, and Laksanya P. Dewan

Abstract Pipelines are responsible for transporting close to 90% of the crude oil, natural gas, and natural gas liquids across the globe. Leakages are a huge risk to any midstream oil and gas organization leading to huge losses from a financial, environmental, and brand perspective. Although organizations have an integrity management program to help detect pipeline ruptures and leakages, engineering-driven maintenance by itself is not enough to predict future incidents. Hence in these scenarios, data and analytics-driven methodologies can help organizations predict anomalies in advance which can have a huge benefit in terms of timely maintenance of those specific cross sections of the pipelines and thereby prevent leakages and incidents.

Keywords Pipeline Integrity · Maintenance · Geohazard Conditions · Anomaly Prediction · Physics Models · Neural Network · Artificial Intelligence · Statistical Methods

1 Introduction

Oil and gas pipelines across the world span over a variety of rough terrain stretches like mountains and forests to name a few. The pipelines tend to experience inclement weather conditions over the years, not to ignore intra-day temperature fluctuations. As a result, they stand a higher risk of suffering an anomaly in structure or function from weather-driven geo-hazards (also causing corrosion) than third-party damages.

S. Samanta (✉)
A&I Business Analytics, Tata Consultancy Services, Kolkata, India
e-mail: sugato.samanta@tcs.com

K. K. Ganapathi
A&I Business Analytics, Tata Consultancy Services, Hyderabad, India
e-mail: kirankumar.ganapathi@tcs.com

L. P. Dewan
A&I Business Analytics, Tata Consultancy Services, Bengaluru, India
e-mail: laksanya.dewan@tcs.com

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_2

Pipeline routes are particularly challenging. The presence of steep slopes, heavy rains, and soil movements cause major and minor incidents on a regular basis, environmental impact, and long periods of service disruptions. The essence of pipeline integrity is to maintain stress/strain within the allowable limits based on the existing operating conditions. Heavy rainfall and landslides present significant hazards to buried pipelines because of the propensity to cause large deformations induced by longitudinal stresses above allowable levels [1]. The occurrence of these incidents each year causes a millionaire inversion, more importantly, irreversible and irreplaceable damages to the environment and life around.

In most cases, the probability of pipeline rupture is usually higher due to anomalies like dent, ovality, and metal loss (corrosion) despite using high-grade stainless steel in the pipeline as per ASME B31.8 or CSA Z662-11 standards [2]. Traditionally, oil and gas firms rely on integrity management programs and an engineering-driven maintenance team to detect the possibility of an incident. These methods lack continuity in monitoring frequency and experience delays as harsh terrains and bad weather conditions tend to hinder the maintenance activities across various sections of the pipeline.

2 Literature Review

In recent years, there has been some research in this area to utilize data science-driven algorithms to predict anomalies and develop a digital twin of the equipment in question. Rex et al. [pg. 13–15] have tried to assess the susceptibility or hazard posed by landslides fall into qualitative and quantitative methods. The qualitative methods have been subdivided into Geomorphic Analysis which is based on field observation and aerial-based imagery. The other qualitative method is based on Weighted Parameter Analysis using Geographical Information System (GIS) where the evaluator selects and maps parameters (such as slope, geology, and drainage density) that influence the stability of slopes based on personal experience and assigns a weight to each parameter in accordance with its relative contribution to slope failure. This is then used to create the hazard map, and the quantitative methods are based on Statistical Analysis and Geotechnical Models which use multivariate analysis and probabilistic models to estimate the spatial probability of landslide occurrence [1].

For a gas transmission and distribution pipeline system, there are guidelines laid down by the American Society of Mechanical Engineers regarding the inspection of piping systems intended to operate and the testing criteria [pg. 30–33] which are important factors specifically in the welding portions of the pipelines which are prone to higher stress and strain [2]. Similarly, the measurements from strain gauges from National Instruments are an important parameter to understand the impact of the force/pressure applied [pg. 1–2] and the corresponding Poisson effect [3]. Feng et al. [pg. 2–4] mention a calculation method for pipeline bending strain which was relevant to this exercise along with a possible method of modeling and computation of dent [4]. The pipelines mentioned throughout the study are made of stainless steel

conforming to US material standards [pg. 6, 7, 46, 47] and governed by material design guidelines for minimum strength and burst values [5].

The criteria of using von mises were inspired by several documents including a research material by Bereiša et al. [pg. 1–3] to determine the criteria for a thick wall pipe with closed valves based on principal stresses [6]. Finally, inclinometers are an important instrument to measure soil displacement and the horizontal displacement is derived from Slope inclinometers for landslides by Timothy D. Stark [pg. 3] [7] and Geokon instrument user manual [pg. 74–85] [8].

3 Material and Methodology

3.1 Defining the Solution Using Data and Analytics

With the onset of technology and tools such as PIG (Pipeline Inspection Gauge), SCADA (Supervisory control and data acquisition), LiDAR (Light Detection and Ranging), and other scientific instruments along with the availability of drones have expanded the velocity, veracity, volume, and variety of the data collected. This propels the scope of Analytics and Business Insights and opens doors to a specific domain within the vast world of analytics. Since the pipelines pass through rough terrain, it is imperative that specific features like the correlation between soil pressure, rainfall, strain, and groundwater pressure based on historical data are understood; and whether there are any pattern of similar events and/or incidents which has happened in the past. This also means that analyzing the readings from the scientific instruments using Physics-based formulae and their association to the business problem is key to developing a successful solution.

The solution involves a two-step process methodology—(1) Smart Pipeline Analytics System (SPAS) which incorporates the development of a novel ensemble model of physics and machine learning, to predict the occurrences of defects taking into consideration the existing operating conditions. The machine learning model self-learns, expanding its intelligence basis historical data patterns and the correlations drawn from various hypotheses. (2) Early Warning System (EWS)—A risk scoring methodology based on Z-Scores. It helps generate risk profiles providing real-time pipeline health reports, easing out the maintenance using data and analytics. This is usually a precursor to the earlier method wherein post risk profiling, the logical step would be to determine the exact cross-section of the pipeline which could be impacted due to defects and anomalies.

3.1.1 Novel Ensemble Technique

Here engineering meets artificial intelligence. It involves preparing a list of variables combining the previously mentioned parameters. These broadly include material

properties and sensor readings. There are also a set of artificially generated variables based on a deflection in readings, cumulative effects e.g. aggregated rainfall and temperature differences to name a few. These are incorporated in AI/ML models—

Artificial Neural Network

A Deep Learning algorithm that uses mathematical models to find relationships between the input variables and output by finding patterns in the data. The relationships are based on weights and hence the algorithm chooses to classify the anomalies from non-anomalies based on patterns.

Support Vector Machine

A supervised learning model that can be used to separate anomalies from non-anomalies to train the model and predict future anomalies. After training the models, their internal parameters are tweaked to ensure their fine-tuning for optimized results.

As mentioned in the below representation (Fig. 1), the approach is to break the overall problem statement into subcomponents and solve:

Step 1: Calculations for scientific instrument readings based on physics-based formulae involving stress, strain, etc.

Step 2: Establish a relationship using statistical analysis between variables like rainfall, inclinometer readings, etc.

Step 3: Introduction of novel ensemble technique using machine learning or deep learning to combine physics and statistics and solve the stated hypothesis.

3.1.2 Data Required

There are multiple data sources and types which are required to develop the ML/AI model as mentioned below, detailed variables are depicted in Fig. 2. All the datasets will be available with the operation and maintenance team(s) of the organization. Instruments like strain gauges are usually installed every five hundred meters to record the strain caused on the pipeline due to flowing materials. Readings from strain gauges are captured within GIS (Geographic Information System) of the organization. GIS is usually used to capture details related to the topology of the area through which the pipeline has crossed. In this context the strain gauge readings contain data variables like microstrain, delta, reading date, and SMYS (Specified Minimum Yield Strength). During the course of regular maintenance, maintenance personnel uses PIG's to understand the characteristics of the materials within the pipeline. The horizontal and vertical curvatures of the pipeline are recorded by the sensors present within the PIG and the inertial data is analyzed offline. The main intention of the maintenance engineers is to calculate the deviation in both curvatures since these deviations could potentially trigger anomalies if not rectified at a suitable time. Maintenance personnel also keep records related to the locations where historically an incident has occurred for future references. The pipeline incidents dataset contains details of the cross-section of the pipeline where past incidents had occurred.

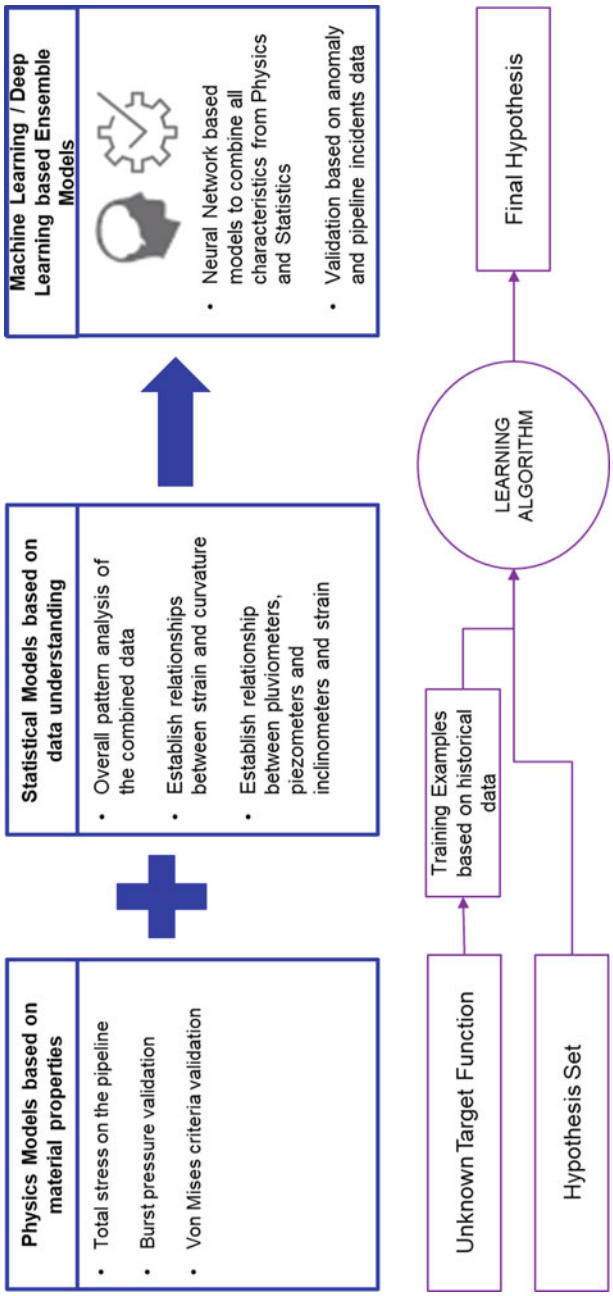


Fig. 1 Model development hypothesis

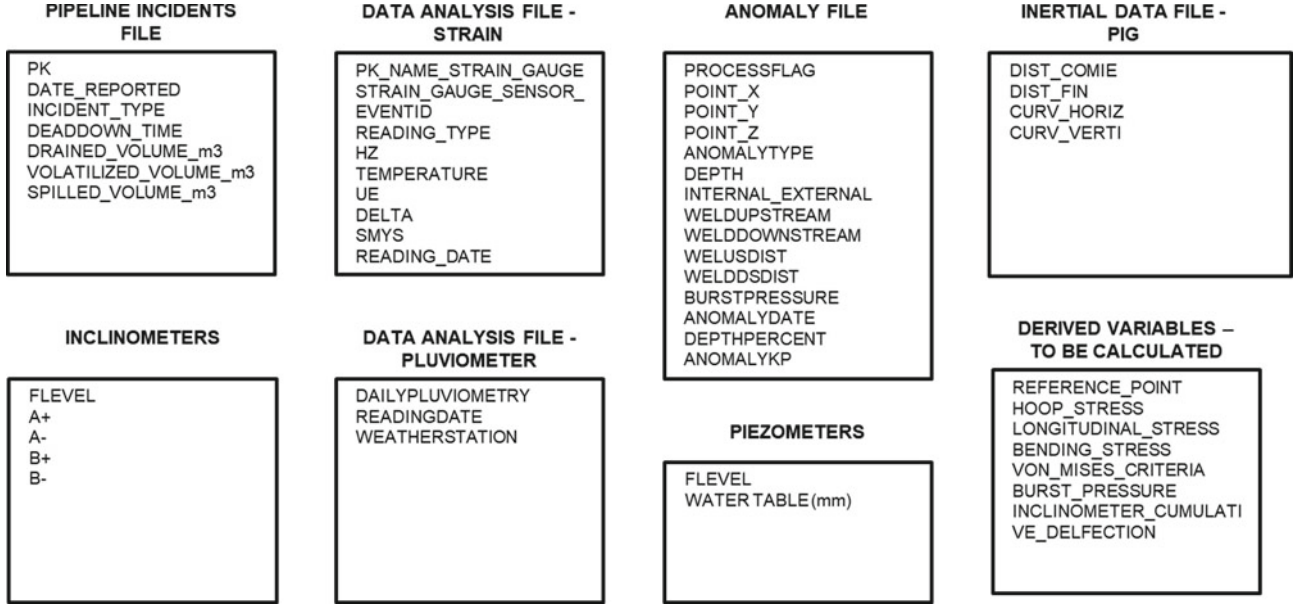


Fig. 2 Data variables

Similarly other instruments like inclinometer, pluviometer, and piezometer are used for various measurements for the upkeeping of the pipeline. Inclinometer measures the soil displacement (in mm). The soil displacement can be caused due to several reasons like the impact of the weight of the pipeline, heavy rainfall, and natural causes. Piezometer measures the groundwater level of the area where the pipeline is placed to ensure that there is not a sudden change in the water pressure upwards which might displace the pipeline from its original location. Also instrument like a pluviometer is a standard measuring tool of the total rainfall received within the region. The weather stations are usually installed every 2 km of the pipeline in regions that receive heavy rainfall. The readings are taken twice a day to determine the average rainfall received per day.

Finally, it is very important to understand the material characteristics of the pipeline and its changes over a duration of time. Pipeline material properties like Young's Modulus, Pipe Type, Yield, Tensile, Thickness, and Diameter will change depending on the type of the pipeline and its usage. Owing to weather conditions and other factors the pipeline is also prone to corrosion and changes in material strength at the welding points resulting in dent and ovality. These must be referenced as per geographical coordinates to make the information usable to the maintenance personnel.

The derived variables are based on calculations as explained in the next section.

Other factors related to pipeline properties which were important to the model building exercise have been mentioned:

PIPE TYPE—API 5L X70 PSL2.
EXTERNAL DIAMETER—14 inches/355.6 mm.
THICKNESS—0.21 inches/5.56 mm.
YIELD—70,000 psi/483 MPa.
TENSILE—82,000 psi/565 MPa.
YOUNG'S MODULUS—210,000 MPa.
POISSON RATIO—0.3
INTERNAL PRESSURE—8 MPa.
BURST PRESSURE—2590 psi/18 MPa.

3.1.3 Building the Analytics Base Tables and Variables for AI/ML Models

As mentioned in Sect. 3.1.1, the first step would be to develop the physics-dependent hypothesis and then proceed with the statistical analysis and finally merge both approaches using an ensemble technique. In case of physics-dependent hypothesis, it is important to understand the factors affecting the material characteristics of the pipeline and thereby its behavior. From an engineering standpoint, stress and strain exerted on any material will result in a change.

As depicted in Fig. 3, Stress or Hoop Stress in this context deals with the vertical force applied on a cross-section of the pipeline due to the flow of oil/gas/water through the pipeline. Strain relates to the amount of deformation material experiences due

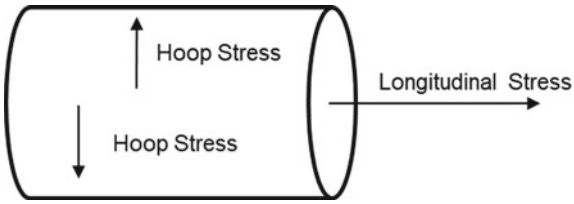


Fig. 3 Stress factors on a pipeline

to an applied horizontal force. This value is positive for elongation and negative for compression. Both combined influence the material characteristics and contribute to the factors responsible for pipeline leakage or rupture.

From a stress perspective, the main calculations revolve around determining Hoop Stress and Von Mises criteria.

Consideration for **Hoop Stress = (2 * S * T)/D** in MPa

$$SMYS(S) = \frac{Delta\ MicroStrain\ \times 10^{-6} \times Young's\ Modulus}{\%SMYS\ from\ Strain\ Gauge} \times 100\%$$

T => Pipe Thickness, D => Pipe External Diameter

Consideration for **Longitudinal Stress** = (Young's Modulus * Delta Microstrain from strain gauge) in MPa.

Total stress which can be sustained by the pipeline is based on Von Mises criteria for plastic deformation. This combines the Hoop Stress and Longitudinal Stress. The calculation of Von Mises criteria is defined as:

$\boldsymbol{\sigma_e = [sqrt((\sigma_1 - \sigma_2)\ \hat{}\ 2 + (\sigma_2 - \sigma_3)\ \hat{}\ 2 + (\sigma_3 - \sigma_1))] * [1 / sqrt(2)]}$ where σ_1, σ_2, σ_3 are the principal component stresses [6].

If the above criteria are to be replicated in the current scenario with the data available, then it can be replaced as

$\boldsymbol{\sigma_e = sqrt(\sigma_H\ \hat{}\ 2 + \sigma_L\ \hat{}\ 2 - \sigma_H\ \sigma_L).}$

where $\boldsymbol{\sigma_H}$=> Hoop Stress

and $\boldsymbol{\sigma_L}$=> Longitudinal Stress.

Hence mathematically the Von Mises criteria can be denoted as

sqrt (Longitudinal Stress ^ 2 + Hoop Stress ^ 2 − Longitudinal Stress * Hoop Stress).

Strain is the other component which can impact the material properties. Since a pipeline is a cylinder with welding done after every few meters. Also, the pipeline is not straight and is curvaceous at several points over the full length. Hence there is the involvement of bending strain at the curvatures leading which must be accounted for.

The bending strain is calculated as **Dk/2** where D => diameter of the pipe in mm, k => total curvature based on vertical curvature (k_v) and horizontal curvature (k_h). Both the vertical and horizontal curvatures are captured as a percentage in the PIG

during regular inspection of the pipeline [4].

$$k = \sqrt{k_v^2 + k_h^2}$$

Usually, the output from the PIG run will provide data related to Horizontal Strain (HStrain%) and Vertical Strain (VStrain%), and hence the basic method of calculating bending strain is not required. The total bending strain can be calculated as **TStrain = sqrt((VStrain%)^2 + (HStrain%)^2)** and has been used as an input variable for the model.

Barlow's Method for maximum allowable pressure [5], mentions the total pressure which can be applied before the pipeline bursts out. This in simple terms can be calculated as **P = 2St/D** where S => SMYS (if yield is considered), t => nominal wall thickness, D => pipe diameter. In real conditions the cylindrical pipeline has two layers and hence some thickness is also present, which must be considered as part of the calculations.

Burst Pressure = 2 * Tensile * (Thickness − Depth)/((External Diameter − 2 * Thickness − Depth) * Safety Factor) We are assuming safety factor = 1.

Using the above formulations, the following Physics dependent hypothesis were checked:

Hypothesis 1: Bending strain will be calculated based on the curvature data for every point where reading is available from PIG and its value must be less than the microstrain readings from strain gauges [Bending Strain < με] for the same pipe section. This is to ensure that the strain exerted by the flow is within the allowable limit.

Hypothesis 2: Von Mises criteria value should be less or in the neighboring range of tensile value at any point ($\sigma_e < 565$).

Hypothesis 3: Calculated Burst Pressure at any point would be less than Burst Pressure set by OEM standards. In mathematical terms it is denoted as Calculated Burst Pressure = (2 * σe * Thickness)/External Diameter and eventually this burst pressure must be less than 18 MPa for such pipeline types.

Hypothesis 4: Calculated Burst Pressure at any point would be validated against the Burst Pressure from anomaly table caused due to External–Internal Metal Loss.

Hypothesis 5: The overall stress in the pipeline is related to soil pressure, rainfall, and other external factors.

The following statistics dependent hypothesis were also checked:

Hypothesis 1: There will be a correlation between rainfall and pressure of the water layer based on historical data. If there is rainfall, then the water layer will increase and thereby the water pressure.

Hypothesis 2: There will be a correlation between inclinometer, rainfall, strain, and water pressure. Changes in values of the inclinometer are based on soil movements and thereby affecting strain.

Considerations for the hypothesis: (1) Correlation between change in inclinometer readings (referred as Δ Inclinometer) with change in strain readings (referred as Δ Strain), (2) Correlation between change in inclinometer readings (referred as Δ

Inclinometer) with change in pluviometer readings (referred as Δ Rainfall), and (3) Correlation of Δ Inclinometer with change in piezometer readings (referred as Δ Piezometer).

Δ Strain => Changes in strain values for month on month,

Δ Piezometer => Changes in piezometer values for month on month.

Δ Rainfall = SUM (Rainfall from Pluviometer as observed on D')—SUM (Rainfall from Pluviometer as observed on D) where D => current month where observations are available for rainfall and D' => preferably next month where observations are available for rainfall.

Δ Water Pressure per month = SUM (Pressure from Piezometer as observed on D')—SUM (Pressure from Piezometer as observed on D) where D => current month where observations are available for piezometer and D' => preferably next month where observations are available for piezometer.

For inclinometer readings [7, 8], the important factor is to measure the soil displacement between readings which are important criteria to determine the changes.

Cumulative_A is calculated as => DA at a certain FLEVEL = 20.5 and then the next values of DA are added up with the previous level. A similar approach has been taken for Cumulative_B using DB. The assumption is the majority of the displacement will happen at the top soil level ($L\sin\theta \sim 0.5$ m). As shown in Fig. 4, the cumulative displacement is the measured basis on the vertical elevation post which the deviations are calculated.

However since the recorded readings for the inclinometer are in the form of displacement (A+ , A−, B+ , B−), hence the final cumulative displacement has to be calculated. This is specifically for inclinometers where the manufacturer type is 'Geokon'.

Calculated digit change for A axis SA = (A+ − (A−))/2.

Calculated digit change for B axis SB = (B+ − (B−))/2.

Deflection A not corrected CA = M * RINT * SA.

Deflection B not corrected CB = M * RINT * SB.

RINT => absolute reading interval in feet or meters (considered as 0.5) and M => 0.05 for mm since the probe constant is 20000.

Deflection corrected for angle DA = (CA * cos(ZZ)) − (CB * sin(ZZ)).

Deflection corrected for angle DB = (CA * sin(ZZ)) + (CB * cos(ZZ)).

ZZ => correction angle (usually zero degree).

Final cumulative deflection = sqrt(DA^2 + DB^2).

Net_Cumulative_Displacement = sqrt(Cumulative_A^2 + Cumulative_B^2).

Δ Inclinometer = (Net Cumulative Displacement at Current month − Net Cumulative Displacement at Previous Month) at every level.

The Δ Inclinometer has been compared with the Δ Strain values for the corresponding months to find any pattern with soil related strain changes.

Finally, the following relationships were observed based on the correlation values between inclinometer, strain, pluviometer, and piezometer using standard correlation techniques from a statistical view:

$$\Delta \text{Inclinometer} \propto \Delta \text{Strain}$$

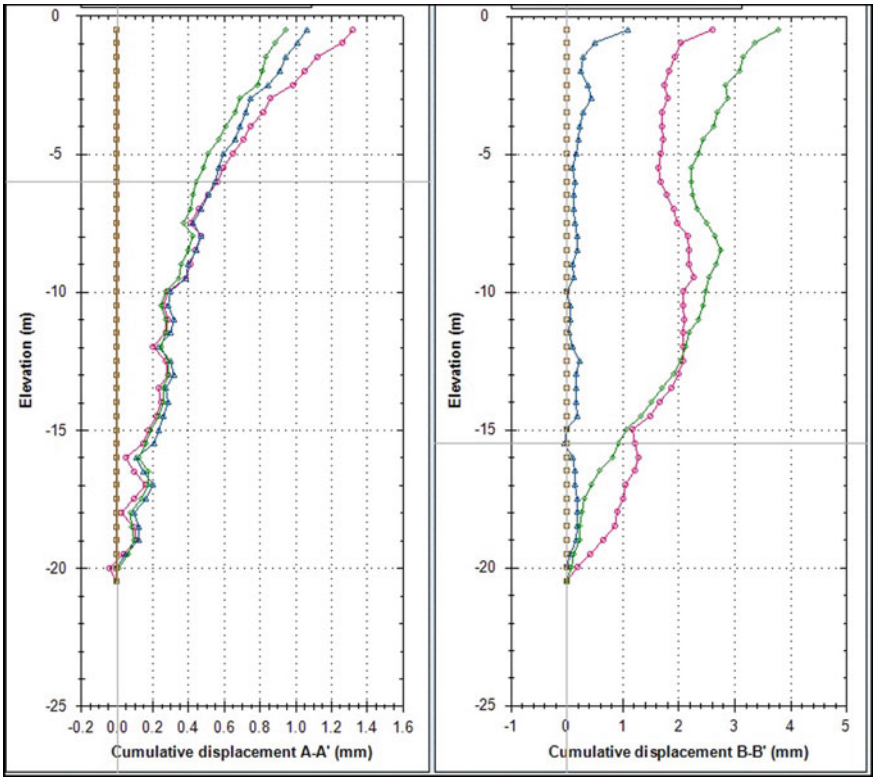


Fig. 4 Cumulative displacement for inclinometer

$$\Delta \text{Inclinometer} \propto \Delta \text{Rainfall}$$

$$\Delta \text{Inclinometer} \propto \Delta \text{Peizo}$$

3.1.4 Ensemble Model Based on Combination of Physics and Statistics

Step 1: Ensemble Model—

Below are the variables finally considered in the model development exercise:

Temperature, Micro Strain, Delta Strain, Longitudinal Stress, Hoop Stress, Von Mises, Pressure at Point, Bending Strain, Daily Rainfall, Inclinometer (4 in number), Piezometer (2 in number).

Techniques which were leveraged: Logistic Regression, Decision Tree, Support Vector Machine, Neural Network.

Final selected modeling technique: Support Vector Machine, Neural Network.

Prediction Outputs:

As it can be observed in Fig. 5, both the machine learning models have predicted similar outcomes for the validation data set. There were subtle differences in the model output when SVM had a better accuracy rate than Neural Network (Fig. 6) in marginal cases. However, if data sets related to LiDAR, video, or other data sources are added to the model, then it is advisable to use Neural Network models.

Finally, a simulation model can be developed to help the field engineers determine the probability of an anomaly occurring using various what-if analysis scenarios

Support Vector Machine

Predicted Output for Validation Set

	No Anomaly	Anomaly
Actual	305	10
Predicted_SVM	307	8

Predicted Output

	No Anomaly	Anomaly
Predicted_SVM	32	66

Neural Network

Predicted Output for Validation Set

	No Anomaly	Anomaly
Actual	305	10
Predicted_NN	307	8

Predicted Output

	No Anomaly	Anomaly
Predicted_NN	66	32

Fig. 5 Prediction outputs

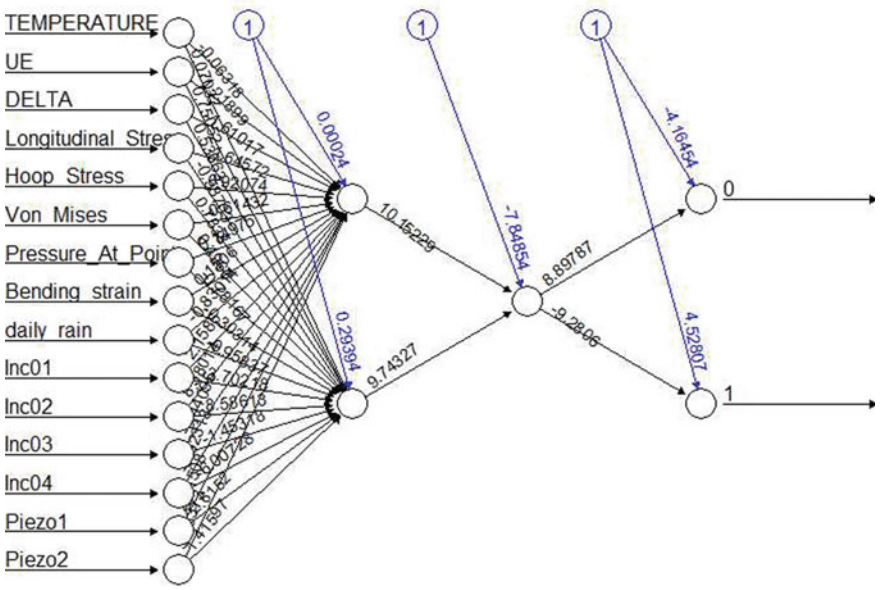


Fig. 6 Neural network representation

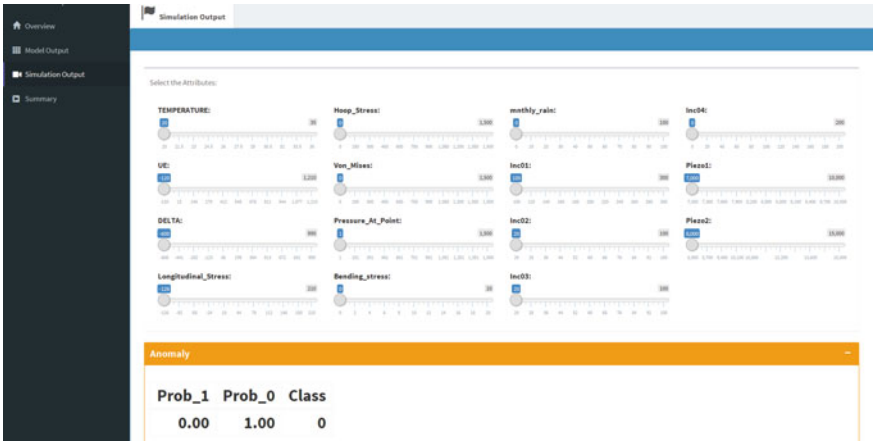


Fig. 7 Scenario simulation

(Fig. 7). The maintenance team will have the ability to change the values of the various scientific instruments as per present operating conditions and knowledge of the domain to visualize in run time the various values of the parameters leading to the model predicting an anomaly.

Step 2: Early Warning System (EWS)—

The Early Warning System is an approach which re-calibrates readings and generates outputs based on the Z-Score methodology in statistics. This results in setting up thresholds which buckets the output, also known as Risk Scores, into three risk bands namely High, Medium, and Low. It helps generate a risk profile namely a Pipeline Health Report which provides the real-time health of any given section in the pipeline. The high-risk prone zones are flagged red and demand immediate attention (Fig. 8). The self-learning ML models can classify more than 80% of the incidents.

To arrive at the final model, and thereby risk profiling there were several investigations which were conducted:

Uneven distributions of anomaly basis on the available data, leading to issues with misclassification of various algorithms (Fig. 9). Hence the anomalies had to be synthesized to enable the data distribution between the development and the validation set.

Dimensionality reduction techniques (Fig. 10) to converge on the key drivers leading to anomalies, it can be observed that inclinometer readings combined with bending strain and delta as observed from the strain gauge are the most contributing factors.

Using various modeling techniques (Fig. 11) to determine the best fit case, tree-based models were discarded due to overfitting. Logistic regression was the preference for risk assessment since the sensitivity of regression was better suited to drive subtle changes in model parameters.

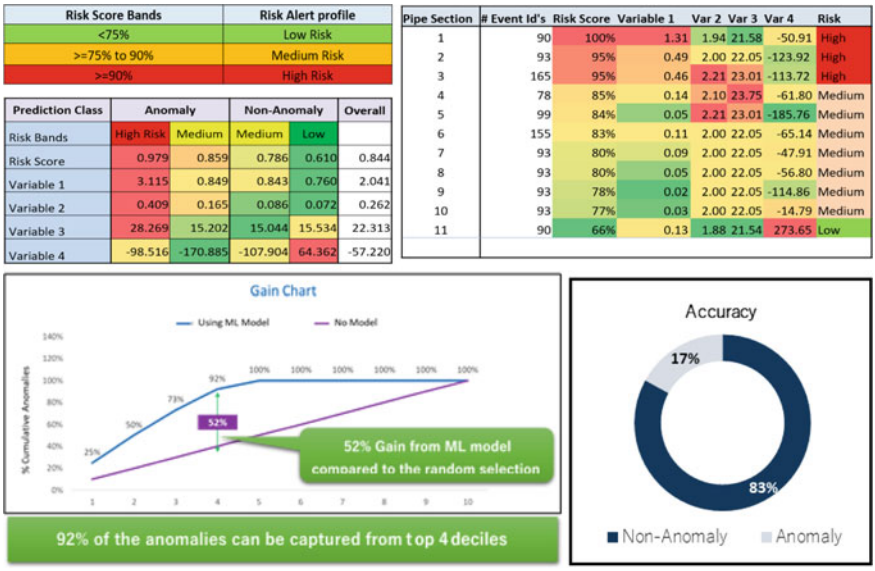


Fig. 8 Pipeline risk profile and health report

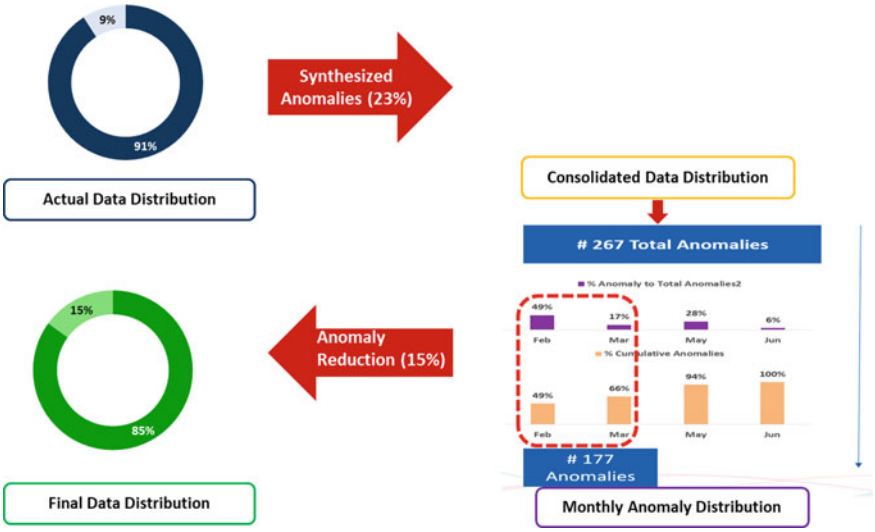


Fig. 9 Anomaly distribution

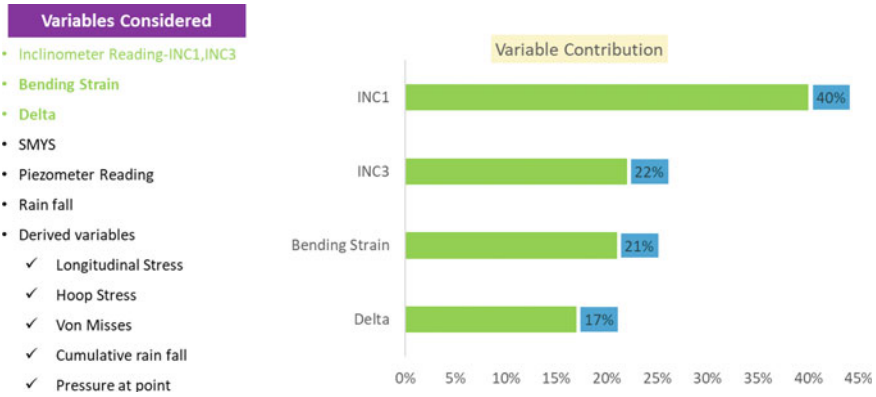


Fig. 10 Key drivers for anomaly

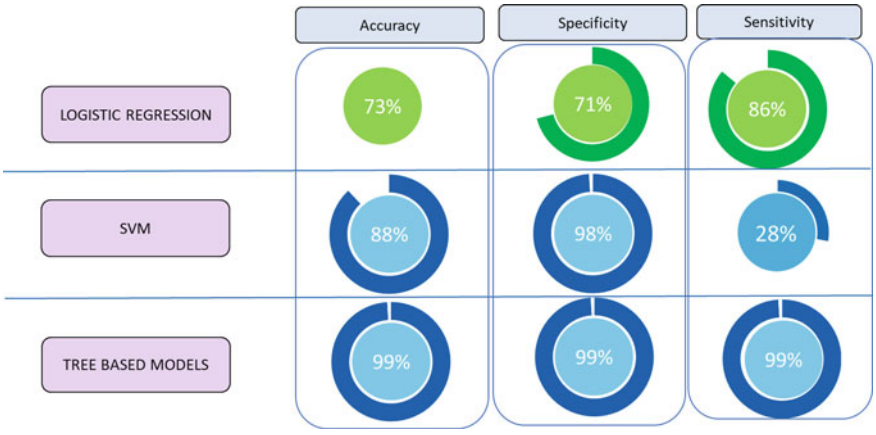


Fig. 11 Model characteristics

Final selection of logistic regression model due to consistency. As it can be observed in Fig. 12, based on multiple iterations and fine-tuning we were able to arrive at the model having significant variables related to inclinometers (INC 1 and INC 3) out of four, delta from each strain gauge and overall strain (Von Mises value).

4 Results

The basis on the usage of Machine learning and Deep Learning techniques, anomalies for a particular section of the pipeline can be predicted with an accuracy of close to 82%. This can be of immense help for the maintenance teams to carry out preventive maintenance well in advance and stop the anomaly to get converted into an accident.

Logistic Regression (Model :N)

```
Call:
glm(formula = Anomaly ~ DELTA + SMYS + INC1 + INC2 + INC3 + INC4 +
    `Total Strain` + PZ1 + PZ2 + PZ3 + `Rain fall` + long_stress +
    hoop_stress + von_mises + pressure_at_point + cumulative_rain,
    family = "binomial", data = train_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.55775  -0.66185  -0.08456  -0.01155   2.13786

Coefficients: (3 not defined because of singularities)
                      Estimate        Std. Error z value Pr(>|z|)
(Intercept)       0.669211084198   200.582790250353   0.003  0.99734
DELTA          -126.464324083010   171.080804309228  -0.739  0.45978
SMYS         295091.454808274808 399188.685600002005   0.739  0.45977
INC1             -1.088338800993     1.045049146339  -1.041  0.29768
INC2             -0.469470866248     0.499077173886  -0.941  0.34687
INC3              0.058976510670     0.048492365026   1.216  0.22391
INC4              0.003610854214     0.004989729788   0.724  0.46928
`Total Strain`   -3.799582093451     0.867944515632  -4.378 0.000012 ***
PZ1              -0.545537452755     1.446346019934  -0.377  0.70604
PZ2              -0.369093160335     1.086241636092  -0.340  0.73402
PZ3               0.975094385909    18.920440518566   0.052  0.95890
`Rain fall`       0.001705865139     0.008433454763   0.202  0.83970
long_stress                   NA                 NA      NA       NA
hoop_stress                   NA                 NA      NA       NA
von_mises         0.000000013031     0.000000004045   3.222  0.00127 **
pressure_at_point             NA                 NA      NA       NA
cumulative_rain   0.002350526571     0.003084703713   0.762  0.44606

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 703.12  on 798  degrees of freedom
Residual deviance: 496.39  on 785  degrees of freedom
AIC: 524.39
```

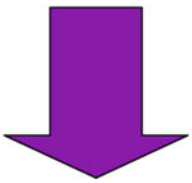


Fig. 12 Regression model output

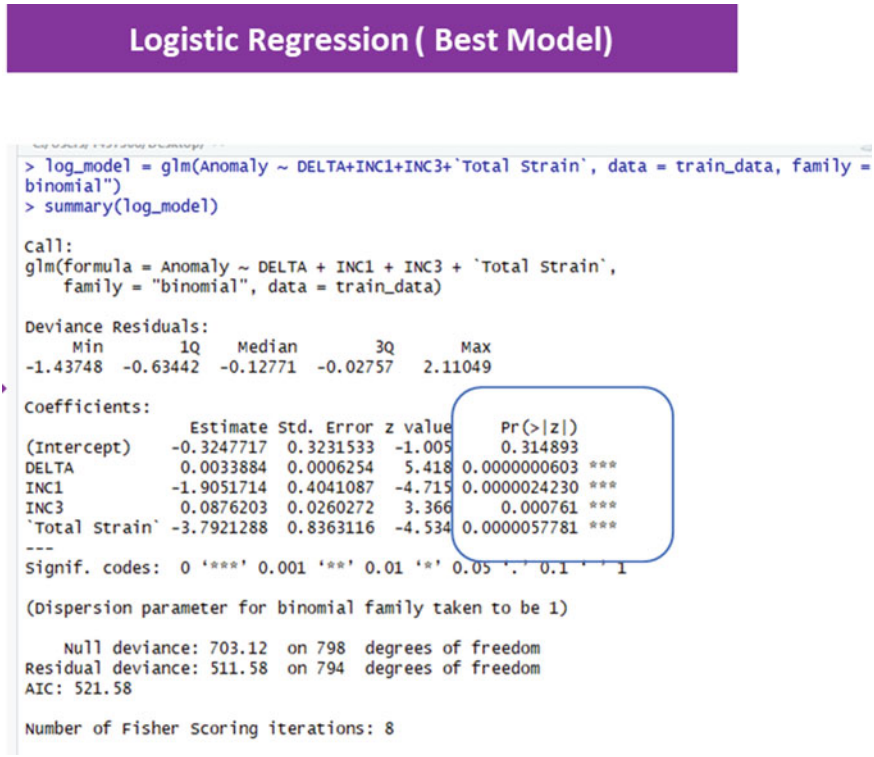


Fig. 12 (continued)

This can also reduce the cost of a pipeline failure by 80%. These include the cost of repairing, part replacement, service, third-party damages, and many more. The ability to know the exact location in the pipeline, which might be in rough terrains or not locations which cannot be frequently visited by the maintenance team can help reduce the overall maintenance costs and eases the deployment of the maintenance team. Additionally, the model can be fine-tuned and scaled to add additional variables related to LiDAR, optical fiber, drone images, etc. Also, there are opportunities to include the concepts of reinforced learning to have self-learning capabilities every time it is executed.

5 Conclusion

It is important to look beyond conventional techniques involving only engineering specifically when it is unable to predict future incidents caused due to geographical conditions which are highly unpredictable. In such instances an ensemble approach combining physics and data analytics as detailed above is of importance to solve

critical business problems. The benefits to such programs are not only limited to the organization but also extend to the ecosystem in which they operate. The incidents not only cause monetary damages but also environmental and human life is at stake in certain situations. Hence, efficient usage of data and analytics in today's world can have a huge contribution to nature and life around which are irreplaceable and priceless.

References

1. Baum RL, Galloway DL, Harp EL (2008) Landslide and land subsidence hazards to pipelines, prepared in cooperation with the U.S. department of transportation, pipeline research council international, and DGH consulting, Version 1 U.S. Geological Survey 1164:13–15
2. Gas transmission and distribution piping systems 2004: ASME 831.8-2003:30-33
3. Measuring strain with strain gages. National Instruments Corporation (2020) 1–2
4. Qingshan F, Rui L, Hong Z (2016) Modeling and calculation of dent based on pipeline bending strain, China. J SensS, Hindawi 2016:8126214, 2–4. https://doi.org/10.1155/2016/8126214
5. Standard pipe and line pipe, US steel tubular products, Rev 2A 9/12:46-47
6. Bereiša M, Žiliukas A, Leišis V, Jutas A, Didžiokas R (2005) Comparison of pipe internal pressure calculation methods based on design pressure and yield strength. MECHANIKA. 54(4):1–3
7. Slope inclinometers for landslides by Timothy D. Stark University of Illinois, Urbana-Champaign, 2008: s10346-008-0126-3:3
8. Inclinometer Readout User Manual Geokon, GK-604D, 2013:C(7):74–85

Chapter 3
Network Optimization of the Electricity Grid to Manage Distributed Energy Resources Using Data and Analytics

Sugato Samanta

Abstract The conventional electricity grid is laden with various challenges due to aging infrastructure, growing power demand, and the rising community of consumers who have turned to produce their own electricity thereby being converted to prosumers. Developed countries are committed to achieve net-zero carbon emissions by 2050, and hence the generation and distribution companies must make the best use of low carbon generation to develop a sustainable zero-carbon electricity network. In the next decade, we will see millions of homes and businesses embrace electric vehicles and use battery storage in combination with renewable sources of generation. However due to the increased generation at downstream and end-user premises, there are situations of grid imbalance that are being handled by the transmission and distribution companies by use of specific load balancing software. This chapter intends to also investigate the problem from an optimization angle thereby helping the organizations with grid optimization and forecasting overall energy requirements and pricing.

Keywords Distributed energy resources · Grid optimization · Smart grid · Optimization · Goal programming · Electricity utilities · Transmission & distribution · Solar energy

1 Introduction

Electrical grid optimization refers to the determination of the necessary network requirements by maintaining the supply of required electricity to the demand sites at the lowest cost. It is mainly related to the transmission and distribution side of the value chain. As detailed in Fig. 1, in the present-day scenario the utility organizations deal with situations like growing power demand, providing electricity efficiently, and integrating the distributed energy resources as part of the main grid supply. Hence

S. Samanta (✉)
A&I Business Analytics, Tata Consultancy Services, Kolkata, India
e-mail: sugato.samanta@tcs.com

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_3

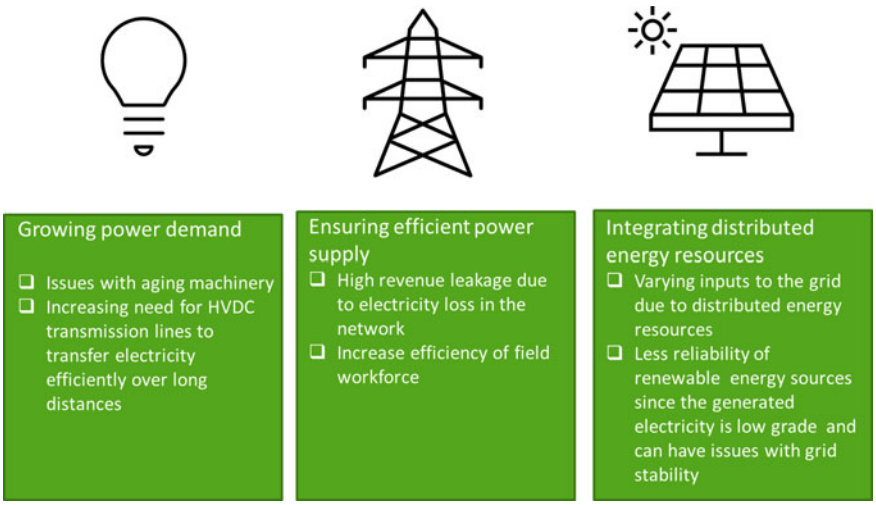


Fig. 1 Challenges faced by electricity grids

the organizations must continuously innovate to keep up with the disruptions in the energy marketplace.

The three trends of decarbonization, digitization, and decentralization are each presenting new challenges for distribution companies, whose revenue plans have long been based on simply selling more kilowatt-hours. As per the Energy Data Taskforce report [1] in the UK, related to the digitization of the energy network, the fundamental pillars of delivering a digital energy network are: Promoting Data Transparency, Improving Network Operations and Innovation, and Open Market and Regulations.

The regulators across the globe are bringing in innovation to the electricity market by introducing new entrants related to renewables along with the conventional organizations to improve grid connectivity and lower carbon footprint. These "Digital disruptors" who are aggregating solar, storage, and other resources located on customer sites to participate in energy auctions. This is exemplified by companies like Stem in the United States and Sonnen in Germany, which are both using cloud-based software applications to manage customer interactions within utility markets and between each other. In these cases, the distribution utility grid is providing the necessary platform for these transactions.

Similarly, various Oil & Gas companies like Shell, bp, ExxonMobil, etc. who had erstwhile invested in renewables like offshore wind farms or large-scale solar farms to offset their carbon footprint, are now more aggressive to include the generation of clean energy as part of their energy business portfolio. This has been largely driven due to the Covid-19 pandemic wherein it is now clear that the oil & gas companies will change themselves to become energy companies to run a sustainable business.

Smart grids [2] use advanced sensing, communication, and control technologies, to monitor and manage the transport of electricity from different sources to

meet the varying demands of end-users. Despite these advancements, the emergence of Distributed Energy Resources (DERs) driven by environment-friendly customer choices, regulatory pressures, and increasing energy prices are impacting supply and demand scenarios. While distributed generation was providing 87,300 megawatts (MW) of power per year in 2014, it is projected to grow to 165,500 MW by 2023, according to the Global Distributed Generation Deployment Forecast published by Navigant Research [3]. Here as shown in Fig. 2, analytics-driven by data and algorithms are much more efficient in understanding the customer generation and consumption patterns, futuristic load forecasts, and predict equipment failures to drive lower operations costs.

DERs are small-scale power generation sources which include solar photovoltaics, energy storage, Electric Vehicles (EVs) and charging infrastructure, and Combined Heat and Power (CHP) [4]. By leveraging distributed sources, consumers who have become prosumers can export the generated electricity back to the grid or consume that electricity based on their needs, thereby reducing the reliance on conventional sources of electricity.

The generated electricity can also be stored during the day and be used during peak hours. However, this depends on the pricing provided by the utility supplier in that region. If the per unit selling price is substantially more than the per unit price during the peak period, exporting electricity is more beneficial to the consumer. However, if the selling price is not significantly higher compared to per unit prices, consumers benefit by using the stored electricity. Due to the dependency on pricing incentives, conventional load forecasting methods are unable to forecast the load or supply at the grid accurately.

For instance, load forecasting becomes difficult with the growing adoption of plug-in electric vehicles and plug-in hybrid electric vehicles. This is because multiple vehicles recharge from recharging points in the same location, causing a spike in the load at that node or zone. If this load is not constrained, it can lead to a brownout or blackout in that zone. Hence, managing peak demand is key to better planning throughout the electricity system. Electric utilities have invested in smart meters and consumer education yet struggle to manage peak loads. Grid optimization techniques can be used to balance supply and demand at grid levels in near real time, while optimizing transmission costs based on network constraints.

Distribution Network Operators (DNO) can no longer play the only role of transferring electricity from the National Grid of a country to the end B2B (Industrial & Commercial) and residential consumers and maintain the distribution network. Since the flow of electricity has reversed to account for household generation at residential properties, DNOs must take into consideration the energy generated at the community level and ensure that the grid frequency and reactive power are within the threshold limit. Hence, they will have to elevate their role to convert themselves to a digital business driven by data. As such data fundamentals like sharing of knowledge & data, data catalog, data reliability, and thereby data governance are key drivers for the DNOs to become Distributed System Operators as shown in Fig. 3.

In the present scenario T&D companies rely on network balancing systems to determine the impact of congestion (Defn—Congestion is a state of the electrical

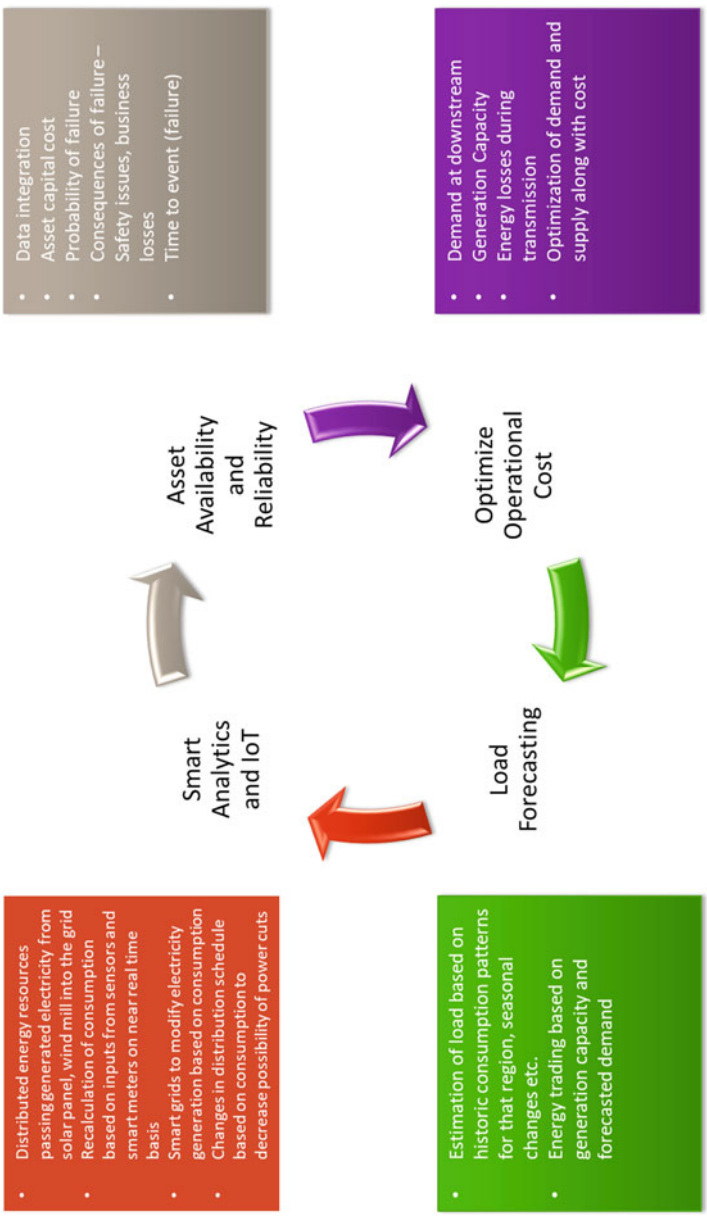


Fig. 2 Why analytics is crucial for transmission & distribution

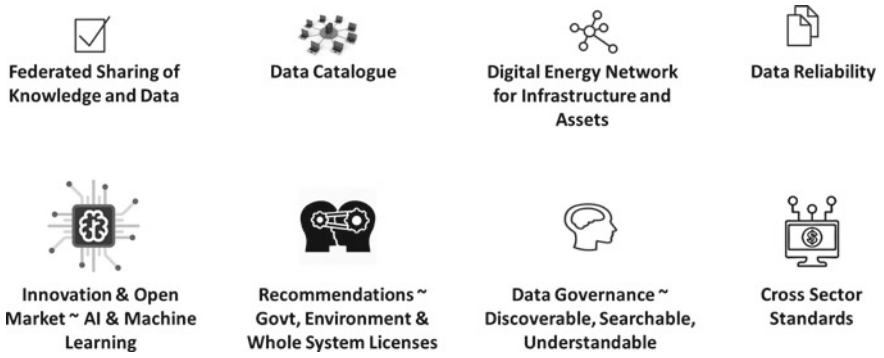


Fig. 3 Key transformation drivers for distributed network operators

network when transmission constraints impact the capability to transfer electricity from low-cost generation sites to the demand sites resulting in meeting the demand from higher cost electricity generation sites). This is an important parameter primarily in energy trading activities since the relevant input and output of the electrical grid needs to be balanced as there is no mechanism for storing electricity. Below (Fig. 4) is a sample representation of how the ISO (independent system operator) balances flow during energy trading sessions to meet the demand at the end user and supply at the generation level [5].

To supplement the above scenario, DERs pose to a threat to the grid since they can transfer electricity at any point in time resulting in a situation of having microgrids to tackle the problem via network optimization. To develop a network optimization solution, we need to first formulate the problem statement encompassing grid-specific parameters and constraints. The first step is to determine the minimum cost of the grid during transmission and distribution while considering the impact of DERs once they start supplying electricity back to the grid. There can be two possible scenarios—(a) DERs export the generated electricity back into the grid or (b) DERs consume less electricity from the grid. In either scenario, the supply and demand parameters can be changed based on the electricity needs in the overall grid in real-time, using digital technologies such as load balancing. Initially load balancing happens at the microgrid level and is then extended to the entire network.

The parameters measured in Mega Watts (MW) include electricity supply from the generation sites, electricity demand at the load sites, transmission losses, and electricity generated by DERs. The transmission constraints comprise maximum and minimum transmission capacity in MW.

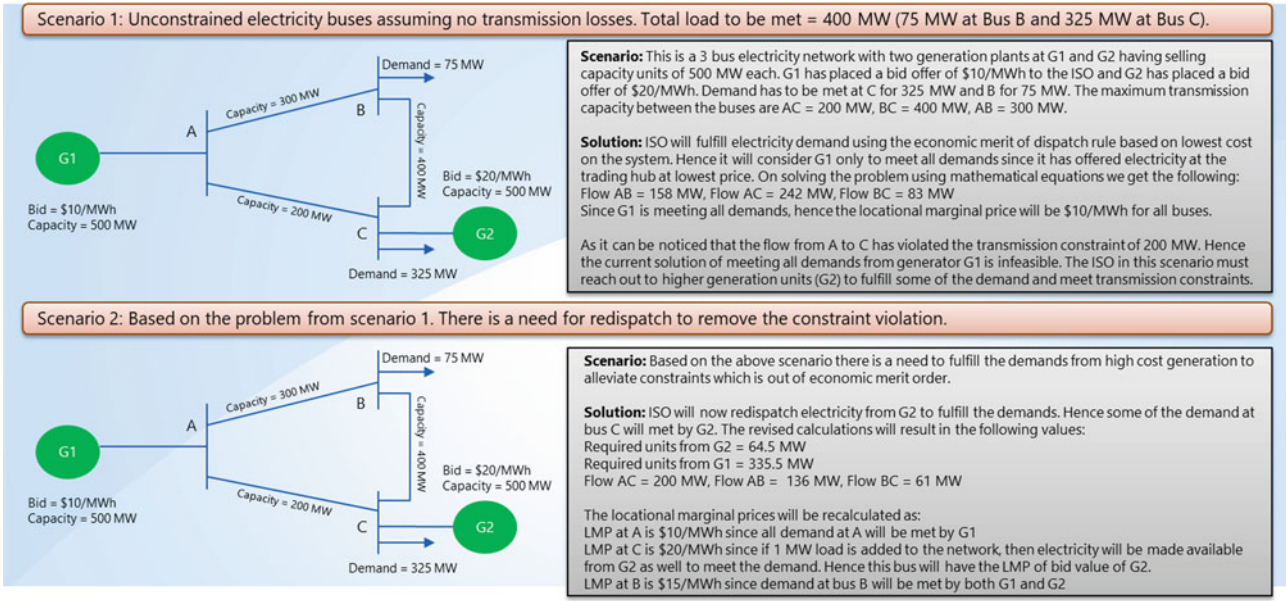


Fig. 4 Illustration of grid dispatch based on congestion

2 Literature Review

Due to Paris Treaty on climate change, there has been a disruption in the utility industry where the organizations must set goals related to increase of the footprint of renewable energy in their overall generation mix to continue with their license to operate. Additionally, the growing population of Millennials and the younger generation as mainstream residential homeowners are more conscious of the energy source and topics like global temperature control leading to the increasing use of rooftop solar panels. Hence there is a possibility of an impact on the conventional grid which has led to research in grid modernization and the impact of Government regulations on the utility companies. In a fifty-six-page document, the Energy Task force for the UK has laid down clear guidelines which the utility organizations must follow. Energy Data Task Force along with Offgem UK has laid guidelines for the future energy marketplace [pg. 16–32] [1]. Like the utility organizations, the grid is also undergoing infrastructure changes to accommodate interdependencies between the utility companies and the end users. This is clearly depicted in The Future of the Grid Evolving to Meet America's Needs [pg. 9] [2]. With the sustained growth of the renewable energy mix in the overall generation portfolio in most of the countries as per Navigant Research, the utility organizations must prepare themselves for the disruption [3]. The change is primarily in the areas of load transfer and balancing [pg. 7–9] of Preparing for Distributed Energy Resources thought leadership paper [4]. This disruption in load distribution has a major effect on the ISO in terms of scheduling distribution based on price and congestion parameters to the required areas (refer to dispatch fundamentals of NYISO as an example) [5]. The other major problem posed by DERs is the inability to forecast exact generation due to fluctuating weather conditions as mentioned in Integrating Renewable Electricity on the Grid by APS Panel [pg. 10–12] [6]. Hence the need of the hour to improve the transmission and distribution infrastructure which requires an overall modified distribution network architecture as demonstrated by UKPN in their future strategy document DNO tools and systems learning [pg. 39–40] [7]. Similarly, there have been huge investments in grid modernization initiatives and energy credit systems across various countries to stimulate house owners to move to a cleaner energy source which can be observed as part of the overall strategy by UKPN in the document titled Towards a net zero energy future [pg. 10–14] [8].

3 Methodology

Utilities (electricity or water) have an interconnected network for electricity transmission or water distribution from generation to consumption as shown in Fig. 5. Normally, the generation unit is responsible to produce electricity from sources like thermal, nuclear, hydel, or renewable energy sources wherein generation must be the balanced basis on the demand further downstream. This is critical since electricity

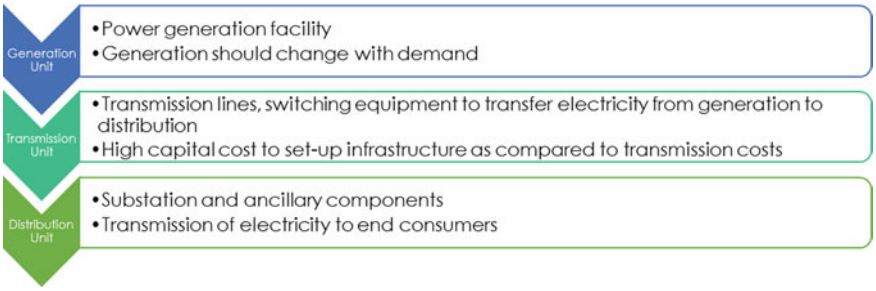


Fig. 5 Analytics interventions across the electricity value chain

cannot be stored on a mass scale and hence whatever is produced must be consumed. The transmission unit consists of high voltage electricity lines, primary transformers, etc. to transmit the electricity. Finally, the distribution unit is where the consumption happens, and this controls the flow of electricity to our homes and offices.

With the advent of distributed energy resources (DER) the electric grid is undergoing a behavioral change where end consumers are generating electricity from renewable sources—wind, solar, and agricultural biomass and send excess electricity back to the grid (consumer to prosumer) for credit. Energy utilities must adjust with changing times and focus on the reliability of the grid since the quality of electricity generated can de-stabilize the grid due to over-voltage, under-voltage, frequency fluctuations, etc. Predictive analysis of the expected participation levels of a prosumer will be useful to support distribution operations by temporarily switching a subset of customers to that prosumer during a planned outage or to manage the peak load expected based on a weather forecast [6]. Similarly, an analysis of the capacity level for each demand will support the utility's ability for energy trading in the wholesale market. Below is an overall depiction of the flow of electricity through the electric grid.

Electricity demand is the highest in urban and suburban areas where renewables from remote sources can help to meet this demand without increasing carbon emissions. However, the additional power currently must be distributed over infrastructure designed and installed to meet much smaller needs. Congestion on existing lines inhibits growth, and as urban areas expand and merge, the area over which power distribution needs to be coordinated growth.

Since the majority of the DERs that can produce more electricity which can be used to meet demand at load sites are located away from congested urban areas, there is a necessity to lay down new transmission lines. These lines would carry the generated electricity from the renewables like wind or concentrating solar power (solar thermal energy driving conventional generators) to the demand sites. However costly conversion in form of AC-DC-AC would still create problems for the electrical grid.

Advanced smart grid software designed to support DER management and optimize grid operations and planning are being tested. It works with a real-time network

model, based on an accurate geo-database and incorporating data from operational systems such as a supervisory control and data acquisition (SCADA) system and outage management system (OMS) to determine the DERs which are causing problems to the grid in terms of voltage or harmonics levels and cut off from the grid.

3.1 Defining a Network Optimization Solution to Build an Agile Grid

Electrical grid optimization refers to the determination of the necessary network requirements by maintaining the supply of required electricity to the demand sites at the least cost. It is mainly related to the transmission and distribution side of the value chain. In the last decade electricity markets have moved from regulation to de-regulation enabling the trading of electricity as a commodity. Power transmission and distribution are considered as services which are remunerated by the users of the corresponding systems. Hence transmission systems need to be transparent and hence under the responsibility of transmission system operators (TSO).

The relevant input and output of the electrical grid needs to be balanced since there is no storage mechanism of electricity available till now. Although many utility companies are conducting pilots for electricity storage batteries, however the feasibility of the same is not yet confirmed.

There would be transmission losses across the grid since transmission of electricity across the overhead transmission wires, circuit breakers, transformers, etc. does heat up the lines and devices resulting in losses. Long distance transmission using HVDC (high voltage direct current) results in lesser losses as compared to AC transmission; however, it involves costly AC-DC-AC conversion.

The business model for TSO's is governed by regulations, hence information flow and exchange must adhere to GDPR, and other regulations as set down by the Government. In terms of information exchange, there are a few areas which are of importance: electricity handover points/dispatch points from one operator to another and supplier switches. Similarly, from the DNO perspective information related to asset management and outage prevention strategy, growing customer interactions with the service provider, situational awareness for grid balancing, and distributed energy resources marketplaces will generate huge volumes of data which can act be the baseline of a data hub if shared between the system operators.

The building block of the data hub would be the variables [7] captured by the DNOs (Fig. 6) on an everyday basis, where some of them can be shared between operators. Most of the variables related to distributed generation like DER generation capacity and values, electric vehicle charging load, generation mix at LV in terms of solar, wind, etc. are captured by the distribution companies, but not always shared as a dedicated data exchange API with the transmission operator and other DNO's resulting in situational outages (brown cuts) and reactive power at the supply points.

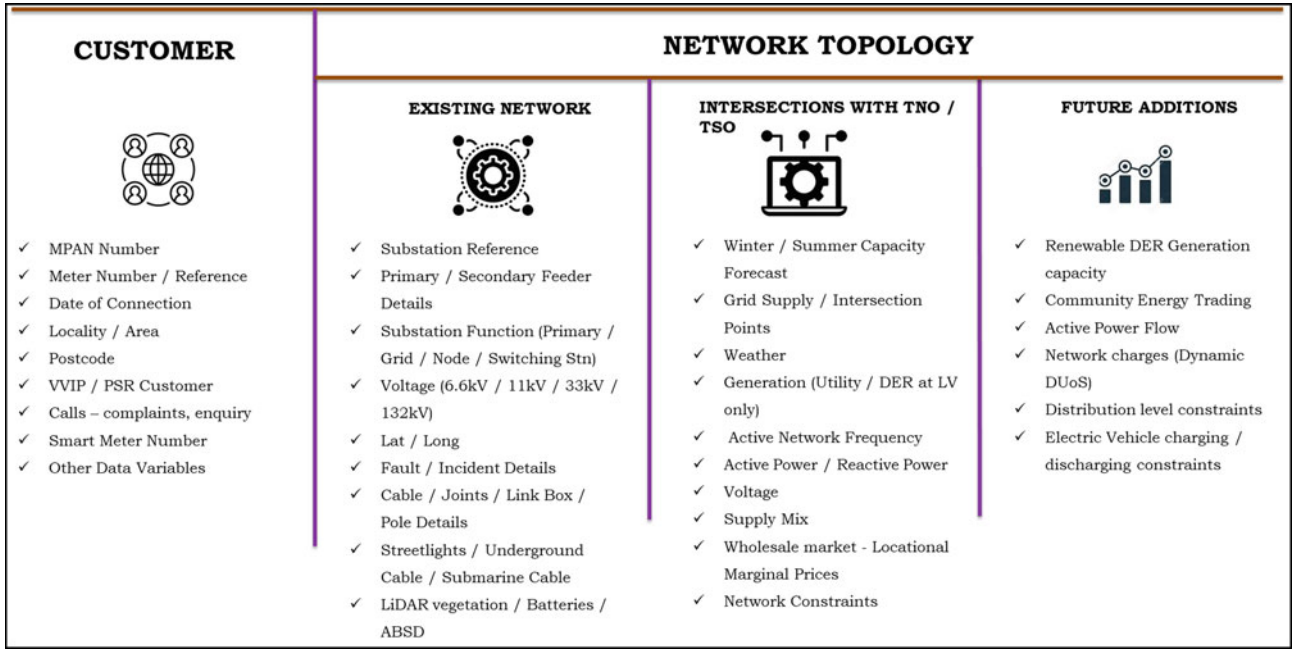


Fig. 6 Data variables handled by a DNO

To move towards delivering a digital energy network (Fig. 7), DNOs must undergo a digital transformation programme and change the way they address their strategy and daily operations. The key tenants will revolve around the following: Data Transparency, Digitizing the Energy Network, Technology Transformation, Transforming Customer Experience, Using Data & Analytics to take business decisions, and creation of the Digital Workforce.

To digitize the network, it is important to know the power flow in real-time across the network from generation to end consumption which is a challenge with DER's ability to produce their electricity. This combined with concepts of Active Network Management, Dynamic Time of Use, dynamic load profiles, and forecasting is crucial to maintain the balance of the grid [8]. As the energy ecosystem is going through an evolutionary phase with the advent of IoT-enabled smart grid, smart meters, and electric vehicles coupled with the inception of the micro grid, paving the genesis of the marketplace for data and analytics to drive decisions beneficial for the entire ecosystem.

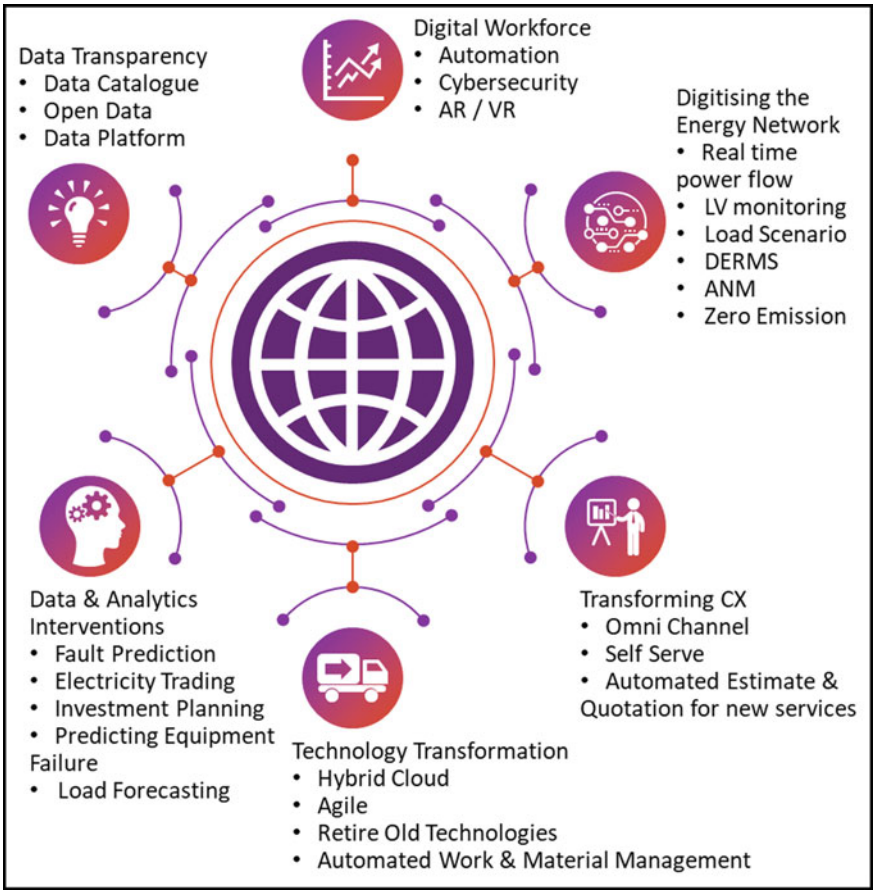


Fig. 7 Key drivers to deliver a digital energy network

3.2 Defining the Problem

Conventional load forecasting algorithms are unable to provide the level of information required to balance supply or electricity generation with the demand or load. Even network stabilization software such as Active Network Management (ANM), which can determine which DERs are causing grid de-stabilization in real time, are not capable of determining the flow and its consequent impact on the grid.

The algorithm discussed in this section offers a grid optimization solution that can help maintain supply and demand in the overall grid by considering the impact of DERs. At the same time, it optimizes the cost of transmission while maintaining transmission constraints.

Let us consider a scenario based on a 'minimum cost flow problem'. The scenario starts with the intermediate network points or the micro-grid. In the case of the overall electricity grid, it starts from the generation point.

The parameters would constitute Electricity supply from the generation sites in MW (Mega Watt), Electricity demand at the load sites in MW, Transmission losses in MW, and Electricity generated by the Distributed Energy Resources in MW.

The transmission constraints would constitute Maximum transmission capacity in MW, Minimum transmission capacity in MW.

Using the above information, the objective function needs to be defined. This would be in the form of an equation.

F(x) = min (cost of electricity supply).

Based on Supply = Demand + Transmission Losses and Transmission constraints being honored.

3.3 Defining a Solution for the Problem

This scenario is based on the "Minimum Cost Flow Problem" which deals with finding the set of arc flows to minimize a linear cost function subject to the constraints. The function can be written as

$$Minimize \sum_{(i,\ j)} a_{ij} x_{ij}$$

Subject to the constraints

$$\sum x_{ij} - \sum a_{ij} x_{ij} = s_i$$

$\mathbf{b_{ij} \leq x_{ij} \leq c_{ij}}$

where a_{ij}, bij, cij and s_{ij} are scalars with the following terminology.

a_{ij} is the cost of (i, j),

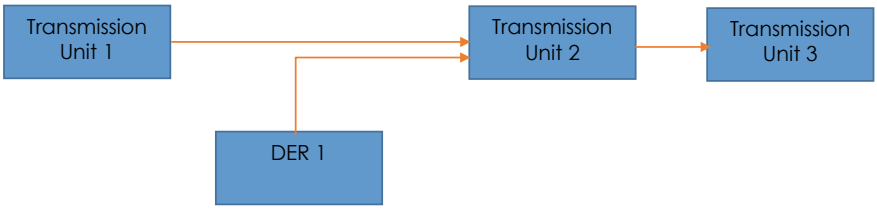


Fig. 8 Sample 3-bus network

b_{ij} and c_{ij} are the flow bounds of (i, j),

[b_{ij}, cij] is the feasible flow range of (i, j),

s_i is the supply of node i (when s_i is negative then it is the demand of i).

The flow vector satisfying all constraints is deemed to be a feasible solution otherwise it is termed as infeasible. This mathematical programming model is known as a linear program. However linear program only holds good if the solution determined by the equations is feasible. In such scenarios where the result is an infeasible one, we will move to another mathematical model called goal programming.

The programming technique can also be applied to a sample three-bus electrical network (Fig. 8) which is a subset of the entire electrical grid. In case of an actual network, it will start from the generation point to the final load point. The entire network model (schematic diagram) must be made available to have all the network participants in place.

The sample network consists of three transmission units and one DER, alternatively these transmission units can also be replaced by distribution units.

Transmission units 1, 2, and 3 are responsible for passing/transferring electricity from the generation plant to the demand sites. Transmission happens at different voltage levels like 33 kV or 11 kV. The DERs who are prosumers can supply their excess electricity back to the grid at these voltage levels. It has been assumed that the quality of electricity will be at par with conventional sources with respect to voltage levels, harmonics, and frequencies and they are fit to transfer electricity to the grid. This excess electricity from prosumers combined with conventional electricity sources can be used to meet the load at the demand sites. In the above network model, there would be transmission constraints between transmission unit 1 and transmission unit 2 and between transmission unit 2 and transmission unit 3. To formulate the problem scenario let us consider the following conditions:

Cost of supplying electricity from transmission unit 1 (TU1) to transmission unit 2 (TU2) is $ x/MW.

Cost of supplying electricity from transmission unit 2 (TU2) to transmission unit 3 (TU3) is $ y/MW.

Demand at load sites nearest to transmission unit 3 (TU3) is z MW.

Maximum flow of electricity between TU1 and TU2 is a MW.

Maximum flow of electricity between TU2 and TU3 is b MW.

Transmission losses is 5% of the flow from TU1 to TU2.

Transmission losses is 6% of the flow from TU2 to TU3.

Based on the above conditions we will start building the equations to have constraints, functions, and others in place.

Objective function

min: + x*(Flow TU1_TU2) + y*(Flow TU2_TU3).

Flow constraints upper bound equations

Flow TU1_TU2 <= a.

Flow TU2_TU3 <= b.

Flow constraints lower bound equations

Flow TU1_TU2 >= 0.

Flow TU2_TU3 >= 0.

Bus/Zone level constraints

Flow TU2_TU3 <= z.

Balancing equations

Flow TU1_TU2 − 0.05*Flow TU1_TU2 + Supply DER1 >= Flow TU2_TU3.

Flow TU2_TU3 − 0.06*Flow TU2_TU3 >= Demand T3.

Execution of the above equations in an LP solver will result in the following: (1) Flow values of electricity between transmission unit 1 and transmission unit 2, (2) Flow values of electricity between transmission unit 2 and transmission unit 3, and (3) understand whether the linear program was able to meet the constraints and provide a feasible solution. If the solution is infeasible then the equations must be written again using goal programming standards to incorporate the deviation values along with the actual values.

This would be developed using Operation Research (OR) technique using Mixed Integer Linear Programming/Constraint Programming. However, if the linear programming technique does not result in a feasible solution, then other techniques like goal programming or genetic algorithm must be leveraged.

Goal programming has been considered because it is a technique designed to solve problems in which there is more than one objective/multiple goals (Multi-criteria Decision Problems). The objective function in this case is to minimize the deviational variables unlike linear programming which works only with the actual variables.

In case of goal programming, there are two priorities in terms of meeting supply/demand constraints and lowering of costs. The objective function will be as Minimize total deviation = $d_1^- + d_2^- + d_3^+ + d_4^-$

Step 1—The objective function is the minimization of deviation

The deviation is measured as negative (under achievement of the target) and positive (over achievement of the target).

'v' is measured as the negative deviation of demand/supply requirements.

'u' is measured as the positive deviation of time taken for of assignment.

Step 2—Calculate the overall optimized costs

The objective is to minimize negative deviation (underachievement) of 'v' (thereby meet the demand or not overshooting the demand) and minimize positive deviation (overachievement) of 'u' (thereby containment of cost). In simple terms if 'v' is positive then demand is not achieved and if 'u' is positive then the cost has exceeded the optimal value.

Since demand is marked as a priority hence the priority or weight is high for 'v' (hence the multiplication factor of 2 in the objective minimization).

/* Multi Objective Function */

/* Minimize negative deviation (v's) for Demand and positive deviation (u) for time */

Objective Function: min(2v + u).

/* Constraints */

Dispatchers will be able to know if they have been able to meet their primary and secondary goals which can be represented by the demand and supply at various nodes of the electricity network. If not, then by how much they have missed their goals or the deviation from optimal values.

4 Results

Utility companies are investing in innovation and technology to meet the needs of the next generation consumer and changing environment. Currently, the eco-system of utility companies consisting of Generation, Transmission, and Distribution companies are regulated by Independent System Operators or Energy commissions, hence the supply/demand aspects are balanced. However, with the advent of concepts like DER, peer-to-peer trading, and micro-grids, the prosumers are playing an active role in the system thereby limiting the dependency on the overall grid and thereby the conventional utility company. Due to this evolution, utility organizations are facing challenges of how to incorporate them into the grid without causing imbalances. Network stabilization software like Active Network Management is useful to determine the DERs which are causing de-stabilization to the grid in real-time but are not capable of determining the flow and its consequent impact on the grid. This is where network optimization will come into the picture and help T&D utilities to determine demand and supply efficiently.

Network optimization is especially important in the transmission and distribution business chain. This can also be implemented with some changes to water distribution. The main crux of the business problem is to maintain proper demand and supply of energy or water in the network and understand supply deficiency if any. To deliver the digital energy network, transmission, and distribution network operators must work closely in collaboration with the ecosystem of local communities and new market entrants to have a seamless flow of electricity across the grid.

5 Conclusion

Cost-effective DERs are expected to displace 320 GW of centralized generation from 2014–2023. The grid optimization solution will enable utilities to understand the overall supply and demand at the grid at any point in time and optimize generation accordingly. The utility supplier can also submit efficient sell or purchase bids with lower hit margins. By analyzing electricity flow from DERs at the micro-grid level, trading algorithms can be used to optimize the grid with respect to supply and demand at a minimal cost for transporting electricity. This can also be used to route appropriate amounts of electricity to different regions, as needed, in near real-time. Until a fully automated smart grid is implemented, the grid optimization solution will be critical to improving grid stability and reliability and energy trading margins, as well as reducing overall operating expenses.

References

1. A strategy for a modern digitalised energy system by energy data taskforce UK (2021):16–32. https://es.catapult.org.uk/report/energy-data-taskforce-report/#:~:text=The%20Energy%20Data%20Taskforce%20identified,gaps%20and%20maximise%20data%20value%3A&text=Open%20Markets%3A%20Achieving%20much%20better,location%20and%20service%20value%20data
2. The future of the grid evolving to meet America's needs—final report: an industry-driven vision of the 2030 grid and recommendations for a path forward (2003):GS-10F-0103J:9. https://www.smartgrid.gov/document/future_grid_evolving_meet_americas_needs_final_report_industry_driven_vision_2030_grid_and.html
3. Navigant, The annual installed capacity of distributed generation is expected to double by 2023 (2014). https://www.utilitydive.com/news/navigant-distributed-generation-will-nearly-double-by-2023/341619/
4. Preparing for distributed energy resources by Schneider electric (2012):998-2095-05-29-12AR0_EN:7-9. https://www.se.com/in/en/download/document/998-2095-05-29-12AR0_EN/
5. Power trends 2014 by NYISO (2014):39–44
6. Integrating renewable electricity on the grid by APS panel (2007):10–12
7. DNO tools and systems learning by UKPN 2014 (2014):D2:39–40
8. Towards a net zero energy future by UKPN 2021 (2021):10–14

Chapter 4
Enhancing Market Agility Through Accurate Price Indicators Using Contextualized Data Analytics

Surekha Deshmukh and Nagalakshmi Subramanian

Abstract The volatility of the price of power being an inherent characteristic of the Electricity Trading Market, Utilities need to address this volatility strategically influencing key operational decisions in the day ahead, spot market, wholesale market, etc. The latest experience of Power Utility Ecosystem involved in the trading business while dealing with market-driven uncertainties, face major challenges introduced due to adaptation of distributed energy recourses (DER) at a significant percentage of penetration into grid. Utilities across the globe, while embarking on the green energy path, rely on the advanced digital enablers in increasing the visibility of DER penetration in achieving the Net Zero Carbon goal. The key influencers to price volatility being random, unforeseen, the accurate price-forecasting is always a very important differentiator for Utilities in taking up market-driven decisions, balancing the grid, and maximizing portfolio expectations. Electricity Market Operators and Electricity eco-system across the globe, face challenges in view of inaccurate predictive models, creating an impact on market participation, frequency control, dispatch strategies, scheduling downstream decisions, negative prices, and other techno-commercial implications. The potential of data analytics tools can be acknowledged to assess and harmonize data of load demand, generation availability, infrastructure constraints, power flow constraints, market constraints, double sides or single-sided auction bids, etc. The data-driven insights would enable Utility to unlock the market corridors, build flexibility of participation, accommodating a greater number of business players, increasing market agility, and best cost of power to the end customer. This chapter aims to present the global challenges of Power Utility belonging to electricity trading business in developed countries like USA, Europe, Australia, UK geo and how the blend of the tool with data and AI can empower the decisions of utility on keeping tradeoff between the profit, revenue, and risk in view of the supply chain, changing generation mix, consumption and load patterns, regulatory compliances and innovation and incentivization at customer engagements at the downstream of the Utility value chain.

S. Deshmukh (✉) · N. Subramanian
Analytics & Insights, Tata Consultancy Services, Pune, India
e-mail: surekha.deshmukh@tcs.com

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_4

Keywords Contextualized data analytics · Price indicators · Market agility · Analytics · Infrastructure · Digitization

1 Introduction

The electricity Utility industry across the globe is undergoing reform in terms of **Digitalization, Decarbonization, and Deregulation**. With strategic government initiatives, regulatory policies, and technology adoption, Utilities are being able to embark on the journey of creating value to the potential of data flow, with increased competition and flexible innovative opportunities. This transformative journey has enabled greater visibility, greater accountability, and greater flexibility to innovate, operate, and gain competitive advantage.

A. Digitalization

Having stated as a legacy industry, over last two decades the Utility industry has transformed into smarter enterprise, well equipped with digital interventions and tools and technology-adoption for the entire value chain and all sectors of business. Not only grid infrastructure is becoming smart, but also Utilities have started leveraging the potential of access and usage of data from various sources, facilitating up-stream, downstream coordination, and integration in the best way, optimizing operations, and maximizing revenue. Globally more than 60% of Advanced Metering Infrastructure (AMI) is already penetrated at junctions of the grid, empowering the two-way DATA flow in real-time, with secure accessibility since more than 1 decade. [1, 2]

With the trend of digital transformation and wider embodiment of Supervisory Control and data Acquisition (SCADA), Phasor Measurement Unit (PMU), Sensors, IOT, RTU, Relays, Drones, etc., the Utility and Energy Analytics Business is expected to grow with CAGR of 24.9% till 2026, as a span of a forecast [3].

B. Decarbonization

Green Energy Initiative, being one of the Sustainable Development Goals (SDG), Utilities across the globe, aim to invest largely in the renewable energy projects as Distributed Energy Resources (DER) along with possible generation mix portfolio and contribute to reduce the greenhouse gas emission and achieve ZERO Carbon Dream [1]. Around 50% of low carbon generation is expected from DER sources by 2050 [1]. The global annual expected investment in low carbon generation is expected to grow by 64.86% and the rest of the digitization of the grid by 50% as per Suitability Development scenario by 2030 [1, 2].

This opportunity of green energy generation has introduced operational challenges to Utilities to handle the uncertainties involved. The weather-specificity introduces uncertainty as the intrinsic nature of DATA of renewable energy sources as wind speed, solar insolation, power availability, etc. Utilities are seen to be investing in analytics solutions in building strong forecasting tools for accurate load management, uncertainty management, and generation-mix as well as storage optimization. The prosumer analytics is best suited for newer revenue streams [4, 5].

Today E-mobility is the modern, popular, and commercially viable choice of commute over petrol-diesel-based locomotives. It is anticipated that the Electric-vehicle number on road will get doubled in coming decade [1]. This fast-growing segment also provides stakeholders the vital information regarding number and type of vehicle, charging pattern, frequency of charging, electricity-consumption, along with system integration data, power transacted to and from the grid, etc. With the research and deployment of advanced battery technology, the market of Electric Vehicle is aggressively progressing, maintaining affordability [4, 5].

There is a huge scope of contextualized analytics in the EV domain in properly integrating diverse data useful for stakeholders as vehicle owners, OEMs, Vendors, Distribution Utilities, Retailers, City officials, etc. This segment has opened innovative business opportunities and revenue models for stakeholders.

C. **Deregulation**

This step has provided operational autonomy, accountability, authorization, and authentication to Utilities to focus on maximizing operational aspects in terms of improved accuracy, efficiency, reliability, quality, and continuity of supply chain, with the trade off between ROI and risk.

The establishment of the Power Trading Market is the ramification of the restructuring and deregulatory policies adapted globally to accelerate the business in the Energy Sector, trying to nullify the slothful economic impacts of the monopolistic power sectors.

Electricity as a traded commodity exhibits dynamic characteristics of volatility, uncertainty, and risk as a function of market-driven gaming, congestion. Utilities as well as market operators prefer to build accurate predictive models to address the challenges of market balancing as well as economic viability. Knowing the accurate price of power is the key requirement of market players to participate in short-term, long-term market, bidding strategy, and risk assessment as well as management. This chapter orchestrates the challenges as well as the potential of data propensity in enhancing the accuracy of the model in predicting the price of power by exploring the price of power of the market.

2 Literature Review

The Electricity industry ecosystem handles the humongous amount of data, originating from different sources across the value chain, with diverse specifications. The business stakeholders unlock the potential of data in enhancing visibility. Along with levering the data, applying data mining techniques to develop Artificial Intelligent (AI), ML algorithms have become an imperative practice in achieving business KPIs, ensuring operational excellence, and techno-commercial benefiting along with safety, reliability, and resiliency of the grid. Has highlighted the techniques of data mining along with key challenges in terms of handling the sufficiency, correctness, missing data, etc. and approaches to improve. Have explained the use of clustering

techniques such as K-Means, EBK Means, and important details of deciding the number of clusters to begin with, also applying the Elbow method to decide the required number of clusters for a given set of samples. Has a coverage of significant number of literatures in demonstrating the use of data mining tools for electricity-related time series parameters, mainly electricity demand. Have applied data mining tools for the electricity market to forecast the day ahead price of power and effectively handle uncertainly randomness in the price variation.

The electricity market deals with complex variations in load demand in real-time. While fulfilling the load demand in real-time, there are chances of congestion of transmission corridors, while fulfilling generation-load balance. Hence congestion management becomes a critical task for a market operator to ensure continuity of market operations with minimum impact on the price of power. In papers, authors have provided a detailed approach of congestion management, along with techniques to forecast short-term transmission congestion. The accurate forecasting of the congestion event along with the time of event would enable the actions of market splitting, etc.

3 Data-Flow in Utility Value Chain

The Utility Value chain is multifaceted, engaging diverse energy resources, assets, processes, practices, regulatory compliances, revenue models, and people, with interesting schemes of rebates and incentives to keep up the momentum of ensuring stable, secure, and reliable operation of power Grid. The performance index of any Utility gets governed by Customer experience as the catalyst of business growth.

The utility industry has Generation, Transmission, Distribution, Retails, Aggregators, Suppliers, Customers and Prosumers, Market-operators, Power-traders, Network and system operators, Regulatory agencies, Investors, etc. are very important stakeholders as depicted in Fig. 1. The energy ecosystem generates diverse data of diverse nature, at diverse locations, in real-time spanning end-to-end value chain. Different stakeholders need such data to make real-time, short-term, and long-term decisions, that are techno-commercially viable.

To provide real-time operational decisions for secure and stable power flow, Utilities seek data as Voltage, Current, Active Power flow, Reactive Power flow, Frequency, Load-demand, Available-generation, etc. in real-time. Along with these

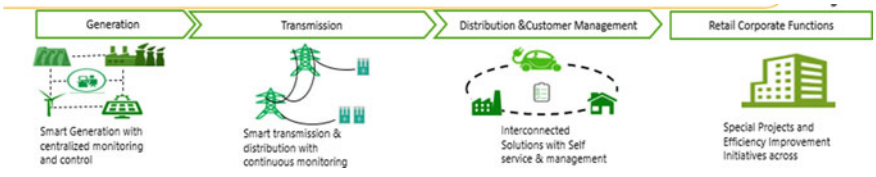


Fig. 1 Representative utility value chain

fundamental quantitative attributes, the qualitative attributes such as interruptions, harmonics, flickers, surges, and Power Quality (PQ) events are the essential DATA to assess the performance of the system and project the action plan and investments.

To increase the operational lifetime of network-assets, Utilities depend on data such as normalcy or anomalous characteristics, aging, and deterioration. This data has enormous potential of revealing the context to deep dive and draw insights on asset health, aging, risk of failure, criticality, restoration complexities, etc., to improve the life of assets and network along with optimizing the time, cost, and efforts [4, 5]. Such a strategic look-up approach supports the Utilities to ensure the improving maintenance metrics as minimum Downtime, Mean Time to Repair (MTTR), Minimum time to fail, (MTTF), and Reliability indices as business KPIs.

The customer is the vital stakeholder of the Utility Value Chain, having wider capacities to influence the business opportunities as well as revenue flow. The active engagement of customers in system balancing along with trading of power, under the schemes of rebates and incentives has changed the competitive appetite of distribution and retailers in the energy business. Utilities are proactive in leveraging the Data-driven insights for the best resource management through Demand Response programs, Prosumer-initiatives, Peer-to-Peer trading opportunities, etc. The increased customer satisfaction, customer-visibility toward own consumption behavior, the strategy of minimizing the bills, and energy conservation practices is very well capitalized globally [6].

4 Handaling Market data Volatility and Coherency

The economics of power generation, transmission, and distribution of the power to the end customer with reliability and quality is the prime responsibility of the Utility value chain along with market operators, independent system operators, and regulatory commissions.

The focus of Utilities is to encourage competition among power providers. Usually, electricity is traded in two ways: (i) bilateral contracts and (ii) pool or exchange trading. The hourly bids submitted by producers and consumers are matched by the market operator to set the spot price in the pool trading market.

In the bilateral contract system, a certain amount of power is agreed to be transferred through the network between seller and buyer at a specific fixed price.

The price of power is volatile and depends on various factors, including demand variation, generation variations, fuel price volatilities, fuel constraints, the volume of market participation, renewable energy participation, etc. Being a time series parameter, the magnitude of price of power is a function of time of day, type of day, type of month, weather, and seasonal variations. Due to its highly random, volatile nature, the handling market data is always a challenge for market players.

Table 1 provides the sample of structured, unstructured, time series as well as stochastic data, with variety and volume, in the space of the Electricity Market, to

Table 1 Sample electricity market data

Price data	Day-ahead and real-time market location based marginal price, time-weighted/integrated real-time LBMP, balancing market (Hour-Ahead) advisory prices, ancillary service prices, reference bus LBMP, price correction logs
Power grid data	Outages and outage schedule, constraints, interface flows
Load data	Load forecast, zonal load commitment, real-time actual and integrated real-time actual load, load and security constraint, unconstraint forecast data—monthly data postings, and current hourly loads
Reports	Daily energy report, NYISO capacity report, real-time events, generator, load and TO information, operational announcements, dispatcher notices, bid data, operating studies, system reliability & impact studies, generator and load names, subzone information, balancing market advisory summary, major emergency events, generation economic dispatch analysis, notice of discretionary
Time series graphs	Day-ahead, hourly advisory and real-time zonal LBMP, zonal actual versus forecast load, interface flows, and zonal load versus LBMP
Zone maps	Day-ahead, hourly advisory and real-time LBMP, and load flows
Market access	Bidding and scheduling, settlement data exchange

be maintained, accessed, and utilized. Based on the geo-specific regulations and power-grid structure, the details may vary.

- For illustration purposes, here are a few data variables like—LBMP ($/MWh), Marginal Cost of Losses ($/MWh), Marginal Cost of Congestion ($/MWh), Regulation ($/MWh), 10-Min Spinning Reserve ($/MWh), 10-Min Non-synchronous Reserve ($/MWh), 30-Min Operating Reserve ($/MWh).

The huge database of statistically distributed nature is the key characteristic of the price of power. Hence based on the timeline of prediction of the price of power, the relevant set of influential factors is required to be filtered and normalized.

The examples of influential factors include historical electricity prices, demand, weather conditions, and transmission congestion load, system operating conditions, reserve imports, fuel prices, time indices, etc. The predicted Electricity price is used in day-ahead profit maximization, bilateral contracts planning, and investment recovery confirmation as a function of short-term, medium-term, and long-term actions.

5 Leveraging Data Analytics in Improving Accuracy of Price-Prediction Models

Data analytics is a logical process, used to search through large volume of data in order to extract useful data to make certain decisions for the development of businesses, through the chain of explorations, correlation, mapping, classification,

and clustering, as represented in Fig. 2. The time frame decides the volume of data, for example, a few electricity markets operate on a settlement time of an hour, 15 min, 5 min, etc. There are different bidding strategies and auctions, implemented by power trading exchanges.

For the purpose of developing predictive model for the price of power, it is required to define for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction [7].

A. Clustering and Classification of Days and Months with Price of Power as Key Reference-Attributes

As discussed in the previous section, Market participants need to select the type of variables, influencing the price and its variation too. The universal framework of data treatment is presented in this chapter, and it can be replicated to other market variables too.

In this chapter, K-means and K-NN techniques are presented for clustering and classifying days and months to create covariant groups to encompass the span of the data landscape, enabling the increased accuracy of prediction models, along with optimizing the number of models to be developed to predict the price of each day, each month over each year. While churning the data, the correct choice of k is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user.

The data of electricity price values of the International Electricity Market in the US for the last three years is used to develop the predictive model. The type of days and months are clustered, before developing a predictive model. The following sections provide exercises and outcomes.

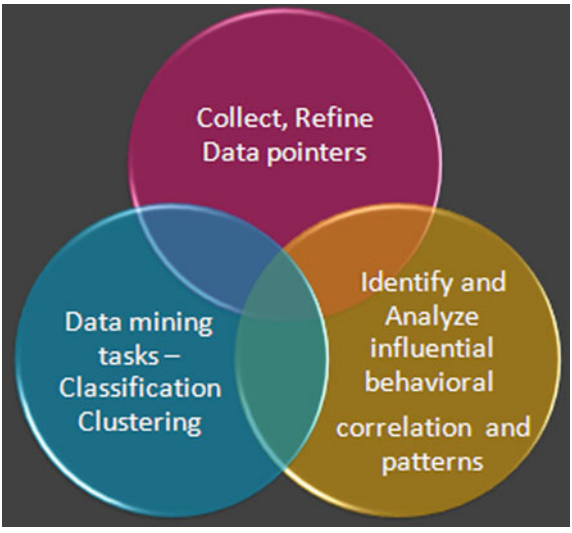


Fig. 2 Data analytics processes

- **Determining Number of Clusters**: Using three years of hourly data of price of power of US Utility, the classification and clustering is performed on 26,280 samples. The attributes identified are days and months.
- **The Elbow Method**: The clusters are defined such that total intra-cluster variation is minimized. The plot of the Sum Square of Errors (SSE) for each value of k is shown in Fig. 3. The plot looks like an arm, then the "elbow" on the arm is the best value of k corresponding to the acceptable small SSE and the optimum number of clusters. The elbow usually represents where SSE starts to have a diminishing trend by increasing k (number of clusters) [8]. As presented in Fig. 3, for a type of day category, the 3 classes are sufficient for clustering the similar power-price values of all days.
- **The NbClust method**: The NbClust method evaluates 30 indices for determining the optimal number of clusters. For the task of clustering the days, 17 indices are supportive to the 3 clusters based on the covariance of price values, as shown in Fig. 4.

B. **Clustering of Days and Months**

Once the number of clusters is decided, using the k-means algorithm, the likelihood of mapping is assessed [9]. Figure 5 shows the 3 clusters of the price of power of the US market as a case study.

Cluster 1—The LBMPs of Monday, Tuesday, and Wednesday form a first cluster.
Cluster 2—The second Cluster has price values of Thursday and Friday.
Cluster 3—The third group includes all prices of Saturday and Sunday.

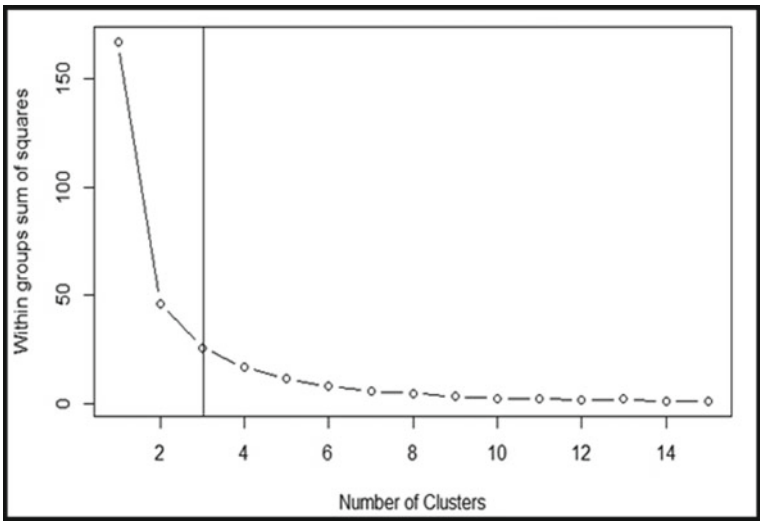


Fig. 3 Trend of Clusters verses SSE

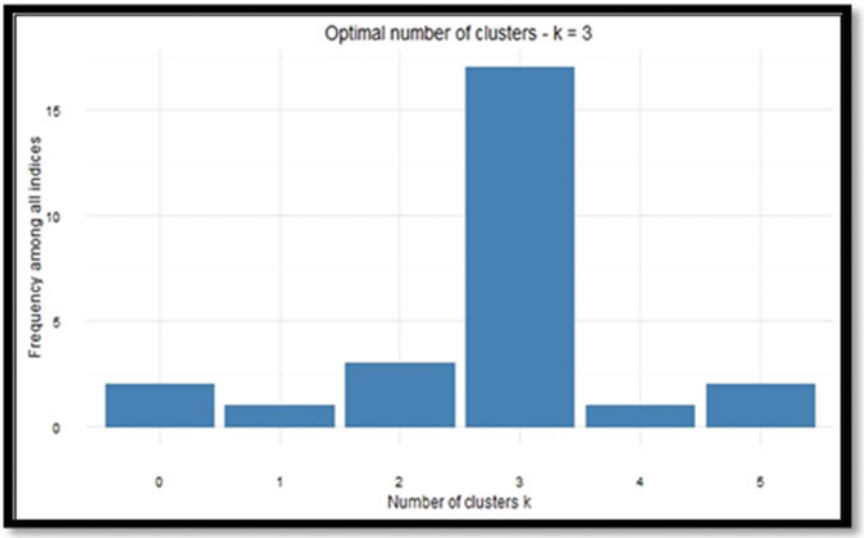


Fig. 4 Optimal number of clusters for market-price

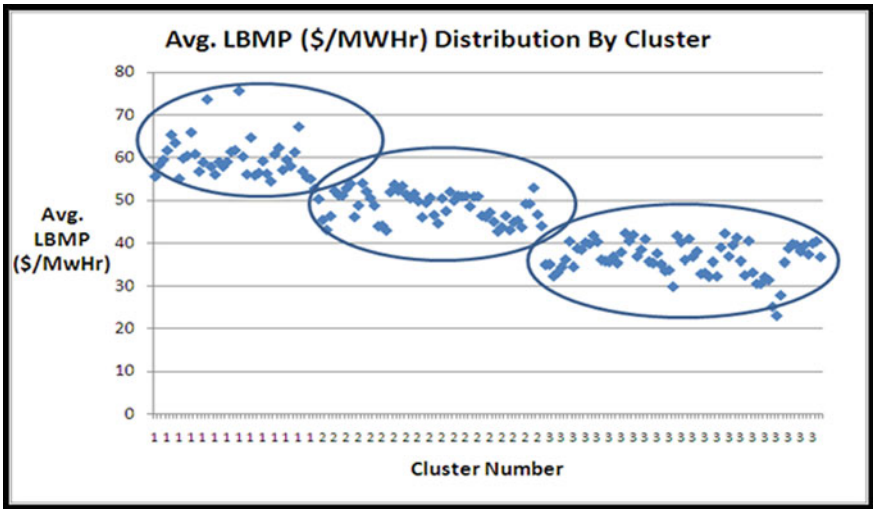


Fig. 5 Price distribution by Clusters (Type of Day)

The box plot depicts the standardized way of displaying the distribution of data based on the five indices: minimum, first quartile, median, third quartile, and maximum. For the US market, it can be observed that the behavior of prices from January and April to December is more or less homogenous, and the price variation

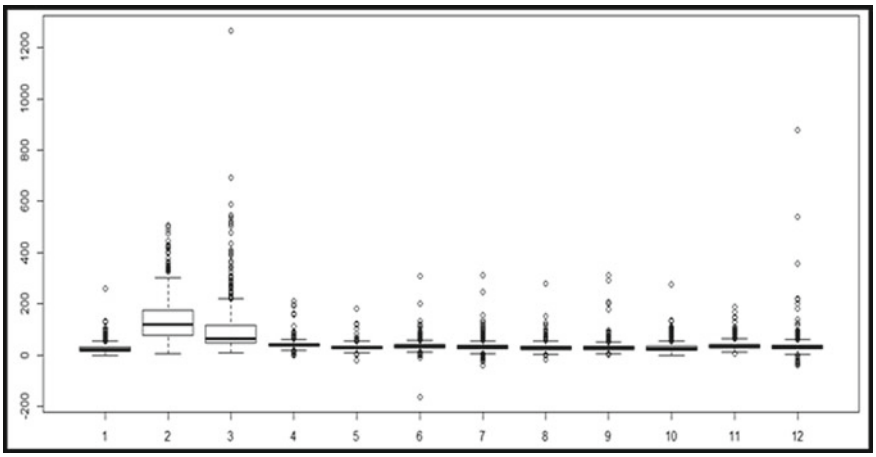


Fig. 6 Price distribution by Clusters (Type of Month)

between February and March is in good coherence. Thus, month as an attribute is grouped into two sets as shown in Fig. 6.

C. Use of ARIMA for Price-forecasting

The Auto-Regressive Integrated Moving Average (ARIMA) model has three types of parameters: the autoregressive parameters ($\varphi_1,\ldots, \varphi_p$), the number of differencing passes at lag-one (d), and the moving average parameters ($\theta_1,\ldots, \theta_q$). A series that needs to be differenced d times at lag-1 and afterward has orders p and q of the AR and MA components, respectively, are denoted by ARIMA (p, d, q) [10–12].

Data Preparation

The forecasting model is designed and developed for each of the three groups of datasets having a varying volume of historical data. Considering the above notations of datasets, the groups comprised as below.

- **Model 1**—Designed and build using $m_{(y)}$ dataset and tested for $m_{(y+1)}$dataset
- **Model 2**—Designed and build using $m_{(y,\, y-1)}$ dataset and tested for $m_{(y+1)}$ dataset
- **Model 3**—Designed and build using $m_{(y,\, y-1,\, y-2)}$ dataset and tested for $m_{(y+1)}$ dataset

With the following group and dataset notifications the below figure would provide the way the data was arranged to support the development of the forecasting model.

Dataset Notations

$m_{(y+1)}$	Hourly data of month to be forecasted
$m_{(y)}$	Hourly data of same month, but of prior year
$m_{(y, y-1)}$	Hourly data of same month, but of prior 2 years
$m_{(y, y-1, y-2)}$	Hourly data of same month, but of prior 3 years

The number of observations would vary with the number of days in each month. For any given month having 31 days the dataset $m_{(y)}$ would comprise of 744 observations, $m_{(y, y-1)}$ would comprise of 1488 observations, and $m_{(y, y-1, y-2)}$ would comprise of 2232 observations.

For any given month having 30 days the dataset $m_{(y)}$ would comprise of 720 observations, $m_{(y, y-1)}$ would comprise of 440 observations, and $m_{(y, y-1, y-2)}$ would comprise of 2160 observations.

For any given month having 28 days the dataset $m_{(y)}$ would comprise of 672 observations, $m_{(y, y-1)}$ would comprise of 1344 observations, and $m_{(y, y-1, y-2)}$ would comprise of 2016 observations.

As shown in Fig. 7, the data space will vary based on the inclusion of previous years inputs.

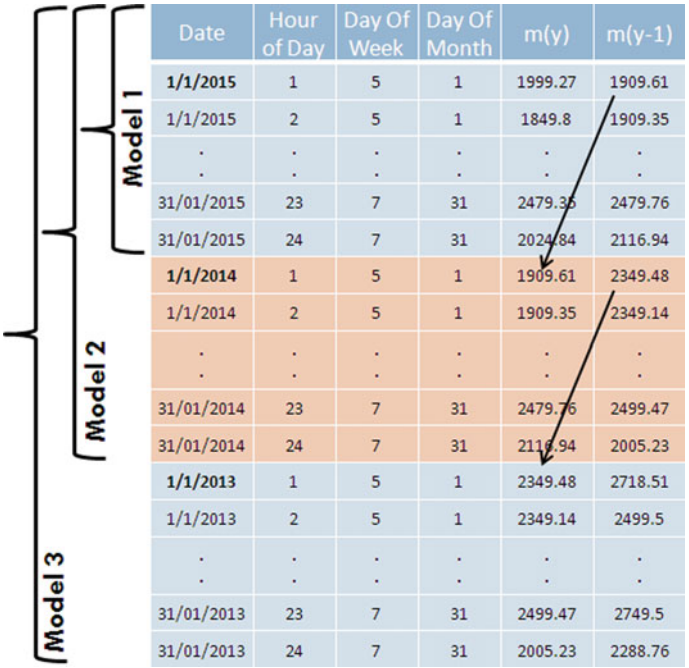
The performance characteristics of models are shown in the consolidated window in Fig. 8, wherein it is important to observe the trend of change of time series of the actual price of power values, getting followed by the predicted time series with a higher level of correlation, having correlation coefficient magnitude of more than 0.97. The prediction accuracy of the predictive models is evaluated in terms of Mean Absolute Percentage Error (MAPE). ARIMA model provided a satisfactory MAPE of around 5% for different days.

The predicted price values are used as key indicators while assessing volatility and market participation by all the market players. One important decision of congestion is elaborated in the next section.

6 Data-Reliant Congestion Management

The range of market applications, wherein the accurate price of power is the key differentiator in taking up operational decisions delves around all market players including Policy Makers, Investors, generation utility, Suppliers, Transmission Utility, Power Traders, Open access Industries, etc.

Better congestion management through market splitting allows full utilization of available transfer capability simultaneously maximizing the trade. The market splitting mechanism avoids e-bidding and makes it possible to integrate auctioning of transmission capacities within the bidding mechanism of the exchange hence acting



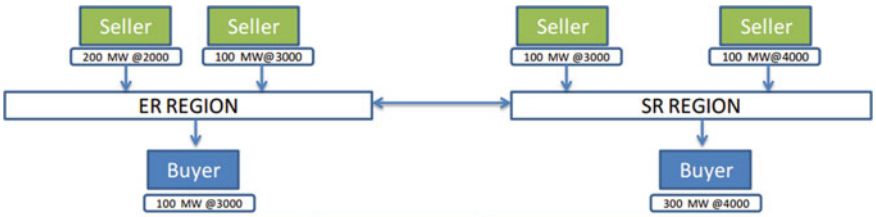
Date	Hour of Day	Day Of Week	Day Of Month	m(y)	m(y-1)
1/1/2015	1	5	1	1999.27	1909.61
1/1/2015	2	5	1	1849.8	1909.35
. .	. .	. .	. .	. .	. .
31/01/2015	23	7	31	2479.35	2479.76
31/01/2015	24	7	31	2024.84	2116.94
1/1/2014	1	5	1	1909.61	2349.48
1/1/2014	2	5	1	1909.35	2349.14
. .	. .	. .	. .	. .	. .
31/01/2014	23	7	31	2479.76	2499.47
31/01/2014	24	7	31	2116.94	2005.23
1/1/2013	1	5	1	2349.48	2718.51
1/1/2013	2	5	1	2349.14	2499.5
. .	. .	. .	. .	. .	. .
31/01/2013	23	7	31	2499.47	2749.5
31/01/2013	24	7	31	2005.23	2288.76

Fig. 7 Data preparations for ARIMA model

Fig. 8 Predicted values verses actual values

as a powerful platform for integrating energy and transmission markets. Implicit auction of transmission capacity through market splitting reduces the procedural complexities related to managing price bids and transmission capacities concurrently [13, 14].

The complete bid process from accepting bids to collecting funds and issue of requests to dispatch centers is completed within a few hours in real-time. The fundamental premise is that exchange handles transmission capacity in a market-oriented way. With this, there is neutral and fair day-ahead congestion management. The exchange system secures that the day-ahead plans send the commodity in the right direction, i.e., from low-price areas toward the high-price areas and transactions are netted out in one area [15, 16].



	Quantity (MW)	Price (Rs./MWhr)
ER Seller-1	200	2000
ER Seller-2	100	3000
SR Seller-1	100	3000
SR Seller-2	100	4000
SR Buyer	300	4000
ER Buyer	100	3000

Fig. 9 Sample case of congestion

A. Case Illustration

Two regions have been considered i.e. East Region and South Region. Four Sellers and Two Buyers participate in a 15-min time block and are taken with the following Bid Scenario, as presented in Fig. 9.

The congestion was reported by Dispatch center (Market Operator) from East Region to South Region corridor, and the power flow is constrained to 100 MW. Due to flow constraint, the system will "Split" the market into two regions i.e. Deficit (SR Region) and Surplus region (ER Region) and will again run the calculation chronology for both the regions separately considering the flow constraint and will derive the price of power as well as the volume of power transaction.

This scenario results in the decrease in the price of power for ER Surplus region. As shown in the example, the area clearing price (ACP) would be 2500 Rs/kWh whereas there is an increase in the price of power for the SR deficit region, with ACP for the SR region would be 4000 Rs/kWh. The electricity market is vulnerable to such price variations. The representative price effect is shown in Fig. 10.

A better price forecast will lead to more effective bids with lower risk and higher profit. Participants will be able to create better plans to maximize their own benefit and to protect themselves against price increases.

7 Unlocking Techno Commercial Benefits to Utility

Not only accessing data but also unlocking the potential of building contextualized analytics is an imperative action for Utilities, to unfold the dynamics of the power grid to enhance the visibility of techno-commercial opportunities. The figure represents electricity market players and the benefits offered with improved price indicators.

	Price (Rs./kWh)	0	999	1000	1999	2000	2999	3000	3001	3999	4000	4001	6000	8000	10000	20000
ER-Surplus Region	ER Seller-1	0	0	0	0	-200	-200	-200	-200	-200	-200	-200	-200	-200	-200	-200
	ER Seller-2	0	0	0	0	0	0	-100	-100	-100	-100	-100	-100	-100	-100	-100
	ER Buyer	100	100	100	100	100	100	100	0	0	0	0	0	0	0	0
	Net (Buy-Sell)	100	100	100	100	-100	-100	-200	-300	-300	-300	-300	-300	-300	-300	-300

Flow Towards SR of 100 MW

	Price (Rs./kWh)	0	999	1000	1999	2000	2999	3000	3001	3999	4000	4001	6000	8000	10000	20000
SR-Deficit Region	SR Seller-1	0	0	0	0	0	0	-100	-100	-100	-100	-100	-100	-100	-100	-100
	SR Seller-2	0	0	0	0	0	0	0	0	0	-100	-100	-100	-100	-100	-100
	SR Buyer	300	300	300	300	300	300	300	300	300	300	0	0	0	0	0
	Net (Buy-Sell)	300	300	300	300	300	300	200	200	200	100	-200	-200	-200	-200	-200

Fig. 10 Price change during congestion

Fig. 11 Price Indicator for Utility Market

The game of offerings stated in Fig. 11 is described in the following sections, having a potential influence on more stakeholders.

- **Generation Utility and Investors—Profitability and Confidence—**Huge investments made in power generation capacity are certainly recommended for the evaluation of the trade-off between risk and return. The profitability of such investments is directly dependent on the future price of electricity. Due to uncertainty and volatility, in the future price of electricity, such investment decisions carry huge risks. An accurate price forecasting tool can be instrumental in risk assessment and hedging techniques while mitigating investment risks. For generation Utilities, profit can be maximized by optimal scheduling of their dispatch and have an appropriate bidding strategy in place, anticipating market participation and dynamics.
- **Transmission Organization and System Operators—Demand on the Transmission Network—**To cater to the responsibility of providing full transfer capacity, the Transmission network as well as System Operators need to reiterate the planning models, based on new participants in the market, tariff structures, bidding options, etc. Price-forecasting models improve long-term planning and short-term congestion management.
- **Down-stream Operational decisions of Load Dispatch Centers—**Such a tool can also help Regional and State Load Dispatch Centers whose primary responsibility is to ensure grid discipline, facilitate load scheduling and dispatch functions, and energy accounting and settlement. The price prediction model can enable understanding the behavior of the market in terms of price sensitivities, which in turn initiates resource optimization at the network level.
  - **Power Traders /Load serving enterprises—**For the load serving enterprises, it is beneficial to use price forecasts to develop strategies and negotiation positions for entering long-term and short-term contracts with customers. The predictive model can leverage the knowledge of market clearing prices of the previous day, and previous month to provide a forecast of the next day's price,

which becomes crucial input in formulating bidding strategies. The very short-term forecasting model is useful in the spot market. In view of opportunities as peer-to-peer trading, prosumerization, the price indicator is the key catalyst.

– **Regulatory Body and Policy Makers—**The price indicators will enable the strategic decisions in applying amendments to accommodate the market changes as well as to ascertain the commercially viable trading practices. Thus, deploying contextualized analytics solutions enable Utilities to leverage the potential of structural as well as regulatory reforms as a proactive justice to the objectives of governments and regulatory commissions, keeping pace with Suitability Development Goals, along with commercial gains and customer delight.

8 Conclusion

The era of transformation in terms of digitalization, decarbonization, and deregulation has opened greater opportunities in the market, unlocking the potential of data accessibility and analytics capabilities, and encouraging cross-functional usage. The power-trading opportunities made Utilities to take proactive steps toward achieving the best business outcomes. The energy ecosystem, while generating, transmitting, distributing retailing as well as trading electricity has become adoptive to the rapidly changing technologies along with keeping communities safer, making operations more agile, and ensuring the end customers are happier.

Although these data-driven programmes have helped to improve operational efficiencies, the market-driven risks are to be assessed and managed. With high penetration of renewables behind the meters, they have created a huge impact on the price indicators and hence need to be studied scrupulously. With the trend of increased data flow, across the value chain, the data-driven intelligent tools will be an integral part of decision making.

References

1. https://www.mckinsey.com/industries/oil-and-gas/our-insights/global-energy-perspective-2021
2. https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/modernizing-the-investment-approach-for-electric-grids
3. https://www.mordorintelligence.com/industry-reports/utility-and-energy-analytics-market
4. https://ieeexplore.ieee.org/document/6681858
5. https://energyinformatics.springeropen.com/articles/10.1186/s42162-018-0007-5
6. https://www.cleanenergywire.org/factsheets/why-power-prices-turn-negative#:~:text=Negative%20power%20prices%20on%20the,such%20as%20Christmas%20or%20Pentecost
7. Mlambo N (2016) Data mining: techniques, key challenges and approaches for improvement. Int J Adv Res Comput Sci Softw Eng 7(3):692–697
8. Bholowalia P, Kumar A (2014) EBK-means: a clustering technique based on elbow method and K-means in WSN. Int J Comput Appl 105(9):17–24

9. Pham DT, Dimov SS, Nguyen CD (2005) Selection of K in K-means clustering. Manufacturing engineering centre. Cardiff University, Cardiff, UK 219(1):43–54
10. Martínez-Álvarez F, Troncoso A, Asencio-Cortés G, Riquelme JC (2015) A survey on data mining techniques applied to electricity-related time series forecasting. Energies 8:13162–13193
11. Voronin S, Partanen J (2013) Price forecasting in the day-ahead energy market by an iterative method with separate normal price and price spike frameworks. Energies 6(11):5897–5920
12. Kekatos V, Veeramachaneni S, Light M, Giannakis GB (2013) Day-ahead electricity market forecasting using kernels. Innovative Smart Grid Technologies Conference, IEEE PES (pp 24–27)
13. Zhou Q, Tesfatsion L, Lin CC (2010) Short-term congestion forecasting in wholesale power markets. IEEE Trans Power Syst 1–14
14. Neuhoff K, Hobbs BF, Newbery D (2011) Congestion management in European power networks: criteria to assess the available options. DIW Berl Ger Inst Econ Res 3:1–21
15. Løland A, Ferkingstad E, Wilhelmsen M (2012) Forecasting transmission congestion. J Energy Markets Fall 5(3):65–83
16. Wang Q, Zhang C, Ding Y (2014) Congestion management strategies of real-time market. J Power Energy Eng 2:227–232

Chapter 5
Infrastructure for Automated Surface Damage Classification and Detection in Production Industries Using ResUNet-based Deep Learning Architecture

Ankit Pandey

Abstract One of the most important tasks in production processes in manufacturing industries is to ensure the high quality and standard of the product. In this chapter, we try to address the problem of surface damage classification and segmentation to automate the manual inspection of images not only in manufacturing industries but for other industries as well in general. We propose two-phase learning approach where we develop two models (1) Residual Network (ResNet)-based damage classification model and (2) ResUNet, a combination of ResNet and U-Net, based damage segmentation model for accurately segment defects in images. In phase 1, ResNet-based model discard images of product with no defects (filtering phase) and pass only the defective product images to phase 2. We train ResUNet model on filtered images containing only defects for segmentation in Phase 2. Also, we experimented with image augmentations techniques and observe that using intelligent image augmentations techniques helped in getting better results. We present the details of our experiments on an open-source industrial steel sheet surface image data. The results showed that the highest classification accuracies (binary cross entropy) for training and testing set obtained by the ResNet were 91% and 78%, respectively, and highest segmentation accuracies (combined binary cross entropy and intersection over union) by ResUNet were 89%, and 72%, respectively.

Keywords Damage detection · Classification · Segmentation · Deep learning · Convolutional neural network · Residual Network (ResNet) · U-Net

1 Introduction

The objective of manufacturing industries is to make the real-time industrial production processes more efficient to deliver high-quality products and to be cost effective.

A. Pandey (✉)
A&I—Business Analytics, Tata Consultancy Services, Noida, India
e-mail: pandey.ankit5@tcs.com

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_5

Furthermore, early and timely identification of surface damages in product is crucial otherwise it can have adverse impact on the quality, brand value and performance of products. Complex machines are involved throughout in production line and can induce surface defects to a product. For example, in production process of steel sheets, several machines touch flat steel ranging from heating and rolling to drying and cutting which can lead to defects such as surface cracks, patches, scars, scales and scratches rendering the product to be damaged. It is of utmost importance for production industries to timely detect such damages and in-turn ensure that best quality product reaches to consumer.

Manual inspection of images captured at different stages of production line is time consuming, incur high-level inspection cost, error-prone and lots of low-level details of product images go unnoticed. In addition, images captured at various camera angles, out-of-focus images, varying lighting conditions, etc. makes it much more difficult and time-consuming for manual inspection. To maintain the high-quality standard of products and growth, intelligent visual inspection systems are becoming essential in production lines to augment the manual quality control procedures.

Automated surface damage classification and segmentation systems can enable production industries to maintain the product quality standard and reduce manpower cost and resources. In the recent years, we see many advancements in machine vision using deep learning-based techniques to classify and segment the objects in images with high accuracy automatically and precisely. Recently, researchers have proposed and used deep convolutional neural network (CNN)-based algorithm that have shown outstanding performance in image classification, object detection and segmentation tasks. Lots of recent research have also formulated the surface damage detection problem as classification, object detection and segmentation problem for classifying, localize and segment the defects from images. This has many advantages as compared to traditional machine learning-based algorithms that involve lots of hand engineering features and are less accurate as well. On the other hand, convolutional neural network-based algorithms require no hand engineering of features and are proven to surpass human-level performance thereby saving both time and cost for industrial applications such as surface damage detections. Furthermore, He et al. proposed fully convolutional neural networks (FCN) and applied successfully to solve the challenging problem of semantic image segmentation.

In this chapter, we formulated the problem of surface damage classification and detection as two-phase learning approach where we train residual network ResNet (CNN based architecture) for classification of defective vs non-defective product in phase 1 and a combination of residual network ResNet and fully convolution neural network based on UNet for surface damage segmentation is proposed in phase 2. Furthermore, we also propose intelligent image augmentations techniques that can help in getting better results with limited amount of image data. The remaining chapter is organized as follows. In the next section, we discuss state-of-the-art research that has been conducted in the area of object detection and segmentation techniques which can be applied for surface damage detection and show how the proposed approach is different from that of earlier research. In Sect. 3, the data

which has been used for the purpose of the study has been described followed by methodology and results in Sects. 4 and 5 followed by the conclusions in Sect. 6.

2 Literature Review

Before the rise of neural networks and deep learning, researchers used much simpler classifiers like linear classifiers over hand engineered features to perform complex and challenging computer vision tasks. With the recent advancement of computer vision using deep learning, researchers have proposed and created many deep learning architectures that has made remarkable achievements in computer vision fields such as image classification, object localization, detection, semantic and instance segmentation, image captioning, etc. The applications of these fields are also enormous ranging from manufacturing industries, health care, surveillance systems to other applications such as autonomous vehicles and surface damage detection [1, 2].

Recent studies also suggest that there has been a significant rise in using computer vision and deep learning-based techniques specifically for surface damage classification and detection in industrial applications [3–5]. Very Deep Neural Networks are difficult to train because of Vanishing/Exploding gradient problems. In theory, as you make a Neural Network deeper, it should only do better and better in terms of performance and accuracy. Training of a deeper neural network almost always helps. But in practice, as the number of layers increases, train error tends to decrease for a while and then tends to go back up. This is because of Vanishing gradient problem. With respect to image classification task, in 2015, He et al. proposed ResNet: "Deep Residual Network for Image Recognition" a powerful architecture that enables training of very deep neural networks, sometimes even networks of over 100 layers. Residual network helps solving this above problem [6]. In ResNet, the train error will only go down even if you train with more than 100 layers. This in turn allows training of very deep neural network without hurting of performance. The idea presented by inventors is to use "skip-connections" also known as "residual block" in a network that allows to take the activation from one layer and feed it to another layer even much deeper. Stacking such multiple residual blocks together allows to train very deep neural network and in turn learning very complex features at the deeper layers as shown in Fig. 1. In this chapter, we propose to use powerful ResNet architecture specifically for surface damage classification task in phase 1 of our approach that filters out the non-defective images more effectively, leaving out only images containing defects.

Authors proposed sliding-window detection approach using Convolutional Neural Network classifier for detecting objects such as damages [7]. In this approach, authors propose to pick a window or kernel of varying sizes and slide this window through every region of an image, thereby generating lots of cropped images. Finally, these cropped images are passed into the CNN-based classifier and have it classified as presence or absence of a surface damage. The huge disadvantage of this approach is computational cost, i.e. cropping out so many images and running each of them

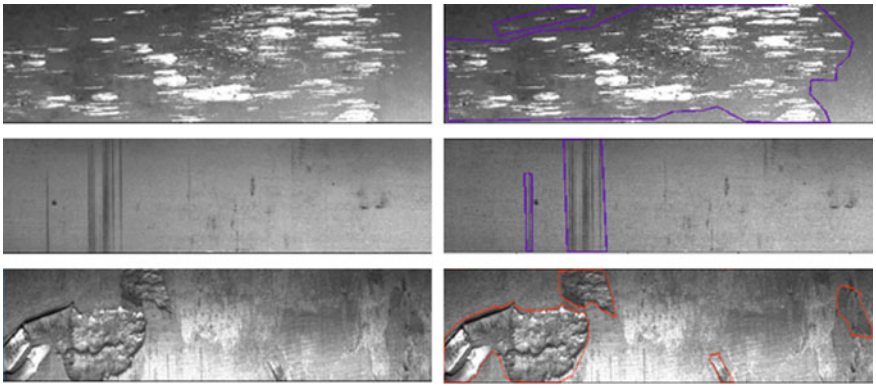


Fig. 1 Raw images and annotated images

through CNN independently is much more expensive task and makes sliding-window detection unfeasibly slow.

In other studies, authors proposed using fully convolutional implementation of sliding-window detection to detect objects [8, 9]. The authors proposed implementing sliding-window detection in a single forward pass-through CNN over entire image instead of cropping out multiple images in sliding window and running CNN sequentially. This helps reduce computational cost to a larger extent, thereby performing object detection much more efficiently. However, this method would fail in detecting small objects such as surface damages which can occur in any aspect ratio instead of always detecting objects in fixed square shape window.

In object detection and segmentation literature, region proposals CNN (R-CNN) algorithms proposed by Ren et al. [10] are also very influential in Computer Vision. Basically, these methods use traditional rule-based segmentation algorithm such as "watershed algorithm" to propose regions in the first step and further classify all the proposed regions using CNN. The huge disadvantage of using these algorithms is huge computation cost as region proposal step is still quite slow. Later to that, Cha et al. [11] proposed Faster region-based CNN wherein the authors proposed to use convolutional network even for region proposal step.

In more recent study, Joseph Redmon et al. proposed a novel algorithm called "YOLO: You only look once" for unified real-time object detection for bounding box predictions. This algorithm proposed placing a coarser grid of size n x n over an image and the idea is to apply image classification and localization algorithms to each of grid cells to output a bounding box around a detected object. This method allows CNN to "output bounding boxes of any aspect ratio" as well as output much more precise coordinates of object in turn detecting objects of any sizes. However, the algorithm would only produce bounding box and not the exact masks for an object.

In a similar application to ours, Zhenqing et al. proposed a Fully Convolutional Network (FCN) based algorithm U-Net for concrete crack detection method. The

idea of the FCN-based segmentation algorithm is to classify every pixel of an image into some category and finally group the pixel according to the category they belong. The advantage of using FCN based algorithm is that they are faster to train and has high efficiency and robustness. In this chapter, we proposed a surface damage segmentation method based on U-Net which in-turn uses Residual Network (ResNet-50 which has deeper network layers) as an encoder. The advantage of combining the U-Net with ResNet-50 as an encoder is to gauge the characteristics pattern of surface damages efficiently and accurately and finally produce the accurate segments of detected damages. Furthermore, only very few studies leverage the importance of image augmentation step to enhance the accuracy of a surface damage detection systems. We further demonstrate that leveraging advanced image augmentation techniques further improve the performance of system and can detect the surface damage locations for images captured under various conditions such as varying lightning conditions, camera angles, viewpoint variations and background clutter. During this section, you tap into a type of nostalgia, getting people to associate and attach to what you're saying by recounting familiar scenes from the history of the issue.

3 Dataset Description

For this study, we evaluate the performance of the proposed methodology on an open-source steel sheet surface damage dataset. The dataset consists of around ~6000 grey scale high-resolution images containing images of both defects and no-defects. The size of the typical image in a dataset is 256 (height) by 1024 (width) pixels which is consistent throughout the data. The dataset provided is also labelled and defects are classified into four categories namely transverse crack, longitudinal crack, scars and scales. Each image may have no defects, a defect of single category, or defects of multiple categories. Furthermore, subject matter expert labeler has also precisely segmented the defects, i.e. the exact location for each defect class is also provided to us. Run length encoding scheme is used for providing the defect segments in which each pixel is marked for each defect class. The training annotations which provide defect categories and segment for defects in the dataset are used directly for training and evaluating proposed deep neural network models for defect classification and segmentation task in this study. Figure 1 shows some of the raw images sampled from the dataset and the labelled image showing the defect segments.

4 Methodology

The damage classification and detection tasks can be defined in several ways. Generally, the task is to identify the type of defect present over surface of steel images and predict the precise location of defect. One way to approach this problem is to first classify the defect and then predict the bounding box coordinates over a defect. The

initial exploratory analysis of steel surface defect image data revealed that shapes of defects are either too long or bending which makes it difficult to predict the bounding boxes. Also, instead of training multi-label classification model for steel defect classification and then localization, we proposed two-phase learning approach.

In phase 1, we train a deep neural network-based binary classification model to distinguish between defective and non-defective images. Further, we pass only the defective images in the next phase 2 of our approach. In the phase 2, we trained a deep segmentation model to exactly predict the defect segments in an image. The details of our two-phase learning approach are presented in next section.

4.1 Two-Phase Learning Approach

Traditional computer vision-based approaches for tasks such as image classification and object detection heavily relied on hand-engineered feature that consumes lot of time and effort designing features. Recently, convolutional neural network-based algorithm has shown remarkable performance on these tasks. Inspired by CNNs, we proposed to use Residual Networks also called ResNet, in filtering phase 1 of our approach. The choice of using residual networks is attributed to the fact that they are highly accurate, and even more than 100 layers can be easily trained without being suffering from vanishing gradient problems. This step helps us discard any images that contain no defects. The architecture of the phase 1 classification model is shown in Fig. 2.

The dataset of steel surface images is collected from high-resolution cameras at the time of visual inspection. The acquired dataset is manually labelled by expert

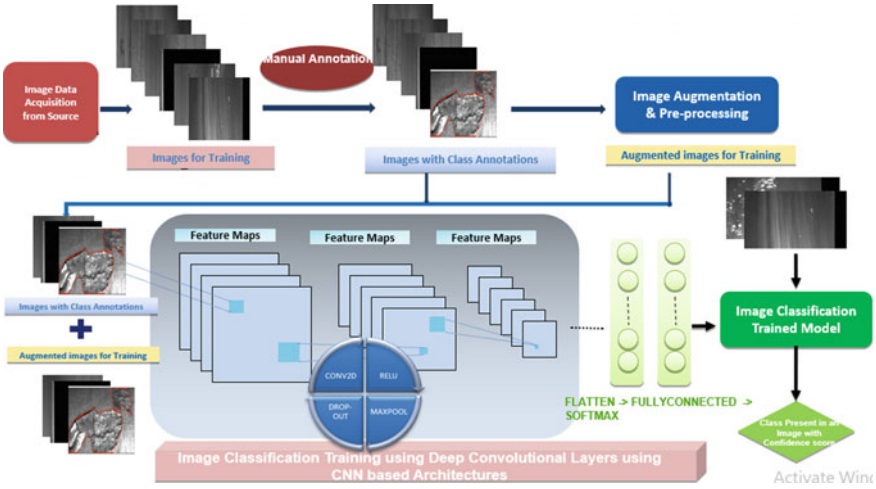


Fig. 2 Automated defect classification framework

labeller for the type of defects and location of defect. This labelled dataset with raw images and annotation is further passed to intelligent image augmentation and pre-processing block. The images are pre-processed in batches which includes image normalization, resizing of images for residual network model input, removing noises using Gaussian blur, histogram normalization technique, etc. Also, based on the characteristics of defects, we proposed to use specialized intelligent image augmentations methods such as random cropping, blurring, colouring and horizontal/vertical flipping to address challenges such as probable variations in appearances. Images with class annotations and augmented images are then passed to deep residual network model containing 50 layers followed by 2 fully connected layers and a final soft-max layer to output the class. This model is trained for around ~100 epochs/iterations on around 6000 images which learn to classify defective vs non-defective images present in the data. In the process of training the model, the distribution of data flowing in the network will gradually shift sometimes also called as covariate shift, thus slowing down the training process. To alleviate this problem, we used BatchNorm technology to force the data back to more standard distribution. So, Residual Network with BatchNorm solves the problem of changing distributions with vanishing gradients and speeding up the training process. In the final step of phase 1, we pass the real-time images as test data to identify the presence or absence of defects in an inference module with a model predicted confidence score for visual inspection as shown in Fig. 2.

Once the images are filtered leaving only the defective images, the next step is to identify and predict the precise location or segments of defects present over a surface of steel images. Inspired from Full Convolutional Network (FCN), we proposed a full convolution network based on encoder and decoder architecture called Unet with ResNet as encoder network to extract features. The encoder consists of Residual Network that down samples the original image to feature maps to represent the discriminative characteristics of defects and background. In addition, features obtained from down sampling techniques have a larger receptive field which in turn gather information from over a broader context. This helps in learning to predict the objects with significant variations in sizes. Furthermore, the decoder network up samples the feature maps back to original size of an image with increased number of channels. Each channel decodes the information of each type of defect separately. The automated object segmentation framework based on ResUNet architecture is shown in Fig. 3.

5 Results

We train ResUNet model on filtered images containing only defects for segmentation. We evaluated the results of our experiments on an open-source industrial steel sheet surface image data. The results showed that the highest classification accuracies (binary cross entropy) for training and testing set obtained by the ResNet were 97%, and 95%, respectively, and highest segmentation accuracies (combined

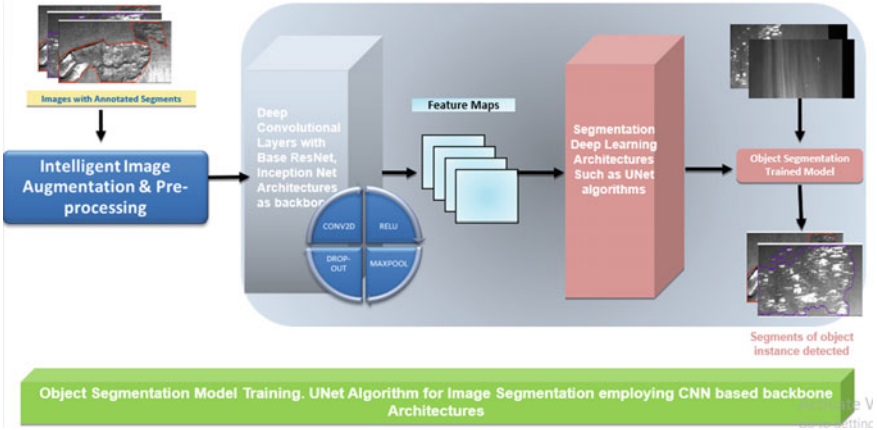


Fig. 3 Automated defect detection framework

binary cross entropy and intersection over union) by ResUNet were 75%, and 72%, respectively. Also, results showed that applying specialized image augmentations techniques helped in getting better results and gave a bump of around 3% in overall accuracy. The training and validation loss and accuracy plots are shown in Fig. 4. The figure also depicts the phase 2 training and validation IOU loss (intersection over union).

Model prediction results of some of the segmented images are shown below in Figs. 5, 6, 7, 8, 9 and 10 for instance. The results show that model can classify and segment the defect with good accuracy.

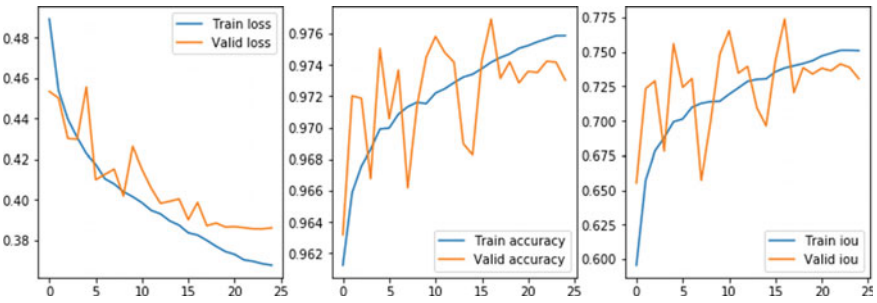


Fig. 4 Visualization of model training and validation loss and accuracy for phase 1 and phase 2 (last plot in the above figure)

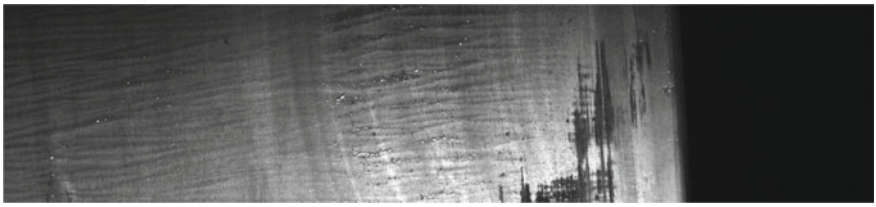


Fig. 5 Sample input image containing defect

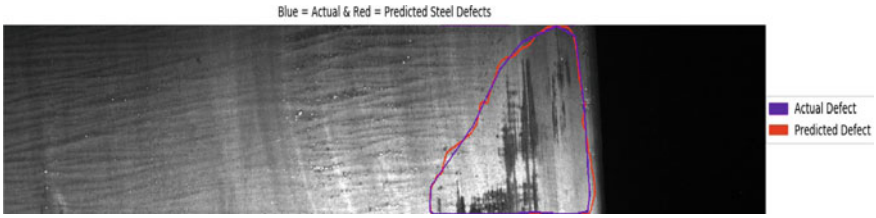


Fig. 6 Model prediction superimposed with actual

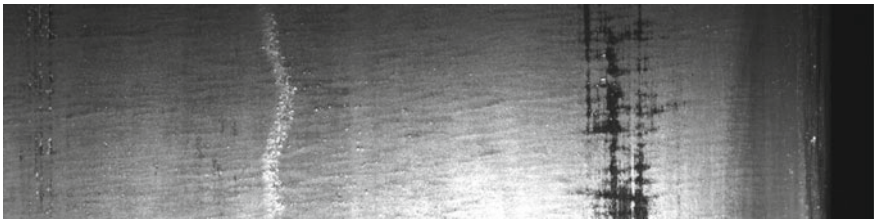


Fig. 7 Sample input image containing defects at different locations

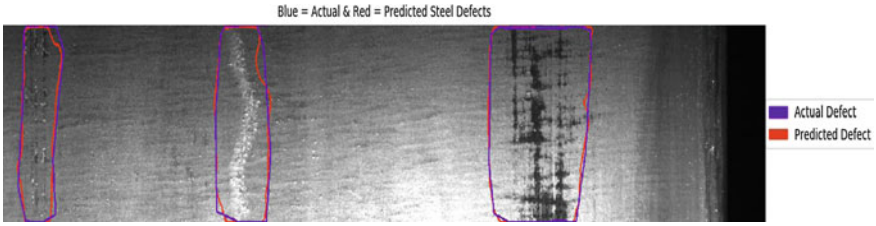


Fig. 8 Model prediction superimposed with actual

6 Conclusion

Visual inspection is impacting manufacturing industries and lot of research has been conducted in this area due to rise of deep learning. In this chapter, a novel two-phase

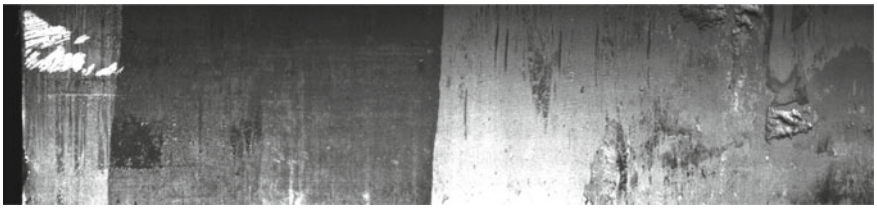


Fig. 9 Sample input Image containing defects at different locations

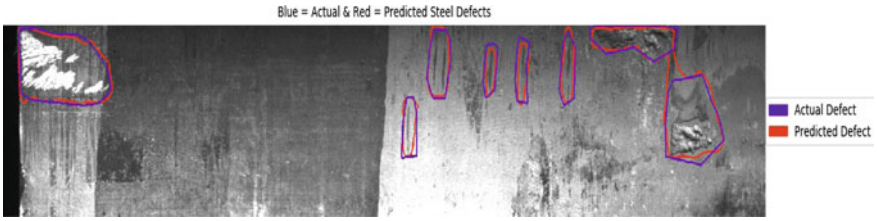


Fig. 10 Model prediction superimposed with actual

deep learning-based approach is presented to accurately perform both defect classification in filtering phase 1 and segmenting defect precisely in phase 2 for steel surfaces in production industries. In this study, we trained two models (1) Residual Network (ResNet)-based damage classification model and (2) ResUNet, a combination of ResNet and U-Net, based damage segmentation model for accurately segment defects in images. The major challenges addressed present the ways to combine different deep neural network architecture for the task of visual inspection and can be trained in end-to-end fashion without the need for any human intervention. Also, the study addressed the challenges of various probable variances in the appearances of defect with different shapes and sizes using advanced augmentation techniques.

References

1. Nhat-Duc H (2018) Detection of surface crack in building structures using image processing technique with an improved Otsu method for image thresholding. Adv Civ Eng
2. Kim B, Cho S (2018) Automated vision-based detection of cracks on concrete surfaces using a deep learning technique. Sensors
3. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-V4, inception-Resnet and the impact of residual connections on learning; AAAI: Palo Alto. CA, USA
4. Dai J, Li Y, He K, Sun J (2016) R-fcn: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems. NIPS, Barcelona, Spain
5. Kampffmeyer M, Salberg A, Jenssen R (2016) Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, Las Vegas, NV, USA

6. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. MICCAI (3)
7. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA
8. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD Backpropagation applied to handwritten zip code recognition. Neural computation
9. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the conference on computer vision and pattern recognition (CVPR), Boston, MA, USA
10. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the international conference on machine learning, Lile, France
11. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition

Chapter 6
Cardiac Arrhythmias Classification and Detection for Medical Industry Using Wavelet Transformation and Probabilistic Neural Network Architecture

Rajan Tandon

Abstract One of the most important tasks in medical technology industry is to achieve a maximum correctness in the detection of cardiac arrhythmias disorders. In this paper, author tries to address the problem of extracting the features of Electrocardiogram (ECG) signals and then classifying them in order to automate the cardiac disorders fast and provide timely treatment to patients. Author proposes the combined approach for automatic detection of cardiac arrhythmias: discrete wavelet transformation (DWT) method in order to extract the features of ECG signals (selected from MIT-BIH database) and probabilistic neural network (PNN) model for signal classification. The feature extraction part during ECG signal analysis plays a vital role while analyzing cardiac arrhythmias. ECG signals were decomposed till 5th resolution level using wavelet transformation technique in order to calculate the four important arithmetic parameters: (mean median, mode, and standard deviation) of normal and abnormal ECG signals of persons. Then these parameters were provided as an input to PNN classifier with two distinct outputs, i.e., normal and arrhythmic patients which clearly differentiate between the normal signals from abnormal signals. Accuracy of the proposed work was assessed on the basis of feature extraction and the results show that the recommended classifier has some great possibilities in detecting arrhythmic signals. This approach gives sensitivity, specificity, and selectivity are 99.84, 99.20, 99.40, respectively, and overall accuracy is 99.82% as compared to the approaches discussed in existing literature. For experimental results, the MIT-BIH databases were used.

Keywords Discrete wavelet transform · Probabilistic Neural Network architecture · Cardiac arrhythmias · MIT-BIH database · Ventricular tachycardia · Supraventricular arrhythmias

R. Tandon (✉)
Tata Consultancy Services, Noida, India
e-mail: Rajan.tandon@tcs.com

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_6

1 Introduction

The mortality rate is too high due to sudden cardiac deaths globally. In order to reduce this, health care professionals always seeking the best ways to detect heart diseases on time and provide an accurate treatment to the patients [1–3]. In United States of America, every year around 450,000 persons are dying due to sudden cardiac death. So, this life-threatening arrhythmia is the biggest challenge in front of health care professionals. Although many researchers have already proposed several techniques in order to classify cardiac arrhythmias which we already discussed in the above paragraph [4–6]. The problem statement of this work is to develop a system for classifying ventricular tachyarrhythmia in order to provide an accurate and timely treatment to arrhythmic patients.

ECG is a tool which is very popular in every country for detecting any type of abnormalities of the heart. Electrocardiogram (ECG) method employs the recording of electrical signals which is produced by the heart. This task is extremely important and challenging in order to detect any kind of heart disorders [7]. The most popular heart disorder is nowadays is cardiac arrhythmias. Usually, arrhythmias are divided into two parts: Ventricular arrhythmias (when abrupt heart rate feels in lower portion of heart termed as ventricles) and another is supraventricular arrhythmias (when abrupt heart rate feels in upper portion of the heart termed as arteries) [8]. In order to detect and analyze these disorders, the graphical representations are required which are obtained from recorded ECG. Experienced healthcare professionals are required to analyze these graphs which clearly diagnose the particular arrhythmia from ECG signals [9, 10]. Sometimes, this process desires long time to diagnose the correct arrhythmia which results delay in treatment of patients.

In this research work, author formulated the problem of analysis and classification of ECG signals in order to determine cardiac arrhythmias. If any abrupt changes of heart rate are found in the ventricles and arteries is termed as cardiac arrhythmias. Furthermore, the author differentiates between the ECG signals of normal and abnormal patients (arrhythmic). The rest paper is organized as follows: Sect. 2 deals with other researchers' work in this field, methodology is presented in the next section of paper, i.e., Sect. 3, and then Sect. 4 comprises the experimental, which is then followed by conclusion in Sect. 6.

2 Literature Review

Several techniques are proposed by numerous researchers in order to reduce the severity of cardiac arrhythmias. So, health care professionals require an automated system to determine arrhythmias in very short span of time. Yang et al. develop an imaging method for reconstructing the activation sequences of ventricular arrhythmia [11]. Authors of this paper have also solved the ECG inverse problem with respect to frequency domain and time in order to encode the phase data of imaging. They

used cellular heart model for generating focal ventricular tachycardia [12–14]. Their experimental results have shown that spatial sparse frequency (SSF) is very useful imaging tool in order to identify the focal entrants of ventricular tachycardia. The correlation coefficients by using SSF have shown is 88%. The SSF technique is also very effective in detecting macro circuits in 3D ventricular space.

In order to balance the statistical performance measures, authors proposed the study of ECG signal analysis on the basis of convolutional neural network (CNN) architecture. The data set was used in this study was taken by MIT-BIH [15]. Some data mining techniques are used to enhance the positive predictive values in few instances.

A classification method is proposed for determining arrhythmias by using PNN based on three features: Rate of the heart, regressive coefficients, and spectral dimensional entropy by Patidar et al. [16, 17]. The main motive given in this work is to build up a PNN-based technique in order to improve better diagnosis of cardiac disorders. Their results have shown that the combination of these three above-mentioned features is highly useful in determining cardiac arrhythmias.

In another study, Sahoo et al. proposed diagnostic tool for cardiac disorders on the concept of wavelet transformation for filtering the components of noise for increasing the features of ECG signals [18]. Hilbert transform method is also used with adaptive thresholding technique for determining the R peaks efficiently. Their experimental results have shown that this method is very effective in analyzing R peaks. Authors of this paper have achieved 99.71% sensitivity and predictability is 99.72%.

Asadi et al. proposed a novel wavelet-based algorithm in order to detect the occurrence of events in ECG signal when used with multi-layer perceptron [19]. PNN is used in this work for classifying ECG signals. Firstly, they used DWT for the removal of noise and artifact, the multi-leads are used to find the QRS complexes in order to determine morphology of signals. Authors of this paper have used three classifiers, i.e., MLP, SVM, and PNN; the proposed algorithm have shown the accuracies such as 97.42%, 98.24%, and 97.42, respectively. The results were tested on 40 samples of MIT-BIH arrhythmia database.

A novel technique is proposed by Chetan et al. in order to detect cardiac arrhythmias on the basis of signals achieved by multi-lead ECG [20]. This method works on nonlinear features which are calculated on signals once they have gone through the method of DWT. The nonlinear features are calculated on sub-band signals of ECG, and their performance is measured on the basis of multi-layer perceptron, radial basis, and probabilistic neural network approaches. Their method has achieved an accuracy of 98.76%.

3 The Solution

3.1 Discrete Wavelet Transformation

This is very accepted method in signal processing area in order to compute complex scientific applications [21]. This method employs numerous applications in every area like artificial intelligence, engineering, mathematics, signal processing field, etc. [22]. In order to extract signals, DWT method is used in this proposed work. In the very first step, the experimental signals are used for feature extraction through low pass and high pass filters as mentioned formula:

$$Y[s] = (X^{*}g)[s] = \sum_{k=-\infty}^{\infty} X[k]g[s-k]$$

Till the fifth level, decomposition method was performed for improving the resolution of frequency. In the second step, filter bank analysis concept is used to produce the filter outputs of DWT [18]. The equations used in this analysis are as follows:

$$Y_{low}[g] = \sum_{k=-\infty}^{\infty} X[k]g[2n - k]$$

$$Y_{high}[h] = \sum_{k=-\infty}^{\infty} X[k]g[2n - k]$$

where $Y_{low}[g]$ represents low-pass filter and $Y_{high}[h]$ represents high-pass filter [22]. Once this filter analysis method is complete, the authors obtain the values of arithmetic parameters (mean, median, mode, and standard deviation) using MATLAB programming based on extracted features.

3.2 Probabilistic Neural Network

PNN concept is basically used in classification and especially popular in solving pattern recognition problems [23, 24]. The operations performed by PNN are managed into a multi-layer feed-forward network that employs it's well-known four layers: input, output, pattern, and summation layer as shown in Fig. 1.

PNN method has numerous functions in signal processing area. For training and testing the classifier, selected experimental signals are used. Tables 1 and 2 have clearly depicted the dissimilarities between normal signals from abnormal signals. These statistical values will give to PNN classifier for generating the results. Finally,

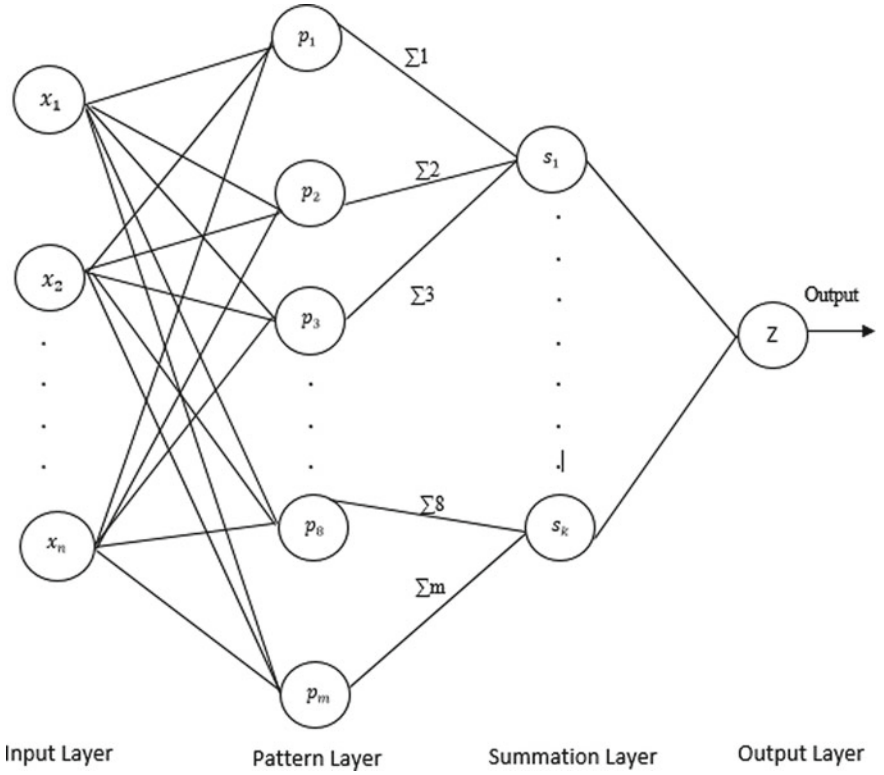


Fig. 1 Architecture of probabilistic neural network

the output of classifier is measured based on sensitivity, specificity, selectivity, and accuracy.

4 Experimental Outcome

For producing the experimental results, the author of this work has selected the experimental ECG signals from MIT-BIH database. Feature extraction work of ECG signals (abnormal and normal) has been done by using the concept of DWT. Figure 2 shows the decomposed ventricular tachycardia (VT) arrhythmic signal and normal signal.

As soon as the wavelet decomposition process has been completed, the following arithmetic parameter values have been measured in order to properly classify the selected ECG signals.

- Mean: These are the regular summation values at every level of decomposition.
- Median: These are the common frequency values while decomposition.
- Mode: These are the most repeated values on every level of decomposition.

Table 1 Arithmetic parameter values of Abnormal Signals (AS) after decomposition

Abnormal ECG signals	Mean	Median	Mode	Standard deviation
AS1	0.3471	0.3500	0.4600	0.4929
AS2	0.0074	0.1450	0.3150	0.6942
AS3	0.2714	0.2300	0.2450	0.4590
AS4	0.1865	0.2630	0.1880	0.8300
AS5	0.4266	0.3450	0.2450	0.7845
AS6	0.5237	0.4752	0.5682	0.5678
AS7	0.1278	0.3671	0.2783	0.6782
AS8	0.2618	0.3727	0.4182	0.5682
AS9	0.2568	0.2638	0.5719	0.7720
AS10	0.1268	0.3782	0.3600	0.6709
AS11	0.2678	0.4567	0.3781	0.6792
AS12	0.2980	0.3789	0.6791	0.9864
AS13	0.3678	0.6781	0.3786	0.7664
AS14	0.2679	0.3678	0.3678	0.5667
AS15	0.3678	0.4679	0.4178	0.6781
AS16	0.2678	0.3678	0.4789	0.6791
AS17	0.3678	0.3901	0.4199	0.6900
AS18	0.2789	0.3789	0.4900	0.5789
AS19	0.3789	0.3891	0.4568	0.6810
AS20	0.2567	0.5681	0.5789	0.3789

- Standard Deviation: These are the array of variations in data set values.

In this paper, the author has selected 40 ECG signals for computing the following arithmetic parameter values on every wavelet decomposition level. Tables 1 and 2 show arithmetic parameter values of abnormal ECG signals (AS) and normal ECG signals (NS), respectively.

These values are calculated by using MATLAB software. For better understanding to the readers of this paper, authors made a classification based on these values which shows a differentiation between arrhythmic patients and normal persons. Classification has been done by using the values of every arithmetic parameter like mean, median mode, and standard deviation as presented in Figs. 3, 4, 5, and 6, respectively.

Figure 3 represents the statistical analysis of arrhythmic patients and normal persons by using the mean values.

Figure 4 illustrates the statistical analysis of arrhythmic patients and normal persons by using the median values.

Figure 5 depicts the statistical analysis of arrhythmic patients and normal persons by using the mode values.

Figure 6 displays the statistical analysis of arrhythmic patients and normal persons by using the standard deviation values.

Table 2 Arithmetic parameter values of Normal Signals (NS) after decomposition

Normal ECG signals	Mean	Median	Mode	Standard deviation
NS1	1.0015	1.0160	1.0240	0.0480
NS2	1.0618	1.0640	1.0560	0.0339
NS3	1.0036	1.0040	1.0120	0.0481
NS4	1.1658	1.1720	1.1920	0.0869
NS5	1.0879	1.0840	1.0640	0.0432
NS6	1.0467	1.0972	1.0276	0.0738
NS7	1.0789	1.0682	1.0628	0.0678
NS8	1.0876	1.0456	1.0567	0.0278
NS9	1.0456	1.0567	1.0234	0.6789
NS10	1.0297	1.0346	1.0578	0.4567
NS11	1.0456	1.0467	1.0245	0.0456
NS12	1.0235	1.0358	1.0789	0.0356
NS13	1.0678	1.0437	1.0782	0.0726
NS14	1.0463	1.0678	1.0536	0.0762
NS15	1.0876	1.0826	1.0783	0.0278
NS16	1.0378	1.0378	1.0638	0.0378
NS17	1.0789	1.0287	1.0672	0.0288
NS18	1.0267	1.0678	1.0467	0.0256
NS19	1.0367	1.0367	1.0378	0.0277
NS20	1.0378	1.0456	1.0876	0.0378

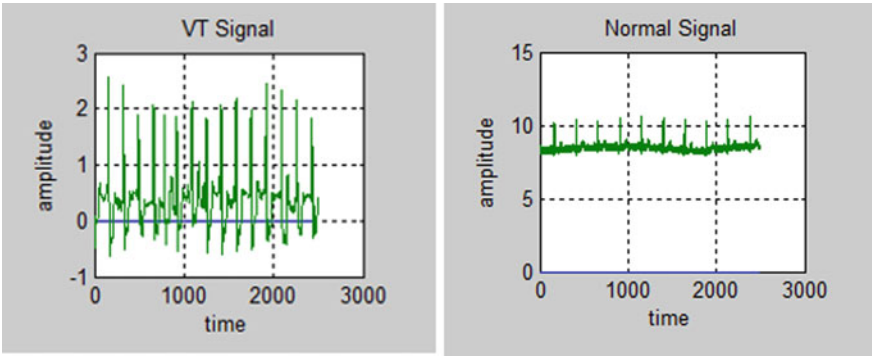


Fig. 2 Decomposed ECG signals

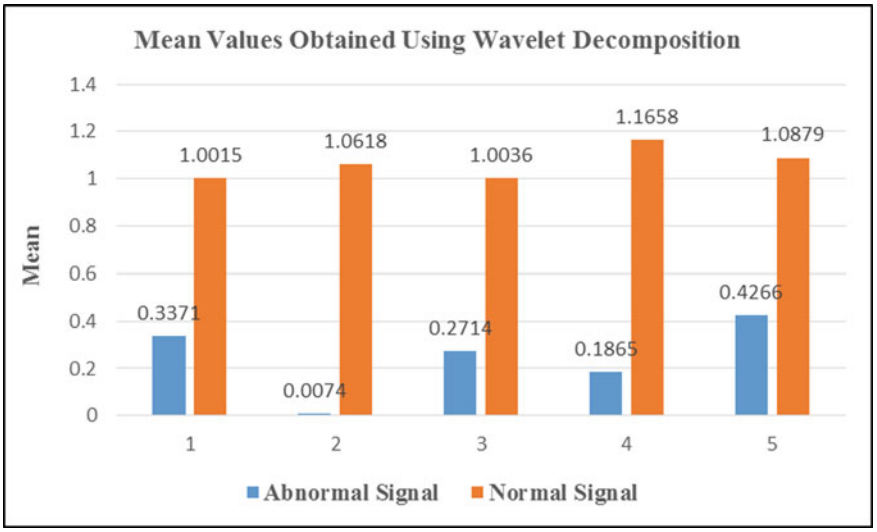


Fig. 3 Analysis using mean values

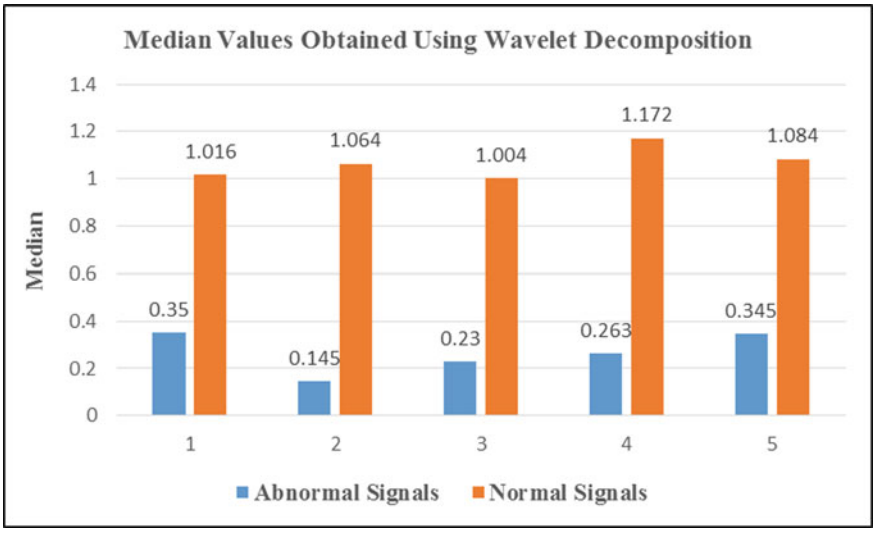


Fig. 4 Analysis using median values

5 Results and Discussion

Total of 40 ECG signals have been used in this paper for cardiac arrhythmias detection. 20 normal signals and 20 abnormal signals have been taken from MIT-BIH

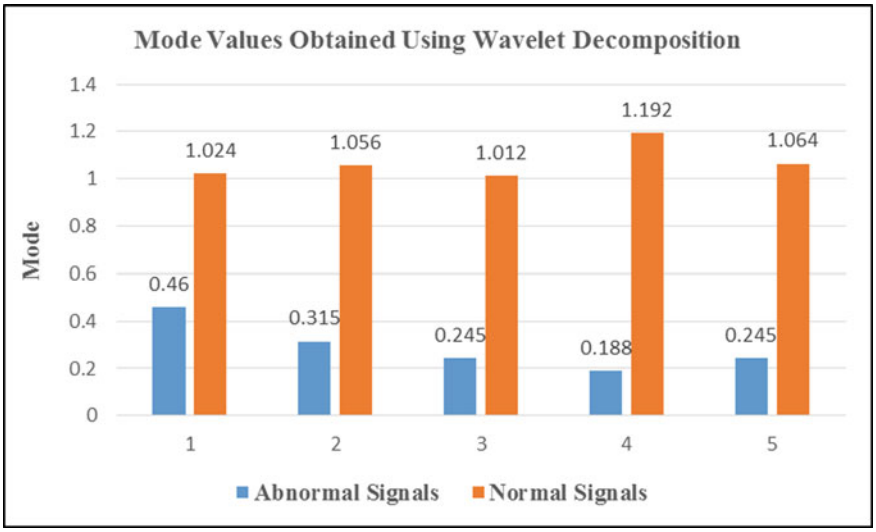


Fig. 5 Analysis using mode values

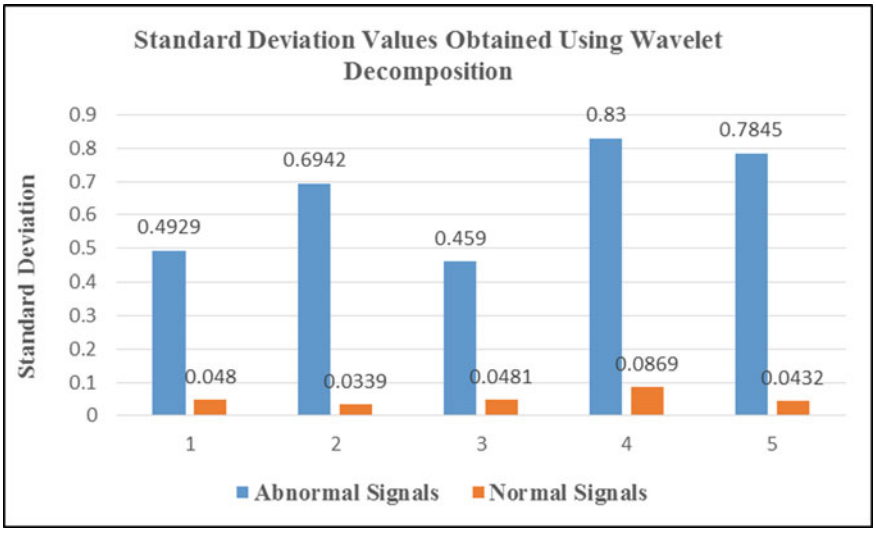


Fig. 6 Analysis using standard deviation values

database for presenting an accurate differentiation between normal and abnormal subjects. Wavelet transformation concept has been chosen for feature extraction work. Once feature extraction has been completed by using the low-pass and high-pass filters of transformation method, the arithmetic parameter (mean. Median, mode, and standard deviation) values are computed on MATLAB software. Tables 1 and 2

Table 3 Comparison between our results and other researchers' results

Authors	Sensitivity	Specificity	Selectivity	Accuracy
Sahoo et al. [18]	99.71%	NA	NA	99.72%
Lek-uthai et al. [21]	99.72%	99.19%	NA	NA
Asadi et al. [19]	NA	NA	NA	97.42%
Chetan et al. [20]	NA	NA	NA	98.76%
Our method	99.84%	99.20%	99.40%	99.82%

represent the arithmetic parameter values of normal and abnormal signals, respectively. Comparison of our results with other researchers' results has shown in Table 3. By using the statistical values, classification has been performed and shown in Figs. 3, 4, 5, and 6 of normal and abnormal signals. For better understanding of patient's heart condition, the doctors analyzed the classifier's results. In order to recognize the efficiency of the proposed method, confusion matrix concept has been used and generates output in terms of sensitivity, specificity, selectivity, and accuracy.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TP} + \text{FP}} * 100\%$$

$$\text{Selectivity} = \frac{\text{TP}}{\text{TP} + \text{FP}} * 100$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Cases}} * 100$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively.

6 Conclusion

Accurate detection of arrhythmias is extremely subtle job in medical industry. This process needs an accurate ECG recording in order to levy correct arrhythmia for providing timely treatment to patients. The proposed work in this paper is to be done for diagnosing any abrupt changes in the heart rate for reducing unexpected cardiac deaths over the globe. The author of this work has used the combined approach of DWT (for feature extraction of ECG signals) and PNN (for classification purpose of ECG signals). Based on extracted features, the values of mean, median, mode, and standard deviation are calculated on MATLAB software. Tables 1 and 2 clearly indicate the differentiation between normal and abnormal signals. Table 3 depicts classifier results which generate by using confusion matrix concept for comparing the author's results with other researchers of this field. Our proposed work for detection

of arrhythmias will definitely help in improving the efficiency of medical industry. Author has achieved the sensitivity, selectivity, specificity, and overall accuracy are 99.84%, 99.20%, 99.40%, and 99.82%, respectively.

References

1. Srivastava R, Kumar B, Alenezi F, Alhudhaif A, Althubiti SA, Polat K (2022 Mar) Automatic arrhythmia detection based on the probabilistic neural network with FPGA implementation. Math Probl Eng 22:2022
2. Mathunjwa BM, Lin YT, Lin CH, Abbod MF, Sadrawi M, Shieh JS (2022) ECG recurrence plot-based arrhythmia classification using two-dimensional deep residual CNN features. Sensors 22(4):1660
3. Gupta V, Saxena NK, Kanungo A, Gupta A, Kumar P (2022) A review of different ECG classification/detection techniques for improved medical applications. Int J Syst Assur Eng Manag 4:1–5
4. Pandey SK, Janghel RR (2021) Automated detection of arrhythmia from electrocardiogram signal based on new convolutional encoded features with bidirectional long short-term memory network classifier. Phys Eng Sci Med 44(1):1 73–82
5. Mohapatra SK, Mohanty MN (2021) ECG analysis: a brief review. Recent Adv Comput Sci Commun (Formerly: Recent Patents on Computer Science). 1;14(2):344–59
6. Ullah A, Rehman SU, Tu S, Mehmood RM, Ehatisham-Ul-Haq M (2021) A hybrid deep CNN model for abnormal arrhythmia detection based on cardiac ECG signal. Sensors 21(3):951
7. Sai YP (2020) A review on arrhythmia classification using ECG signals. In: 2020 IEEE International students' conference on electrical, electronics and computer science (SCEECS). IEEE, pp 1–6
8. Liu Z, Yao G, Zhang Q, Zhang J, Zeng X (2020 Oct) Wavelet scattering transform for ECG beat classification. Comput Math Methods Med 9:2020
9. Sarvan Ç, Özkurt N (2019) ECG beat Arrhythmia Classification by using 1-D CNN in case of class imbalance. In: 2019 Medical technologies congress (TIPTEKNO)) IEEE, pp 1–4
10. Marefat M, Juneja A (2019) Serverless data parallelization for training and retraining of deep learning architecture in patient-specific Arrhythmia detection. In: 2019 IEEE EMBS International conference on biomedical & health informatics (BHI). IEEE, pp 1–4
11. Yang T, Pogwizd SM, Walcott GP, Yu L, He B (2018) Noninvasive activation imaging of ventricular Arrhythmias by spatial gradient sparse in frequency domain—application to mapping reentrant ventricular tachycardia. IEEE Trans Med Imaging 38(2):5 25–39
12. Saraswat S, Srivastava G, Shukla S (2018) Classification of ECG signals using cross-recurrence quantification analysis and probabilistic neural network classifier for ventricular tachycardia patients. Int J Biomed Eng Technol 26(2):141–156
13. Lu W, Shuai J, Gu S, Xue J (2018) Method to annotate arrhythmias by deep network. In2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, pp 1848–1851
14. Saraswat S, Shahi P (2018) Probabilistic neural network approach for classifying ventricular Tachyarrhythmias. In: 2018 Fifth International conference on parallel, distributed and grid computing (PDGC). IEEE, pp 290–294
15. Lin CH, Kan CD, Wang JN, Chen WL, Chen PY (2018 Sep) Cardiac arrhythmias automated screening using discrete fractional-order integration process and meta learning based intelligent classifier. IEEE Access. 17(6):52652–52667
16. Patidar V, Srivastava R, Kumar B, Tewari RP, Sahai N, Bhatia D (2018) arrhythmia classification based on combination of heart rate, auto regressive coefficient and spectral entropy

using probabilistic neural network. In: 2018 15th IEEE India council international conference (INDICON). IEEE, pp 1–4
17. Teeramongkonrasmee A, Somboon P, Lek-uthai A (2017) Performance of a QRS detector on self-collected database using a handheld two-electrode ECG. In: 2017 10th Biomedical engineering international conference (BMEiCON). IEEE, pp 1–5
18. Sahoo S, Biswal P, Das T, Sabut S (2016 Jan) De-noising of ECG signal and QRS detection using Hilbert transform and adaptive thresholding. Proc Technol 1(25):68–75
19. Asadi F, Mollakazemi MJ, Atyabi SA, Uzelac IL, Ghaffari A (2015) Cardiac arrhythmia recognition with robust discrete wavelet-based and geometrical feature extraction via classifiers of SVM and MLP-BP and PNN neural networks. In: 2015 Computing in cardiology conference (CinC). IEEE,pp. 933–936
20. Chetan A, Tripathy RK, Dandapat S (2015) Cardiac arrhythmia classification from multilead ECG using multiscale non-linear analysis. In: 2015 IEEE UP section conference on electrical computer and electronics (UPCON). IEEE, pp 1–4
21. Lek-uthai A, Somboon P, Teeramongkonrasmee A (2016) Development of a cost-effective ECG monitor for cardiac arrhythmia detection using heart rate variability. In: 2016 9th Biomedical engineering international conference (BMEiCON). IEEE, pp 1–5
22. Perlman O, Katz A, Amit G, Zigel Y (2015 Sep 23) Supraventricular tachycardia classification in the 12-lead ECG using atrial waves detection and a clinically based tree scheme. IEEE J Biomed Health Inform 20(6):1513–1520
23. Desai U, Martis RJ, Nayak CG, Sarika K, Seshikala G (2015) Machine intelligent diagnosis of ECG for arrhythmia classification using DWT, ICA and SVM techniques. In: 2015 Annual IEEE India conference (INDICON). IEEE, pp 1–4
24. Jannah N, Hadjiloucas S (2015) Detection of ECG Arrhythmia conditions using CSVM and MSVM classifiers. In: 2015 IEEE signal processing in medicine and biology symposium (SPMB). IEEE, pp 1–2

Chapter 7
Investor Behavior Towards Mutual Fund

Devamrita Biswas

Abstract This study explores how mutual fund advisors help in planning the unmet needs of the clients and suggesting them different investment options. It helps to 'demonstrate and analyze' the 'Selection Behaviour' of customers towards 'Mutual Funds' and depicts the risk associated with those. Financial Planning along with all the aspects of Risk Management by means of 'Life Insurance', 'Health Insurance', 'Emergency Planning' is discussed. Further, the discussion moved into details for 'Mutual Funds', its mechanism of working, types, 'Risk-Reward' parameters and Ratios, Systematic Investment Plan, 'Rupee-Averaging', 'Compounding effect', Accounting and Taxation part. Client Interaction part is thoroughly highlighted where common investor biases are depicted. Later, mechanisms to overcome those biases by means of 'Systematic allocation of assets' and 'Risk-Profiling' have been highlighted.

Keywords Risk-profiling · Birds-eye view · Rupee averaging · Compounding effect · NAV · Systematic investment · Risk-reward mechanism

1 Introduction

The focus area of this chapter is the study of investor behavior during normal times. COVID-19 and economic crisis is being taken out from the scope of current research. Hence, our study will revolve around post-2008 Financial Crisis, how market survived, how investor kept switching funds, and how Net Asset Value (NAV) always set a new benchmark by breaching all the past records. Trend among investors demography-wise and fund-wise is being analyzed since 2008. NAV comparison among peers, portfolio evaluation and suggestion of suitable portfolio, current-investment scenario, and trends have also been highlighted briefly. Though number of Annual Maintenance Contract (AMC) operating in India has significantly increased, however lack of awareness persists among investors. '**Birds-eye view**' on Mutual Funds market and in-depth analysis of '**Financial Portfolios**' totally depends upon

D. Biswas (✉)
Project Manager, BFSI SEAL US-SE 1.6—Group 6, Tata Consultancy Services, Kolkata, India
e-mail: Biswas.Devamrita1@tcs.com

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_7

the ‘Current economic market. It does not include any artifact related to COVID-19 pandemic market that fall in Feb’20–Apr’20. It followed general artifacts since Financial Crisis of 2008.

2 Literature Review

As per publication from Cambridge University Press on ‘Social Transmission and Investor Bias’ by Hodrick et al. [1], Conversion rate from one investment strategy to another is convex in realized returns. Here, authors suggest that ‘High Variance’ and ‘More Skewed’ strategies dominate overactive strategies and are directly depended on their social networks. Further, the paper points out at ‘Idiosyncratic volatility anomaly’, which is a rather complex manifestation of historical-market anomalies related to ‘excessive extrapolation on firm growth’, ‘over-investment tendency’, ‘accounting accruals’, or ‘investor underreaction to earnings news’. Therefore, Idiosyncratic Volatility is defined as f(Corporate Selective Disclosure), where anomaly>>‘less sophisticated investor base’.

Doukas et al. [2] in their journal ‘Divergence of Opinion and Equity Returns’ have highlighted dispersion in forecast of analyst is just mirror image of ‘Short Selling Constraints’, ‘Optimism in Analyst’s forecasts’, ‘Herding’ of analyst’s behavior. Baltussen et al. [3] in their journal ‘Unknown Unknowns: Uncertainty About Risk and Stock Returns’ have clearly shown ‘Vol-of-vol’ effect is distinct from (combinations of) at least 20 previously documented return predictors, as well as survives many robustness checks and holds in the United States and across European stock markets.

‘Does Risk-Neutral Skewness Predict the Cross-Section of Equity Option Portfolio Returns?’ by Bali et al. [4] highlights ‘strong negative relation’ between ‘risk-neutral skewness’ and ‘the skewness asset returns’, which is consistent with a ‘positive skewness preference’. The returns are not related nor explainable by ‘well-known market’, size, book-to-market, momentum, short-term reversal, volatility, or option market factors.

Research by Bailey et al. [5] shows that the individual investors increasingly invest in mutual funds to invest in the equity market rather than trading individual stocks. As per Huan et al. [6]: ‘Individuals hold 47.9% of the market in 1980 and only 21.5% in 2007. This decline is matched by an increase in the holdings of open-end mutual funds, from 4.6% in 1980 to 32.4% in 2007’. ‘Hence, it is increasingly important to understand how individual investors hold and trade mutual funds. Traditional portfolio choice models imply a simple investment strategy based on well-diversified, low expense mutual funds and minimal portfolio rebalancing. Index funds, and other equity funds with low fees and low turnover, are cheap, convenient vehicles for individual investors to implement such a strategy’. The extent to which individuals consuming mutual funds is a critical positive function of ‘the rationality and effectiveness with which investors approach capital market’.

Research Objectives:

To find out common behavioral biases for common investors and how short-sightedness of investors impacts while investing in Capital Market. The aim of this research is to get a holistic view of different risk appetites along with different investment products for different investors. It does not say what is right or what is wrong. The sample is not biased as it is surveyed among office executives, academic professionals, different shops, high net-worth individuals (HNIs), etc. Portfolio pattern will also give an idea of how capital market changes its pattern from time to time.

3 Materials and Methods

Common trends were after visiting to the clients, which always highlighted some biases towards their investment approaches. Bias leads to financial losses for the investors like Herding Bias (30 percent), Over-confident Bias (22%) Confirmation Bias (20%) that have been analyzed after visiting individual as well as Corporate Investors [7].

(A) **Data Source**

Individual clients:

The clients who were visited personally includes HNIs, Non-Resident Indians (NRIs), officers from corporate sectors, and insurance companies in Kolkata. These visits and interactions happened in 2019 where suggestions were provided after analyzing their portfolios, for suitable investments based on their goals. Investors often get biased and keep switching around funds copying others, whereas Risk-averse investors are not willing to participate in Capital Market, and mostly prefer Debt Funds. Young people with age between 25 and 40 with moderate financial strength kept trying Equity investments without even knowing the market and without having any financial objectives. Investor's behavior totally is towards Equity Market post Pandemic, due to more return and more risk coverage. Datasets were collected from Financial Portfolios and Association of Mutual Funds in India (AMFI) website for the analysis and to understand the Risk-Return mechanism for Mutual Funds.

Corporate and Retail Visit:

Corporate experience in IT companies is totally different. People between the age of 26 and 35 are not risk-averse, as they have already invested their money in shares. A personal meeting with two C level executives of approximately 30 years of age helps to understand their risk-loving approach. A recommended to take 'Emerging Blue-chip Fund' which gives a decent return of around 14% across years and 'Health Guard Fund' which is a combination of Health-Insurance and Life-Insurance for his family, due to volatility of corporate jobs. Those emerging Blue-chip fund and Health Guard Fund are from two reputed AMCs will secure both investment and health insurance to cover mortality risks.

(B) **Description of Dataset**

Data sets are built upon collecting information from Individual Investors and Corporate Investors. Portfolios-based data are collected from HNI as well as NRI client, which makes the significant part of this research.

Individual Investors

a. A query from an existing client of a Small AMC was regarding the returns he can expect in the HDFC Small Cap funds post 5 years and the reasons for NAV values not improving much. The response was that around 15–18% returns can be expected based on the historical trends from Value-Research data. Besides, he explained that there is no need to worry about NAV value as it depends on the market. Concept of accumulation of units, rupee-averaging, and compounding rates were also explained.
b. It was suggested to one of the investors and his colleagues in schools to go for SIP of 'Mirae Asset Emerging Blue-chip Fund' which has return of around 18–21.37% over the 5 years, and 19.20% over the last 3 years, i.e., performing really well in last few years.
c. Another NRI client who is not interested in Mutual Funds as it is subject to 'Market Risk' and had a long history of loss in recent years. It was recommended to him to pursue 'Fixed Deposit' and go for 'Term Insurance' only, as he is already a customer of 'Star Health' Health-Insurance Premium.
d. Similar scenario was faced by Government employee who does not want to take risk. The recommendation was to go for 'Reliance Liquid Funds' where face value of 'Reliance ETF Gold BeES' is Rs. 100/unit, along with Fixed Deposit.
e. While visit to another investor, who had a newborn child. He was suggested to go for 'Term Insurance' and 'Health Insurance' for his family. Also advised him to put money in 'Aditya Birla Sun life Focused Fund', i.e., large-cap funds which has average return of 14.05% which he can use for education for his child after 15 years.

From the previous interactions with clients, it was found that there are a lot of biases prevailing among the investors. Methods of investment proposition and how bias affects investment horizon, portfolio, investment strategies are mentioned below

- **Projection Bias:** Sometimes investors use recent past to project distant future while ignoring distant past which is not a good practice. It is advised to carry out 5–10 yearss of financial analysis that should be done before investment.
- **Herd Mentality Bias:** Investor has a belief that other investors must have better information than him and tend to follow the same investment pattern or go for same portfolio of their acquaintances. Looking at the short-term performance could be fatal and might often lead to bubbles and crashes.
- **Ownership Bias:** Sometimes investors do not invest more with current portfolio as he is contented with their present investments. It gives rise to 'Endowment effect' which leads to arrogance and self-optimism. The investors think that their portfolio is optimum, while the truth is that the investment is a long-term process and should be carried on satisfying future financial goals.

- **Loss-Aversion Bias:** Investors are more worried about the loss of principal amount. They do not prefer to take risk in 'Long Term' funds or they seek redemption seeing the short-term market scenario, which is fatal. Also, we have observed that losing money gives much more pain to investors than pleasure of gaining.
- **Gambler's Fallacy:** Sometimes investors predict random events 'in the Financial Market' on the basis of 'Past trends' or without any concrete information, which is fatal. They tend to trust that if something happens more frequently now, will happen less frequently in the future. The perception of 'Balancing Nature' is not applicable to Financial Markets.
- **Winners Course:** Sometimes investors go for huge investments to make sure that they are able to win 'Competitive Bid' for their social position and status—even if it is a financial loss.
- **Anchoring:** Investors have an inborn tendency to believe in preliminary information while making decisions. Suppose, the investors wait for the right price to sell their stocks, while that expected price level is no longer relevant today.
- **Confirmation Bias:** Investor starts to pursue his own perception and belief, therefore starts prioritizing information according to that belief, i.e., 'Cognitive Bias'.
- **Worrying:** Sometimes a trader purchases a stock for some financial goals which did not work out well, and they simply switch to another stock-owning to that position based on their inductive hypothesis and self-belief which ultimately leads them no-where.

(C) **The History**

2008 Financial Crisis

By 2007, Lehman was revealing huge numbers—with $19.3 billion income and a record $4.2 billion overall gain. For a few reasons, including banks defaulting on the unsafe advances and unsustainable subprime contracts, the lodging market started to crash in 2006—at the same time, courageous, Lehman Brothers kept expanding (multiplied, truth be told) a lot of the land pie as much as $111 billion in resources and securities in 2007.

By September 2008, Lehman declared a normal '$3.9 billion misfortune' in its last quarter, however, the actual figure is close to '$5.6 billion misfortune', which includes 'poisonous' resources too. Lehman professed to have helped its liquidity to around $45 billion, diminished home loans by 20%, and decreased its influence factor by somewhere 'in the range of 7'.

In the main seven-day stretch of September, Lehman's stock dropped radically—about 77%. Financial specialists' questions were developing, and they went on to build up an association with 'Korea Development Bank' to seek help and turn off business land resources.

When Moody's rating revealed to Lehman that he would need to surrender a 'greater part' of the stake of his organization to financial specialists to keep up its appraisals and hence made the stock overpriced. As a result, their stock price went

down by 42% on September 11, 2008—leaving Lehman with just $1 billion in real money when the week was finished. Furthermore, by September 15, 2008, Lehman Brothers company's stock further fell down by 93% from its standing only three days earlier.

(D) **Data Preparation**

Datasets for the research is collected from multiple sources like—a survey conducted by the authors to gather the investment details, from Financial Portfolios of NRI, HNI and Office Executives, and Association of Mutual Funds in India (AMFI) website. Portfolio of Individual and Corporate Investors was analyzed along with the Mutual Fund Performance since 2008 and has reached to the consensus that NAV values improved over the years, and those who have invested in Mutual Funds are in a profit now. Post-re-election result of new Government who gained majority, 'SENSEX increased 1550 points and hits 40000'. Investors and Corporates expect that market will increase steadily and Mutual Fund performance will be boosted by recent improvement in the market. Tight liquidity due to 'NBFC Crisis' and sharp increase in 'global oil price' and 'trade war' are still a concern for market.

1. **Data Cleaning and Wrangling:**

After removing unnecessary data (funds that are not valid), Local Grouping, Temporal Grouping, and Grouping by specifics are done. Next point is to analyze the portfolios of the HNIs keeping for decades (2008 onwards).

2. **Outcome of the Lehman Brothers Crumble**

Over 6 million positions were lost, joblessness rose 10%, 'the Dow Jones Industrial Average (DOW)' dropped by a shocking 5,000 points, as indicated by ABC News. This trend gradually expanded its influence at global level and crushed economies in nations like 'Latvia', 'Hungary', and 'Lithuania' (also the European Union). Indeed, even Pakistan looked for a bailout after the emergency from the 'International Monetary Fund (IMF)', and 'Iceland' confronted an emergency when authorities reported that the legislature had no assets to prop up real banks in the nation.

3. **SENSEX Performance Scenario post-2008:**

The Sensex hit in the market at 21000 marks elementarily in 2008 and NIFTY closed at 4899 on January 22, 2008, at a loss of 310 points. However, it crashed to 2763 in February 2008, due to crisis of 'US Subprime Mortgage' which is due to decrease in home prices and the housing bubble saw a bizarrely expansive number of subprime contracts endorsed for individuals who battled with credit and pay. At the point when the Fed started raising financing costs again and again, those credits turned out to be progressively costly and the borrowers got themselves unfit to pay it off.

4. **Problems to Dataset (As per Market Scenario):**

Investors were expecting a steep rise in price of NIFTY post-election due to majority of BJP though there is a sluggish market condition due to trade war between India and China and trade policy of Trump which imposed tariff on Chinese goods and vice versa.

4 Experimental Result and Evaluation

Article 'Investor behavioral bias and news inattentiveness proxies, though crude', depict that 'behavioral effects are at work in the mutual fund decisions of many investors' and have a significant effect on their performance. Also, 'the bias and inattention to news proxies are themselves' correlated in interesting ways that allow us to 'identify and study stereotypical investors'.

The Experiment results show the contributing five factors identified using factor analysis which also explain over 75% of the variance of the behavioral factors and other investor characteristics. The intuitive combinations of investor characteristics that comprise these five factors relate to mutual fund trading habits and performance in an interesting and consistent manner. Research shows that "the top-quintile narrow-framing investors have average mutual fund returns that are 2.16% lower than those in the bottom quintile, while top-quintile disposition effect investors have average returns that are 0.89% lower than those in the bottom quintile. In contrast, behavioral biases do not appear to affect the performance of index fund holdings".

From historical analysis of sensex, it has been showcased that from 2009 onwards, market started moving upwards and increased to around 6000 in September 2010. Since 2013, the market moved heavily from 6200 to 10600 due to clear majority of one political party in the Loksabha election and huge growth in economy that resulted in decrease in the price of commodities, as shown in Fig. 1.

From January 2008, it took around two years to regain the principal amount invested and it has been assumed that no fresh investments have been made during this period. There will be huge benefit for long-term investors with return of 3X in past ten years.

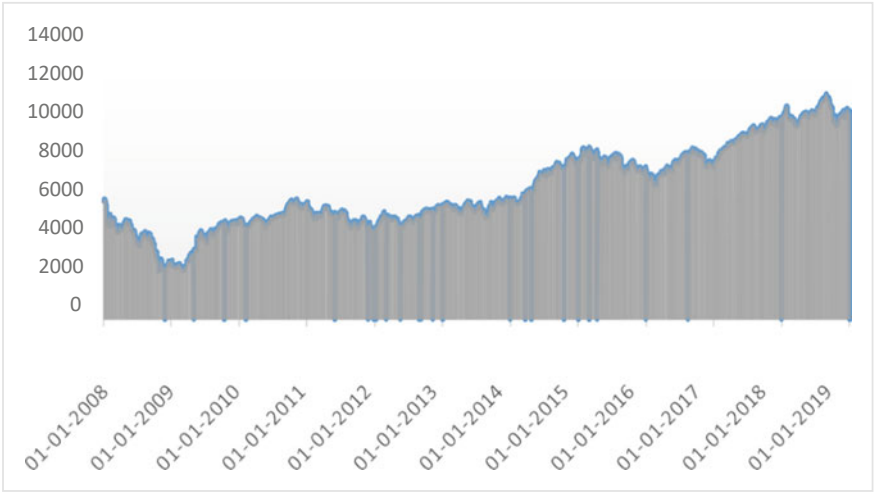


Fig. 1 NIFTY closing price since 2008

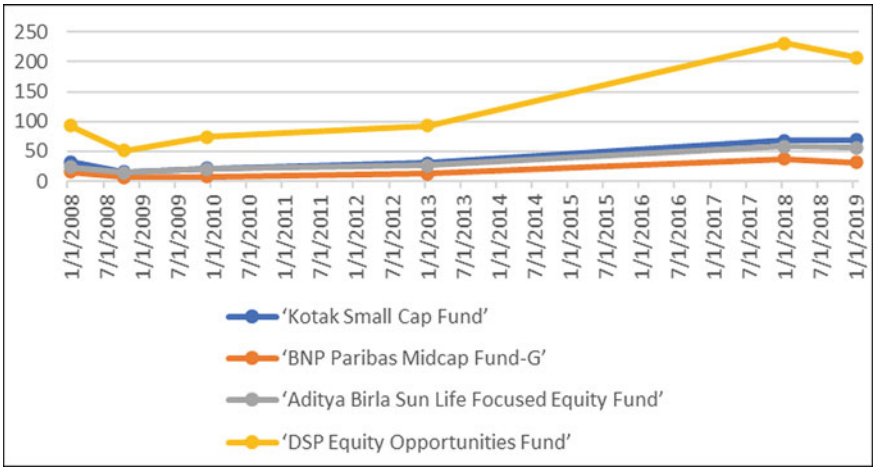


Fig. 2 Peer comparison of NAV since 2018

(a) **Review for 'AXIS Small Cap Fund'**

'AXIS Small CAP Funds' and 'ICICI Prudential Small-cap Direct Plan-G' having YTD 7.28% and 12.26%, respectively, are well ahead of benchmark level and also doing better than 'Kotak-Small-Cap Fund' which has return of −6.91% YoY in Q4FY19, i.e., below Market average.

(b) **Data Insights**

NAV fluctuation of Mutual Funds in 5 years gap, as depicted in Fig. 2, is analyzed, so that risk versus return scenario can be compared. Exit load is not loaded into calculation.

1. Its return since launch is 14.91%. Alpha is −0.28 which shows underperformance compared to standard benchmark. Beta value 0.8 shows 20% less volatile than market.
2. Recent report shows that its performance is not that good in comparison to its competitors. 'HDFC Small-Cap Regular Growth Fund', 'Axis Small-Cap Fund' are recommended which are performing far better than 'Kotak-Small-Cap Fund'.

(c) **Review for 'BNP Paribas Midcap Fund':**

On 5-year basis, return from 'BNP Paribas Midcap Fund' is around 13.87%, i.e., well below its competitors in the market like 'DSP Midcap Regular Plan', 'HDFC Midcap Opportunities Fund', 'Invesco India Midcap Fund', as shown in Table 1.

'Kotak Emerging Equity Scheme' is leading in the market with 18.69% return on 5-year basis.

1. Its Alpha is −3.55 means its underperformance is around 3.55% and Beta is 0.91, which means it is less fickle than market. It should be >1.

Table 1 Mutual fund NAV fluctuation in 5-year gap

NAV							
Type of fund	Fund	1/1/2008	10/1/2008	12/1/2009	1/1/2013	1/1/2018	1/1/2019
Small-cap	'Kotak small cap fund'	32.17	15.38	21.31	30.71	68.57	69.58
Mid-cap	'BNP Paribas midcap fund-G'	15.92	6.89	7.96	12.62	37.83	31.23
Large-cap	'Aditya Birla Sun life focused equity fund'	23.71	14.41	20.16	26.51	58.48	56.64
Large and mid-cap fund	'DSP equity opportunities fund'	93.32	51.52	74.2	92.73	231.4	207.75

2. It is recommended to go for 'DSP Midcap Regular Plan', 'HDFC Midcap Opportunities Fund', and 'Invesco India Midcap Fund'.

(d) **Review for 'ABSL Sun life Focused Equity Fund':**

It is a 'Large-cap' fund performing fairly well in the market. Its 5-year return is around 11.87% which is an average performance. Its competitors like 'Mirae Asset Large Cap Regular Plan' and 'Reliance Large Cap' are performing well above the market average, i.e., giving around 16.09% over 5-year basis and 16.62% return on 3-year basis. 'Reliance Large Cap fund' and 'SBI Blue-chip fund' performing quite good giving around 14% return on 5-year basis, i.e., well above market-benchmark

1. Alpha is around −1.97, i.e., under-performing, around 2% below benchmark and risk adjusted return is not great. Beta value is 0.85, i.e., 15% below benchmark value of market volatility.
2. It is recommended to go for "Mirae Asset Large Cap Fund-Regular Plan, Reliance Large Cap fund, and SBI Blue-chip fund".

After analysis, it is being found that the portfolio of NRI client as per Table 2 and Fig. 3 is 'Moderately Aggressive' profile as major percentage of investment is in 'Small Cap Funds', i.e., 56% where investment in Mid-cap funds is 24% and Large Cap fund is around 20%.

(e) **'Indian Mutual Fund Industry':**

- Average Asset under Management (AUM) had a sum value at ₹24.58 Lakh Crore, i.e., INR 24.58 Trillion till March 2019 which has increased by 1.32 Trillion YoY.
- 'Asset Under Management (AUM)' as on '31st March 2019' is at ₹2,379,584 Crore.

Table 2 Percentage allocation of the mutual funds

Types of funds	% of Allocation	Total (Rs.)
Small cap fund	56	823,456,780
Mid-cap fund	24	356,789,456
Large cap fund	12	178,958,900
Large and mid-cap fund	8	114,467,890

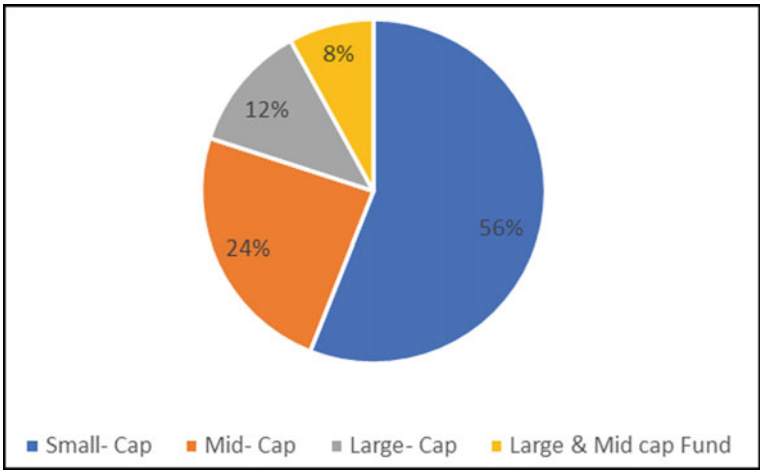


Fig. 3 Percentage allocation of mutual funds

- AUM of 'Indian Mutual Fund Industry' has rose >5½ fold, i.e., from ₹4.17 Trillion on 31 March to ₹23.80 Trillion on 31st March 2019 and increased 3 times compared to ₹8.25 trillion as on '31st March 2014'.
- AMFI data also tells us that AUM of the Industry have crossed the benchmark level of ₹10 Trillion.
- Sum of all Folios or accounts currently stands at 8.25 Crore till March 31, 2019, while the count of folios under 'Equity, ELSS and Balanced schemes', where the topmost investment is from 'Retail segment' is at 6.93 crore (69.3 million). This is 58th consecutive month experiencing the rise in the count of folios.

(f) **Industry QAAUM Growth:**

- It is being observed that there is a sharp decline in QAAUM in Q3FY19, but in Q4FY19, QAAUM increased by 4%.
- SIP Contribution moves ahead of Q4FY19 in comparison with Q3FY19 as per Table 3.
- The growth trend of Top AMC in Q4FY19 is being highlighted below.

Table 3 AMC ranking versus growth rate

AMC ranking	Growth rate
Top 5	−1 to 7%
Next 10	−1 to 15%
Rest	−29 to 29%

(g) **Trend of highest investment in 'Equity Category':**

- As per SEBI data, folio count in this category increased '2.7% QoQ and 1.56% YoY in Q4FY19'.
- Total 21.5 lakhs folio is a new addition to this quarter, whereas 16.4 lakh is added in this '**Equity Category**' (including ELSS), other ETF and Income category which stays at the highest position among all.

(h) **Share of Assets in small town:**

- Investors beyond top 30 cities (B30 locations) tend to rely more on distributors rather than direct online transfer. Smaller towns of the countries accounted for 15.4% of total industry AAUM that increased 15.1% YoY in Q4FY19.

(i) **Net Inflow from investors:**

- This Mutual Fund industry registers net inflow of Rs. 23,000 Crore in Q4FY19 which has decreased both YoY and QoQ.
- As Mutual Funds remain attractive to retail investors much more, maximum inflow was seen for 'Equity Funds'.
- Net inflow from 'ELSS and Equity' stood at Rs. 23036 Crore in Q4FY19 versus Rs. 27642 Crore YoY.
- Tight liquidity due to NBFC Crisis and sharp increase of global oil price and trade war still a concern for market. After establishment of stable NDA Government market expectations are high and we can expect good prospects.

(j) **Sector-wise pattern analysis:**

Top 5 sectors as highlighted in the below graph depict that together those contribute 50% of total Equity AUM.

- AMCs continue to look for BFS sector for reduction of NPA and good recovery of earnings.
- Petroleum sector witnessed the highest YoY inflow growth due to attractive valuation.
- Due to fragility in rupee, we can observe the highest exposure to Software Sector.

(k) **Fund-wise performance analysis:**

- 'Mid-Cap and Small-Cap' funds did not perform much as their performance was below the expected average level. In some cases, they gave negative return in FY19.
- Large-Cap Fund ruled in the 'Equity Segment'.
- 4.9% to 6.1% on an average return was generated by the aggressive hybrid funds in last 1 year Which Was Modest Compared to Long-Term Averages.
- Return of aggressive hybrid funds fell well below the return of debt or equity funds.

(l) **Performance of Debt Funds:**

- 'Long term Debt Funds' are in the top in the category of debt funds in the last 6 months.
- 'Gilt fund performed the best in last 6 months.'
- Purchases of 'Open Market Notes' by the RBI and 'a rate cut' by the 'Monetary Policy Committee' in Mar-19 'outweighed losses due to fiscal slippage anxieties'.

(m) **Investment Statistics in Mutual Fund:**

An approximate 2.62 'Systematic Investment Plan (SIP)' accounts are used currently by Mutual Fund investors. Till March 2019, AUM is approximately Rupees 8055 Crore. According to AMFI data, '9.13 Lakhs SIP accounts on an average each month added by Mutual Fund Industry in FY 2018–19 as per Table 4 with expected SIP Size of per SIP account is ₹3070.'

Category wise AAUM Break-up:

- "**Liquid and Money-Market Funds**" constitute Debt-Categories which accounts for 54% of Mutual Funds assets-which depicts huge "**Risk-Averseness**" of investors.
- Equity Category contributes nearly 42% of the asset base.
- Share of ETF still below 5%.

State-wise AAUM Break-up:

- **Maharashtra** continues to have largest contribution in the investments of Mutual Funds.

Table 4 SIP contribution over months

Month	SIP contribution ₹ core		
	FY 2018–19	**FY 2017–18**	**FY 2016–17**
Total during FY	**92,693**	**67,190**	**43,921**
March	8,055	7,119	4,335
February	8,095	6,425	4,050
January	8,064	6,644	4,095
December	8,022	6,222	3,973
November	7,985	5,893	3,884
October	7,985	5,621	3,434
September	7,727	5,516	3,698
August	7,658	5,206	3,497
July	7,554	4,947	3,334
Jun	7,554	4,744	3,310
May	7,304	4,584	3,189
April	6,690	4,269	3,122

- **Top 5 states** as highlighted in the graph having share of 70% of Mutual Funds Assets.

5 Results and Discussions

'Risk Profiling' as well as proper 'Asset allocation' is essential. Along with that, investors should keep in mind that they should not blindly follow the 'Yesterday's Winners' and, they should not get influenced by recent past, completely forgetting the distant past to remain free from Bias before investment. Attachment or staunch belief to some information, i.e., valid today about particular stock is hazardous. Moreover, as per Fig. 4, investors should also know stocks of big companies do not necessarily perform well every time. Emotions should be managed while investing as well as investors should not seek point of view of other investors. Concept of 'Model Portfolio' for Individual Investor where we segregate three different types of Portfolios like—Conservative Portfolio, Aggressive Portfolio, and Moderate Portfolio.

There is no particular % or limit for investment portfolios. It depends upon one's risk appetitive, income, and financial awareness. It is evident that from age 45, equity% should decrease and debt% should increase in the portfolio of investors because of 'mortality risk' and for stabilizing their money. For the benefit, model portfolio and allocation of funds are presented for different segments of people:

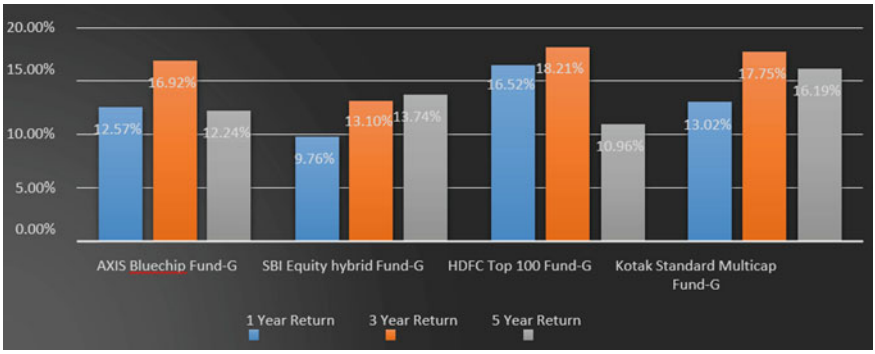


Fig. 4 Comparison of return of mutual funds

(a) **Model Portfolio:**

For every distinct individual investor, there should be a model portfolio. A Certified Planner helps in that. Though there is existence of no perfect 'Model' as per Fig. 5, as this depends on investors, risk present in those circumstances and very subjective. Theoretically, as per Table 5, model portfolio is illustrated for investors based (in different categories).

This percentage of allocation is subject to vary based on circumstances, risk-appetite, investment horizon, client profile, and their natures.

(b) **Recommendations in SIP:**

Based on Risk-factors and contribution to SIP, we have divided as per Tables 6 and 7 investors into three separate categories:

1. Conservative (Rs 1000–4000, Rs 4000–10000, Rs 10,000 above)

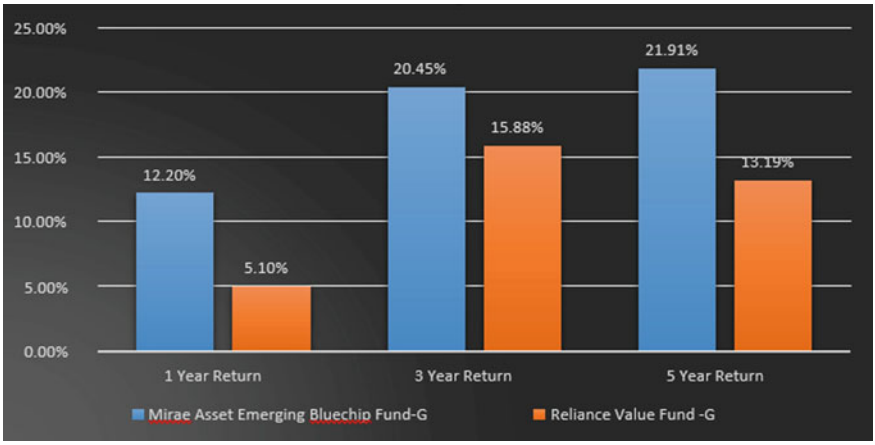


Fig. 5 Return comparison in 5 years

Table 5 Model portfolio

Category	Model portfolio
'BPS Call Centres'	50% of the corpus should be in 'Diversified Equity Schemes', 20% in 'Sector funds', 10% in 'Gold-ETF', 10% in 'Liquidity Schemes', and rest in 'Diversified Debt-Schemes'
'Young-married, single-income family with 2 school-going kids'	35% of the corpus should be in 'Diversified Equity Schemes', 10% in 'Sector Funds', 15% in 'Gold ETF', 30% in 'Diversified Debt Schemes', rest in 'Liquid Schemes'
'Single income family with grown up children and yet to settle down'	"'35% in 'Diversified Equity Schemes', 15% in 'Gilt Fund', '15% in Gold-ETF', '30% in diversified Debt Fund', 10% in 'Liquid Schemes'"
'Couples in Seventies with no family support'	'15% in Equity Index Schemes', '10% in Gold ETF', '30% in Gilt Funds', '30% in Diversified Debt Funds', '15% in Liquidity Schemes'
'Young well settled unmarried independent individual'	'80% can be invested in Equities', 10% in Gold-ETF, and rest 10% in diversified 'Debt-Schemes'

Table 6 Weightage percentage allocation in conservative portfolio

For conservative	Profile	(5 years)
SIP amount	**Scheme name**	**Percentage**
1000–4000	'AXIS Blue-chip-G'	50
	'HDFC Top 100 Fund-G'	50
>4000–10,000	'HDFC Top 100 Fund-G'	20
	'AXIS Blue-chip-G'	30
	'SBI Hybrid Equity Fund-G'	50
>10,000	'Kotak Standard Multi-Cap Fund-G'	10
	'HDFC Top 100 Fund-G'	25
	'AXIS Blue-chip-G'	50
	'SBI Hybrid Equity Fund-G'	15

2. Moderate (Rs 1000–4000, Rs 4000–10000, Rs 10,000 above)
3. Aggressive (Rs 1000–4000, Rs 4000–10000, Rs 10,000 above)

(c) **Schemes Suggestion:**

As per Table 8, the below schemes are suggested:

i. 'HDFC Top 100 Fund-Growth'
ii. 'Mirae Asset Emerging Blue-chip Fund'
iii. 'Axis Blue-chip Fund'

Table 7 Weightage percentage allocation in moderate portfolio

For moderate	Profile	(5 years)
SIP amount	**Scheme name**	**Percentage**
1000–4000	'AXIS Blue-chip-G'	65
	'HDFC Top 100 Fund-G'	35
>4000–10,000	'HDFC Top 100 Fund-G'	35
	'AXIS Blue-chip-G'	40
	'SBI Hybrid Equity Fund-G'	25
>10,000	'Kotak Standard Multi-cap Fund-G'	20
	'HDFC Top 100 Fund-G'	30
	'AXIS Blue-chip-G'	35
	'SBI Hybrid Equity Fund-G'	15

Table 8 Weightage percentage allocation in aggressive portfolio

For aggressive	Profile	(5 years)
1000–4000	'AXIS Blue-chip-G'	50
	'Kotak Standard Multi-cap Fund-G'	50
>4000–10,000	'HDFC Top 100 Fund-G'	15
	'AXIS Blue-chip-G'	20
	'Kotak Standard Multi-cap Fund-G'	30
	'Mirae Asset Emerging Blue-chip Fund'	35
>10,000	'Reliance Value Fund'	15
	'HDFC Top 100 Fund-G'	35
	'SBI Hybrid Equity Fund-G'	10
	'AXIS Blue-chip-G'	10
	'Mirae Asset Emerging Blue-chip Fund'	30

iv. 'SBI Hybrid Equity Fund'
v. 'Kotak Standard Multi-Cap Fund'
vi. 'Reliance Value Fund'

(d) **Valuation:**

As valuation is high now due to upliftment in the market as 'NDA Government has returned with thumping majority,' it is recommended not to go for lump sum investment. It is advisable to go for 'Systematic Investment Plan (SIP)'. If investors are looking for 'Mutual Fund', one should watch out for its expense ratio as well as AUM of the fund; it also implies heftier fees for the fund house and vice versa for

the smaller AUM. Here, also Moderate as well as Aggressive Portfolio is analyzed along with analysis of its assets returns. Debt funds rely on their AUM to manage its return and dividends to investors. Large Fund size generally implies lower expense ratio/investor, reflected in fund returns.

Large AUM affects small-cap and mid-cap as small-cap companies always look for hefty growth potential companies as well as these funds inhibit cash inflow beyond a certain level as AUM crosses threshold value. For mid-cap companies, main target is the liquidity factor, and accordingly AUM growth can be accommodated by fund managers.

6 Conclusion and Future Scope

(a) **From the perspective of AMC:**

- AMC and Brokers should look for the emerging cities like '**Jharkhand, Bhubaneswar, Bhopal, Lucknow, Patna, Ludhiana, Nagpur, Kanpur**' which have greater investment potential, and therefore the awareness program should be started in these cities.
- In '**West-Bengal**' net AUM percentage increased in cities like '**Burdwan, Asansol, Siliguri, Durgapur, Howrah**', hence we should look for more clients in those areas along with increasing awareness program for investors.
- It has been found that '**Maharashtra, Karnataka, and Gujarat**' have already saturated with AMC. AMC branches should be increased in these areas and more campaigning should be there in these states.
- It is being advised that AMCs should never neglect North-eastern states like **Assam, Arunachal Pradesh, Sikkim, and Tripura**. There are much **lesser number of branches of AMC in these states**.
- **Chhattisgarh, Orissa, Jharkhand** are basically '**Industrial Belt**' as well as these states have plenty of natural resources like Coal, Uranium, etc. But, as it can be seen from the above data, investments are pretty low which have a good prospect of emerging market of Mutual Funds. So, Mutual Fund companies and Fund-Managers start aiming training and campaigning these states so that more investors gain interest there.
- '**Madhya Pradesh**' has a lot of prospects in various sectors like 'Mining', Tourism and as well as it has good connectivity due to its Geographical position, so it can be a prosperous destination for corporates. But we can see from the data that Market—Penetration is not up to the mark. So, Mutual Fund companies can get "Prime Mover" advantage if they target better campaigning in these states and make them aware about SIPs.
- '**Kerala, Telangana, Andhra Pradesh, and Punjab**' have a very 'high immigration rate', which ensures lots of inflow of Foreign-Currency which makes a prospect of good emerging market for investments.

(b) **From the Investor Perspective:**

- Investor should not look for short-term perspective. They should concentrate on '**Long-Term**' perspective. All the biases mentioned earlier should be eliminated based on awareness program.
- **Financial goal** should be clear and viable.
- 'Investment is a continuous process' and investors should pursue investing seeing '5–10 years' performance and also take recommendations from Research Analyst seeing long-term perspective.
- Risk should be taken 'to achieve maximum return' as Mutual Funds are always subject to market risk.

Investor should invest money in Mutual Funds, Stocks, and Bonds keeping a certain amount as liquid money for Emergency planning. If the person is not financially literate enough, he can seek advice from "Financial Advisors", "Research-Analysts", "Websites" like "Value Research", "Morning Star", and" Groww" app is beneficial. In this research, it is highlighted mainly on the investment on Mutual Funds, on biases of investors, and how to eradicate that. For that, AMC should conduct awareness program to eradicate bias among common investors and spread awareness on the beneficial signs of 'Mutual Funds' in different parts of India and emerging cities like Jharkhand, Bhopal, and Lucknow, etc. Along with that, awareness of investors on the market risks is very much needed and they should be free from all the biases.

References

1. Jiang G, Xu D, Yao T (2009) The information content of idiosyncratic volatility. J Financ Quant Anal 44(1):1–28. https://doi.org/10.1017/S0022109009090073
2. Doukas JA et al (2006) Divergence of opinion and equity returns. J Financ Quant Anal 41(3):573–606. JSTOR
3. Baltussen G, Van Bekkum S, Van der Grient B (2018) Unknown unknowns: uncertainty about risk and stock returns. J Financ Quant Anal 53(4):1615–1651. https://doi.org/10.1017/S0022109018000480
4. Bali T, Murray S (2013) Does risk-neutral skewness predict the cross-section of equity option portfolio returns? J Financ Quant Anal 48(4):1145–1171. https://doi.org/10.1017/S0022109013000410
5. Margarida A (2019) How biased is the behavior of the individual investor in warrants? Res Int Bus Financ 47(C):139–149
6. Huang D, Schlag C, Shaliastovich I, Thimme J (2019) Volatility-of-volatility risk. J Financ Quant Anal 54(6):2423–2452. https://doi.org/10.1017/S0022109018001436
7. https://www.icraonline.com/mfanalytics.html; https://www.valueresearchonline.com

Chapter 8
iMask—An Artificial Intelligence Based Redaction Engine

Shobin Joyakin

Abstract The industrial use cases in Computer Vision and Artificial Intelligence have seen a lot of progression in the last few years. The advancements in the processing power of machines and the availability of data have its own share for many successful implementations in Computer vision technology. The tightened regulations and restrictions like GDPR on the usage of personal data of customers have increased the complexity of how we ship, store, and process these data in images. In these circumstances, the Personally Identifiable Information (PII) of customers like Date of Birth, SSN, PAN card#, Mobile number, Aadhaar number, etc. that are shared in image format as additional proofs for applications need to be safe-guarded from identity theft and misuse. iMask is an Artificial Intelligence-based framework/tool that has a great potential to identify such Personally Identifiable Information from variety of multiple images automatically using Computer vision technology to either encrypt or redact them from original image. These redacted images can be shipped outside its controlled environment across geographies and use them in non-production like environment for any further backend processing. This automated approach to remove the personal information from images ensures no to very less manual intervention for masking them.

Keywords PII data · Redaction · Computer vision · Artificial Intelligence · Personally identifiable information

1 Introduction

One of the North American-based Insurance companies was tasked with digitizing its 400 K odd historical Medical lab reports for a study of the reports. This study would enable them to build to an AI-based Life insurance product. As part of the digitization project, the requirement was to build an ML-based OCR solution that can extract the contents from image documents into a structured format. The objective is to develop

S. Joyakin (✉)
TATA Consultancy Services, America, Columbus, OH, USA
e-mail: meet_shobinjoy@yahoo.co.in

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_8

and test the ML-based OCR solution in a dev like region. It was observed that the customer was spending manual efforts to mask these historical image documents (non-searchable PDFs) using PDF editing software as the local Data privacy act did not allow them to share the PII data of customers outside their geography. We noticed considerable delay in getting the masked images for our extraction process as the manual masking process took time and was erroneous. The delay in getting the masked documents impacted the project timeline. This delay made us to rethink on how we could help customers automate their manual efforts and meet our project deadline by building an AI solution to mask the PII data automatically.

Computer vision technology enables a machine to scan through different kinds of images, videos, live recordings and understand the different patterns hidden in them. These patterns help a machine understand and classify what is being fed into it. However, for a machine to derive valuable insights with much better accuracy we need to train them with huge amount of training images that are annotated manually. Some of the popular free image annotation tools like labeling could be used to annotate the images with desired customized labels. However, annotating textual data using tools as these could be challenging if the images that contain the textual data have different templates. Manual annotation of the texts in variety of image templates would be time-consuming process and may not yield good accuracies.

Here we are trying to leverage few basic principles of image processing in Computer vision that would help us detect texts and identify the coordinates of textual information in images. These texts are extracted from its coordinates in the image and with the help of basic tools of Natural Language processing techniques we will identify if the text contains PII data. The goal is to identify specific PII information from images with texts and redact them with good accuracy using available open-source tools with less spend on cost.

2 Literature Review

Research on "An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition" [1] by Baoguang Shi, Xiang Bai, and Cong Yao has neatly detailed on using neural network-based approach called CRNN (Convolutional Recurrent Neural network) for recognizing sequence like objects in images. It's interesting to see how we could use Neural-network architectures to automate a lot of manual efforts. Most neural network need a lot of data to be fed in to train the network with an exception to transfer learning where we have pre-trained models available on similar dataset. Here we are exploring other techniques in computer vision when we do not have enough labeled dataset.

3 Methodology

The high-level design for the framework consisted of the below procedures as shown in Fig. 1.

For Text detection, some of the possible solutions are.

a. Sliding window technique
b. Single shot techniques (SSD, YOLO)
c. EAST (Efficient accurate scene text detector) [2]
d. OpenCV's inbuilt contour detection.

For Text recognition, the possible solutions are.

a. CRNN (Convolutional Recurrent Neural Network)
b. Tesseract OCR or any other OCR techniques [3].

Problems as these do not have any labeled images with PII data nor any pre-trained models for us to reuse for Machine training. So, the approach taken to solve the solution is to leverage the capability of OpenCV to detect text in image documents. The detailed design implemented by this redaction engine is shown in Fig. 2.

Below steps were performed to effectively capture the texts from image documents.

A. **Image Pre-processing** [4]:

Here an attempt is made to clean the images with certain degree of noise so that the texts in the images can be easily identified. Below procedures were followed for the cleaning.

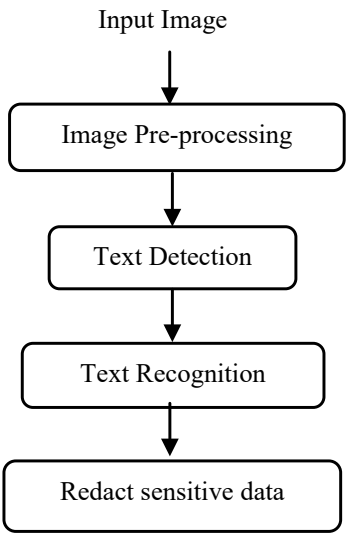


Fig. 1 High-level design

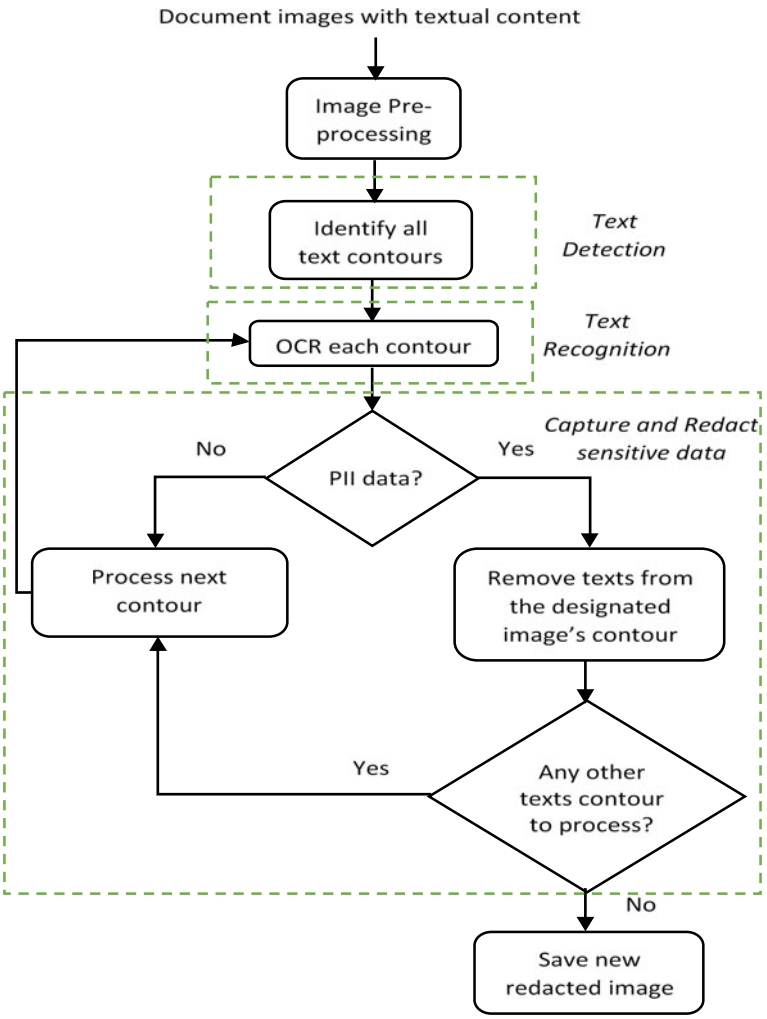


Fig. 2 Detailed design

i. **Grayscale conversion**:

Convert the image into a grayscale to standardize the input into a single channel as shown in Fig. 3.

ii. **Image blurring:**

A two-dimension low-pass filter (LPF) or kernel can be convolved on the image to remove noise or blur the image. This works by removing high-frequency contents like edges from the images. Some of the common blurring techniques used are.

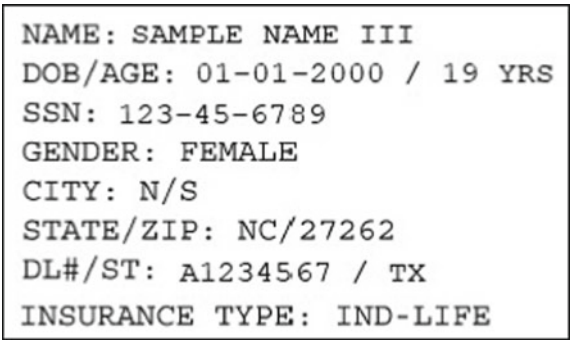


Fig. 3 Grayscale conversion sample output

- **Averaging**:
  Convolve the image with a normalized box filter and replace the center element of the image by the average value.
- **Median**:
  Compute median of all pixels in the image under the kernel and replace the center element of the image by the median value.
- **Gaussian**:
  The values of the Gaussian kernel don't have equal coefficients but rather consist of values based on the standard deviation mentioned while creating the filter. It is good at removing Gaussian noise in data.

  We will focus on smoothening the image using Gaussian filter as shown in Fig. 4.

iii. **Morphological operation (black hat)** [5]:

One of the morphological operations called black hat is used to reveal black region in a bright background of the image. A sample output is shown in Fig. 5.

iv. **Image gradient** [6]:

The Sobel operator is a relatively inexpensive computation for discrete differentiation operation. This is more resistant to noise. The direction of the derivatives can be

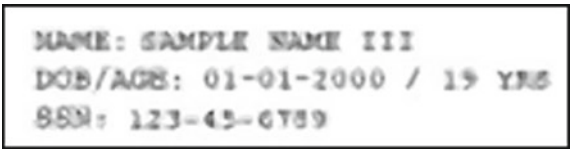


Fig. 4 Blurred sample output

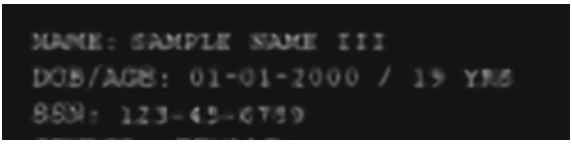


Fig. 5 Blackhat sample output

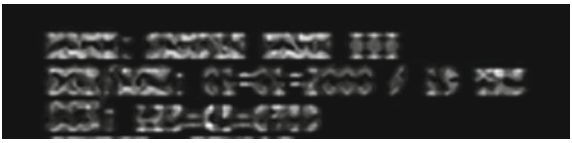


Fig. 6 Image gradient sample output

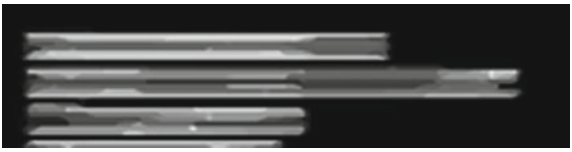


Fig. 7 Closing operation sample output

chosen in either x or y direction based on the texts in the image. This will show the vertical and horizontal changes in the gradient as shown in Fig. 6.

v. **Morphological operation (closing)**:

This operation involves dilation followed by erosion. This will help join nearby characters that are part of the same word into a single object in the image as shown in Fig. 7.

vi. **Thresholding**:

Threshold the image using OTSU's binarization. Here an image is considered as bimodal before automatically calculating a threshold value thereby minimizing the weighted within class variance as shown in Fig. 8.

vii. **Erosion and Dilation**:

For the final cleaning of the word components, a series of iterative erosion and dilation of the image is performed to remove any noise blobs and break apart any components that were incorrectly connected during closing operation as shown in Fig. 9.

B. **Text Detection using OpenCV**:

We will leverage *findcontours* functionality in OpenCV to detect texts from the images. Contours are curves joining all the continuous points along the boundary having same color or intensity in an image. We will remove all redundant points and compress the contour by only saving the four edges of the contours. This will help

Fig. 8 Thresholding sample output

Fig. 9 Erosion and dilation sample output

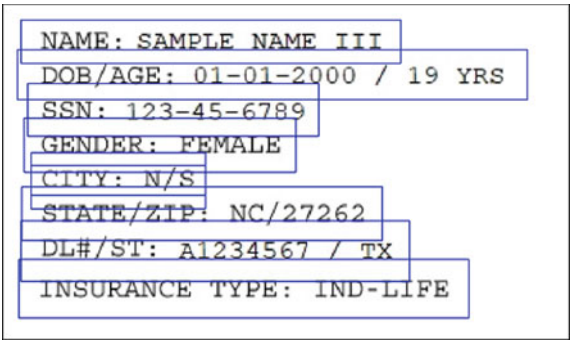


Fig. 10 Text detection sample output

to identify the Region Proposals (RPs) or Region Of Interest (ROI) that contain texts in the image as shown in Fig. 10.

C. **Text Recognition using OCR** [7]:

Recurrent Neural Network (RNN) is a branch in deep neural network that was built to handle sequences of objects. Long Short Term Memory (LSTM) is a kind of RNN-based model that is able to learn long-term dependencies of sequences in text that a vanilla-RNN was not able to do accurately [8].

Each of the region proposals is passed through a neural-nets-based Optical Character Recognition (OCR) engine called Tesseract (Version 4.0+) built on LSTM [9]. This version of OCR engine is focused online recognition and not just classical character patterns unlike its predecessors. Thereby the model would be able to predict the next characters as it reads the patterns. Hence, it outperforms its legacy models by many times. Current version of Tesseract is trained on over 100 languages including Indian languages (like Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Sanskrit, Tamil, Telugu, Urdu).

"An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition" [1] by Baoguang Shi, Xiang Bai, and Cong Yao has neatly detailed on using neural-network-based approach called CRNN (Convolutional Recurrent Neural network) for recognizing sequence like objects in images. This is another approach we could use for text recognition; however, we chose the Tesseract-based solution as it is:

a. Light weight computer vision solution.
b. New training of models is not required unless OCR performance is very poor.

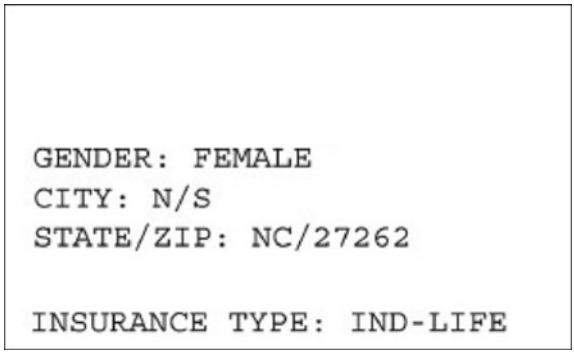


Fig. 11 Redacted image sample output

c. Building Deeper neural networks is not required as we are using existing deep network models called LSTM.
d. Do not need GPU machines.
e. Lexicon free.

D. **Capture sensitive data using Text processing**

Natural language processing of texts based on context enables us to create entities for each word using NER (Named Entity Recognition) technique. This could help us identify the sensitive data given a NLP model is trained to recognize the new entities like PII data. Spacy NER Annotation tool [10] is one such tool that can be used to train new entities in a text document.

Here we will try a simpler approach as we are working on template-based forms and word pattern recognition using regular expression (regex) would be a better option. Regular expression is a classical approach that can pick specific word patterns from text documents. We can also build complex regular expressions that could read text with possible errors in OCR output and still identify sensitive data in the image. The output from OCR engine is fed to the word pattern recognizing function that will flag the contour if it contains any sensitive information.

E. **Image redaction**

The contours that are flagged for sensitive information will be overlaid with a white rectangular box and the image can be saved as final as shown in Fig. 11.

4 Results

Some of the notable findings of this work is:

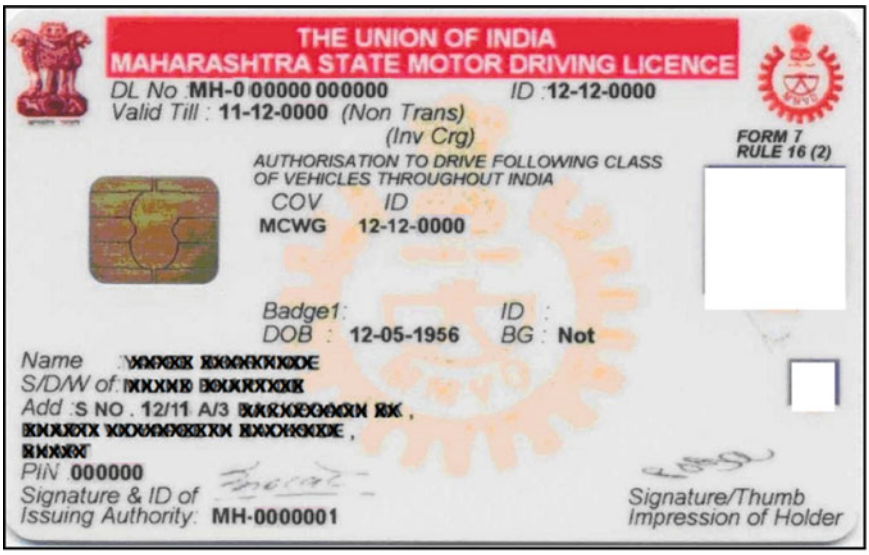
The current technique can also be used to mask sensitive handwritten texts from images as well provided the labels like Name, DOB are typed words. The detected contours of typed text should include the handwritten text and the OCR should be able

to extract the typed labels accurately so that the word patterns can easily recognize the PII data.

Based on testing done, there are no restrictions to the color of the image on which sensitive data are to be removed. Basic preprocessing like rescaling of images, converting to grayscale, removal of watermark can help standardize the images. We can identify sensitive information if the data is present below a label as well using further customization of filters.

We were able to expand the solution to mask the PII data from other images like Passport, Driving License, Hospital records, etc. Some sample input and outputs are shown in Fig. 12a–d.

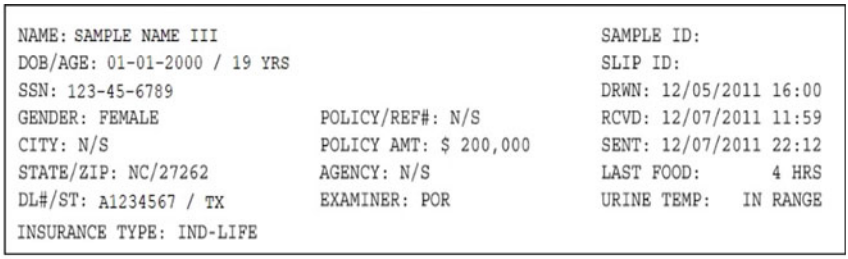
a) Before masking:



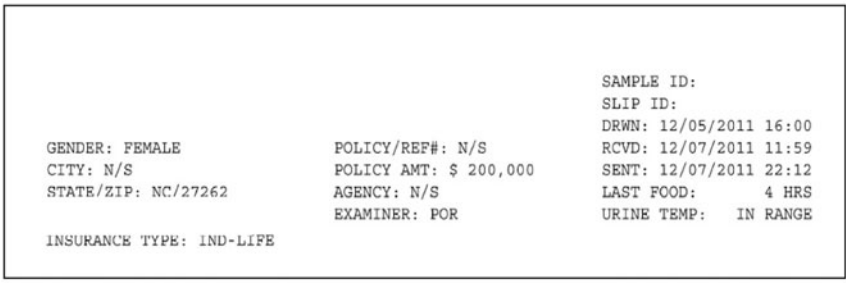
After masking:

Fig. 12 **a** Comparison of an Indian Driving License before and after masking. **b** Comparison of a Medical Lab report before and after masking. **c** Comparison of a Medical Lab report before and after masking. **d** Comparison of a Discharge Summary before and after masking

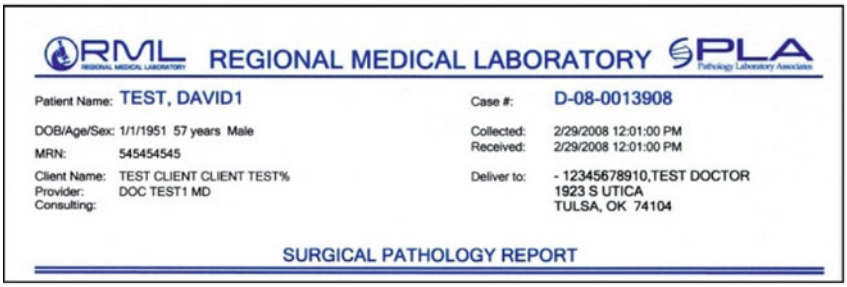
b) Before masking:



After masking:



c) Before masking:



After masking:

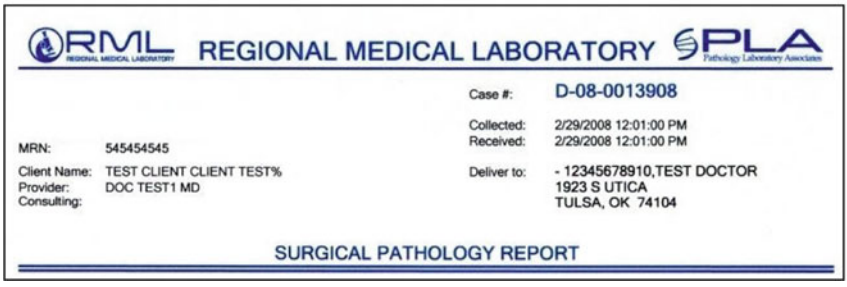
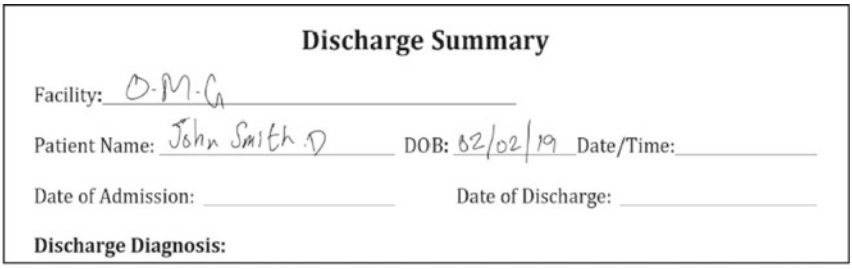


Fig. 12 (continued)

d) Before masking:



After masking:

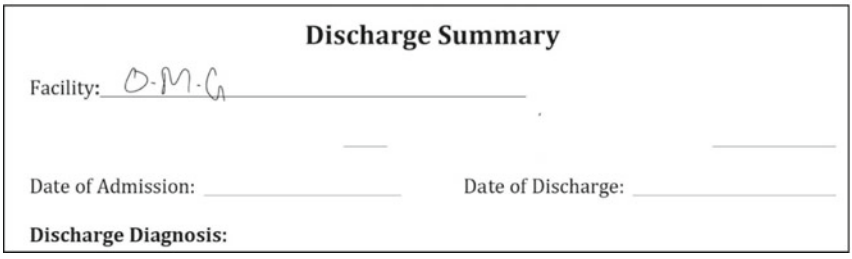


Fig. 12 (continued)

Evaluation Metrics:

In Table 1, the results from iMask are compared against EAST model for a set of different images to redact the sensitive data. It is found that the technique used in iMask is performing way better than the pre-built model in EAST.

Current evaluation of the metrics was done manually based on the texts that were captured in the region proposals and the effectiveness of the PII redaction. Two parameters were used to evaluate the effectiveness of the tool.

(a) **Redaction Precision**:

This evaluates out of all data redacted from Region Proposals (RPs), how many PII data were redacted. The basic idea is to reduce False Positives in the redaction, i.e., we do not want any non-PII data to be redacted.

$$Precision = \frac{PII\ data\ in\ the\ redacted\ RPs}{All\ the\ redacted\ data}$$

Example If three data labels were redacted from the image (like Name, SSN, Bill number) and only two of them contained PII data (Name, SSN), then the precision of the tool will be $2/3 = 0.67$.

(b) **Redaction Recall**:

Table 1 Comparison chart

Document Type	Test File name	Image size (W x H)	Find Contours technique in OpenCV (iMask) (Scores: 0–1)			EAST model (Scores: 0–1)
			Text detection accuracy (by word block) using Find contour in OpenCV	Redaction-Precision	Redaction - Recall	Text detection accuracy (by word block)
Medical Lab report	8_Medical_report.pdf	5100 × 6600 (B&W)	0.95	1.00	1.00	< 0.10
	9_Medical_report.pdf	5100 × 6600 (B&W)	1.00	1.00	1.00	< 0.40
	10_Medical_report.pdf	5100 × 6600 (B&W)	1.00	1.00	1.00	< 0.40
Pathological report	0_RML.jpg	800 × 1101 (Color)	0.80	1.00	1.00	< 0.20
	1_RML.jpg	492 × 640 (Color)	0.83	NA*	0.00	< 0.20
	2_RML.jpg	800 × 1040 (Color)	0.76	1.00	1.00	< 0.20
	3_RML.jpg	800 × 1040 (Color)	0.76	1.00	1.00	< 0.20
	4_RML.jpg	800 × 1101 (Color)	0.81	1.00	1.00	< 0.20
Driving license	11_India_Driving_License.jpg	1016 × 644 (Color)	1.00	1.00	1.00	< 0.80
	12_India_Driving_License.jpg	800 × 509 (Color)	1.00	1.00	1.00	< 0.80
	13_India_Driving_License.jpg	448 × 285 (Color)	0.96	NA*	0.00	< 0.80
	14_India_Driving_License.jpg	526 × 339 (Color)	0.88	1.00	1.00	< 0.80
	15_India_Driving_License.jpg	1186 × 622 (Color)	0.88	0.50	1.00	< 0.30
Passport	16_Indian_Passport.jpg	768 × 1056 (Color)	0.71	1.00	1.00	< 0.80

(continued)

Table 1 (continued)

Document Type	Test File name	Image size (W x H)	Find Contours technique in OpenCV (iMask) (Scores: 0–1)			EAST model (Scores: 0–1)
			Text detection accuracy (by word block) using Find contour in OpenCV	Redaction-Precision	Redaction - Recall	Text detection accuracy (by word block)
Discharge Summary	5_Discharge Summary.pdf	5100 × 6600 (B&W)	0.10	0.50	1.00	< 0.10
	6_Discharge Summary.pdf	5100 × 10,200 (B&W)	0.67	1.00	1.00	< 0.10
	7_Discharge Summary.pdf	5100 × 6600 (B&W)	0.75	0.67	1.00	< 0.10
	17_Discharge Summary.pdf	5100 × 6600 (B&W)	0.88	0.67	1.00	< 0.40
	18_Discharge Summary.pdf	5100 × 6600 (B&W)	0.90	1.00	0.50	< 0.40
Total Performance			**0.82**	**0.90**	**0.87**	**< 0.40**
NA*	No PII data were redacted					
Note	**Precision**	Out of all Redacted Region Proposals (RP), how many RPs have only PII data (Reduce False positives)				
	Recall	Out of all PII data in the image, how many PII data were redacted (Reduce False negatives)				

This evaluates out of all PII data in the image, how many PII data were redacted. The basic idea is to reduce the False Negatives in the redaction, i.e., we do not want to miss any PII data from redaction.

$$Recall = \frac{PII\ data\ in\ the\ redacted\ RPs}{All\ the\ PII\ data\ in\ the\ image}$$

To validate the results, we can save the redacted RPs as an image into another folder temporarily as we redact through the images.

Example If three PII data labels were present in the image (like Name, SSN, DOB) and only two of them were redacted (Name, SSN), then the recall of the tool will be $2/3 = 0.67$.

In most situations, it is recommended to improve the Recall value, but the best decision is taken through a discussion with the business stakeholders.

Some of the sample outputs from iMask tool and EAST model are compared below.

(i) **Driving License**

(a) Text Detection output using OpenCV technique used in iMask is shown in Fig. 13

Here all the texts were captured within the contours including adjacent words. This helps the tool to easily redact PII data from the images.

(b) Text Detection output based on EAST model is shown in Fig. 14.

Here almost all the texts were captured; however, additional techniques are required to connect two adjacent words for redaction.

(ii) **Medical lab report**

(a) Text Detection output using OpenCV technique in iMask is shown in Fig. 15

Here the text detection output is very good as it was able to detect all the labels along with its values.

(b) Text Detection output based on EAST model is shown in Fig. 16.

Here the results are below par and we are not able to detect most texts accurately. The technique used by iMask tool is a good option,

i. If we don't have enough labeled images to train a new model to detect the text's contours.

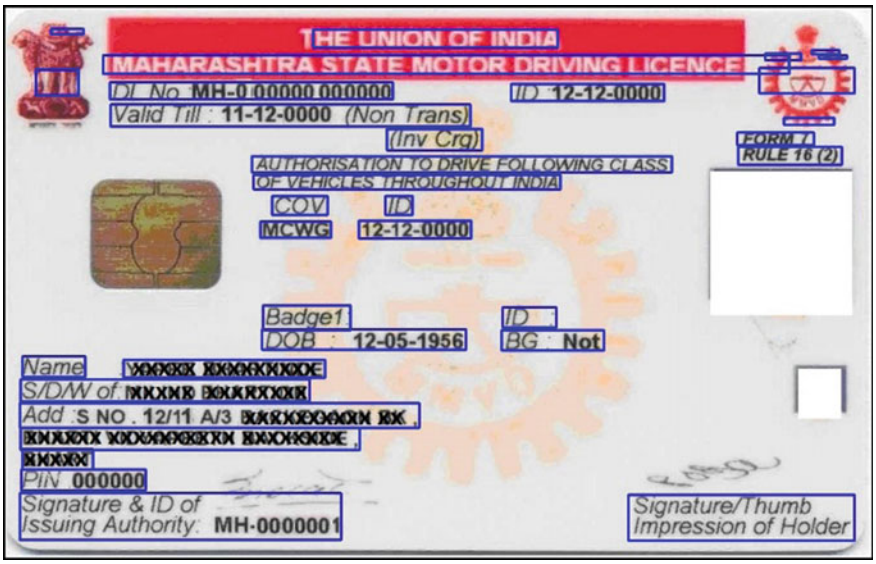


Fig. 13 Text detection by iMask on driving license

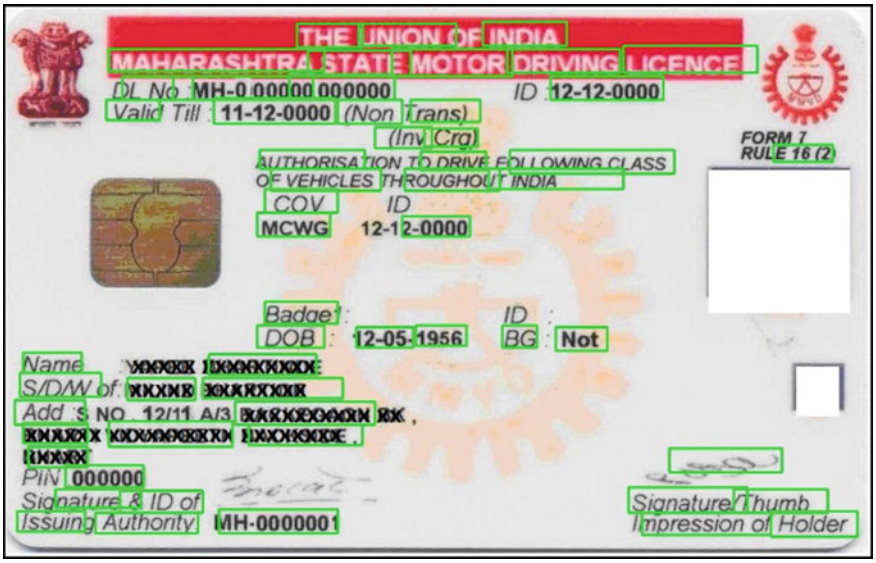


Fig. 14 Text detection by EAST model on driving license

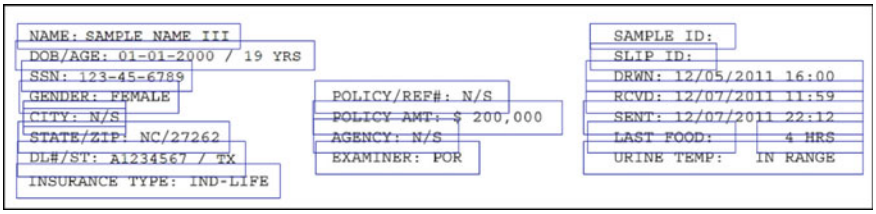


Fig. 15 Text detection by iMask on medical lab report header

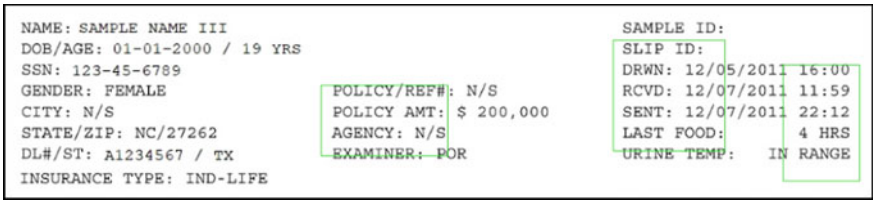


Fig. 16 Text detection by EAST model on medical lab report header

ii. If we don't have an existing model that was pre-trained on labeled images with contents having similar textual features as our images.
iii. If we don't have high-end machine with GPU capability to train models. As models like EAST needs GPU for model training.

The iMask tool was built on contour detection using OpenCV due to unavailability of labeled images.

This technique comes with few constraints as well.

i. Poor Image quality due to low resolution can impact the OCR output and the word pattern cannot be recognized to identify if a text contains PII data. This can cause False Negatives (Sensitive data being missed from redaction). For such scenarios, retraining OCR with bad images or building text recognition models like CRNN could help if we have labeled text data.
ii. Labels of PII data are required for the sensitive information to be captured in the images. Example: "DOB", "Data of Birth", "SSN", etc. needs to be present in the image to identify them using word patterns. Further fine-tuning of the regex pattern is required based on the contents in the image. If labels are not present, NER technique can be used to identify the sensitive data. Additional training on NLP models like spacy using a NER annotator would be required for this to give expected results.
iii. The current technique expects the labels of the sensitive data like "Name", "DOB", "Age", etc. to be in close proximity of its value and other textual data needs to be little apart so that the computer vision's customized filters will not capture them as sensitive data. If the non-sensitive text is present too close to the sensitive text in the image, then the kernel we defined in the tool will remove the non-sensitive text from the image as well. As the kernel thinks it is part of

the same word. This is termed False positives. This is mostly noticed when the images are having way different resolutions. Separate kernels would be required for such images. Image classification models built on CNN layer can be built upfront to identify an input image before passing it through iMask tool to reduce this issue.

iv. It is assumed that all words would retain a uniform distance between them and across labels there would be more distance.

5 Conclusion

This framework will enable to identify sensitive text information in the images faster and with better accuracy. We are relying on Tesseract's OCR technology for extracting texts from the images, as there has been great advancement with the latest 4.0+ version that introduced LSTM-based sequence modeling for recognizing words. For template-based forms, regular expression is a good start. This can be enhanced with Entity recognition models in NLP to identify sensitive data, and they can be removed from the image.

The future scope of this work will include redacting the sensitive data from an image when its labels are not present. This will enhance the tool to redact sensitive data from other documents like Aadhar card, PAN card, etc. that do not have a label for the sensitive data. The contours identified by this technique can be used to train a new model to identify PII data as future robust solution. Similar redaction of sensitive data can be done on document images like Bank statements, Payslips, and other financial documents, other country's Driving license, Passport, Visa, etc.

References

1. Baoguang Shi, Xiang Bai and Cong Yao - An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition https://arxiv.org/pdf/1507.05717.pdf
2. https://www.pyimagesearch.com/2018/08/20/opencv-text-detection-east-text-detector/
3. https://github.com/tesseract-ocr/tesseract/wiki
4. https://docs.opencv.org/4.0.0/d2/d96/tutorial_py_table_of_contents_imgproc.html
5. https://www.pyimagesearch.com/2015/11/30/detecting-machine-readable-zones-in-passport-images/
6. https://en.wikipedia.org/wiki/Sobel_operator#Alternative_operators
7. https://www.learnopencv.com/deep-learning-based-text-recognition-ocr-using-tesseract-and-opencv/
8. https://github.com/tesseract-ocr/tesseract/wiki/TrainingTesseract-4.00
9. https://colah.github.io/posts/2015-08-Understanding-LSTMs/
10. https://github.com/ManivannanMurugavel/spacy-ner-annotator

Chapter 9
Intrusion Detection System Using Signature-Based Detection and Data Mining Technique

J. Visweswara Iyer

Abstract In a corporate network, the security resources have upmost importance to the organization. In order to maintain security, different techniques are used such as firewall and intrusion detection system. Some of the most efficient detection techniques to find network intrusion implements artificial intelligence, machine learning, soft computing techniques, and bio-inspired techniques. But very often, some intrusion attempts are able to breach these defense mechanisms. So, a new secure network model is proposed which combines both signature-based detection and data mining technique which act as an efficient detection and response system against intrusion and is well established and evolved system. The proposed model combines signature-based detection and data mining technique to improve the efficiency of the intrusion detection system. In this model, a feature selection methodology is implemented whose objective is to reduce irrelevant features and also to detect features which will assist most to enhance the detection rate of intrusions. This is done based on the score of each features which is calculated in the selection process. This model also implements a recursive feature elimination and combines it with data mining technique such as the decision tree classifier to increase the detection rate of the intrusion detection. This model is implemented on the NSLKDD data set. Using this model, important features were recognized from the data set and the accuracy of detecting intrusion by the system was improved. Understanding and detecting the relevant or important features which correctly identifies an intrusion attempt will lead to the improvement in the design of intrusion detection system.

Keywords Intrusion detection system · Data mining · Statistics · NSLKDD · Decision Tree · Signature based detection · Machine learning

J. V. Iyer (✉)
"VISHNUPRIYA", Pulicot Street, Kollengode, Palakkad 678506, India
e-mail: jv.iyer@tcs.com

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_9

1 Introduction

In recent times, several organizations store their data in different ways with a single objective that is to safeguard their private and confidential data from both the external and internal attack. There also exists another possibility in which some authorized users will try to leak the confidential data of the organization for individual gains. Thus, in present condition, it has become challenge to detect the intruder because of the forged IP and attack packets being created. Methods which were used before such as firewall are unable to detect the real-time attackers which are carried out in the absence of the administrator and without their prior knowledge. A set of hardware and software are combined to form a computer network. The attacks in the software make the valuable data stored or in transit vulnerable. Those users who have prior knowledge about the programming language and the underlying resources can find out the various activities which are carried out by the systems using the log files pretty easily. Thus, they can aid in increasing security. The real problem is encountered when the users who doesn't have any knowledge of the programming language, and the underlying system and gets attacked by the hackers, thus they are unable to find out what the problem is. Even though there are many kinds of attacks, the most difficult one to be recognized is the insider/internal attack. The network security is an area which ensures that every user and his systems are protected from all the malicious and suspicious attacks that include both internal and external attacks. The external attacks by the attackers and the internal attacks by intruders can be identified by the Intrusion Detection System and thus these methods help the organization to protect their systems.

The security of the network in an organization is attained by implementation of various types of security software and hardware systems such as firewalls, Intrusion Detection Systems, and Honeypot. Software or programming-based Intrusion Detection Systems are trained using different training techniques and procedures. Learning incorporated in the Intrusion Detection System during training can be of two types such as conservative or non-evolutionary and evolutionary or progressive. Evolutionary also called as dynamic because the procedure of learning evolves or mutates with every new encounter of intrusion attempt, thus such learning can also be referred to as online learning. Intrusion Detection System utilizing these procedures are referred to as online learning-based Intrusion Detection System. But these online Intrusion Detection System are vulnerable to different learning attacks from the intruders. The intruder with some basic knowledge of the learning algorithm implemented in the defense mechanism will always attempt to modify the classification behavior of the algorithm and try to breach the organizations network security. So, it is necessary for any intrusion detection system to have a subsequent response system to protect the network resource and the infrastructure from further damage whenever the detection engine fails.

Since most of the intrusion detection systems are developed on the basis of the signature-based intrusion detection and these intrusion detection model relay on signature databases to detect intrusion attempts and other attacks, the process of

identifying the intrusions will be a hard or unfeasible job if the newest attack pattern or signature cannot be identified as quickly as it was developed. The other main problems faced by signature-based intrusion detection models are the dynamic update and the rapid improvements of the databases of known attack patterns and the creation of a correct frequency threshold value for intrusion detection. Identifying the relationship between a new packet's signature and the known signatures is contemplated as a hindrance task in identifying new attack signatures. These all drawbacks of signature-based intrusion detection model could lead to the compromise of the security of the network as a result an efficient model is required for the efficient and accurate detection of the intrusions in the network.

2 Literature Review

An Intrusion Detection System's (IDS) main objective is to analyze the network traffic for malicious activity and to generate alerts when such activities are detected. It is a software application which examines the network or a system for suspicious activity or policy violation. Any suspicious activity or breaches are usually reported either to an administrator or collected centrally using a security information and event management (SIEM) system. Although intrusion detection systems analyze networks for potentially suspicious activity, they are also vulnerable to false alarms. As a result, such strategies did not produce great results.

According to [1], a hybrid system which can precisely detect potential attack has been developed based on the very well-established and evolved detection and a response mechanism to pathogens of plants to increase the detection rate. The proposed Plant-based Inspiration of Response in Intrusion Detection System (PIRIDS) model comprises three layers which are divided into Pattern recognition receptors, Guard Agent and Systemic acquired resistance and Hypersensitive response. This proposed system is dynamic enough to include program agents to detect various known attacks. But the major disadvantage of this methodology is that it can be vulnerable to learning attacks. That is an intruder with some knowledge about the learning algorithm used in the system would try to bypass the security measures employed by the system.

In [2], a hypervisor level distributed network security structure is implemented on every computing server of the cloud computing structure. This includes a primary defense mechanism against the known attacks implemented using the signature-based detection. The suggested model also implements the binary bat algorithm which is extended with new fitness functions for identifying the suitable features from the cloud network traffic. The identified features are then fed to the Random Forest classifier in order to detect the intrusion attempts in cloud network traffic and if any such attacks are detected intrusion alarms are produced. The intrusion alarms from various servers are connected together which aims to detect the distributed intrusion 4 attempts and to model new attack patterns. The proposed methodology combines the signature based and anomaly detection techniques and are able to detect both old

as well as new attacks that is both known and unknown attacks. Also, the signature-based detection is done before the anomaly detection which results in minimizing the overall computation cost because the anomaly detection phase only has to analyze the traffic from the network only for unknown intrusion attempts.

In [3], a new framework is proposed which has a feature reducing technique for reducing the irrelevant features and then the supervised data mining models are applied on the network data set for an efficient, well ordered and precise identification of intrusion attempts in the networks. In this chapter, mainly two techniques for feature reduction are implemented, Canonical Correlation Analysis and Linear Discriminant Analysis. The canonical correlation analysis is implemented to find a relationship between two sets of variables by detecting a linear combination of variables that has maximum relationship between them, whereas in case of Linear Discriminant analysis it makes sure that the separation is maximum by finding the ratio of between-class variance to the within-class variance in any particular data. The system is verified over UNSWNB15 data set. This model also used random forest algorithm for classification and regression of network threats. But the overall performance of the system was limited due to the high false positives.

Both [4, 5] propose an intrusion detection system which can be implemented in the energy dependent mobile AD-HOC networks. The model in [4] proposes a challenge-based trust mechanism which is an efficient way to detect malicious or suspicious nodes in a collaborative intrusion detection system. This approach includes an improved sending strategy for the trust mechanism in a collaborative intrusion detection system but is highly vulnerable to learning attacks. While in [5] a selection scheme for the monitoring nodes called Event Triggering-Based Adaptive Selection is developed, which suggested a methodology of tracking the states of the nodes and also provides the procedure when an event is triggered. The proposed event triggering-based adaptive selection performs much better in terms of the network lifetime and stability. Both these models proposed in [4, 5] are constrained by the limitation of resources and energy and can be bypassed by any learning attacks.

In [6], a new model is proposed with an enhanced fuzzy min–max neural net5 work which aims to achieve a better detection rate of intrusion attempts in networks with minimum training and recall time. The fuzzy min-max network-based classifier makes use of the fuzzy sets as pattern classes in which each fuzzy set is a combination of hyper-boxes. This methodology proposes slight improvement in the recall phase of the enhanced fuzzy min-max neural network in order to increase the precision of detection when a number of large-sized hyper-boxes are made. The proposed model implements the Manhattan distance measure which is combined with the current characteristic function of the enhanced FMN which also improves the precision of identifying the intrusion attempts. The system is verified over KDD cup'99 corrected data set.

Zhang and Wang [7] presents an intrusion detection methodology to detect routing attacks or intrusion attempts, which results in removal or blocking of suspicious nodes in the network. This model implements an effective fuzzy clustering-based algorithm for intrusion detection of a MANET implementation in a cloud storage environment.

The architectural model proposed in this methodology is based on the Google App Engine platform. As a result, the implementation of this methodology was limited.

While in [8] a traffic prediction model is proposed which is implemented on auto-regressive moving average (ARMA) utilizing the time series data. This methodology can efficiently and accurately forecast network traffic and was used as a primary defense mechanism in automated industrial networks. These methods failed to propose a centralized solution and was limited in their implementation.

In [9], an intrusion detection model based on the improved genetic algorithm and deep belief network is proposed. The proposed architecture achieves a high detection rate by identifying various types of attacks, through more than one iteration of the genetic algorithm. The improvement in the genetic algorithm is achieved by upgrading the population initiation, improved selection, improved crossover and mutation processes. The improved genetic algorithm is combined with the deep belief networks to generate the best and efficient intrusion detection model to classify the attacks.

Ma and Fang [10] also proposes a similar approach which provides a visual representation of network intrusion detection data in three dimension which helps to have a better understanding of the relationship between the network intrusion detection data and the various kind of network traffic. The proposed methodology uses One hot transformation, Principal component analysis, and machine learning technique such as support vector machine technique to detect the various kinds of threats.

A direct comparison between all these methods is not possible as most of these approaches used a different data-set and systems with different specifications. The approach proposed in [2] is capable of handling high network traffic and detecting various types of attacks. Whereas [1] and [3] showed minimum compromise and high detection rate. Thus, the different challenges faced by these approaches are handling high network traffic, fast detection, minimum compromise, detection of a variety of attack, high accuracy, and minimum false rate. Besides, the authors referred to many literatures for this research [11–19].

Research Objectives

The objective of this chapter is to explain the design and methodology to develop a network intrusion detection system which combines signature-based detection techniques with data mining techniques which aim to secure the underlying network.

This chapter is divided into 05 sections. The rest of the chapter is organized as follows. In Sect. 2, a review of the existing intrusion detection systems is given. Section 3 presents the overall design of the system and the system requirements. Section 4 deals with the experimental setup. Section 4 also deals with the experimental results associated with this work and Sect. 5 brings out the conclusion.

3 Materials and Methods

Data Source

In this system, a data set called the NSL-KDD is used. This NSL-KDD data set comprise 126,620 records in its training set and 22,850 records in the testing set. Both these sets include 41 features which are labeled as normal traffic or specific attack types. These 41 features are further classified into 4 types such as:

- Basic features: which include information such as the protocol_type and duration,
- Time-based traffic features: which includes features which matures over a 2 s time window,
- Content features: which include features having domain information,
- Host-based traffic features: which contains information of the attacks that contains same destination host as current connection.

Description of the Dataset

As stated above the model constructed is tested using the NSL-KDD test dataset which comprise 22,850 records. The test set is tested using predict() methodology in scikit-learn library. The accuracy, f-measure, and precision were calculated for the evaluation process. A 10-fold cross-validation process was also carried out.

The platform on which the system was implemented, and the various dependencies and libraries used are as follows:

- Platform: Python 3.
- System: PC with a powerfull CPU, Minimum 8 GB RAM.
- Operating System: Windows 10.
- Dependencies: OpenCV, Dlib Libraries: Numpy, Pip, Scipy, Matplotlib, Sklearn.

Criteria for feature selection

Out of around 200 features, 52 features were selected with 13 across each attack category. While iterating the accuracy for the predicted outcomes of the learning data, it was found that after a certain number of features the accuracy was constant. The two main reasons for restricting the number of features are as follows. Firstly, due to the over-fitting that is the irrelevant features may wrongly detect correlations between target classes and features which was formed by chance and will prevent the correct modeling of the problem.

Secondly, as the number of features increases the computation time also increases without any classifier improvement. This process begins with a uni-variate feature selection with ANOVA F-test. This is done to perform a feature scoring. In this feature selection, each feature is examined separately to detect the stability of the relationship between the feature and the labels. During this stage, a best subgroup of features are identified which can detect the intrusion attempts, then a recursive feature elimination is performed. This builds a model repeatedly, by keeping aside the feature which repeats and repeating this procedure with the remaining process.

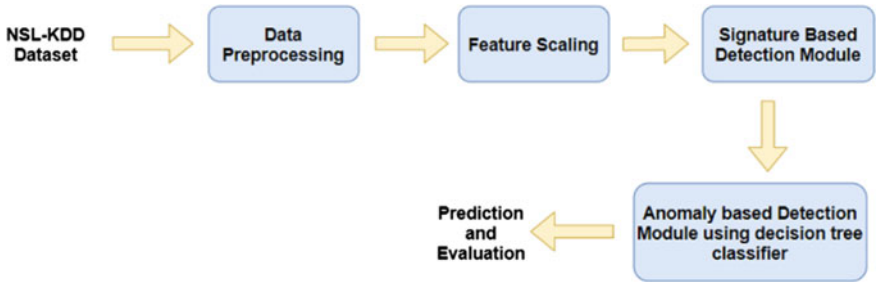


Fig. 1 The proposed model

This process is carried out until all the features of the data set are worn out. The basic idea behind this process is to develop a feature ranking based on their weights, and features which gave a cross-validation score of 0.995 was taken for model training.

Data Preparation

The main objective of the system which combines both signature-based detection and data mining technique for intrusion detection is to improve the efficiency of intrusion detection system by combining both these techniques. It also considers the problems involved in feature selection for the detection.

The design of the proposed system is given in Fig. 1 and described as follows:

1. Data Pre-processing: Data pre-processing operation is carried out because the data set consists of both non-numerical and numerical data and the classifier defined in the sci-kit-learn can handle numerical inputs and non-numerical inputs, so a transformation methodology called the one hot transformation is implemented. This methodology is used to convert every categorical feature to binary values.
2. Features Scaling: Feature scaling is a process which is used to scale some features which may weight too much on the final result as it has a very large value. In this process, the average of each feature is computed, and the mean value of the feature is subtracted from it. Finally, the result is divided by the standard deviation. This ensures that the feature will have a zero average and a standard deviation of one.
3. Signature-based Detection Module: In this process duplicate and irrelevant data are removed. This process mainly consists of identifying or capturing a subgroup of relevant features which can completely detect the given problem and also keeping in mind the minimum deterioration of presentation.
4. Anomaly-based Detection Module: In this process, a decision tree model is developed to split the data set till the instances in every individual leaf node comprise a unvarying class label. The partition is done using information gain. This process is a simple and efficient hierarchical methodology for supervised learning. The decision tree comprises internal decision nodes and leaf nodes. A test procedure is carried out by every single decision node consisting of distinct results marking

the branches. For each input at a branch, a test is performed and depending on the result, one among the branches is selected. In this process, the algorithm begins at the root and terminates when it arrives on the leaf node. This is carried out recursively and the value in the leaf node is considered as the output. Every single leaf node comprises an outcome label which is a numeric value in case of regression and class target in case of classification. A leaf node is used to describe a localized space or territory in which the instance finding in the input region comprises identical numeric values for regression and identical labels for classification.

5. Prediction and Evaluation: The test data set was utilized to make prediction of the designed model and used for the evaluation.

The attack types or labels in the NSLKDD dataset are grouped into four categories, and they are denial of service attack, probing attacks, remote to local, and user to local attacks.

The data set undergoes a preprocessing process which is carried out using OneHotEncoder() and LabelEncoder() defined in scikit-learn library. This is done to convert all the non-numerical data into numerical value and also to represent the categorical feature as a binary feature. This is followed by a data scaling process which is done to scale down those features which may weigh too much on the model. This is implemented using StdardScalar() methodology in scikit-learn library.

Table 1 shows the set of relevant subsets of features selected for each attack types. Thus, by implementing a classification learning algorithm and reduction process to select relevant and important feature, the accuracy of detection of an intrusion detection system can be increased and this also helps to detect the relevant features which play a major role in this improvement.

Figures 2, 3, 4, 5 show the relationship between the features selected for each attack type and the accuracy of the classifier. This proves that the accuracy of the classifier is maximized when the subset of relevant features is identified. We can see that the cross-validation is nearing saturation and doesn't show any improvements over 20 features. The maximum improvement is seen for the 13 features which are selected for the training.

4 Experimental Result and Evaluation

The attack types or labels in the NSLKDD dataset are grouped into four categories: denial of service attack, probing attacks, remote to local, and user to local attacks. The data set undergoes a pre-processing process which is carried out using OneHotEncoder() and LabelEncoder() defined in scikit-learn library. This is done to convert all the non-numerical data into numerical value and also to represent the categorical feature as a binary feature. This is followed by a data scaling process which is done to scale down those features which may weigh too much on the model. This is implemented using StdardScalar() methodology in scikit-learn library. The feature

Table 1 Relevant features selected for analysis

Features selected	Types of attacks
'logged_in', b'count', b'serror_rate', 'srv_serror_rate', 'same_srv_rate', 'dst_host_count', 'dst_host_srv_count', 'dst_host_same_srv_rate', 'dst_host_serror_rate', 'dst_host_srv_serror_rate', 'service_http', 'flag_S0', 'flag_SF'	DoS
'logged_in', 'rerror_rate', 'srv_rerror_rate', 'dst_host_srv_count', 'dst_host_diff_srv_rate', 'dst_host_same_src_port_rate', 'dst_host_srv_diff_host_rate', 'dst_host_rerror_rate', 'dst_host_srv_rerror_rate', 'Protocol_type_icmp', 'service_eco_i', 'service_private', 'flag_SF'	Probe
'src_bytes', 'dst_bytes', 'hot', 'num_failed_logins', 'is_guest_login', 'dst_host_srv_count', 'service_ftp', 'dst_host_same_src_port_rate', 'dst_host_srv_diff_host_rate', 'service_ftp_data', 'service_http', 'service_imap4', 'flag_RSTO'	R2L
'urgent', 'hot', 'root_shell', 'num_file_creations', 'num_shells', 'srv_diff_host_rate', 'dst_host_count', 'dst_host_srv_count', 'dst_host_same_src_port_rate', 'dst_host_srv_diff_host_rate', 'service_ftp_data', 'service_http', 'service_telnet'	U2R

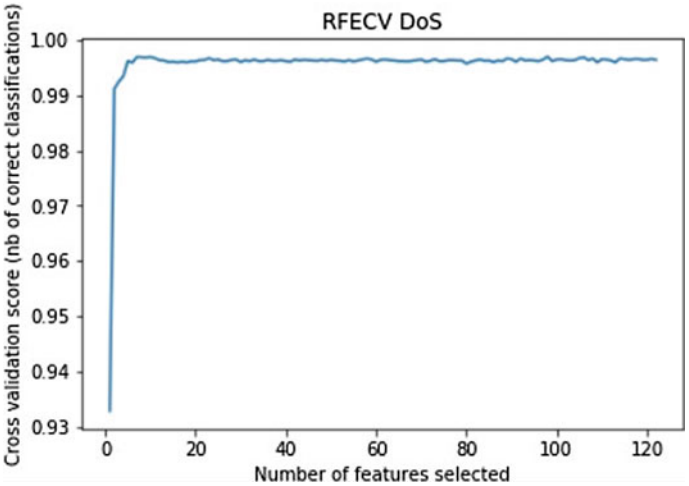


Fig. 2 DoS recursive feature elimination

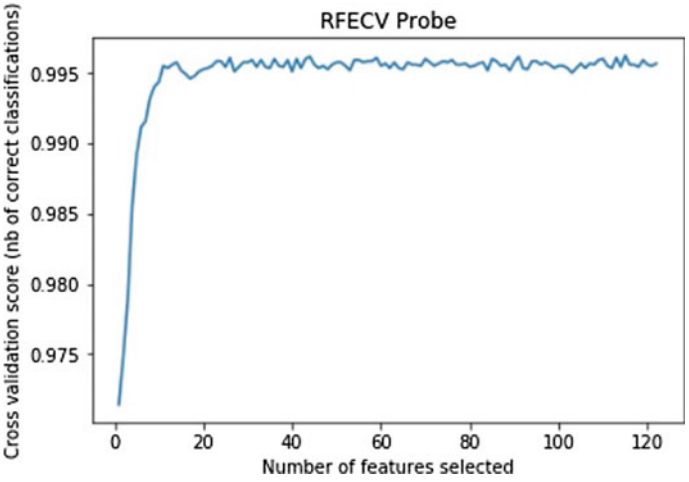


Fig. 3 Probe recursive feature elimination

selection process is carried out to detect a small subset of features to classify each attack types.

Figure 6 shows the experimental model setup for the study. This process begins with a selection process called the univariate feature selection through ANOVA F-test. This is done to perform a feature scoring. In this feature selection, each feature is examined separately to detect the efficiency of the relationship between the feature and the labels. The feature scoring and selection procedure is carried out using SelectPercentile() methodology in sklearn library. Then a recursive feature

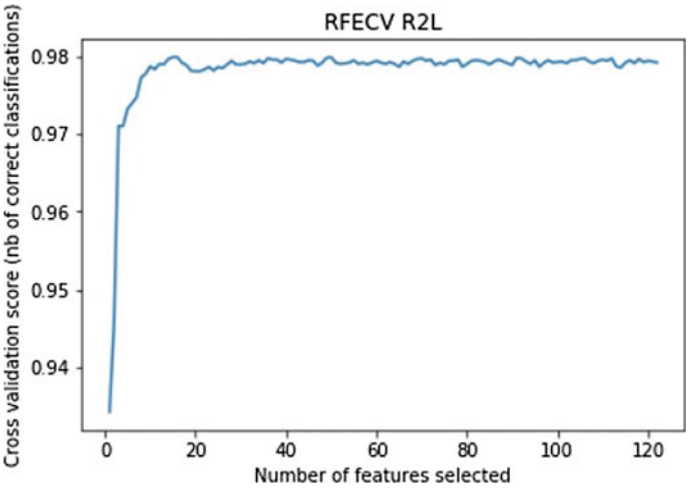


Fig. 4 R2L recursive feature elimination

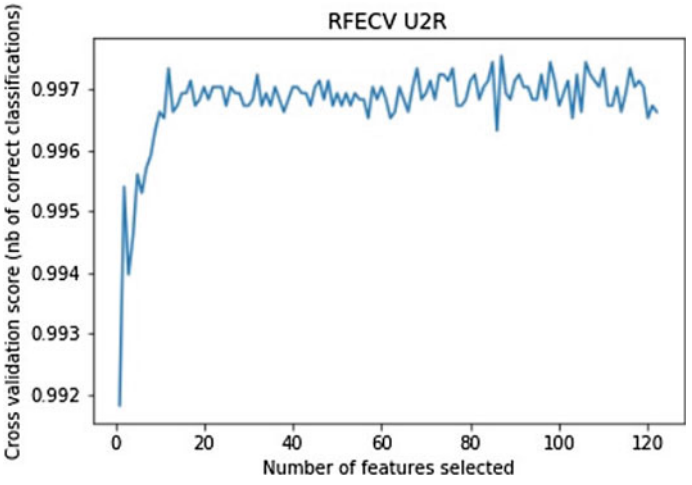


Fig. 5 U2R recursive feature elimination

elimination is carried out using RFE() methodology in feature selection module of scikit-learn library which selects the 13 best features for each attack category.

Now a decision tree model is developed using the DecisionTreeClassifier() methodology of the scikit-learn library to split the data until instances in each individual leaf node comprise a uniform class label. The partition is done using information gain. This process is a simple and efficient hierarchical methodology for supervised learning. In this process, a grid search parameter tuning was carried out

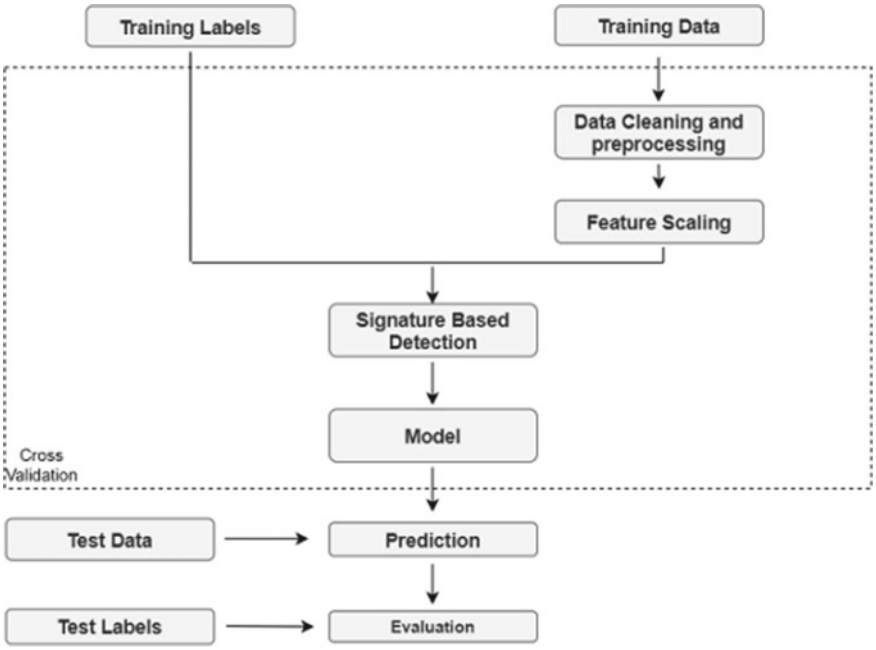


Fig. 6 The experimental setup

to capture the best 13 parameters to fit the model. This also ensures that the overfitting problem is removed. As a result, a tree is constructed from the training data. The features are captured and implemented in a uni-variate manner.

Figure 7 represents the system model built for training the model. The data set is first normalized using a standard normal function and then split into train-test subsets in ratio 4:1. The training data then undergoes a feature selection as explained earlier. With the selected features a decision tree classifier is built and trained. The data is validated against the testing data subset under supervised classification.

Results and Discussions

The proposed system was trained and fine-tuned using the NSL-KDD data set. The model performs quite well by recognizing most of the intrusion attempts. An experimental analysis was carried out to find the accuracy of the classifier after selecting relevant features. The following table depicts the details such as the accuracy, precision, recall, F-measure, and number of features selected by the classifier for each attack types.

Table 2 depicts the details such as the accuracy, precision, recall, F-measure, and number of features selected by the classifier for each attack types. Thus, the proposed methodology has achieved high accuracy rate after selecting the relevant subset of features for each attack types.

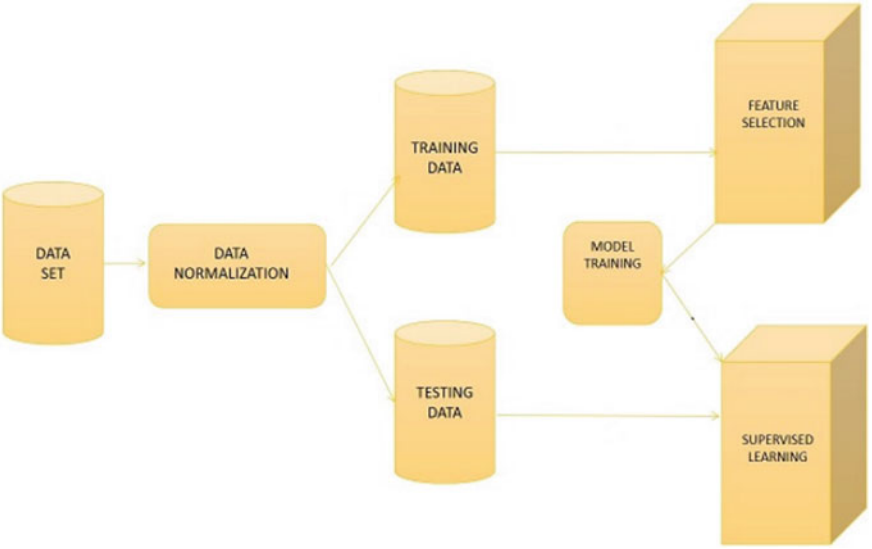


Fig. 7 The system model

Table 2 Performance evaluation

Accuracy	Precision	Recall	F-Measure	No of Features	Types of Attacks
99.9	99.69	99.79	99.74	13	DoS
99.08	98.67	98.46	98.56	13	Probe
97.45	96.68	96.08	96.37	13	R2L
89.65	87.74	89.18	87.49	13	U2R

By implementing a classification learning algorithm and reduction process to select relevant and important feature, the accuracy of detection of an intrusion detection system can be increased and this also helps to detect the relevant features which play a major role in this improvement. Also, the accuracy of the classifier is maximized when the subset of relevant features are identified.

5 Conclusion

During the past few years, there has been significant improvement in intrusion detection system. However, very often new intrusion attacks are able to bypass the security offered by the intrusion detection system. Previous works mainly implemented a signature-based technique or a machine learning algorithm or data mining technique to detect the intrusion attempts which suffers from high false rate and surfaced as

the main drawback. This chapter proposes an efficient system for intrusion detection which combines signature-based detection and data mining technique.

This model implements an efficient feature selection process which implements a uni-variate feature selection followed by a recursive feature elimination and is combined with a decision tree classifier to detect the subset of relevant features and to detect the intrusion attempts. This process is carried out repeatedly until the training dataset is depleted to build the model. The performance evaluation of the proposed model is carried out using various classification criteria, and it has been found that the accuracy of the system could be increased by detecting the relevant subset of a features and using these features to classify the intrusion attacks. The model was implemented using NSL-KDD data sets. Although the system is designed keeping in mind the difficulty involved in working with a live network traffic, due to the limitations of resources the system could not be implemented in real time system with live network traffic.

As a future scope, we would like to analyze the working of the system in a real-time system and also the model could be fined tuned efficiently to incorporate the evolutionary or hybrid attack design techniques in order to make the model as a standard for detecting hybrid attacks and to increase the overall performance.

References

1. Tsai CF et al (2009) Intrusion detection by machine learning: a review. Expert Syst Appl 36:11994–12000
2. Parsazad S, Allahyar SE (2012) Fast feature reduction in intrusion detection data-sets. In: MIPRO 2012 proceedings of the 35th international convention, pp 1023–1029
3. Alazab A et al (2012) Using feature selection for intrusion detection system. In: 2012 international symposium on communications and information technologies (ISCIT), pp 296–301
4. Suthaharan S, Vinnakota K (2011) An approach for automatic selection of relevance features in intrusion detection systems. In: Proceedings of the 2011 international conference on security and management (SAM 11), Las Vegas, Nevada, USA, pp 215–219
5. Wei M, Kim K (2012) Intrusion detection scheme using traffic prediction for wireless industrial networks. J Commun Netw 14:310–318
6. Ghosh P, Debnath C, Metia D (2014) An efficient hybrid multilevel intrusion detection system in cloud environment. IOSR J Comput Eng 16(4):16–26
7. Zhang F, Wang D (2013) An effective feature selection approach for network intrusion detection. In: Eighth international conference on networking, architecture and storage, IEEE
8. Senthilnayaki B, Venkatalakshmi DK, Kannan DA (2015) Intrusion detection using optimal genetic feature selection and SVM based classifier. In: 3rd International conference on signal processing, communication and networking (ICSCN) intrusion, pp 1–4
9. Mukherjee S. Sharma N (2012) Intrusion detection using Naive Bayes classifier with feature reduction. Proc Technol 4:119–128
10. Ma C-X, Fang Z-M (2009) A novel intrusion detection architecture based on adaptive selection event triggering for mobile ad-hoc networks. In: IEEE second international symposium on intelligent information technology and security informatics, pp 198–201
11. Dhanabal L, Dr. S.P. Shantharajah, "A Study on NSL_KDD Daaset for Intrusion Detection System Based on Classification Algorithms," International Journal of Advanced Research in Computer and Communication Engineering, 2015;4(6);446–452.

12. Chen M, Wang N, Zhou H, Chen Y (2017) FCM technique for efficient intrusion detection system for wireless networks in clod environment. Comput Electr Eng 8:401–410
13. Priyanka Dahiyaa, Devesh Kumar Srivastava, "Network Intrusion Detection in Big Dataset Using Spark", International Conference on Computational Intelligence and Data Science, pp. 253–262, 2018.
14. Rupam Kumar Sharma, Biju Issac and Hemanta Kumar Kalita , "Intrusion Detection and Response System Inspired by the Defense Mechanism of Plants", IEEE Access 2019;7.
15. Wei Zong, Yang-Wai Chow, Willy Susilo, "Interactive three-dimensional visualization of network intrusion detection data for machine learning", Future Generation Computer Systems, pp. 292–306, 2019.
16. Wenjuan Li, Lam For Kwok, "Challenge based collaborative intrusion detection networks under passive message fingerprint attack A further analysis," Journal of Information Security and Applications 2019.
17. Rajendra Patil, Harsha Dudeja, Chirag Modi, "Designing an efficient security framework for detecting intrusions in virtual network of cloud computing," Computers & Security, pp. 402–422, 2019.
18. Ying Zhang, Peisong LI, AND Xinheng Wang, "Intrusion Detection for IOT Based on Improved Genetic Algorithm and Deep Belief Network", IEEE Access 2019;7.
19. Nilam Upasani, Hari Om, "A modified neuro-fuzzy classifier and its parallel implementation on modern GPUs for real time intrusion detection", Applied Soft Computing Journal, 2019.

Chapter 10
Cloud Cost Intelligence Using Machine Learning

Saravanan Gujula Mohan and R. Ganesh

Abstract Cloud deployments have brought promise to business services in the entire spectrum of infrastructure such as provisioning, infinite scalability, cost of utility at consumption, etc. Cloud usage data forms a critical data layer to measure usage and apply billing and is made available to cloud users that make the use of cloud services. This chapter walks through methodologies for various optimization applications aided by machine learning using cloud usage data. A framework for cloud intelligence application is recommended to be provisioned as a data analytics and intelligence gathering mechanism that will open opportunities in cost optimization techniques, Intelligent Provisioning, and capacity forecasting mechanisms. For generalization purposes, public utilization datasets from Microsoft Azure are used for Exploratory Data Analysis (EDA) and recommendations. To further emphasize on the applicability of the framework across cloud service providers, framework applicability is showcased across multiple cloud service providers.

Keywords Forecasting · Time series · Cloud cost optimization · Deep learning · ARIMA

1 Introduction

Cloud services provided by public Cloud Service Providers (CSP) are charged at consumption levels; typically, usage data is made available along with cloud cost accountability and lineage are drafted with colorful dashboards based on cloud services utilization. This usage data is typically made available through a host of services, such as AWS and Azure. Intelligence applied on top of this usage data

S. Gujula Mohan (✉)
Tata Consultancy Services, Chicago, IL, United States
e-mail: saravanan.gm@tcs.com

R. Ganesh
Tata Consultancy Services, Chennai, India
e-mail: ganesh.prasathr@tcs.com

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_10

leads us to opportunities which can result in optimal usage and deployment of cloud resources.

2 Literature Review

A. Data Source: Cloud Usage Datasets

(i) Infrastructure Utilization Data
   a. These are the most common datasets made available from CSPs to its users from their IaaS (Infrastructure as a Service)-provisioned VMs. The data may include CPU utilization, memory usage, disk utilization, and network usage covered over time series
   b. Example of datasets for VMs and their data attributes are listed below. Figure 1 shows VM utilization over a period with average, maximum, and 95th percentile of maximum CPU utilization

(ii) PaaS Usage Data
   a. PaaS Usage data will comprise of service usage data along with its attributes related to PaaS service. For example, PaaS Usage of S3 usage may include BucketSizeBytes/day, number of objects/day, request details, data transfer metrics, etc., as shown in Fig. 2.

(iii) Application Usage Metrics
   a. Application usage metrics vary depending upon the type of application as well as the architecture of applications. These metrics will be very useful if layered upon the infrastructure usage and PaaS service usage. Typical application usage metrics should cover up the following:
      i. How many request applications able to satisfy—orders, reports, requests, updates, etc.?
      ii. Metrics that are directly proportional to the cost of running application—time to process orders, requests, etc.

	time_sec	vmid	avg_cpu	max_cpu	P95_max_cpu
3655	0	zzyUrPQWi9i9O0SGP3AvQpZtSQcfhbO7VRHtTgCFOIbEJI...	1.050638	14.365946	1.704018
231045	300	zzyUrPQWi9i9O0SGP3AvQpZtSQcfhbO7VRHtTgCFOIbEJI...	0.966978	3.220826	1.464903
458326	600	zzyUrPQWi9i9O0SGP3AvQpZtSQcfhbO7VRHtTgCFOIbEJI...	1.073010	20.693568	3.019812
685590	900	zzyUrPQWi9i9O0SGP3AvQpZtSQcfhbO7VRHtTgCFOIbEJI...	0.937519	2.095963	1.419127
913099	1200	zzyUrPQWi9i9O0SGP3AvQpZtSQcfhbO7VRHtTgCFOIbEJI...	1.052166	6.617037	1.844468
1140782	1500	zzyUrPQWi9i9O0SGP3AvQpZtSQcfhbO7VRHtTgCFOIbEJI...	0.967635	4.209649	1.575974
1368414	1800	zzyUrPQWi9i9O0SGP3AvQpZtSQcfhbO7VRHtTgCFOIbEJI...	0.867778	6.438934	1.767366
1596035	2100	zzyUrPQWi9i9O0SGP3AvQpZtSQcfhbO7VRHtTgCFOIbEJI...	0.871987	2.164887	1.441794
1823782	2400	zzyUrPQWi9i9O0SGP3AvQpZtSQcfhbO7VRHtTgCFOIbEJI...	0.916802	7.831342	1.825642
2051545	2700	zzyUrPQWi9i9O0SGP3AvQpZtSQcfhbO7VRHtTgCFOIbEJI...	1.029335	1.722658	1.344773

Fig. 1 VM CPU utilization

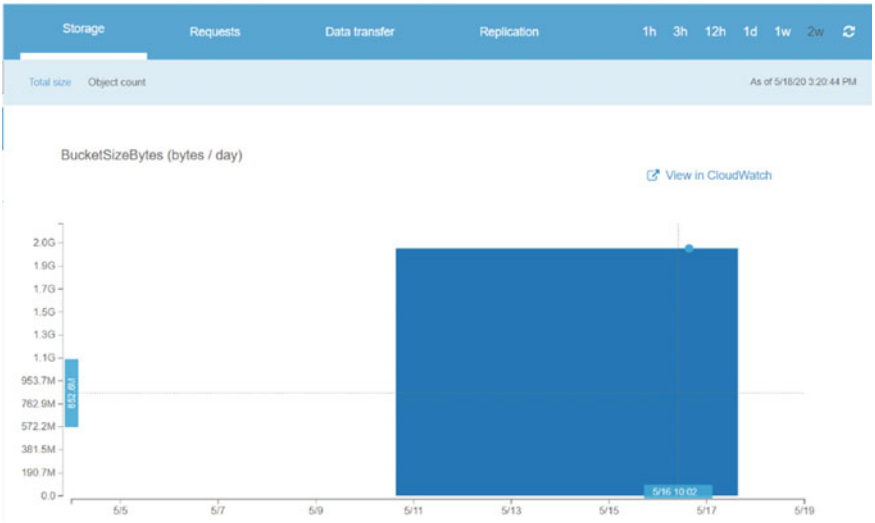


Fig. 2 AWS S3 storage metrics—bytes/day

3 Materials and Methods

A. Data Source

The data source for this study is from Azure VM utilizations publicly available at https://github.com/Azure/AzurePublicDataset [1, 2].

B. Simple Analytics

This method typically focuses on running a simple analysis of usage metrics and gathering intelligence. A sample scenario may be to understand the usage of resources by applying analytical functions like max () over a period of time for which metrics are collected. While this method will provide a simple interface to understand usage metrics, this will provide point-in-time metrics with predefined analytical scenarios. Any change in the behavior of resource under usage needs to be captured and feedback to the scenario to modify function, which may run into highly complex analytics. Knowing Max or Peak utilization as shown below in Fig. 3 around 6000th second, will help to determine the peak utilization metrics requirement for the VM, whereas how VM can be provisioned based on varied behavior and merging other characteristics may need further deep analysis.

This is leading to the second methodology using machine learning, wherein the change in usage behavior can be captured and reapplied to take advantage over various opportunities.

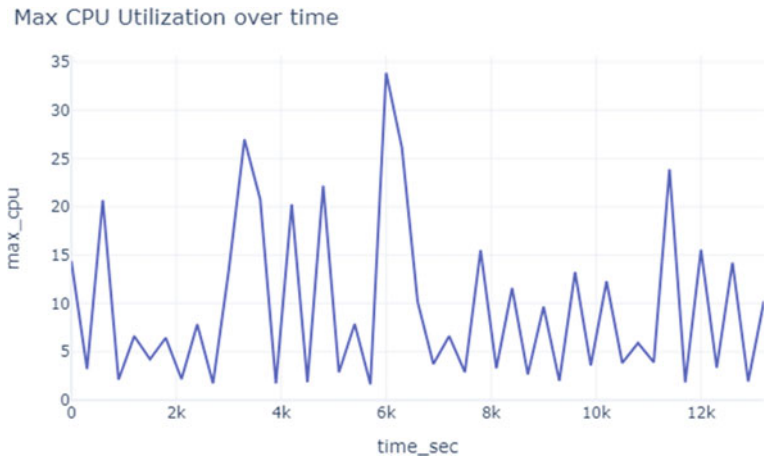


Fig. 3 Max CPU utilization over period for a VM

C. Machine Learning

- Machine learning provides an opportunity to analyze resource utilization or service usage even with the change in patterns of usage, in which patterns can be captured and utilized as applicable intelligence.
- Various techniques available with machine learning including time series analysis [3] with classic algorithms like ARIMA [4–6] and/or customized ML algorithms can be used to analyze usage metrics from VM.
- Figures 4 and 5 depict CPU utilization for VM and with forecasting algorithms applied over an extended period.

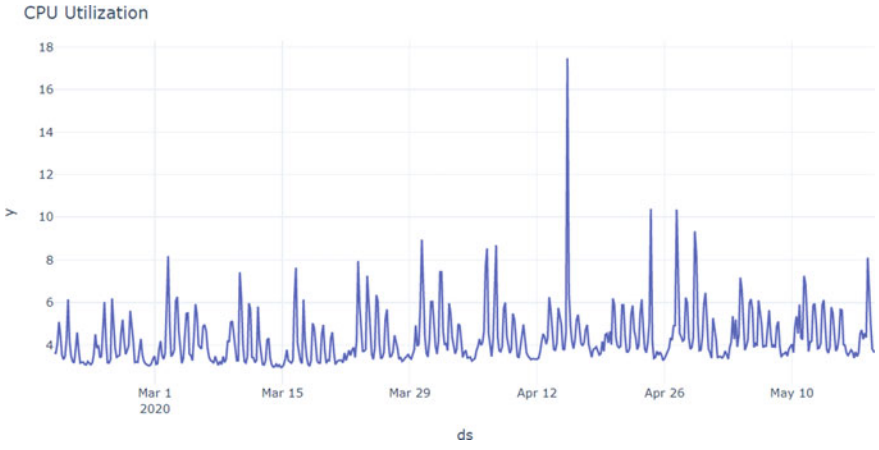


Fig. 4 CPU utilization for known periods

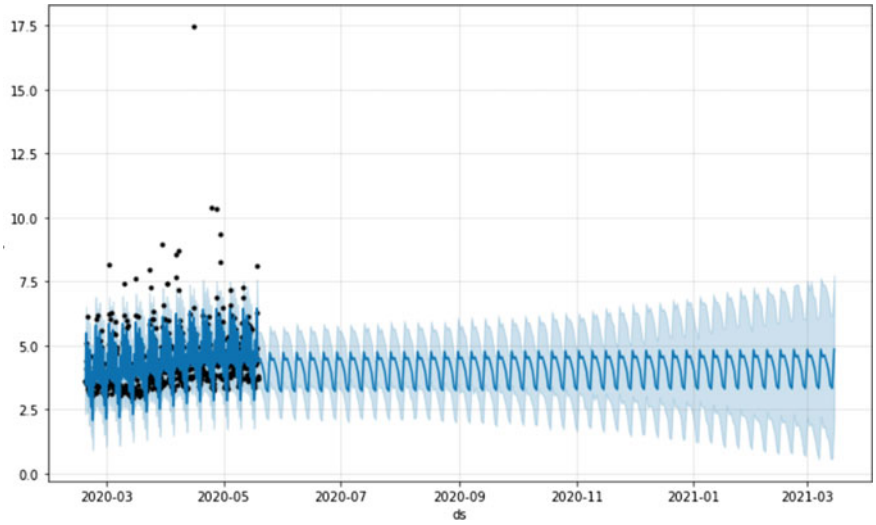


Fig. 5 CPU utilization forecasting for extended periods

D. Forecasting using Time Series Analysis

- VM Utilization Forecast can provide a prediction of VM usage, for example, across various periods.
- The following plot shows various VMs CPU usage forecasted over periods with known observations marked with various trends.
- Various packages like Stan [7] using classic ML algorithms can be purposed to forecast usage of services; one such example is shown below in Fig. 6.

E. Sample Time Series Analysis using ARIMA

Resource usage data can be analyzed with Time Series Analysis (TSA) and used with machine learning and/or forecasting applications. In TSA, resource utilization data points taken over time are expected to have an internal structure like autocorrelation, seasonality, and trending.

TSA [3] will help forecasting in two ways:

1. Unearth key patterns in resource utilization data {Trends, Seasonality, etc.}.
2. Fit a model that is generic enough to be used for forecasting and predictability of resource utilization.

A sample analysis is provided below using public datasets based on resource utilizations in Azure. This public repository contains two representative (anonymous) traces of Virtual Machine (VM) workload of Microsoft Azure collected in 2017 and 2019.

For the analysis, a sample of VM trace from 2019 is used here. Moving through the schema of the dataset available in the repository, for the analysis, CPU readings

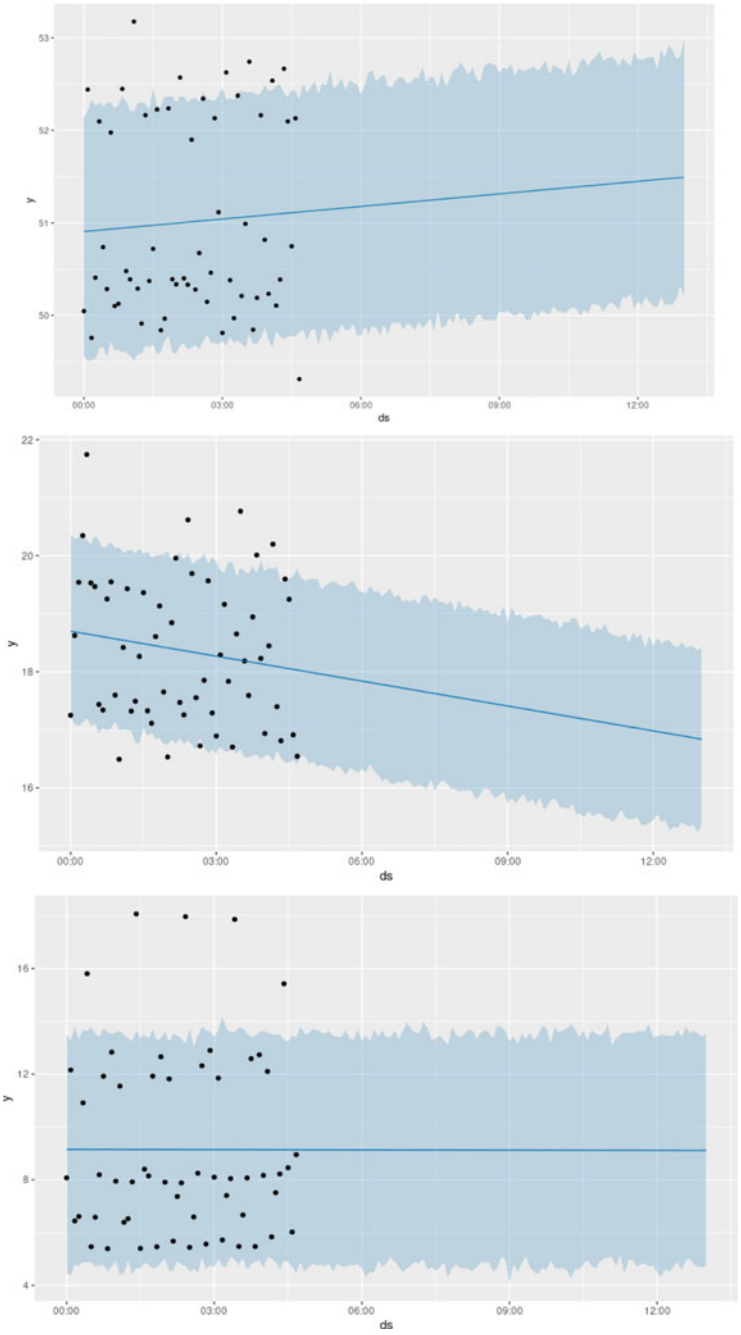


Fig. 6 CPU utilization forecasts for multiple VMs

of the VMs are taken up. The data analysis is completed using Python Pandas API, and forecasting is done with statsmodel.tsa (Time Series Analysis) API.

The steps required for analysis can be briefly listed below:

1. Data Cleanup,
2. Exploratory Data Analysis (EDA),
3. Model Generation,
4. Model Validation, and
5. Model Prediction.

A complete analysis of all the steps above is beyond the scope of this chapter, though, some of the steps are explained as required for the discussion.

Since the data is already cleaned, structured, and labeled, most of the data cleanup can be skipped. At a minimum, for forecasting model, vm_cpu_readings-file-*-of-195.csv.gz can be read using pandas, with timestamp, vmid, and CPU utilization gathered.

A simple plot using matplotlib, and data collected and filtered with one vmid, shows up trends, correlated with the time of collection as shown below in Figure 7.

The data analysis in Fig. 6 shows seasonality along with trends with respect to the time of collection of VM CPU utilization. Another data analysis technique is to see if there is any autocorrelation using autocorrelation_plot. A sample autocorrelation plot is drawn based on the data collected as shown below in Fig. 8.

The autocorrelation plot shows the calculation of seasonal differencing can be for every 50 lags, for the data collected and vmid selected. The seasonality trend through earlier data analysis also shows a similar trend.

ARIMA [5, 6] model is a widely used time series forecasting model on univariate data, following which analysis is done using the ARIMA model from statsmodels.tsa.arima.model (statsmodel), and the results are shown in Fig. 9 [8].

Sample model generation using ARIMA, results, and prediction plots:

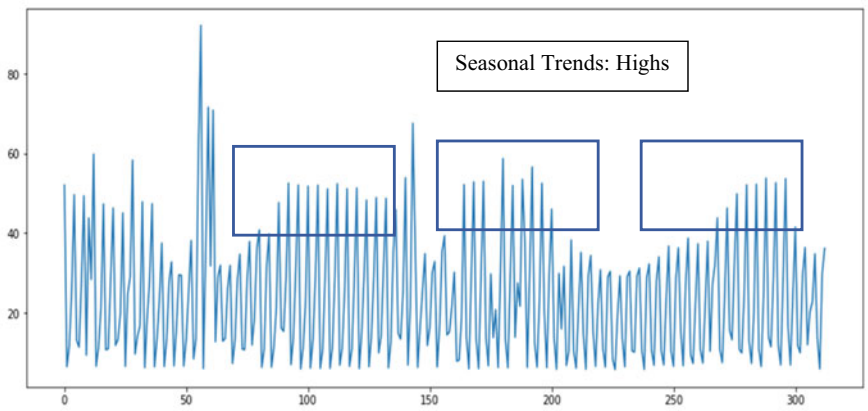


Fig. 7 CPU Utilization of single VM (X), Time (Y)

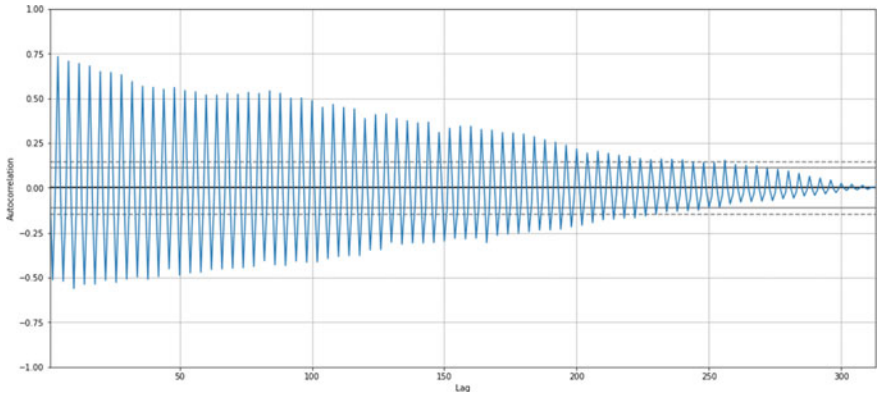


Fig. 8 Autocorrelation plot

```
model = ARIMA(trainx, order=(5,1,0))
modelvar = model.fit()
output = modelvar.forecast()
```

SARIMAX Results

Dep. Variable:	max_cpu	No. Observations:	313
Model:	ARIMA(5, 1, 0)	Log Likelihood	-1183.435
Date:	Mon, 12 Jul 2021	AIC	2378.869
Time:	22:22:03	BIC	2401.327
Sample:	0	HQIC	2387.845
	- 313		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.0279	0.039	-26.223	0.000	-1.105	-0.951
ar.L2	-1.0344	0.061	-16.981	0.000	-1.154	-0.915
ar.L3	-0.8771	0.065	-13.477	0.000	-1.005	-0.750
ar.L4	-0.0950	0.051	-1.858	0.063	-0.195	0.005
ar.L5	0.0585	0.035	1.658	0.097	-0.011	0.128
sigma2	113.9895	5.468	20.847	0.000	103.273	124.706

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	851.29
Prob(Q):	0.92	Prob(JB):	0.00
Heteroskedasticity (H):	0.17	Skew:	0.54
Prob(H) (two-sided):	0.00	Kurtosis:	11.02

Fig. 9 Sample results of ARIMA

In the prediction plot shown in Fig. 10, the blue line plot shows actual utilization and the red shows predicted utilization by the ARIMA model.

By using machine learning and forecasting techniques (TSA), resource utilization can be forecasted and utilized with variegate applications. With the usage data

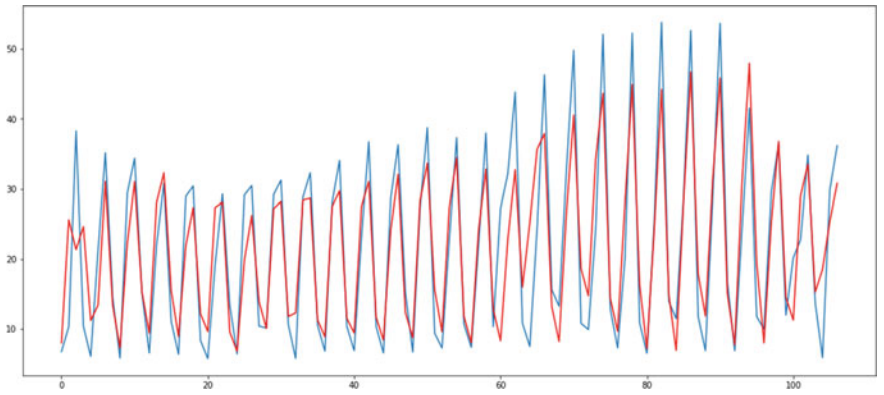


Fig. 10 Prediction Plot; (X) Time Series, (Y) CPU Utilization

and analytics methods discussed so far, the following frameworks are proposed to mine intelligence to effectively manage cloud consumption and proactively provision infrastructure and services.

4 Results and Recommendations

With the usage data and analytics methods discussed so far, the following frameworks are proposed to mine intelligence to effectively manage cloud consumption and proactively provision infrastructure and services.

A. Cloud Intelligence Framework

(i) Intelligent Provisioning

1. Intelligent Provisioning means the ability to provision at demand/point in need with lower cost. Fulfillment of services with low-cost compute without any service or process disruptions.
2. With resource usage metrics and forecasting techniques, along with Spot Instance price history, VMs can be provisioned to fulfill the process or any service without major disruptions in the requests pipeline.
3. Sample scenarios for this Intelligent Provisioning can determine Spark Cluster compute size requirements and match with Spot Instance availability at the cost with the help of price history trends and provision Spark clusters with Spot Instances.
4. A reference architecture is provided for AWS, using AWS Sagemaker as underlying framework [9], in which compute provisioning can be handled intelligently based on the compute parameters recommended by ML model using resource utilization data and cost metrics (Fig. 11).

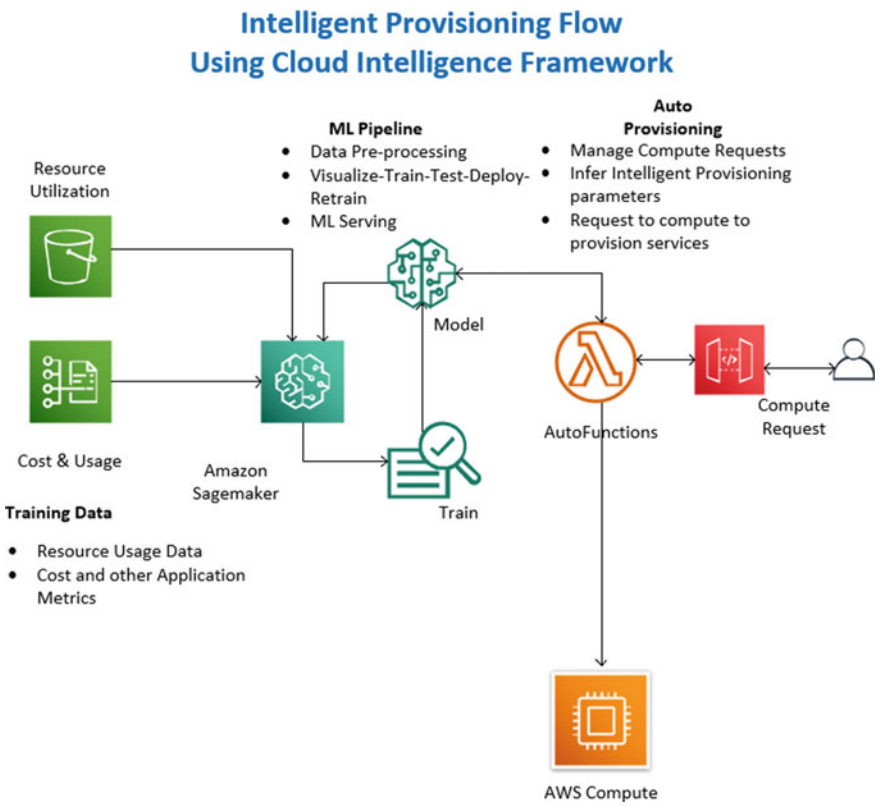


Fig. 11 Intelligent provisioning flow

(ii) Cost Optimization techniques

1. Resource usage analysis using ML can be provisioned to optimize costs by overlaying other metrics.
2. One such technique will be altered between Spot Instances and Reserved Instances where utilization of VM is forecasted to be on the lower side. Here, overlaying metrics can be Spot Instance price history along with VM utilization metrics.
3. Usage analysis with VM along with container spin-up patterns can provide deeper insights into machine availability requirements using which non-critical VMs and zero-utilized VMs can be shut down. This can be achieved by employing auto-scale functions like AWS Lambda or Azure Functions/Logic Apps to act on intelligence provided by Resource usage (Fig. 12).

(iii) Capacity Forecasting

1. Capacity forecasting [10] need not be simple linear answers. Though cloud provides infinite scalability, without capacity forecasting, critical

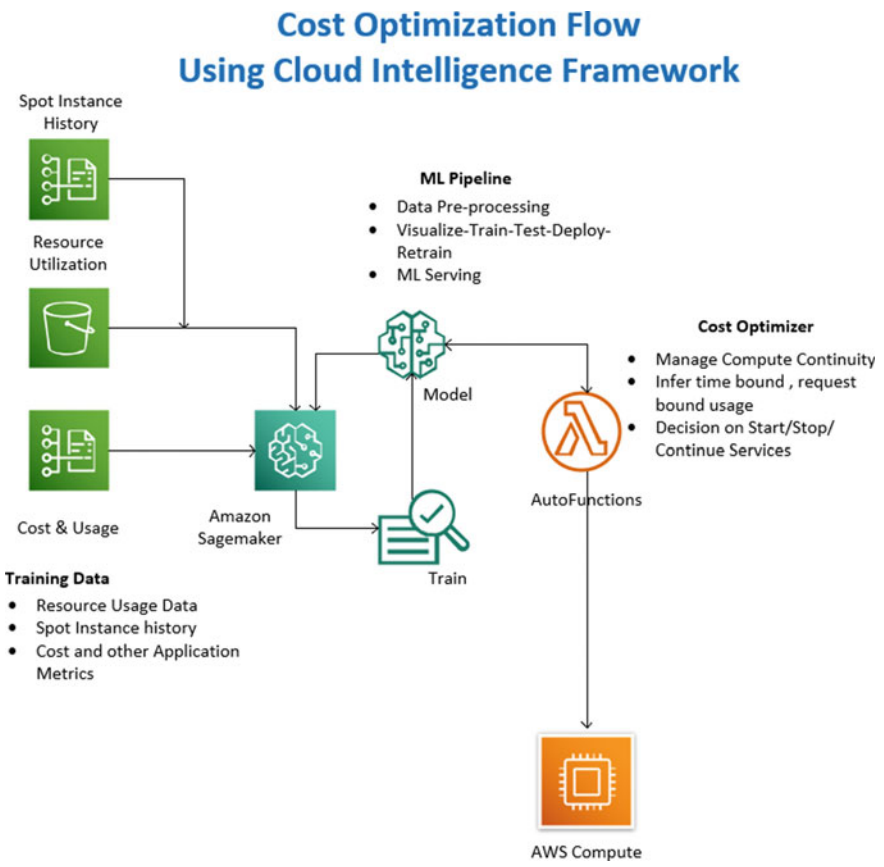


Fig. 12 Cost optimization flow

services will be at risk. With the help of machine learning, a simple classical algorithm analysis like ARIMA [3–7] can provide forecasting required for peak and non-peak hours.

2. In addition, complex usage patterns and optimizations can be unearthed with the capacity requirements known up front and planning with respect to various consumption types, service provisioning can be employed.
3. Cloud intelligence framework, work on raw service utilization metrics and other consumption metrics along with cost, can forecast peak and non-peak resource requirements in line with cost attributes.
4. Forecast data can be stored for trend analysis and can be projected to executive reporting using visual analytics tools like PowerBI (Fig. 13).

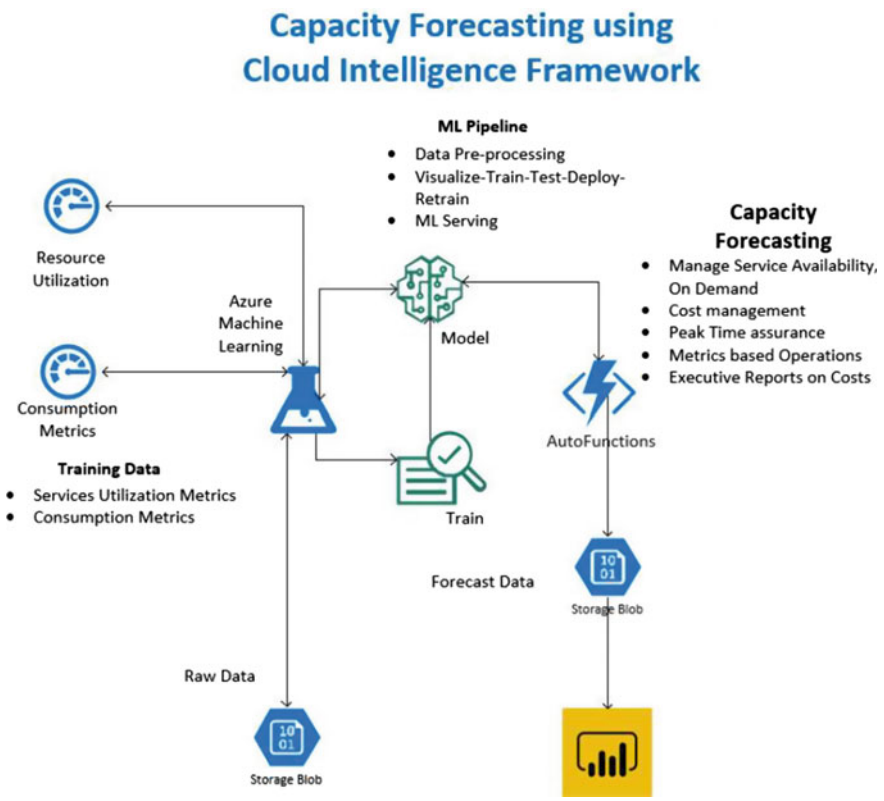


Fig. 13 Capacity forecasting

5 Conclusion

Various analysis techniques and recommended frameworks are discussed in this chapter along with applicable use cases, to acquire and apply various intelligence from cloud usage data using machine learning.

The recommendations can be deployed as customized frameworks or deployed as is with cloud services like Azure Machine Learning studio or AWS Sage Maker [9] with which the ML serving can be compacted and deployed as applications.

A further enhancement to classic machine learning algorithms discussed in this chapter, other situational metrics such as peak demands, seasonal demands as well as target-based requirements (e.g., complex visual recognition may need larger compute), may need deep learning algorithms which can pave the pathway to include artificial intelligence in key decision-making points.

References

1. Hadary O, Marshall L, Menache I, et al This paper is included in the Proceedings of the 14th USENIX Symposium on Operating Systems Design and Implementation Open access to the Proceedings of the 14th USENIX Symposium on Operating Systems Design and Implementation is sponsored by USENIX Protean: VM Allocation Service at Scale Protean: VM Allocation Service at Scale.
2. Romero F, Chaudhry GI, Goiri Í, Gopa P, Batum P, Yadwadkar NJ, Fonseca R, Kozyrakis C, Bianchini R (2021) Faa$T: A transparent auto-scaling cache for serverless applications. In: SoCC 2021 - Proceedings of the 2021 ACM Symposium on Cloud Computing. Association for Computing Machinery, Inc, pp 122–137
3. de Gooijer JG, Hyndman RJ (2006) 25 years of time series forecasting. Int J Forecast 22:443–473
4. Siami-Namini S, Tavakoli N, Siami Namin A (2019) A Comparison of ARIMA and LSTM in Forecasting Time Series. In: Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018. Institute of Electrical and Electronics Engineers Inc., pp 1394–1401
5. Wang J, Tang S Time series classification based on arima and adaboost. https://doi.org/10.1051/matecconf/202030
6. Mehrmolaei S, Keyvanpour MR (2016) Time series forecasting using improved ARIMA. In: 2016 Artificial Intelligence and Robotics, IRANOPEN 2016. Institute of Electrical and Electronics Engineers Inc., pp 92–97
7. Carpenter B, Lee D, Brubaker MA, Riddell A, Gelman A, Goodrich B, Guo J, Hoffman M, Betancourt M, Li P Journal of Statistical Software Stan: A Probabilistic Programming Language.
8. Seabold S, Perktold J (2010) Statsmodels: Econometric and Statistical Modeling with Python.
9. Das P, Ivkin N, Bansal T, et al (2020) Amazon SageMaker Autopilot: A white box AutoML solution at scale. Proceedings of the 4th Workshop on Data Management for End-To-End Machine Learning, DEEM 2020 - In conjunction with the 2020 ACM SIGMOD/PODS Conference. https://doi.org/10.1145/3399579.3399870
10. Zhang J, Huang H, Wang X (2016) Resource provision algorithms in cloud computing: A survey. J Netw Comput Appl 64:23–42

Chapter 11
Mining Deeper Insights from Texts Using Unsupervised NLP

Shobin Joyakin and Sudheer Chandu

Abstract An industry or a business collects huge amount of free form text through internal processes such as typed call center notes, notes typed by document processing teams, etc. or from users through customer feedback mechanism and process them. Many of these free form texts come without labels on what the texts are about, and no supervised learning models could be used to extract insights from these texts. For instance, guest feedback after a hotel stay or feedback after a Restaurant visit might have valuable information much more than just customer sentiments, such as specifics on what was good and what was bad. Manually labeling this information on the texts is erroneous and involves time and effort. Unsupervised learning from machine learning combined with the power of vector algebra can be used to mine deeper insights from these texts. These insights could act as labels for building new supervised models if needed and these predicted insights could be converted to actionable business outcomes through visualizations.

Keywords Data mining · Insights · Text mining · Unsupervised learning · Machine learning · NLP · Visualization

1 Introduction

Businesses are not equipped with techniques that could automatically mine deeper insights from large free form texts. Using natural language processing and machine learning, these businesses have been predicting insights such as user sentiments, custom labels etc., however they have not been able to mine those deeper insights that could be used to make actionable business decisions.

For instance, online platforms like Zomato, Myntra, Olx, Airbnb, Flipkart, Yelp (with crowd-sourced reviews), etc. receive lot of customer feedback for its services or products they sell, or customers have used. These products or services are sold by individual sellers on these platforms. Many at times the services or products

S. Joyakin (✉) · S. Chandu
TATA Consultancy Services, America, Columbus, OH, USA
e-mail: meet_shobinjoy@yahoo.co.in

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_11

offered by these sellers are not at par with the standards set by the online platforms. These degraded services or products affect the overall market performance of these platforms when many customers share negative reviews of their experience. Apart from understanding the negative sentiments from review text and star ratings, there is no mechanism to get automatic insights on the actual reasons for negative reviews from these vast texts and suggest actionable business solutions to reduce such negative experiences.

In addition, industries like banking and insurance, telecommunication and marketing interact with many customers through call centers, document processing teams, etc. The voice calls, if any, are stored and might be transcribed into texts for regulatory reasons. The interaction with the customers could also be just free form text typed by a customer representative. After the interactions are wrapped up between a customer representative and a customer, apart from some generic drop-down lists a customer representative pick from, there are no ways to get insights on the interaction from these texts unless an SME review them. Such insights could help industries take action to automate processes that could reduce customer representative interactions, document processing delays, etc.

Background:

In the case of customer reviews on online platforms or the customer reviews sourced by Yelp, businesses have not been able to get the reasons for negative comments unless a human reads through it and acts. Multi-labels could be applied to these texts; however, it cannot be used for a drill-down understanding of texts.

Here I will deep dive into a solution that will help derive insights from unstructured texts data from restaurant reviews available in Yelp dataset (https://www.yelp.com/dataset). This dataset contains 8 million crowd-sourced reviews from about 161 K businesses. For our experiment, we will narrow our data to negative review text with rating < 3 for all restaurants in the city of Las Vegas in Nevada, USA and see if we can derive actionable insights using a drill-down view from the negative reviews to suggest recommendation to improve a restaurant's rating.

2 Literature Review

An insight engine, also called cognitive search or enterprise knowledge discovery and management, is an enterprise platform that makes key enterprise insights accessible to the users when they need it [1]. It combines search with machine learning capabilities to provide information for users and data for machines. The goal of an insight engine is to provide timely data that delivers actionable insights. The term was first suggested by Gartner and defined as "Insight engines apply relevancy methods to describe, discover, organize and analyze data. This allows the existing or synthesized information to be delivered proactively or interactively, and, in the context of digital workers, customers or constituents, at timely business moments" [2].

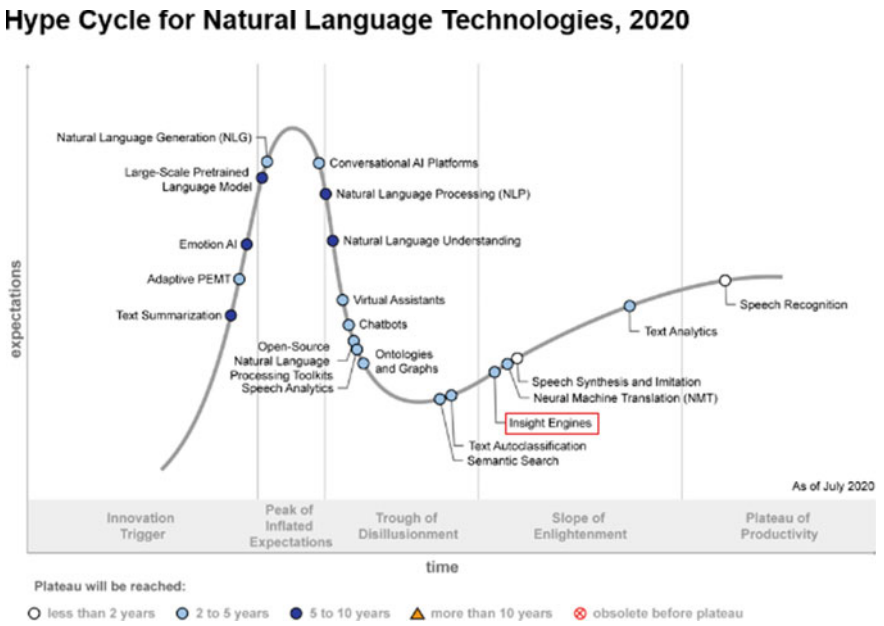


Fig. 1 Hype cycle for NLP tech, 2020

A Gartner report titled "Hype Cycle for Natural Language Technologies, 2020" (ID: G00467879) categorizes "Insight Engine" under "Slope of Enlightenment" with a high benefit and a potential for mainstream adoption in next 2–5 years as shown in Figs. 1 and 2.

3 Materials and Methods

My current study is focused on negative user reviews for restaurants from yelp public dataset; however, similar techniques were also performed on other datasets in insurance domain, and it has yielded excellent results. Below steps were performed after the data was collected.

(a) **Text Cleaning**:

Firstly, any free form text needs to be cleaned to remove unwanted number patterns (SSN, Date, Phone number, etc.), special characters (@, #, %, &, etc.), stop words (most common words in a language), and any other insignificant words. Figure 3 shows word count distribution of all words in the review text before stop words were removed.

Figure 4 shows a sample text pre-processing function which removes non-English words, converts all characters to lower case, and removes all non-alphanumeric char-

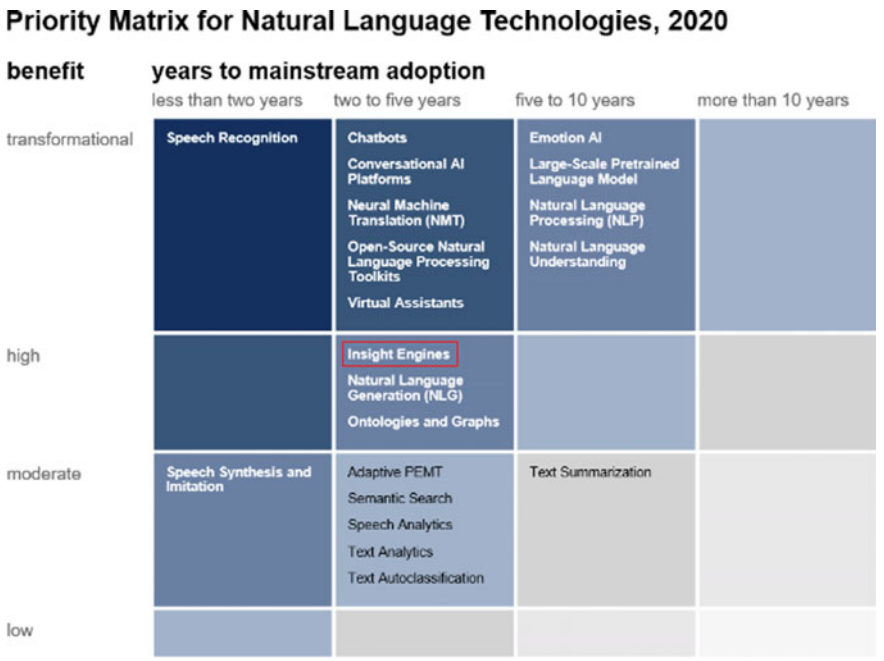


Fig. 2 Priority matrix for NLP tech, 2020

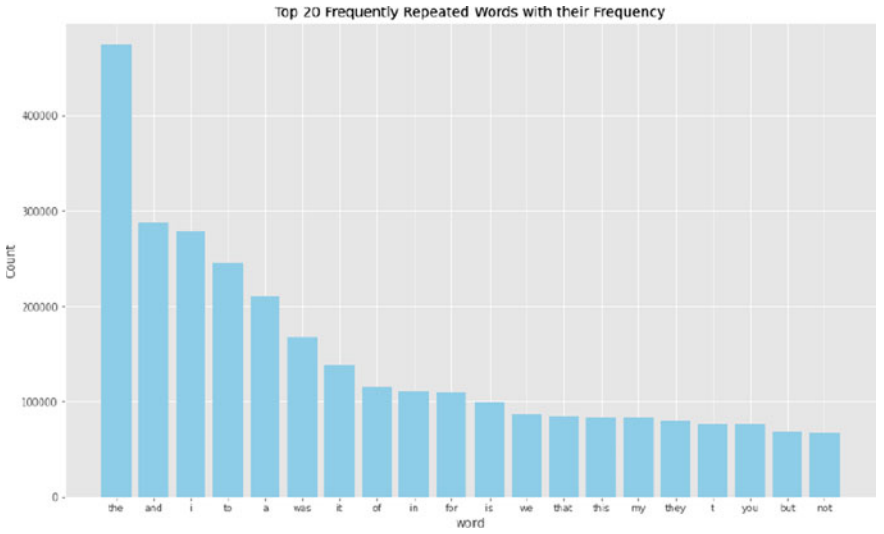


Fig. 3 Word count distribution from review text

```
def preprocess_string(string):
    '''Input:String
        Returns :String after removal of Stop words,punctuation and lemmatizing the string'''
    if (detect_language(string)!='English'):
        return ""
    string=string.lower()
    string=re.sub(r"[^a-z0-9']",' ',string) # Removing all non alpha-numeric characters
    # and replacing with space
    sentence=nlp(string)              #Loading the original string as a spacy sentence
    new_sentence=[word.lemma_ for word in sentence if word.text not in STOP_WORDS]
    new_sentence=' '.join(new_sentence)
    return new_sentence
```

Fig. 4 Sample text pre-processing function

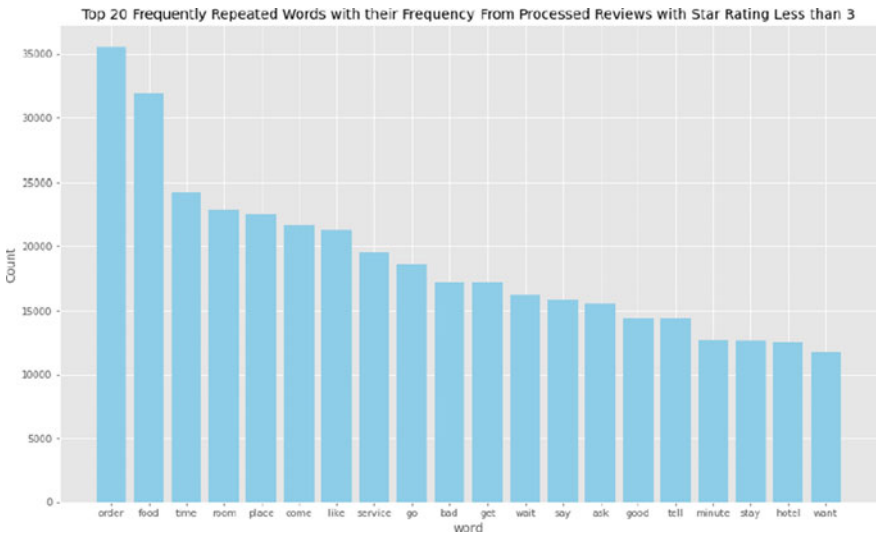


Fig. 5 Word count distribution from processed review text

acters and stop words before lemmatizing the words using SPACY library. This is a configurable function where you can add additional filter criteria as per business context and need. Upon cleaning the word count distribution would look like as show in Fig. 5.

(b) **Text Vectorization**:

Machines can only understand numbers and digits and for it read natural language, the words need to be converted to array of numbers also called vectors. There are various techniques to represent words as vectors. Some of them are Bag-of-Words-based vectors such as one-hot encoding, word frequency using count vectorizer, weighted word frequency using TF-IDF vectorizer, word to index, etc. and context-based vectors such as word2vec, Doc2vec (like word2vec but creates better sentence vectors), GloVe, FastText, etc.

To cluster the customer review texts into similar groups, converting the review sentences to sentence vectors using TF-IDF or doc2vec-based vectors is a good option. In the process, it is also advisable to identify collocations and create multi-gram words either using Gensim library's "Phraser" object or "ngram" parameter in TF-IDF function call. In some instances, implementing unsupervised Doc2vec to create sentence vectors could be very beneficial as it learns the sentence vectors and word vectors based on the context how words were used in the sentences. The word vectors created here in the process could also be used in the later stage of our solution.

Doc2vec—Sentence Vector Representation [3]:

Doc2Vec provided by Gensim library comes with two implementations:

- PVDM
  - Distributed Memory Version of Paragraph Vector (PVDM).
  - This architecture is like the continuous bag-of-words (CBOW) model in word2vec.
- PV-DBOW
  - Distributed Bag-of-Words Version of Paragraph Vector (PV-DBOW).
  - This architecture is the Skip-Gram version of Word2vec.

These two techniques along with few hyperparameters such as word dimension, epochs could be modified to evaluate the model performance using intrinsic measures. This will help pick the model that has learnt the best word vectors. Figure 6a, b compares the two techniques using similarity measure on word "Bad" and "Flu", respectively. We can infer that PVDM model has learned the vectors much better than PV-DBOW as the context-based word vector for "Bad" is closer to vectors for "horrible", "good", "terrible", etc. and word vector for "Flu" is closer to vectors for "Stomach_pain", "diarrhea", etc. using PVDM model.

(a)

Words similar to Bad PVDM(Similar to CBOW)	PVDBOW(similar to Skipgram)
horrible	memei
good	we
terrible	disarray
time	raise_voice
come	lo_behold
food	catch_guard
awful	300_dollar
eat	jumbo_jack
place	save_frustration
go	quick_bite

(b)

Words Similar to Flu PVDM(Similar to CBOW)	PVDBOW(similar to Skipgram)
stomach_pain	mermaid
diarrhea	detailed
feel_nauseous	rex
umc	spotty
stomach_ache	23pm
lose_weight	arbys
upset_stomach	beer_glass
allergic_reaction	presale
starchy	napkin_fork
food_poisoning	haunt

Fig. 6 **a** Intrinsic evaluation—word vector similarity for "bad", **b** Intrinsic evaluation—word vector similarity for "Flu"

(c) **Review Text Clustering** [4]:

Grouping similar reviews together requires us to cluster the sentence vectors using one of the unsupervised learning techniques. We will pick the most common clustering technique called K-means clustering for grouping the reviews. As this is an unsupervised technique, we will run the k-means on k-values ranging from 2 to 15. We could select the max value for the range based on heuristic approach. Some of the metrics that are used to evaluate the best clusters are.

(i) **Scree Plot (Elbow Technique)**
In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a dataset. The method consists of plotting the explained variance as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.
Generally, we consider the distortion (inertia) versus number of clusters plot and pick the elbow. But in Fig. 7, the distortion versus number of clusters plot is a smooth curve. Hence, plotting percentage change in inertia versus number of clusters could help find the best clusters.

(ii) **Silhouette Index**
The silhouette value is a measure of how similar a data point is to its own cluster (cohesion) compared to other clusters (separation).
The silhouette ranges from -1 to $+1$, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering

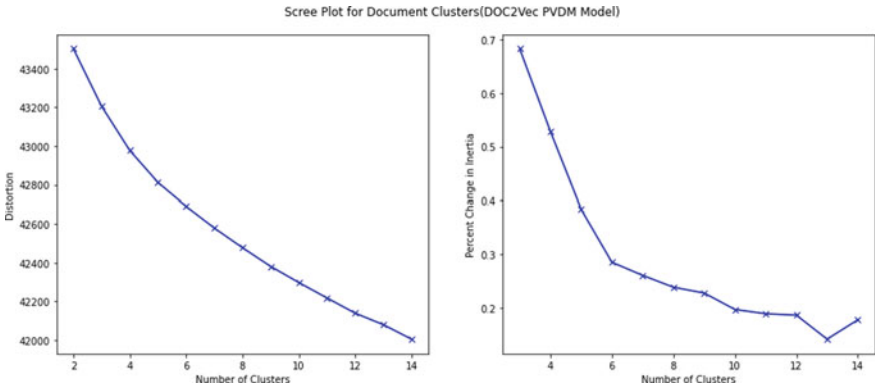


Fig. 7 Scree plot and percentage change in inertia plot for clusters from Doc2Vec PVDM vector representation

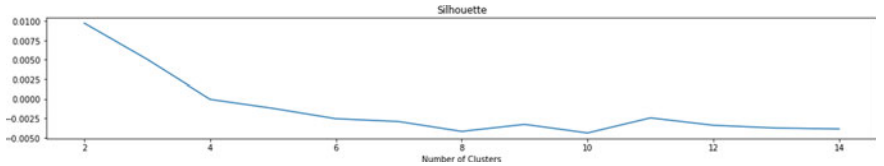


Fig. 8 Silhouette score for clusters from Doc2Vec PVDM vector representation

configuration may have too many or too few clusters. From Fig. 8, best cluster needs to be picked in conjunction with other metrics.

(iii) **Calinski-Harabasz Index**

The Calinski-Harabasz index is a measure of the quality of a partition of a set of data in automatic classification. It is the ratio between the inter-group variance and the intra-group variance.

Greater value of index denotes better cluster. From Fig. 9, best cluster needs to be picked in conjunction with other metrics.

(iv) **Davies-Bouldin Score**

The score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Thus, clusters which are farther apart and less dispersed will result in a better score.

The minimum score is zero, with lower values indicating better clustering. From Fig. 10, best cluster needs to be picked in conjunction with other metrics.

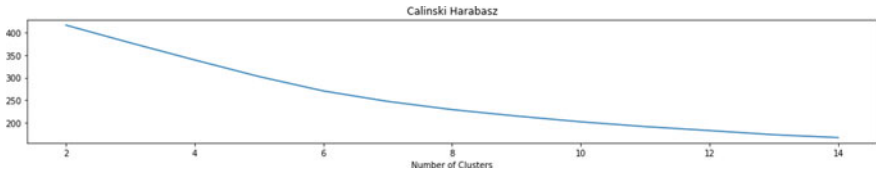


Fig. 9 Calinski-Harabasz index for clusters from Doc2Vec PVDM vector representation

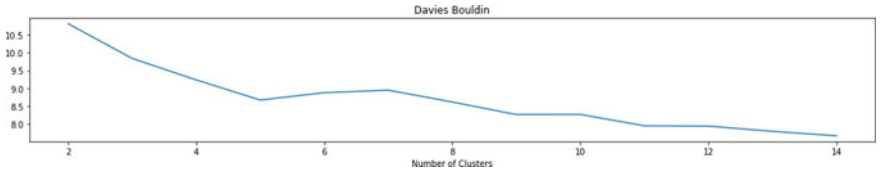


Fig. 10 Davies-Bouldin index for clusters from Doc2Vec PVDM vector representation

(v) **Word Cloud**

This technique involves building wordclouds based on frequency of words seen in each cluster from sentence (review) vectors. This could help visualize the most frequent words in each cluster and label them. However, visualizing wordclouds on range of clusters could be challenging. Hence, this approach is best if we want to compare and finalize between couple of k-values identified from earlier mentioned techniques.

Another wordcloud that we could build is by clustering word vectors from do2vec implementation. These clusters will group similar words together such as "Flu", "Stomach_pain", "diarrhea", etc. which would all belong to the same cluster. This approach will come handy toward later part of our analysis.

Based on the metrics and analysis, K = 6 gave best results for clustering and those were labeled based on the prominent words seen in each of the six clusters. Figure 11 shows the most frequent words that show up in each of the six clusters and their corresponding labels. The distribution of review texts by each cluster is mentioned in Fig. 12.

Fig. 11 Word cloud distribution based on frequency for best cluster (k = 6)

Fig. 12 Review cluster (k = 6) labels distribution

(d) **Other Techniques for Text Segmentation**:

Topic Modeling [5]**:**

This is one of the techniques used in unsupervised learning on natural language processing of texts to extract hidden topics. Here each topic is represented as a distribution of words and each document as a distribution of topics. Some of the topic modeling techniques are LSI (Latent semantic indexing), HDP (hierarchical Dirichlet process), NMF (non-negative matrix factorization), and LDA (latent Dirichlet allocation).

Topic Modeling Using LDA:

Latent Dirichlet Allocation (LDA) is one of the techniques that could be used to identify topics in various text documents. It is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.

The multi-gram words built from the review texts after removing special characters and Stopwords can be passed to the LdaModel() from genism with the number of topic matching the number of clusters (k = 6) to begin with. If the number of topics are known, it can be passed into the ldamodel() function, else multiple topic# could be evaluated. Using metrics such as perplexity score (low value), coherence score (high value), or visual techniques the best topic number can be picked. Figure 13 shows the top words and its corresponding weightage by each topic. Here the number of topic is picked as 6.

This information can be visualized using wordcloud as shown in Fig. 14.

```
[(0,
  '0.071*"order" + 0.043*"time" + 0.035*"wait" + 0.022*"bad" + 0.019*"take" + 0.017*"service" + 0.017*"location" + 0.016*"drive" + 0.015*"get" + 0.015*"place"'),
 (1,
  '0.028*"stay" + 0.021*"say" + 0.021*"tell" + 0.021*"go" + 0.016*"ask" + 0.015*"want" + 0.013*"come" + 0.012*"pay" + 0.011*"day" + 0.011*"get"'),
 (2,
  '0.039*"hour" + 0.021*"arrive" + 0.021*"mcdonald" + 0.018*"server" + 0.017*"las_vegas" + 0.017*"problem" + 0.017*"end" + 0.014*"\'" + 0.014*"feel_like" + 0.014*"read
y"'),
 (3,
  '0.124*"room" + 0.065*"hotel" + 0.027*"night" + 0.021*"check" + 0.020*"casino" + 0.015*"desk" + 0.013*"clean" + 0.011*"floor" + 0.010*"old" + 0.009*"pool"'),
 (4,
  '0.084*"food" + 0.035*"eat" + 0.022*"service" + 0.021*"come" + 0.021*"drink" + 0.019*"restaurant" + 0.018*"bad" + 0.018*"table" + 0.015*"chicken" + 0.014*"cold"'),
 (5,
  '0.037*"good" + 0.032*"place" + 0.025*"rio" + 0.018*"well" + 0.017*"like" + 0.016*"buffet" + 0.013*"strip" + 0.012*"price" + 0.011*"thing" + 0.011*"great"')]
```

Fig. 13 Topics and its top words

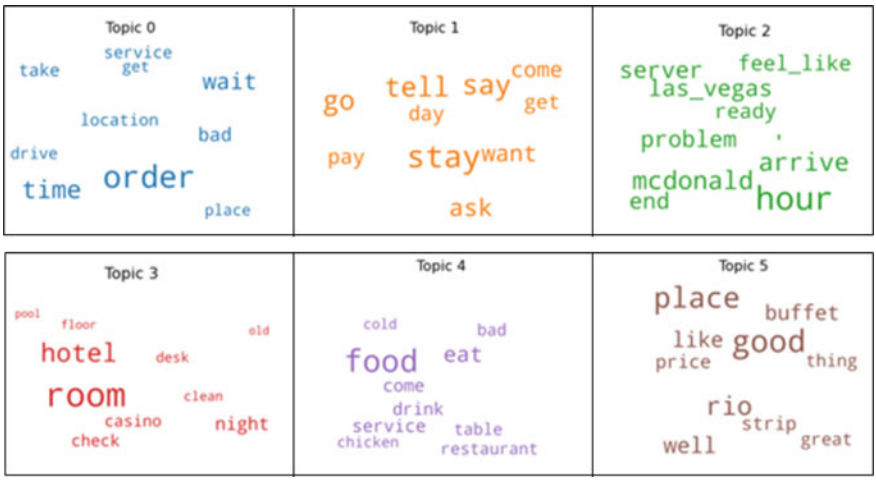


Fig. 14 Wordcloud by topics [6]

Visualization Using pyLDAVIS:

pyLDAvis is designed to help users interpret the topics in a topic model that has been fit to a corpus of text data. The package extracts information from a fitted LDA topic model to inform an interactive web-based visualization. The visualization is intended to be used within an IPython notebook but can also be saved to a stand-alone HTML file for easy sharing.

The figures from 15a till 15f show the distribution of top 30 words by each topic as we select the topic indexed circles on the left. The indices of the topics are ordered by the area in the descending order of popularity, i.e., bigger circles denote more documents are found on that topic in the corpus. The distance between the circles denote approximated topic similarity in two-dimensional space after multidimension scaling.

The red bars on the right denote the frequency of each word given a topic. We can compare this frequency with the overall frequency of the words by comparing it against the blue-shaded stacked bars for the word. The stacked blue bar denotes that some of the words are occurring in other topics as well. By adjusting the lambda (λ) using the slider to a lower value, we can re-rank the words to only show the top unique words by each topic. Adjusting the lambda puts more weight to the ratio of red to blue, i.e., the ratio of frequency of the word given the topic to the overall frequency of the word. We can also hover on top of the words to know which all topics they are prevalent in (Fig. 15a–f).

The above clustering/topic modeling techniques would help us label the review texts in a generic manner and these labels could only provide a high-level understanding as shown below. For further processing, we shall consider the six clusters formed from the text.

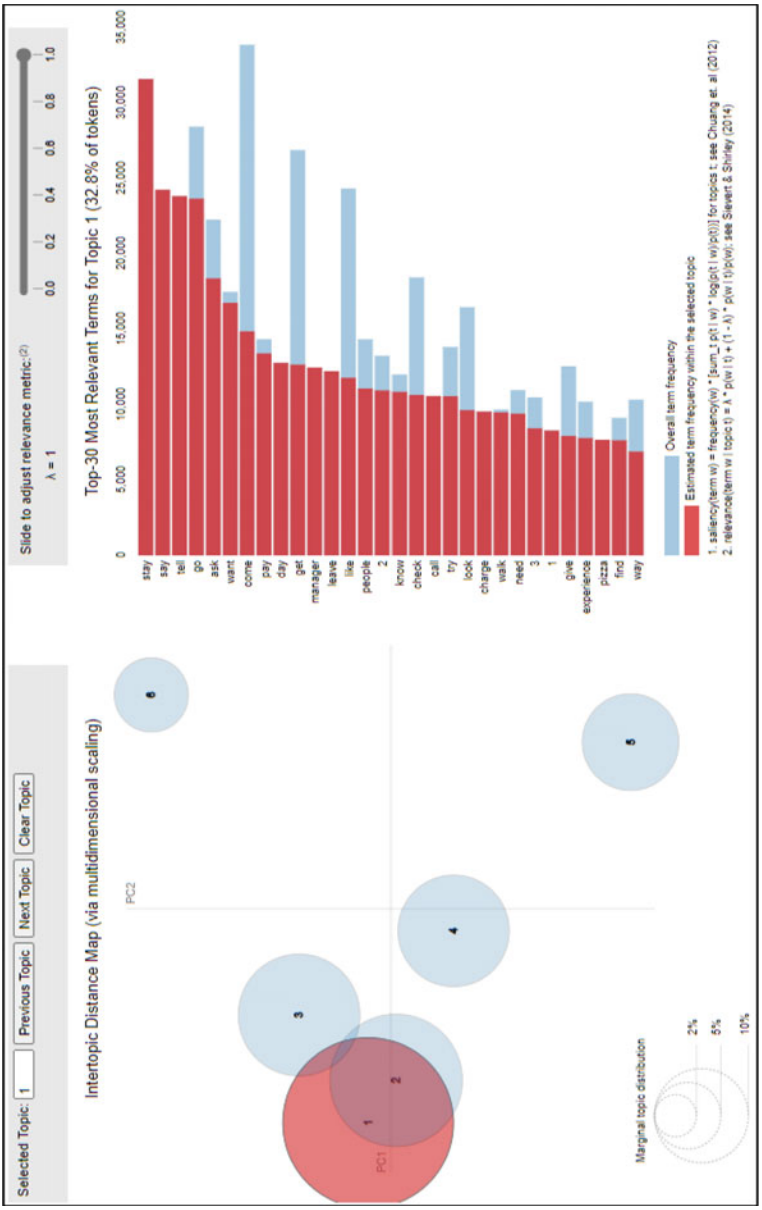


Fig. 15 **a** Topic 1—pyLDAvis view of top words. **b** Topic 2—pyLDAvis view of top words. **c** Topic 3—pyLDAvis view of top words. **d** Topic 4—pyLDAvis view of top words. **e** Topic 5—pyLDAvis view of top words. **f** Topic 6—pyLDAvis view of top words

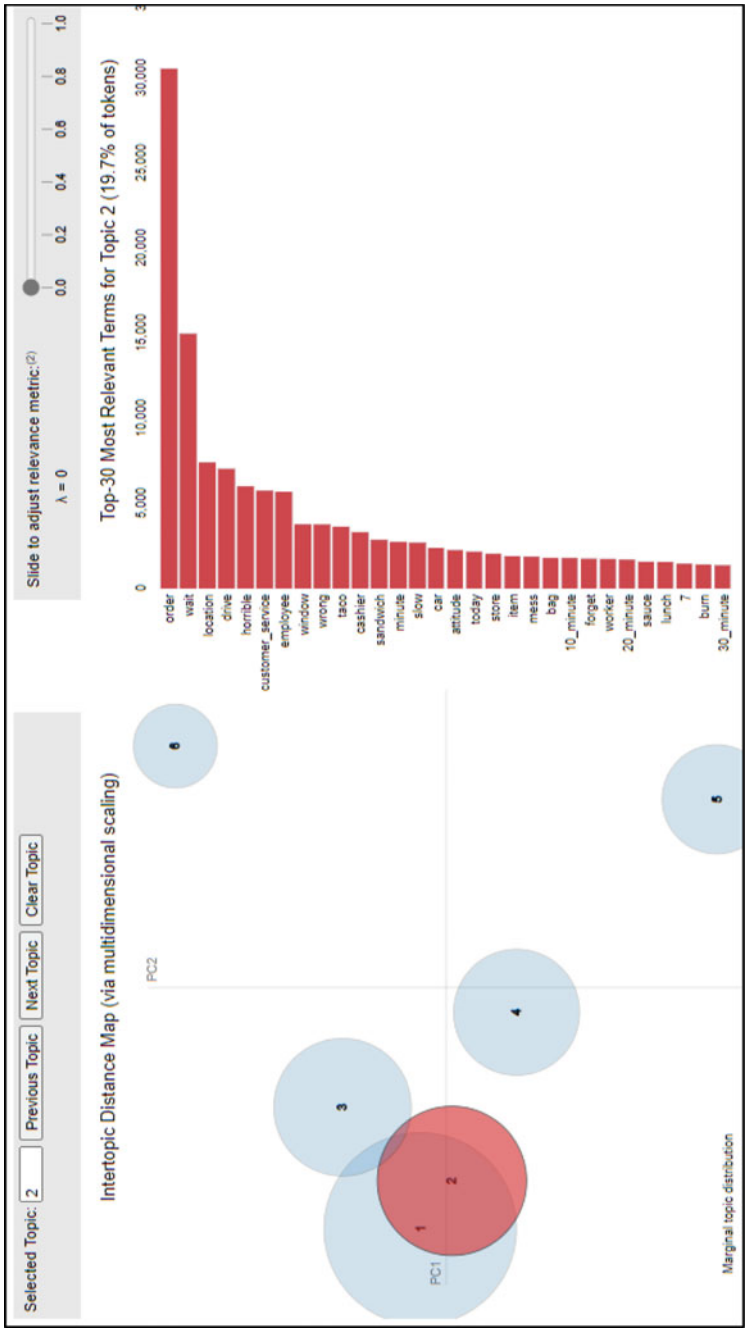


Fig. 15 (continued)

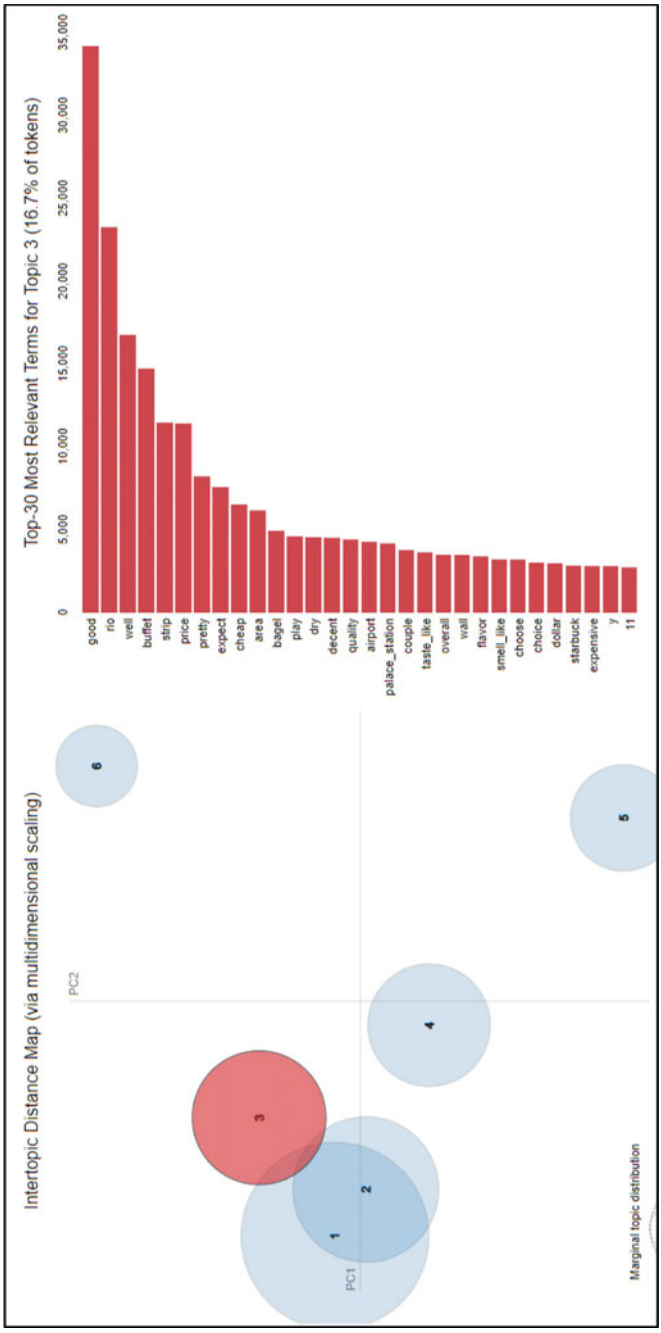


Fig. 15 (continued)

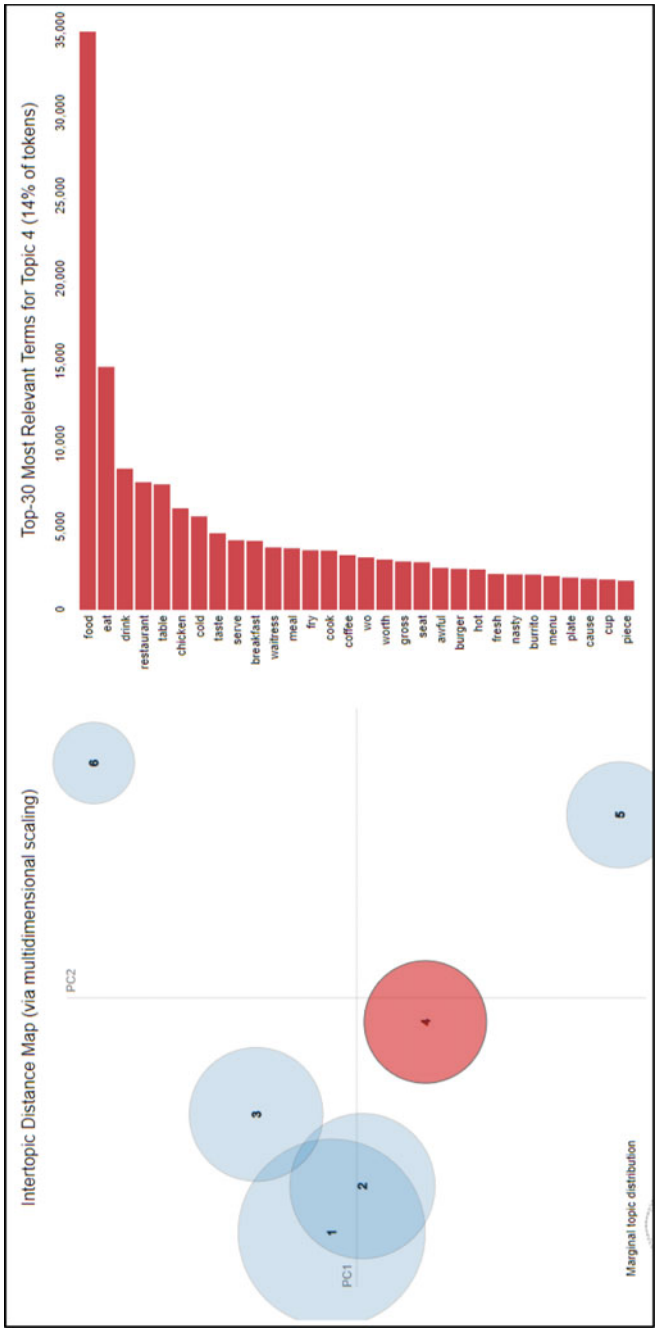


Fig. 15 (continued)

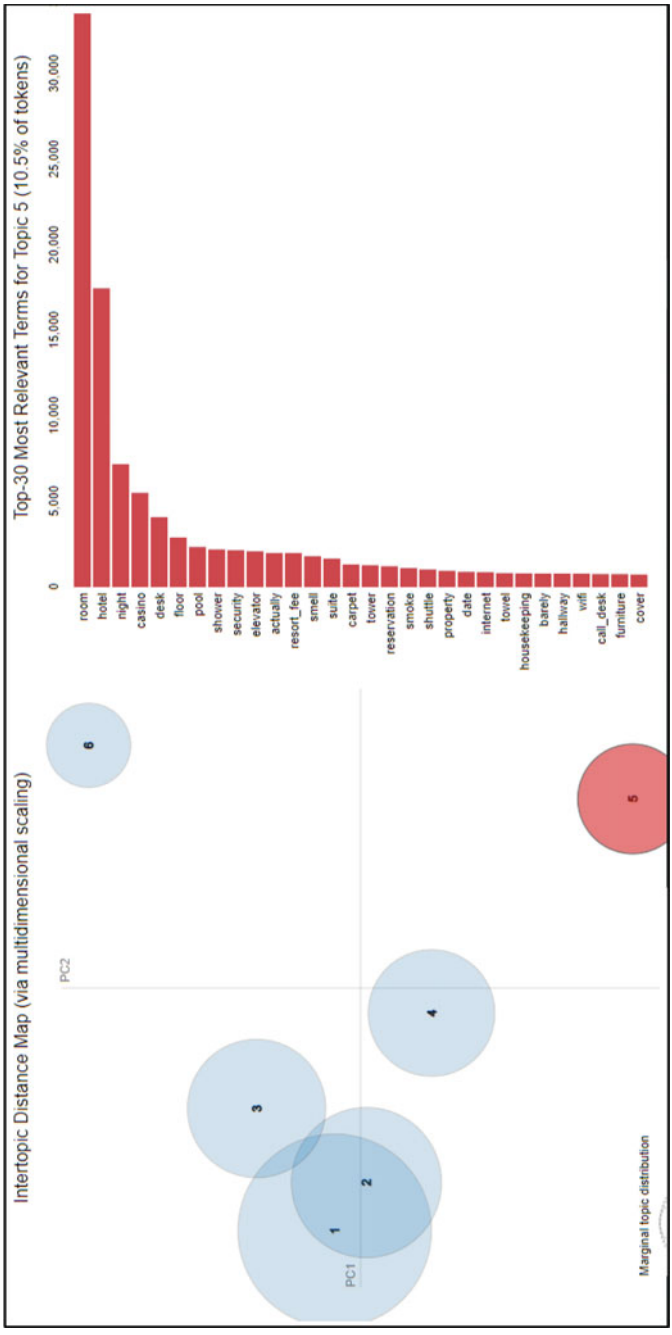


Fig. 15 (continued)

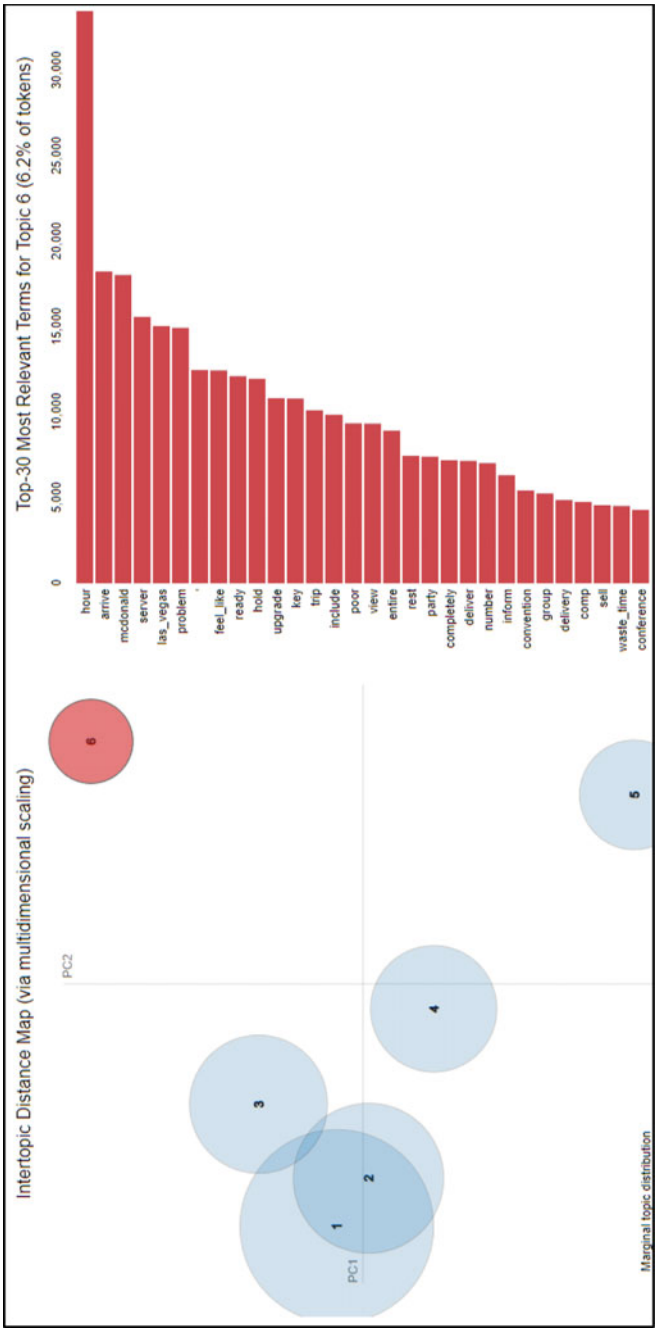


Fig. 15 (continued)

- Cluster 0 → Bad Service (7291 Reviews).
- Cluster 1 → Bad Pizza (3623 Reviews).
- Cluster 2 → Understaff and Wrong Order (11,883 Reviews).
- Cluster 3 → Bad Room (Hotel) and Cold Food (7805 Reviews).
- Cluster 4 → High Price and Bad Food (8981 Reviews).
- Cluster 5 → Slow Service and Dirty Place (9254 Reviews).

There could be review texts that were incorrectly assigned to a cluster or a prevalent topic, however the clusters contained texts mostly on above labels. If we need deeper understanding from the clusters and insights that could help take corrective measures, we will need more specific labels. This will be discussed in the upcoming sections.

(e) **Review Text Sub-clustering**

When review texts are grouped into clusters or assigned to most prevalent topic, some of the clusters or topics could be very large and could be generic. If we need more specific labels, we might end up creating too may clusters/topics with many overlapping. It becomes cumbersome to evaluate texts grouped into too many clusters as well.

A better approach would be to split the already created clusters/prevalent topics of each review into smaller clusters, i.e., sub-clustering the generic clusters to reveal smaller groupings in each. For instance, we can cluster all the review vectors in each of the six review clusters using K-means clustering algorithm with k-values ranging from 2 to 15 to reveal the best sub-clusters for each cluster based on the metrics discussed in Sect. 11.3c.

Figures 16 and 17 show two of the metrics after sub-clustering Cluster 4 (High price and bad food). Based on the metrics, we could decide if it would help us get deeper insights by sub-clustering.

Scree Plot (Elbow Technique):

Silhouette Index

If good sub-clusters exist, it will show up in the metrics and we could label the sub-clusters with specific values such as what kind of food the review talks about, for example, main course, dessert, etc. If further sub-clusters do not exist as show in Fig. 16, then it could be that the review vectors are so close that it can't be split any further or the vectors are too apart and each review in the cluster could be an individual sub-cluster. Nevertheless, it would be a good idea to explore the next section.

(f) **Review Word Clustering**

We have seen how doc2vec technique was used to build vectors for the review texts. In this process, doc2vec also built vectors for each word that occurs in the review text. We had explored this word vectors for intrinsic evaluation of the document vectors. The well-known idiom, "Birds of a feather flock together", applies to the word vectors and document vectors in a similar way. The way clustering document

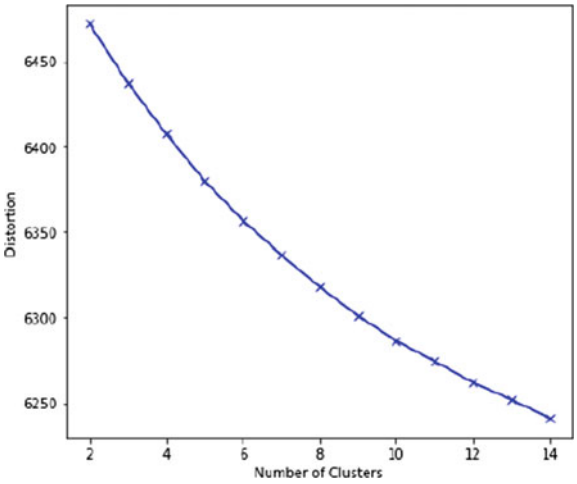


Fig. 16 Scree plot for sub-clusters in Cluster 4 of Doc2Vec PVDM vector representation

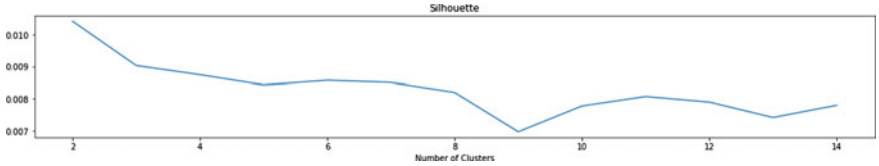


Fig. 17 Silhouette score for sub-clusters in Cluster 4 of Doc2Vec PVDM vector representation

vectors of review texts yielded homogeneous clusters/groups of reviews; clustering word vectors could also form clusters of words used in similar context.

As in Sect. 11.3c, we will run the k-means clustering but on word vectors obtained from Doc2Vec PVDM model for k-values ranging from 2 to 15 and evaluate the clusters using different metrics as shown in Figs. 18, 19, 20 and 21.

Scree Plot (Elbow Technique):

Silhouette Index:

Calinski-Harabasz Index:

Davies-Bouldin Index:

The clusters formed here could be visualized using wordclouds; however, we will use another technique to build the wordclouds. Instead of using word frequency to build wordclouds, we will represent the actual word cluster as shown in Fig. 22. To find the cluster to which a word belongs to, we calculate the Euclidean distance between the word vector and all cluster centroids, we label the word with a cluster number which has least distance between the word vector and the cluster centroids.

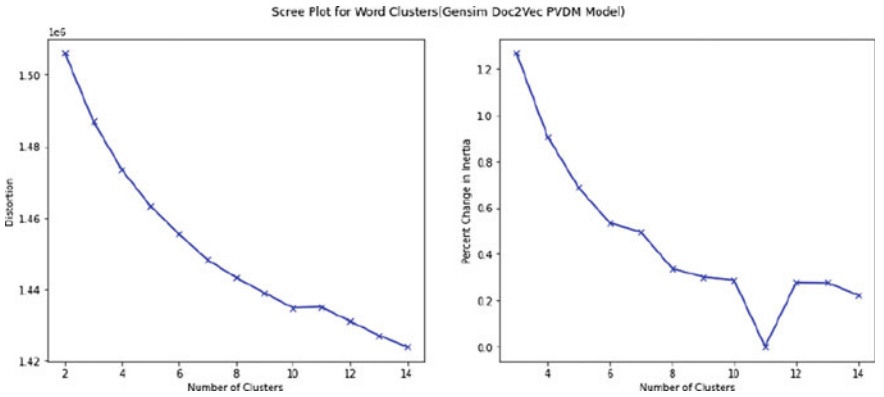


Fig. 18 Scree plot and percentage change in inertia plot for word clusters from Doc2Vec PVDM vector representation

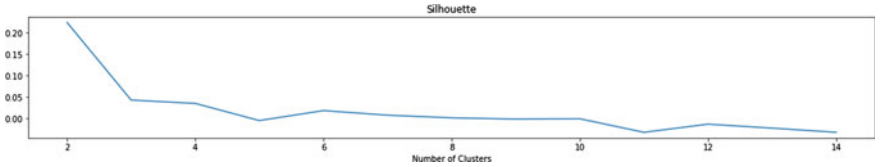


Fig. 19 Silhouette score for word clusters from Doc2Vec PVDM vector representation

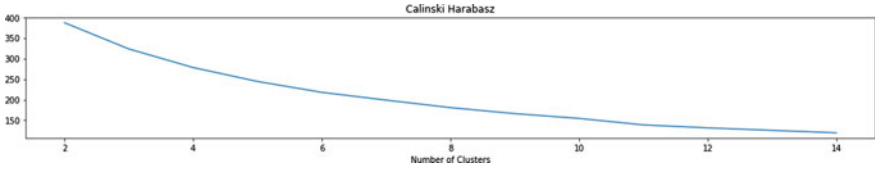


Fig. 20 Calinski-Harabasz index for word clusters from Doc2Vec PVDM vector representation

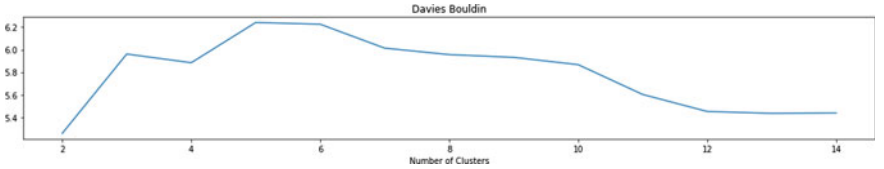


Fig. 21 Davies-Bouldin index for word clusters from Doc2Vec PVDM vector representation

From metrics above we see that K = 8 and 10 could be better values. After comparing the wordclouds, K = 8 is picked as better cluster.

Fig. 22 Word cloud distribution based on word similarity cluster (k = 8)

We see an interesting result where all words used in similar context are grouped together and we could label these clusters based on the words we see in the wordcloud.

- Cluster 0—**Health cluster**: words like diarrhea, calorie, headache, gas, etc.
- Cluster 1—**Party cluster**: words like guestlist, large group, meeting, wedding, etc.
- Cluster 2—**People cluster**: words like bouncer, dealer, waitress, housekeeper, etc.
- Cluster 3—**Facilities cluster**: words like music, blanket, shower, smoking room, etc.
- Cluster 4—**Infrastructure cluster**: words like air conditioner, alarm, furniture, etc.
- Cluster 5—**Restaurants type cluster**: words like cafeteria, manor, buffet, motel, etc.
- Cluster 6—**Food cluster**: words like duck, fruit, broth, big mac, etc.
- Cluster 7—**Restaurant names cluster**: words like Papa John, Domino, etc.

These cluster's centroids are going to be used to find which cluster a word (vector) belongs to. To get deeper actionable insights from review text, our focus will be to identify words from review texts belonging to Cluster 2 (Person), Cluster 4 (Infrastructure), and Cluster 6 (Food). Despite picking negative review comments, there are instances where users might have typed mixed sentiments like Positive, Neutral, and Negative, for instance, as shown below within brackets:

> "I went here for a weekday lunch and ordered a chicken/beef combo plate. *[NEUTRAL]* The meat was somewhat cold and the rice lacking flavour. *[NEGATIVE]* The salad had too much cabbage and not much else, and dressing was dumped on it. *[NEGATIVE]* …
>
> Awful customer service, the food quality has definitely gone down, and just because you are busy during lunch rush hour does not justify serving crap food and having a sour attitude. *[NEGATIVE]* The only two things that are preventing me from giving this place an outright zero(if i could) are the decent roast potatoes and the *possibly* still good lentil soup mentioned in my last review. *[POSITIVE]"*

We will explore how to capture the sentiments of the sentences along with words associated to Person, Infrastructure, and food from the review texts in the next section.

(g) **Sentiment Scoring**

Sentiment analysis is used to analyze the emotion of the text. In other words, it is the process of detecting a positive or negative emotion of a text. Sentiment analysis enables companies to know what kind of emotion/sentiment customers have for them. This can play a huge role because companies can improve their products/services based on the analysis of customer sentiments. While the challenge here is that different people write their opinions in different ways, some people express their opinion straight while some may prefer adding sarcasm to their opinion. Also, some might have both positive and negative opinionss.

VADER (Valence-Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. The processed review sentences can be passed through "NTLK Vader Sentiment Analyzer" to get polarity scores of the following categories [7] :

- Positive.
- Negative.
- Neutral.
- Compound.

The compound score is the sum of positive, negative, and neutral scores which is then normalized between -1 (most extreme negative) and +1 (most extreme positive). The more compound score closer to +1, the higher the positivity of the text.

Here are the advantages of using VADER which makes a lot of things easier:

- It does not require any training data.
- It can very well understand the sentiment of a text containing emoticons, slangs, conjunctions, capital words, punctuations, and much more.
- It works excellent on social media text.
- VADER can work with multiple domains.

We will pass all the negative review texts of the restaurants into a function call word by word to perform the below calculations:

a. Identify the word (vector) closest to the centroids of Cluster 2 (Person), Cluster 4 (Infrastructure), and Cluster 6 (Food) using Euclidean distance. If the distance is lesser than a set threshold value it would be considered part of one of the above clusters.
b. The specific word's frequency in the sentence is counted.
c. Calculate compound score of the sentences where the word is found using Vader and total the scores for all the reviews for the restaurant where the word was found.

Example: The above function will return a dictionary with items like below for each word it finds from the three clusters.

{(dairy,6): [−0.771, 2]}

- 'dairy' word from the review was found in one of the 3 clusters
- This word belongs to Cluster#6 i.e. food cluster
- The total compound score of all sentences containing 'diary' = −0.771
- Frequency of 'dairy' in all the reviews belonging to that restaurant = 2.

4 Result

Figures 23 and 24 show the final visualizations after our analysis on the negative review comments from two restaurants.

A Sample Review Comment for Restaurant 1: Top 10 words identified by our model related to food is shown with an underscore (_), person is shown with an italics, and Infrastructure is shown in bold. There are few more words that are part

of these clusters in this review text but aren't highlighted as they don't show up in top 10 based on total compound score.

> "I went here for a weekday lunch and ordered a chicken/beef combo plate. The meat was somewhat cold and the rice lacking flavour. The salad had too much cabbage and not much else, and dressing was dumped on it before I had time to respond properly with a "yes or no". Oh! and when I asked for the garlic and hot sauce "on the side", the lady working in an assembly line sort of setup said "sure" and dumped hot-sauce in one **corner** of my Styrofoam container, and garlic sauce in the other. On top of that, the hot sauce was REALLY cold! and made my luke-warm pieces of chicken and beef even colder and more unappetizing. When we *reached* the end of the assembly line, the lady informed us they did not take credit cards, and as we took out our debit cards, she let another customer pay (in the 10 seconds it took us!!) and then said in a rude *tone* "So are you ready to pay now?". Awful customer service, the food quality has definitely gone down, and just because you are busy during lunch rush hour does not justify serving crap food and having a sour attitude. The only two things that are preventing me from giving this place an outright zero (if I could) are the decent roast potatoes and the *possibly* still good lentil soup mentioned in my last review."

- "Categories of Negative reviews" mentions the distribution of the review comments by six clusters. (from doc2vec-based sentence vector clustering).
- "Top 10 Negative words related to Food/Person/Infrastructure" captures the individual words found in those clusters along with the frequency of occurrence in the review texts and total/average compound sentiment score of the texts.

Based on the above visualization, the top actionable insight for Restaurant 1 is.

- Most negative reviews received are under category high price and bad food. The restaurant needs to focus on improving the quality of food as most of the food items are highly priced, but customers are not getting a quality food.
- Under food category, they need to focus on improving shawarma, quantity of garlic used, improve flavor of food.
- Under service of staff, the tone of the working staff needs to be courteous.

Working on these recommendations would help the restaurant to receive better customer ratings. If the number of negative reviews is more, we could identify more significant areas of improvements.

The visualization shown in Fig. 24 is derived from negative review comments from another restaurant and these visualizations confirm that using this unsupervised technique lot of words pertaining to food, people, and infrastructure used in negative context could be extracted. There are some scope for improvement which will be discussed in upcoming section.

5 Conclusion

The benefit of Insights engine is to enable organizations to generate insights automatically from existing data. Data-driven decision-making is increasing in popularity

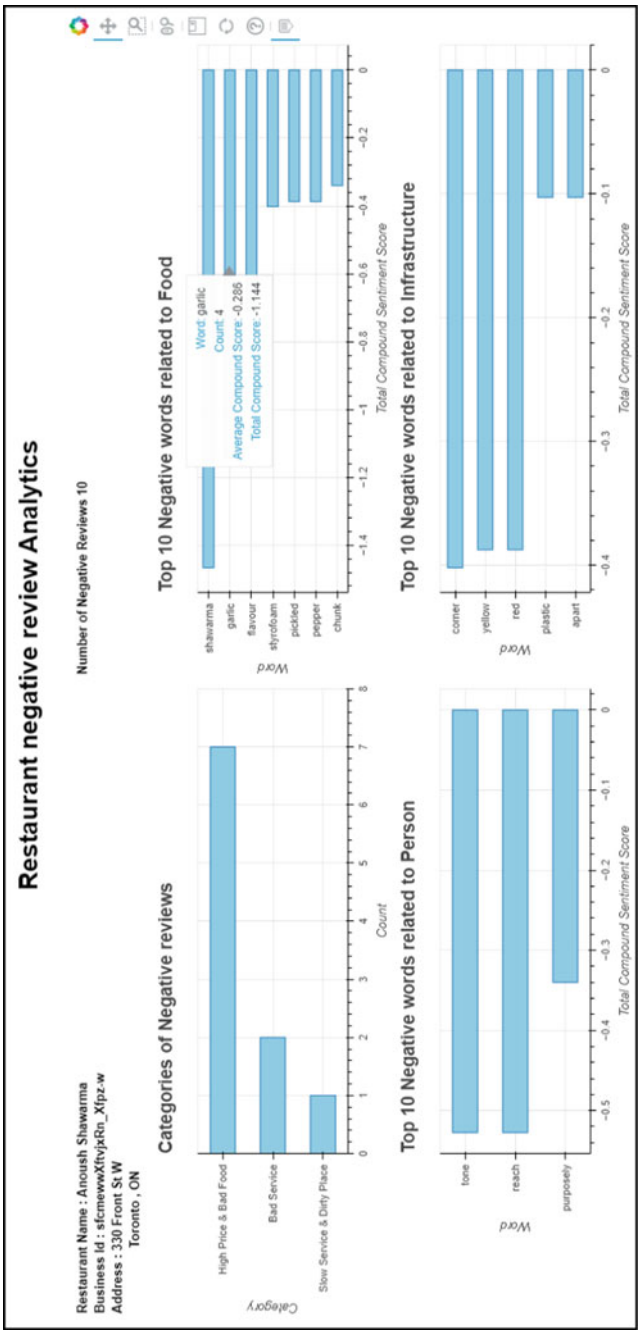


Fig. 23 Negative review analytics using unsupervised NLP on Restaurant 1

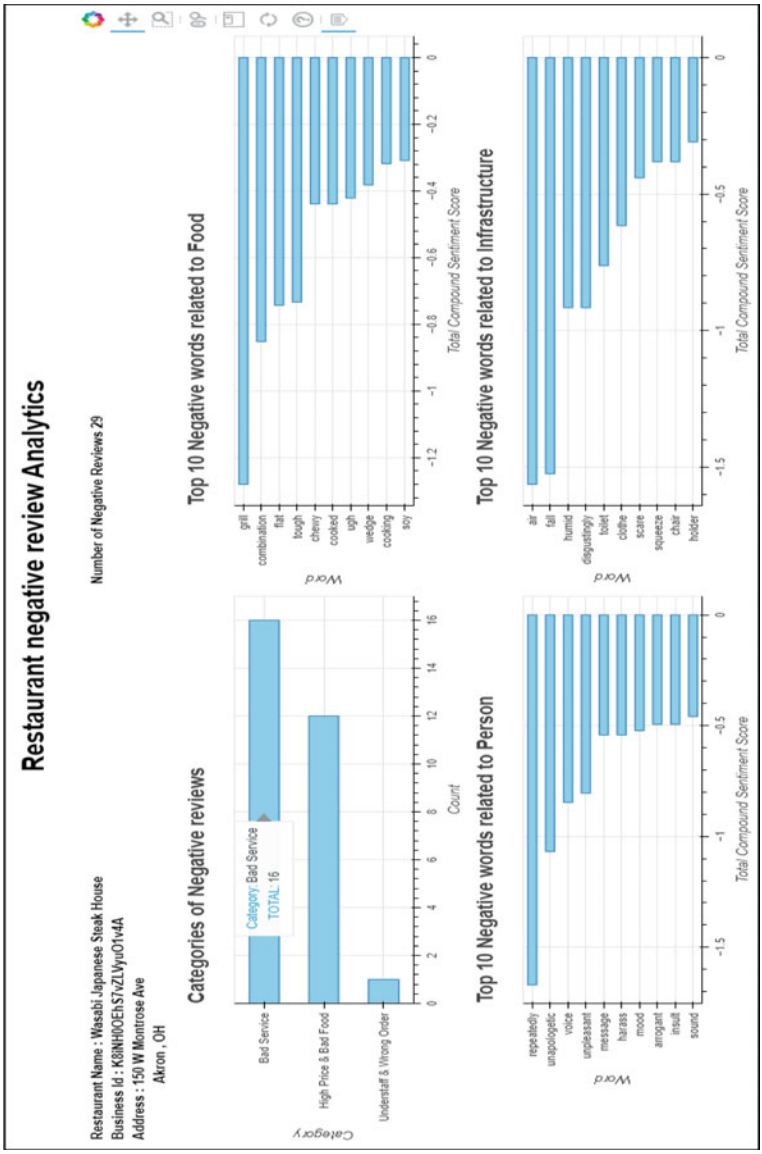


Fig. 24 Negative review analytics using unsupervised NLP on Restaurant 2

among managers and insights engine enable companies to base their decisions on insights extracted from data. Deriving deeper insights from large natural language text is always a challenge especially when they do not come with labels. Most businesses face challenges with unlabeled data, innovative measures need to be taken to label huge amount of data. The techniques discussed in this chapter show many ways that could be used to extract those important labels that could be converted into actionable insights. We were able to achieve this just by using the power of vector algebra on word vectors that has learned the context. Once the cluster centroids are learned and pickled/saved, we could reuse them to make prediction on new texts in a quick manner. This result could also be a base for any future labeling efforts.

This approach is not domain specific as it has been proven to extract deeper insights from other domains (like insurance) with free form text entered by application processing teams, calls logs captured in system, free form text in incident logs, etc. provided we have large volume of long texts.

These techniques have notably given an excellent outcome on natural language texts; however, we cannot evaluate the performance of these models as coming up with an evaluation metric on un-labeled data is not possible without a ground truth to compare with. The labels coming out from these techniques can assist data scientists to label more records to build new Named Entities or Classes. These labels can convert our solution into a supervised learning technique to predict NER (name entity recognition) or multi-classes and it will improve the overall performance of the insights derived from the text. We could also explore attention models to understand the context from multiple sentences.

Acknowledgements I would like to thank my family for their continued support during my research work. I also like to thank Mr. Ishwar Rao, Data scientist and ex-colleague who helped me brainstorm different NLP ideas and techniques. Special thanks to Mr. Mortha Sai Sriram whom I mentored during his internship with us and was able to implement all the ideas I had and put them into a good visualization.

References

1. https://research.aimultiple.com/insight-engine/
2. https://www.gartner.com/en/documents/3987154/hype-cycle-for-natural-language-technologies-2020
3. https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html
4. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics
5. https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
6. https://github.com/bmabey/pyLDAvis
7. https://analyticsindiamag.com/sentiment-analysis-made-easy-using-vader/

Chapter 12
Explainable AI for ML Ops

Sandeep Pathak

Abstract This chapter explains the significance of blending two emerging technologies in AI/ML—Explainable AI (XAI) and Machine Learning Operations (ML Ops) and demonstrates a focused use case that derives value from leveraging XAI to enhance ML Ops. The chapter starts by laying out the "growing pains" problems that enterprises are encountering to scale AI. We highlight the relatively low maturity of post-production ML processes thus exposing enterprises to reputation, compliance, and hence financial risk. We give a historical perspective on the rise of AI. Subsequent gain in mindshare of ML Ops is explained. XAI as a solution is introduced. After a brief explanation on XAI, we delve into the experimental setup and detail the results that demonstrate the potential of using XAI to enhance ML Ops. We end the white paper by reiterating the benefits, opportunities, and market potential and finally some recommendations.

Keywords Machine learning · Explainable AI · XAI · Cloud platforms · Machine learning operations · ML Ops · Solution · Pandemic · Opportunities

1 Introduction

Machine learning and the advanced variants like deep learning, reinforcement learning have shown significant promise in handling large volume and a variety of data and generating predictions. They are able to identify causality, which is beyond human capacity. However, the over-parameterization and lack of understanding of the learning process result in the perception of a "black box" model. We explain the issue and introduce the need for explainability for machine learning.

S. Pathak (✉)
Tata Consultancy Services, Ashburn, VA, USA
e-mail: pathak.sandeep@tcs.com

N. Sharma and M. Bhatavdekar (eds.), *World of Business with Data and Analytics*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-19-5689-8_12

1.1 ML and The "Last Mile" Problem

The rapid adoption of AI has also brought to the forefront the myriad growing pains that mirror the issues impacting enterprise software years ago—development across multiple teams, time to market, product quality, scaling up, monitoring, to name a few. Similar issues impact AI as it strives to scale up and across. There are products addressing the data engineering spectrum of the AI pipeline with promising commercial entrants like Snowflake. Vendors like Dataiku, H2O and cloud platform offerings like SageMaker are helping "democratize" AI by offering "operationalization" of ML models also known as "ML Ops" capabilities and end-to-end UI-based centralized platforms to meet the needs of all stakeholders from data scientists to business analysts. This helps mitigate the "friction" from data cleansing, model training, model management to model deployment, monitoring and feedback.

The focus changes from technical challenges to business value in post-production. Are the models making the predictions and recommendations as expected? How is the notion of "model drift" handled? What about "data drift"—changes to profile of data sets either slow or abrupt or data "poisoning" attacks. Are the model monitoring processes capable of alerting before it is "too late"? The notion of "late" implies models too slow to adapt to changing data profiles, which means lost business opportunities. It also means not realizing the negative impacts of predictions thus exposing enterprises to business and reputational risk. There are many instances like [1, 2] where AI predictions have gone awry (or "rogue") and they were pulled back after negatively impacting the use case.

ML Ops solutions, as of now, focus on improving the ML "experience", democratizing AI to non-data scientists. There is a gap in focus on the "last mile" aspect of the AI/ML lifecycle—ML Monitoring.

1.2 Keeping Tabs on the Model

ML practices offer the notion of measuring "model drift", i.e. a shift in the prediction power of the model. Standard measures like precision, recall, accuracy, AUC etc. can be used to keep track of model performance and can indicate when a model needs to be retrained. But this does not offer traceability all the way back to features, data, and reasoning. There is a very real possibility that the measures will not raise alarms while there has been a shift in the properties of data. The pandemic is perhaps the biggest trigger that has caused upheaval in many patterns in socio-economic behavior and data. Industries like mortgage lending had to keep consistent lending practices but respond to changes in conditions (e.g. how to do home inspections remotely?).

It is thus important that ML Monitoring be expanded to include capabilities that enable "deep" monitoring, i.e. not stop at validating labels but trace deviations back to features and data.

1.3 Explainable AI for Model Monitoring

The branch of AI called Explainable AI (XAI) holds promise here. As the name implies, XAI research and resulting frameworks like—ELI5, LIME, SHAP help determine feature importance and impact in the model results. Reference [3] shows how a XAI framework will indicate which feature contributed how much in the model's predicted result. This is a direct use of XAI. This ability of XAI can be extended to improve ML monitoring. In instances where the standard measures fail to indicate a significant drift in the model, XAI can be used to indicate "feature drift" and help identify a change in the constituency of ML results.

This is even more important in the COVID era where pre-COVID, COVID and post-COVID conditions are resulting in significant changes to data characteristics.

2 Literature Review

We did a review of articles and research papers that illustrate issues in the maturity of AI/ML and the need to make ML production ready, similar to software engineering lifecycle.

2.1 AI/ML Maturity

The survey, [4] by Algorithmia, gives insights into the explosive growth in AI/ML adoption and is fueled by many factors—adopters seeking to assimilate AI into their DNA seek to get benefits in key areas like reducing costs, generate better customer insights and intelligence, reduce friction in customer experience. While there is no dearth of issues in this AI/ML "wish list" that corporations aspire to address, the ML maturity of many organizations needed to bring this wish list to fruition is very much a work in progress. Very few organizations are at the highly mature end of the spectrum.

The survey in [4] also explains high maturity in many ways:

- Models in production for a long duration (years)
- Clear ML strategy alignment between management and analytics teams
- Traceability between investments and benefits from ML model predictions or recommendations, i.e. a good handle on AI/ML ROI
- Mature data and machine learning pipeline
- Established operationalized processes around deployment and monitoring of AI models
- Model governance
- Maturity to buy third-party models and customize/train to enterprise needs (e.g. ResNet).

2.2 Rise of ML Ops

Reference [4] also highlights enterprises that are much behind this high maturity level with only about half of them with any models in production. Also just deploying to production is not enough. It will matter how the deployment is done, how soon is it done, and how the models are being monitored. The "how" and "soon" aspects of ML development are addressed by instituting an extension of DevOps software engineering principles to encompass data and models used in ML; also known as "ML Ops". Reference [5] explains how ML Ops strives to extend the aspirations of software engineering to machine learning, namely:

- Faster turnaround time all the way from data cleansing to model deployment
- Facilitate faster rate of experimentation and adoption
- Governance, assurance of data quality and ethical AI.

DevOps has also matured in monitoring. Reference [6] describes how many frameworks like OpenShift offer monitoring or work seamlessly with tools ideal for monitoring. Open-source tools like Grafana have become popular to create dashboard to visualize metrics. Compared to the DevOps, ML Ops tools landscape is very much a work in progress and not surprising, given the fairly recent spike in AI/ML usage after being dormant for decades. The spectrum of ML Ops space has a few promising open-source candidates on one end and then native support by the leading cloud vendors. The commercials off-the-shelf (COTS) space is still sparse. In the open-source space, Ref. [7] describes frameworks like Kubeflow, ML Flow and recently launched ZenML show promise and are being leveraged. In the cloud landscape the three prominent vendors—AWS, Azure and Google Cloud all have stated support for ML Ops via proprietary offerings incorporated into their platform offerings—e.g. ML Ops via SageMaker on AWS. In the commercial space, there are products being launched at a brisk pace—Neptune.ai, Fiddler.ai that offer monitoring. Also, ML development platforms like Dataiku, H_2O and Databricks focus on enabling the ML development to deployment pipeline, i.e. a subset of the holistic ML Ops end-to-end flow.

Overall, the MLOps market is still in its early stages, with technology solutions emerging only in the last year or two for effective model management. Growing pains and maturity notwithstanding, the market potential for ML Ops as per [8] is pegged to reach $4 Billion (products and services) in a span of a few years, and ML Ops will be an intrinsic part of any organization's AI strategy.

2.3 ML Ops in Postproduction

In the previous section, we talked about the rapid evolution of ML Ops akin to the rapid adoption of DevOps to address similar issues—industrialization of AI/ML to offer rapid turnaround, reduce time to market, reliability and operations predictability.

Machine Learning poses its own unique challenges in terms of "Business as Usual (BAU)". Unlike software where the logic is deterministic, an ML model is highly dependent on data. Significant changes in macro conditions (COVID pandemic) will likely change the characteristics of data. Ref.[8] explains how this will cause models to decay and "drift" away from their initial training and hence expected behavior requiring retraining. Organizations deploying models into production need to extend their notion of ML Ops from deployment to incorporate continuous model monitoring. Monitoring metrics should indicate a trigger to retrain the model and address downstream risks.

3 Materials and Methods

Having set the context of issues in scaling up ML, the need for model explainability; we now detail the experimental setup, the process and outcomes of our experiments in using explainable AI in model monitoring.

3.1 Datasets

For this case study, we decided to use a dataset from Kaggle that contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Data Source—https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

There are 25 variables in the dataset:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- SEX: Gender (1 = male, 2 = female)
- EDUCATION: (1 = graduate school, 2 = university, 3 = high school, 4 = others, 5 = unknown, 6 = unknown)
- MARRIAGE: Marital status (1 = married, 2 = single, 3 = others)
- AGE: Age in years
- PAY_1: Repayment status in September, 2005 (−1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months, 8 = payment delay for eight months, 9 = payment delay for nine months and above)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)

- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- Default_Target: Default payment (1 = yes, 0 = no).

3.2 Explainable AI 101

The rapid adoption of AI has also brought with it a surge in aspects of AI like safety, fairness or bias and interpretability or explainability. The increasingly black-box nature of complex models makes it difficult to test in the traditional notion of software testing, though opinion is divided on this. The model-based decisions still need to enable enterprises to abide by regulations and ensure fairness in decision-making. If the decisions of these models are "explainable" then it brings transparency.

Doshi-Velez and Kim [9] and Gilpin et al. [10] give an introduction to Explainable AI refers to techniques that helps to make AI solutions human understandable. XAI techniques add trustworthiness, auditability, and credibility to our algorithmic solutions. This has taken on significance as standards and regulations are being developed like [11], especially in EU to report explainability and demonstrate fairness in AI.

With the rise and demand of Explainable AI, there have been many frameworks developed to advance our understanding of ML predictions. Reference [12] gives a good overview of the popular frameworks. These frameworks are categorized by their ability to do local (explains each observation/row as required) and global (explains the importance of model features in general) interpretability of ML outcomes. Some libraries are:

- SHAP—SHapley Additive exPlanations: game theory principles to explain individual as well as global predictions.
- LIME—Local Interpretable Model Agnostic Explanations: explain individual predictions
- What-If Tool—Visualizer for Tensorflow data
- ELI5—Explain Like I am 5: for simpler models
- AIX360—AI Explainability 360: new library. Extensible.
- Skater—new library for local and global explanations.

All of these techniques have specific advantages. We chose to work with SHAP considering its accuracy, compute optimization, open-source framework and variety of explainers offered for different machine learning algorithms making SHAP the most popular XAI library currently in use. Note that material is only recently made available, which detail and compare XAI frameworks [12, 13].

SHAP is based on Shapley values using coalitional game theory to distribute payouts from a game. Important work done in [14] takes the game theory approach to ML and uses SHAP to explain the prediction of an instance "X" by computing the contribution of each feature prediction. SHAP offers explainers that can support local as well as global explainability based on Shapely values. The explainers are developed specifically for each or a family of ML model types. SHAP authors proposed KernelSHAP an alternative, kernel-based (can be used with any machine learning algorithm) estimation approach for Shapely values inspired by local surrogate methods. They also proposed TreeSHAP (optimized for tree-based algorithms), an efficient estimation approach for tree-based models.

3.3 Explainability and ML Monitoring

Our proposition is to extend the usage of XAI not only for debugging and exploration but also for real-time monitoring as it can help indicate deviations from model prediction to data reality.

Question becomes, how can we implement it? What are the correct indicators pointing the change or "poisoning" in the test data?

We were able to get a substantial indication about change in the test data using these two approaches:

1. Interpreting Local explainability for similar type of users over time
2. Tracking SHAP loss for each iteration of prediction that takes place.

The flow of our experiment is as follows: Upon model deployment, we will monitor metrics from our model output that suggest how well the model is performing. We will show how the same metric might not indicate changes in data. At this point, SHAP loss metric will be utilized to show how it can play a vital role during model monitoring. We will discuss more about both of these approaches in our "Experiment" section.

4 Exploratory Data Analysis

Before our experiment we took the following EDA steps:

- Missing value check
- Numerical and categorical variables segregation
- Plotting numerical feature distribution
- Bivariate analysis with dependent variable
- Dropping duplicate rows/columns
- Outlier treatment
- Multivariate correlation check.

Here are some highlights from our detailed exploratory data analysis.

Missing values

None of the columns has any missing values and data set is complete.

Target variable Count plot is shown in Fig. 1.
Heatmap is shown in Fig. 2.

Conclusions based on the EDA

- There are some unknown and undocumented categories in variables MARRIAGE, EDUCATION, which needs to be combined with OTHERS
- As per the Data Definition PAY_* variables start with − 1 (e.g., Pay duly). However, there are values − 2 and 0, which also needs to be assumed as Pay Duly

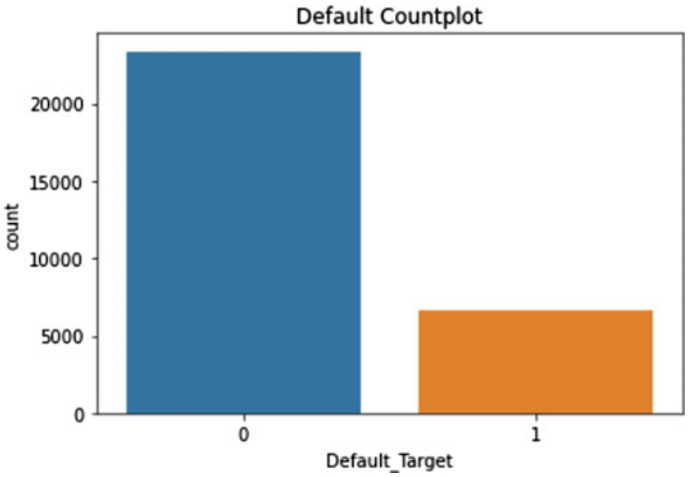


Fig. 1 Target variable countplot

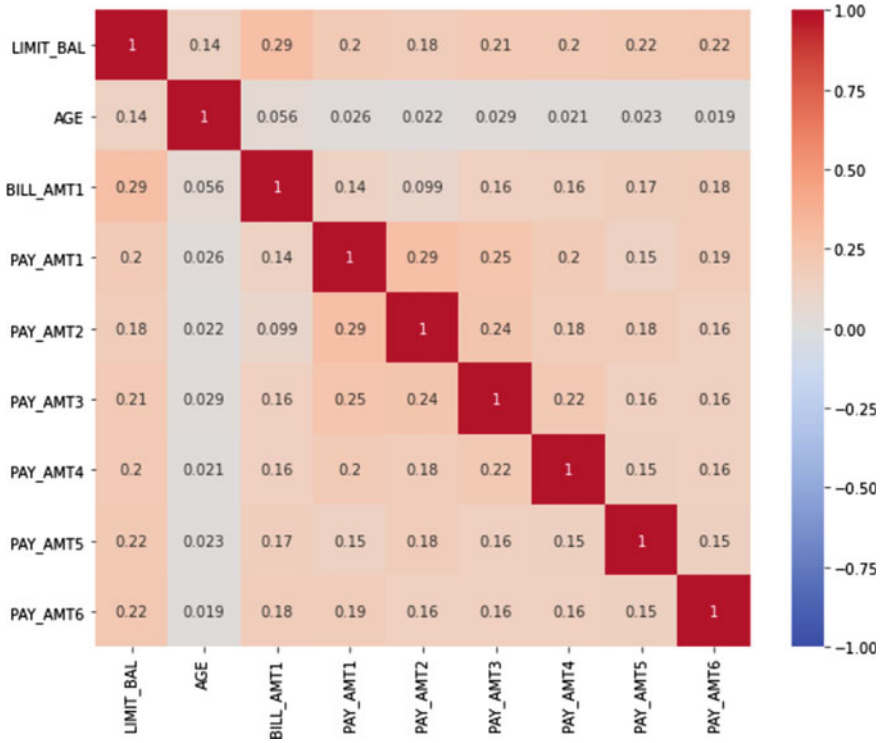


Fig. 2 Heatmap

- Many customers are having very high LIMIT BALANCE compared to rest of the customer base, which will be treated as outliers.
- Heatmap clearly suggests extremely high CORRELATION between variables BILL_AMT*, which either needs to be dropped or converted to a Principal component
- Target variable is imbalanced in the ratio of 1:3.5 (Default: Not Default).

5 Experimental Analysis

We trained and conducted explainability on six different machine learning models ranging from Logistic Regression to Deep Neural Networks. Out of six models, XGBoost had the best results according to confusion matrix. For the purpose of model monitoring, the model performance details of all the other models are not relevant. We will continue with XGBoost model from here on.

This experiment will be explained in steps to be able to clearly showcase each step of model monitoring and how we determine change/manipulation in test data.

```
Cross-Validation score for Xgboost:0.7915160194821584
-------------------------------------------------
Predicting the outcomes using Xgboost
.................................................
Classification report using Xgboost

              precision    recall  f1-score   support

           0       0.81      0.87      0.84      5373
           1       0.58      0.47      0.52      2006

    accuracy                           0.76      7379
   macro avg       0.70      0.67      0.68      7379
weighted avg       0.75      0.76      0.75      7379

-------------------------------------------------
Confusion matrix using Xgboost

[[4687  686]
 [1064  942]]
-------------------------------------------------
ROC AUC score using Xgboost

0.670957906106731
```

Fig. 3 Classification metrics of XGBoost model

Load and train the XGBoost model

After training XGBoost model for the unbalanced data set with the ratio of 1:3.5 (default: not default), we logged standard classification metrics like classification report, confusion matrix and AUC score to determine the results as shown in Fig. 3.

Train Tree Explainer from SHAP using trained XGBoost model

```
explainer = shap.TreeExplainer(best_model)
shap_values = explainer.shap_values(X_test)
```

SHAP Summary plot

```
shap.summary_plot(shap_values, X_test)
```

SHAP summary plot gives an overview of which features are most important for a model and the feature effect. The plot is for SHAP values over all samples. The plot as illustrated in [13] shows the following information:

- Feature importance: features are sorted in descending order.
- Feature impact: the horizontal location shows if the feature had a higher or lower prediction impact (magnitude).

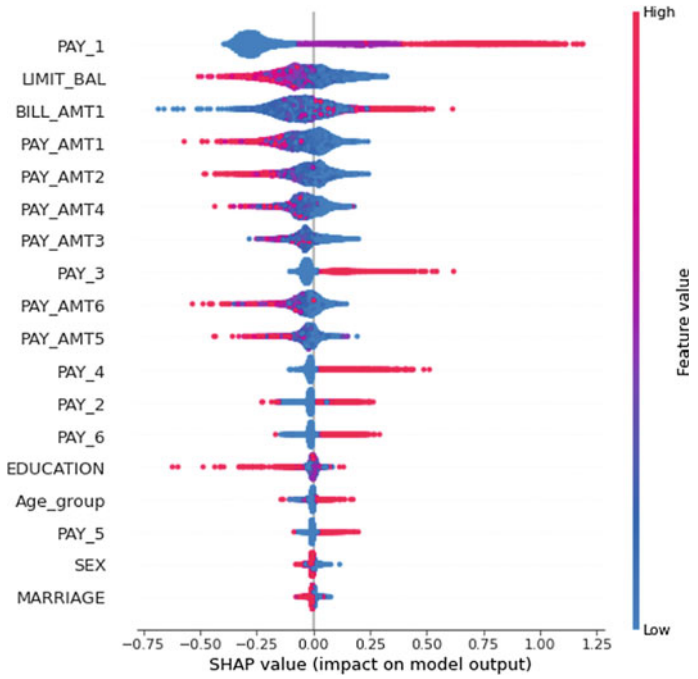


Fig. 4 SHAP plot

- Contribution: the color represents the "push/pull" impact on the prediction, i.e. red represents towards (or higher) the prediction and blue represents away (or lower).

Figure 4 reveals, for example, that a high **PAY_1** pushes the prediction towards **Default.** It clearly indicates that **PAY_1** and **LIMIT_BAL** are the two most important features, which have the highest impact on our predictions.

6 Results

6.1 SOLUTION 1: Local Explanation with One Particular Observation

For this particular step, we will pull out one particular observation out of our test data as that would explain features for a particular customer.

Observation Features: represented in Fig. 5.

Here, the variable **Pay_1 =2,** which indicates that the customer is 2 months behind in his payments. **Bill_Amt1 = 71876,** meaning the customer's outstanding amount.

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	Age_group
7330	240000.0	2.0	2.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	71876.0	3100.0	2400.0	2367.0	2435.0	2500.0	3000.0	0.0

Fig. 5 Observation features

Fig. 6 Explanation summary

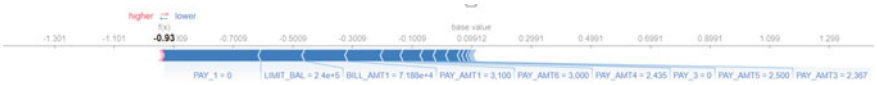


Fig. 7 Explanation summary -2

Our model in normal scenarios predicted that this customer will **DEFAULT,** and rightly so. The local explanation summary plot looks like the following graph shown in Figure 6, where we can understand how each feature is pushing the prediction either towards **DEFAULT** or **NOT DEFAULT**.

In this plot, **PAY_1** is the most contributing feature pushing the prediction towards **DEFAULT** while **LIMIT_BAL** is pushing towards **NOT DEFAULT.**

Introduce data change by swapping Pay_1 with Pay_6 for this observation.

After introducing this change in the observation, we noticed a distinct change shown in Fig. 7. **Pay_1** was still the most significant contributor of the prediction but this time it was pushing the prediction towards **NOT DEFAULT.**

This change in local explainability alerts us to look into the data and investigate further to either inspect the data credibility and quality or to retrain model. This is the first approach among the two to monitor the model using XAI.

6.2 SOLUTION 2: Global Monitoring: Iterating the Model 100 Times, Introduce the Manipulation from the 30th Iteration

The objective of this step is to prove that primary metrics like **AUC & Precision** might still remain similar after completely swapping two features (**PAY_1, PAY_6**). Whereas, **SHAPloss** would picture a completely different story indicating the change in data from the 30th iteration itself.

We ran the model for 100 iterations, bootstrapping 50 observations on each iteration from the test data. From the 30th iteration, the data change was introduced as explained.

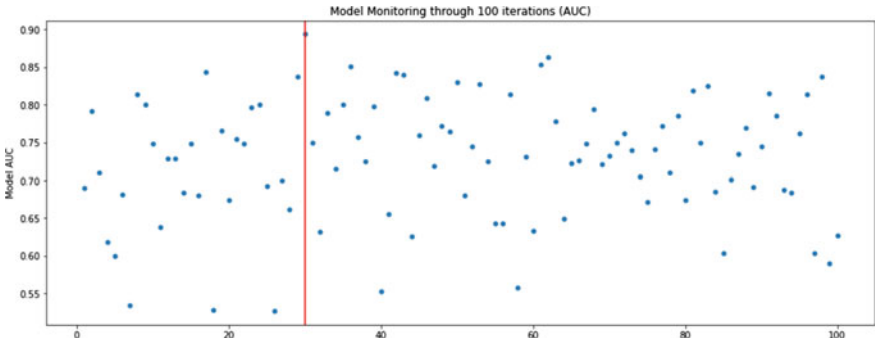


Fig. 8 Model monitoring through 100 iterations

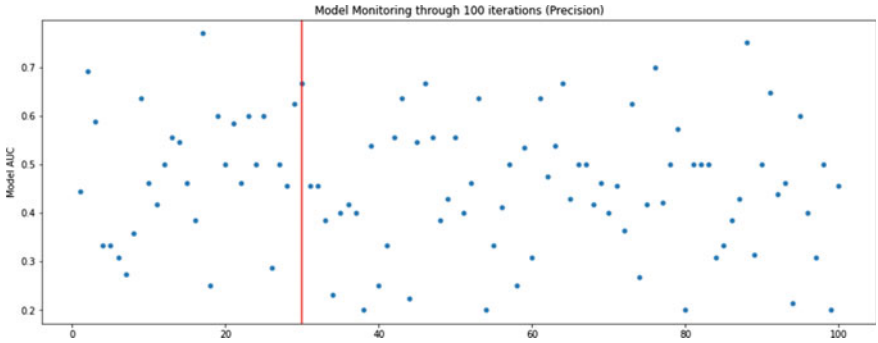


Fig. 9 Plot Precision over the 100 iterations

Plot AUC over the iterations: shown in Fig. 8.

Plot Precision over the iterations: shown in Fig. 9.

Plot SHAP loss over the iterations: shown in Fig. 10.

7 Conclusion

The results from model metrics and distinct change in SHAP loss monitoring clearly demonstrate that while model monitoring would not have indicated that anything is a miss, the XAI monitoring clearly indicates a change in data. This should suggest to the ML Ops team to investigate the test data further to identify either a data quality issue OR the need for retraining the model.

In a PwC survey detailed in [15], over 67% of business leaders indicated that "*AI and Automation will impact negatively on shareholder trust levels in the next 5*

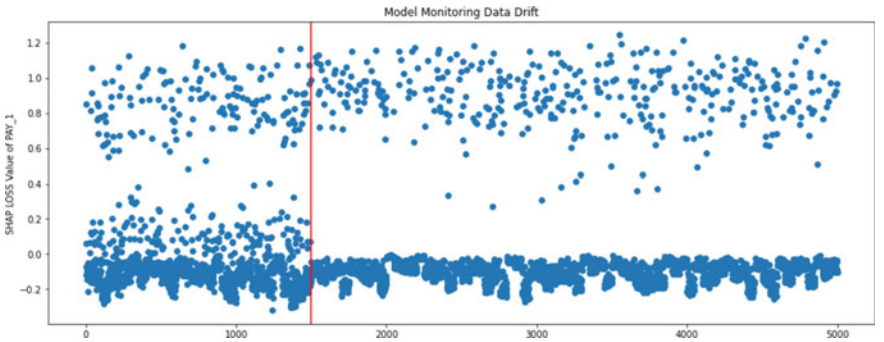


Fig. 10 Plot SHAP loss over the iterations

years". AI, through deep learning and reinforcement learning is becoming increasingly sophisticated and evolving into an algorithmic black box. Enterprises need to be able to place their belief in the outcomes of these complex models. They need to be able to instill trust in their customers and shareholders. The model outputs also need to comply with regulations to demonstrate fairness and lack of bias. This ability needs to be available not only when models are approved but in post-production too.

The need for real-time predictions is increasingly becoming important. In such a context, it is critical that enterprises enhance their ML Ops capabilities with real-time XAI frameworks. The ability to have multiple levels of model behavior validation will prove invaluable. The XAI capability to indicate feature changes will greatly facilitate data quality and validation processes. It will also be an integral part of cyber security where "data poisoning" attempts can be monitored.

Various trends detailed in articles like [16] or analyst reports like [17, 18] highly recommend that readers put XAI at the top of their list of AI/ML topics to understand and gain expertise in. The benefits of XAI in compliance are getting mainstream. The potential in ML Ops needs to be understood and exploited.

References

1. Analytics India (2017) 8 Real Life Examples When Algorithms Turned Rogue, Causing Disastrous Results. https://analyticsindiamag.com/8-real-life-examples-algorithms-turned-rogue-causing-disastrous-results/
2. Simonite T (2020) How an algorithm blocked kidney transplants to black patients. Wired. https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/, 26 October 2020
3. Google (2020). Explainable AI. https://cloud.google.com/explainable-ai/, 2 February 2020
4. Algorithmia. 2020 state of enterprise machine learning (2020). https://algorithmia.com/state-of-ml
5. McKnight W (2020) Delivering on the vision of MLOps: a maturity-based approach. GigaOm. https://gigaom.com/report/delivering-on-the-vision-of-mlops/, 21 January 2020

6. Witzenman M (2017) Monitoring OpenShift: three tools for simplification. Red Hat. https://www.openshift.com/blog/monitoring-openshift-three-tools. 24 August 2017
7. Vizard M (2021) Open source platforms vie with ITvendors for management of MLOps. IT Business Edge. https://www.itbusinessedge.com/blogs/it-unmasked/open-source-platforms-vie-with-it-vendors-for-management-of-mlops.html., 25 January 2021
8. Schmelzer R (2020) The emergence of ML Ops. Forbes. https://www.forbes.com/sites/cognitiveworld/2020/03/08/the-emergence-of-ml-ops/?sh=2949ac8d4698, 8 March 2020
9. Doshi-Velez F. & Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. https://arxiv.org/abs/1702.08608v2. 2017.
10. Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M. & Kagal, L. Explaining Explanations: An Overview of Interpretability of Machine Learning. https://arxiv.org/abs/1806.00069. 2018.
11. European Parliament. EU guidelines on ethics in artificial intelligence: Context and implementation. https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf. 2019.
12. Analytics India. *8 Explainable AI Frameworks Driving A New Paradigm For Transparency In AI.* https://analyticsindiamag.com/8-explainable-ai-frameworks-driving-a-new-paradigm-for-transparency-in-ai/. 2019.
13. Molnar, C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/. 2021.
14. Lundberg, S. & Lee, S. A Unified Approach to Interpreting Model Predictions. Paper Presented: 31st Conference on Neural Information Processing Systems. https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. 2017.
15. Price-Waterhouse Coopers (PwC). Explainable AI: Driving business value through greater understanding. https://www.pwc.co.uk/audit-assurance/assets/pdf/explainable-artificial-intelligence-xai.pdf. 2017.
16. Shaan. Explainable AI is gaining momentum among VCs and startups. https://medium.com/the-blue-stars/explainable-ai-is-gaining-momentum-among-vcs-and-startups-445906869d01. 2020, October 24.
17. Gartner. 5 Trends Drive the Gartner Hype Cycle for Emerging Technologies, 2020. https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020/. 2020.
18. Haranas, M. Gartner's Top 10 Technology Trends For 2020 That Will Shape The Future. CRN. https://www.crn.com/slide-shows/virtualization/gartner-s-top-10-technology-trends-for-2020-that-will-shape-the-future/8. 2019, October 28.