SOLVING BUSINESS PROBLEMS ACROSS SALES, MARKETING, OPERATIONS, AND HR USING AI, ML, DEEP LEARNING SENTIMENT ANALYSIS

Project Submitted to the
SRM University AP, Andhra Pradesh
for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology
in
Computer Science & Engineering
School of Engineering & Sciences

submitted by

Neelofar Shaik(AP21110010047)

Rishitha Kancharla(AP21110010059)

Kavyasri Gullipalli(AP21110010060)
Madhukar Sai Babu Gadde(AP21110010277)

Under the Guidance of

Dr. Rajiv Senapati

Department of Computer Science & Engineering
SRM University-AP
Neerukonda, Mangalgiri, Guntur
Andhra Pradesh - 522 240
May 2025

DECLARATION

I undersigned hereby declare that the project report **Solving Business Problems Across Sales, Marketing, Operations, and HR using AI, ML, Deep Learning Sentiment Analysis** submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology in the Computer Science & Engineering, SRM University-AP, is a bonafide work done by me under supervision of Dr. Rajiv Senapati. This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree of any other University.

Place	:	Date	: June 13, 2025
Name of student	: Neelofar Shaik	Signature	: .....
Name of student	: Rishitha Kancharla	Signature	: .....
Name of student	: Kavyasri Gullipalli	Signature	: .....
Name of student	: Madhukar Sai Babu Gadde	Signature	: .....

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SRM University-AP

Neerukonda, Mangalgiri, Guntur

Andhra Pradesh - 522 240

CERTIFICATE

This is to certify that the report entitled **Solving Business Problems Across Sales, Marketing, Operations, and HR using AI, ML, Deep Learning Sentiment Analysis** submitted by **Neelofar Shaik, Rishitha Kancharla, Kavyasri Gullipalli, Madhukar Sai Babu Gadde** to the SRM University-AP in partial fulfillment of the requirements for the award of the Degree of Master of Technology in in is a bonafide record of the project work carried out under my/our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Project Guide	Head of Department
Name : Dr. Rajiv Senapati	Name : Dr. MuraliKrishnaEnduri
Signature:	Signature: .....

ACKNOWLEDGMENT

I wish to record my indebtedness and thankfulness to all who helped me prepare this Project Report titled **Solving Business Problems Across Sales, Marketing, Operations, and HR using AI, ML, Deep Learning Sentiment Analysis** and present it satisfactorily.

I am especially thankful for my guide and supervisor Dr. Rajiv Senapati in the Department of Computer Science & Engineering for giving me valuable suggestions and critical inputs in the preparation of this report. I am also thankful to Dr. MuraliKrishnaEnduri, Head of Department of Computer Science & Engineering for encouragement.

My friends in my class have always been helpful and I am grateful to them for patiently listening to my presentations on my work related to the Project.

Neelofar Shaik, Rishitha Kancharla, Kavyasri Gullipalli, Madhukar Sai Babu Gadde

(Reg. No. AP21110010047, AP21110010059, AP21110010060, AP21110010277)

B. Tech.

Department of Computer Science & Engineering

SRM University-AP

ABSTRACT

This study seeks to combine the use of machine learning and AI based analytics to enhance decision making in four core areas of business; Human Resources (HR), Sales, Marketing and Public Relations (PR). The HR department uses techniques such as Logistic Regression, Random Forest, Support Vector Machines (SVM), XGBoost, and Deep Learning (Artificial Neural Networks) to predict employee turnover and key factors that influence employee retention. In the Sales department, we use ARIMA, SARIMA, XGBoost and Facebook Prophet for time series forecasting which enhances inventory management and sales trends forecasting. The Marketing department uses techniques such as customer segmentation, clustering includ ing K-Means Clustering, Gaussian Mixture Models (GMM), DBSCAN and UMAP to enhance marketing campaigns and increase customer interaction. The PR department uses NLP and machine learning algorithms such as SVM, Na¨ ıve Bayes and Logistic Regression to analyze customer reviews in the context of sentiment analysis. This study enhances business intelligence and customer engagement and optimizes business processes with data-driven approaches, while combining insights from all departments.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION TO THE PROJECT

In the data driven era, organizations in various industries are leveraging the power of machine learning (ML) and artificial intelligence (AI) to make better decisions and improve performance. This research applies the use of ML techniques in four core business functions: Human Resources (HR), Sales, Marketing, and Public Relations (PR) to develop better workforce management, sales prediction, customer segmentation, and opinion analysis. Through the use of sophisticated predictive models, organizations are able to make data-driven, informed decisions that support growth and enhance efficiency. HR analytics is one of the most important factors in maintaining the employee base by predicting employee attrition which helps the organization identify the key drivers of employee turnover. Employment satisfaction, employee work-life balance and employee compensation are analyzed using methods like Logistic Regression, Random Forest, SVM, XGBoost and Deep Learning (ANN) in this work. The findings are useful to the HR team in formulation of policies that can enhance employee loyalty and decrease turnover rate of the workforce. Another very vital business intelligence component is sales forecasting to assist firms in predicting demand and thus controlling inventories. The time series models (ARIMA, SARIMA, Prophet) and the machine learning regression models (XGBoost) are used in this study to forecast the sales in future from the past data. These models are used by firms to predict the seasonal fluctuations, variations and customer purchasing behavior and hence improve

on their resource planning. The optimization of marketing is performed using customer segmentation with the help of machine learning clustering algorithms such as K-Means, GMM, DBSCAN, and UMAP. On the basis of the spend and interaction behavior, customers can be segmented so that companies can design specific marketing campaigns, personalized offers, and customer loyalty programs that can help in increasing the customer retention. Lastly, PR sentiment analysis uses Natural Language Processing (NLP) and machine learning algorithms (such as Support Vector Machines, Naive Bayes, Logistic Regression, and Random Forest) to analyze customer comments from social media platforms and online reviews. This enables organizations to determine public sentiment, monitor brand reputation, and address customer complaints appropriately. In conclusion, based on the findings from these four departments, this study reveals how organisations can leverage AI and ML analytics to enhance labour stability, enhance sales outcomes, build customer ties, and establish the brand.

Chapter 2

MOTIVATION OF THE PROJECT

In the rapidly changing digital era, data-driven decision-making is an important factor in every sector. This project brings the core departments—HR, Sales, Marketing, and Public Relations—together under one data-driven platform. The driving force of this project is to apply data science and machine learning methods to improve operational efficiency, forecast trends, and streamline decision-making. Through the deployment of different machine learning models, deep learning frameworks, and data analysis methods, this project will offer actionable business insights that facilitate business expansion and workforce effectiveness.

2.1 IMPORTANCE OF DATA-DRIVEN INSIGHTS IN EDUCATION

Data-driven insights are central to education, especially in applied research and capstone projects. As the focus on AI and machine learning increases, students need to work with real-world datasets to close the gap between theoretical knowledge and practical applications. This project illustrates how combining machine learning models across various business functions gives a holistic view of predictive analytics, classification methods, and optimization techniques.

2.2 WHY CAPSTONE ENGINEERING PROJECTS NEED REAL-WORLD DATA ANALYSIS

Capstone projects act as the connecting link between academic education and practical applications. The use of real-world data sets in this project enables students to:

- Operate with high-level, industry-based data.
- Gain insight into data preprocessing, feature engineering, and model performance evaluation.
- Overcome issues related to class imbalance, overfitting, and interpretability of outcomes.
- Have experience with deployment tactics and business intelligence tools.

This project is a case study in the application of machine learning to decision-making in various business sectors, increasing the pragmatic value of capstone projects.

2.3 BRINGING IDEAS TO LIFE IN THE FORM OF SOLUTIONS

Bringing abstract ideas to reality is one of the central elements of this project. Merging AI models into corporate operations guarantees theoretical constructs are turned into real results. This step-by-step transformation process is described in this section.

2.3.1 Identifying Problems Across Business Functions

Every department—HR, Sales, Marketing, and PR—has a different set of challenges. Attrition prediction for employees, forecasting sales, analysis of customer sentiments, and public perception management are key issues resolved in this project.

2.3.2 Applying Machine Learning and AI Techniques

There are various types of algorithms used in this project, such as Logistic Regression, Random Forest, XGBoost, SVM, KNN, CNN, and LSTMs, for extracting insights from structured and unstructured data.

2.3.3 Visualization and Interpretation of Results

Visualizations like Confusion Matrices, Feature Importance plots, and trend predictions facilitate data-informed decisions. These insights close the gap between raw data and actionable strategies.

2.3.4 Deployment and Real-World Implementation

Visualizations such as Confusion Matrices, Feature Importance plots, and trend forecasts enable stakeholders to make data-driven decisions. These insights bridge the gap between raw data and actionable strategies.

2.4 PROVIDING FOR DETAILED DOCUMENTATION AND PRESENTATION

In-depth documentation is crucial for the success of any data-based project. This project Project success is heavily influenced by proper planning and execution. This project uses a systematic pipeline:guarantees:

- Accurate recording of data collection, preprocessing, and modeling procedures.
- In-depth performance evaluation with measurements like Precision, Recall, and F1 Score.
- Detailed visualization reports to facilitate better interpretability of results.
- Organized presentation format that meets industrial standards.

2.5 SUCCEEDING THROUGH EFFECTIVE PLANNING AND PROJECT EXECUTION

Project success is heavily influenced by proper planning and execution. This project uses a systematic pipeline:

1. **Data Acquisition and Cleaning** – Providing high-quality, unbiased data.
2. **Model Selection and Training** – Selecting the most appropriate algorithms for every department.
3. **Performance Evaluation** – Comparing models on main metrics.
4. **Integration and Deployment** – Combining results from all departments into one workflow.

2.6 A SPACE FOR PROFESSIONAL AND PERSONAL DEVELOPMENT

This project offers professional and personal development opportunities. It allows for experiential learning and equips individuals with relevant

skills that can be used in professional life.

2.6.1 Technical Skill Enhancement

Through this project, individuals have practical experience in:

- Machine learning libraries such as TensorFlow and Scikit-learn.
- Deep learning models such as CNNs and LSTMs.
- Preprocessing data techniques such as feature engineering, missing values handling, and data normalization.

2.6.2 Analytical and Problem-Solving Skills

Analytical thinking is needed to handle real-world data. This project improves:

- The skill to comprehend and interpret intricate data structures.
- Hypothesis testing and model selection skills.
- The skill to optimize models for improved performance.

2.6.3 Collaboration and Communication

Because this project involves several business functions, collaboration is essential. Participants enhance:

- Team coordination and collaboration with cross-functional teams.
- Communication skills to report findings to stakeholders.
- The skill to write concise and clear documentation.

2.6.4 Project and Time Management

Effective project timelines and deliverables management is essential. This project assists in:

- Segmenting tasks into manageable milestones.
- Effectively allocating resources.
- Responding to challenges and iterating solutions upon feedback.

2.7 FINAL THOUGHTS ON THE ROLE OF MOTIVATION

Motivation is one of the compelling forces for problem-solving innovation. This project is not only proposing technical solutions but also stimulating business intelligence development, critical thinking, and collaboration. Through solving actual problems in life using AI and data science, it is a holistic learning exercise for future practitioners.

Chapter 3

LITERATURE SURVEY

Previous research has also highlighted different approaches to measuring the effectiveness of human resource management and has stressed the importance of using robust analytical methods. A research work carried out using multi-mode fuzzy logic control was able to review employee competency in de tail by using membership functions and fuzzy set optimizations to improve the overall productivity of the workforce. This approach emphasises the role of data driven decision making in HR management and is in line with our study which focused on machine learning based attrition prediction. Although fuzzy logic is useful for workforce evaluation, our research goes beyond this by combining more sophisticated classification models including logistic regression, random forest and deep learning to enhance the precision of the predictions and assist in effective HR decision making [1]. Previous studies have focused on the importance of knowledge management in the development of HRM innovation and the mutual enhancement between these two components to enhance the value of human capital. Research has shown that the implementation of knowledge management systems is positive in enhancing HRM by improving decision making, employee training and develop ment and overall organizational performance. This paper goes further than this by applying machine learning techniques to HR data to help understand employee attrition and workforce trends. Thus, our approach incorporates predictive analytics to the model, which is consistent with the existing literature and shows how using data-based

approaches it is possible to enhance the effectiveness of HR management and strategic decision making [2]. It has been shown that Human resource management (HRM) is an essential part of enterprise success in the context of adapting to changing economic conditions and dynamic workforce management. Previous studies have established that the complexity of HRM is increasing and that there is a need to improve decision making with data. Research further establishes that HR departments are no longer seen as administrative functions that only process paperwork but rather as functions that use data analytics and machine learning to improve HR operations. Previous research on the integration of decision tree algorithms in HRM has shown how data mining can be applied to evaluating the workforce, predicting attrition and supporting decision making. Our study is a continuation of the previous work which supports the use of advanced machine learning models including logistic regression, random forests, SVM and deep learning in the assessment of HRM and the identification of employee retention strategies and work force optimization [3]. Human Resource Management (HRM) is crucial to the realization of organizational objectives by properly controlling the capabilities of the workforce. Previous research has shown that adopting data-driven HRM practices can increase the effectiveness of the workforce and, therefore, the organization's performance. Research further recommends that HRM should move from traditional administrative roles to embrace technology such as machine learning and knowledge management to create value from human capital.Furthermore, the use of decision tree algorithms in the field of HR analytics has been found to be useful in the assessment of employee performance and turnover. Various works also present the possibility of integrating HRM with lean quality principles and constraint management to demonstrate its criticality in the linkage of

workforce management with business objectives. In line with these studies, our research develops complex machine learning models including logistic regression, SVM, and deep learning to identify potential leaving employees and suggest appropriate retention strategies in a data driven approach to workforce planning [4]. Technology has developed very fast and it has affected the HRM in the organization, thus HR professionals have to use data in managing workers. Previous research has called for the integration of knowledge management with HR innovation to enhance the creation of value from human capital. Other studies have pointed out the effectiveness of machine learning techniques such as decision tree algorithms and fuzzy logic control in identifying employees' competencies and improving the effectiveness of the HR department. Moreover, HRM is no longer viewed as a supportive function but as a critical partner that links workforce planning with organizational goals. New studies show that HR professionals in high-tech firms need to respond to the digital revolution by applying predictive analytics to improve employee participation and loyalty. Our research goes beyond previous studies by applying logistic regression, SVM, and deep learning machine learning models to predict employee turnover, identify critical retention factors, and offer data-informed HR suggestions to enhance the decision-making process in HRM in today's organizations [5]. The development of HRM has resulted in the growth of the application of HR analytics in the improvement of workforce management with numbers. Studies show that HR analytics is crucial in cutting down on the costs of the workforce, improving the selection of candidates, the management of talents, and the level of employees' engagement. Furthermore, the application of machine learning techniques in the HR department has been found to help in the prediction of employee turnover, job satisfaction, and factors that

lead to employee retention. Moreover, the study also establishes that HR digitalization has enabled the use of predictive analytics in decision making. Previous research also reviews different HR analytics tools and their importance in improving the HR functions and the workforce productivity. Moreover, HRM has shifted from being a supportive and administrative department to a strategic partner that supports the organization's strategic goals. This research goes beyond previous work by applying advanced machine learning techniques such as SVM, logistic regression, and deep learning to predict employee turnover to support data-driven HR decision making [6]. Different strategies of pattern recognition are frequently used in various areas, such as fraud detection and predictive analytics in HRM. Previ ous studies have shown that logistic regression models are useful for identifying patterns and anomalies in the datasets, like fake job adverts and employee behavior analysis. The research on novel logistic regression for identifying fraudulent job posts reveals the significance of model selection and feature engineering for improved predictive accuracy. The comparison between logistic regression and linear regression reveals that classification models are better suited for handling categorical HR data in the analysis of employee turnover. Likewise, our research employs logistic regression, alongside advanced machine learning techniques (SVM, Random Forest, XGBoost, and Deep Learning) to forecast employee turnover and retention patterns. Through the use of predictive analytics, HR professionals can improve decision making, workforce planning and talent retention strategies in line with the in creasing popularity of digital HR and data-based approaches [7]. The use of machine learning in HR analytics is not limited to employee retention and workforce planning to areas such as detecting fraud in job postings. Previous research has identified the rising problem of

fake job ads, stressing the importance of developing tools that would help identify scams aimed at job seekers. Studies recommend machine learning classification techniques such as Decision Trees, Sup port Vector Machines (SVM), Na¨ ıve Bayes, Random Forest, and Multilayer Perceptron for distinguishing between the real and the fake job posts. These methodologies are in sync with our research in which we have endeavoured to predict employee attrition using logistic regression, SVM, random forest and deep learning models. While recruitment fraud focuses on classification accuracy, our case applies the same machine learning principles to HR analytics and employee turnover and workforce trends. The use of machine learning for both fraud detection and the decision making process in HR shows the potential of AI based solutions in the context of workforce management and risk prevention in HR practices [8]. This paper highlights recent improvements in natural language processing (NLP) for the recruitment analysis of resume feature extraction. Studies suggest model fusion and Na¨ ıve Bayes-based classification to improve person-job matching. Employee attrition is a critical phenomenon that affects the performance and profitability of organizations. Our research continues from these principles to apply machine learning models to predict employee turnover, to improve the effectiveness of workforce planning and retention strategies [9]. Using machine learning for automated resume screening enhances the process of candidate selection. Studies on KNN, Weighted KNN and SVM KNN for classification, found that Weighted KNN had the highest accuracy (74research is also based on applying similar ML techniques to HR analytics, specifically in the area of employee attrition prediction to help improve workforce management [10]. The conventional methods of sales forecasting models have been improved to consider non-linearity and data inconsistencies. Research includes curve

regression, time series decomposition, and neural networks integration to enhance the prediction accuracy. Our study also employs similar methods in its analysis, including the application of ARIMA, SARIMA, XGBoost, and Prophet for the forecasting of employee at trition, which helps in improving the effectiveness of HR decision making [11]. Sales forecasting has become increas ingly dependent on machine learning models to deal with large scale data and market variability. Researchers compare traditional regression techniques with boosting algorithms and conclude that the intelligent data mining methods outperform in improving the accuracy of the prediction. Our research is in line with these advancements, and we have developed ARIMA, SARIMA, XGBoost and Prophet to improve sales trend forecasting, for data-driven decision making in business planning [12]. Sales forecasting models are used a lot by advanced modern businesses for demand prediction, inventory management and production planning. Research is compared to traditional models (ARIMA, SARIMA) with modern deep learning models (LSTM, Prophet) and SARIMA is claimed to be good at controlling seasonal trends. Our study is based on these findings that we have made, we have used ARIMA, SARIMA, XGBoost and Prophet to improve the accuracy of sales forecasting and for providing data driven decision support in business strategy [13]. Sales forecasting is a key component of business profitability but its accuracy is pre stroved by dynamic factors and customer sentiment. Research reveals limitations of integer based customer ratings that lead to biased product evaluations. Studies suggest sentiment analysis (VADER) to improve customer ratings and forecasting models (ARIMA, SARIMA, LSTM). Our study continues from these findings, using ARIMA, SARIMA, XGBoost and Prophet for sales forecasting to improve business strategy through data driven decision making [14].

Sales forecasting for new product adoption is often modeled using the Bass model however its effectiveness is limited by the absence of historical data. Studies explore analogy-based forecasting methods which use past sales data of similar products to make inferences about future trends. But the research shows that such methods are prone to high absolute percentage errors and that using multiple analogies is better than using a single reference product. Our study is based on these ideas, and it uses ARIMA, SARIMA, XGBoost and Prophet to enhance the accuracy of sales forecasting and offers data-driven insights for business decisions [15]. Promotional sales forecasting is very important in retail for strategy and supply chain optimization. Research is introduced to use interpretable machine learning methods, such as automated weighted K-Nearest Neighbors (KNN) to increase sales prediction precision. Studies show that distance calculations using feature selection based similarity detection outperforms traditional regression trees. Our study builds on these findings by combining ARIMA, SARIMA, XGBoost and Prophet to improve sales trend forecasting and provide data driven insights for inventory and promotional planning [16]. Sales forecasting along with the pricing strategy optimization is very important in the context of inventory management and revenue maximization. Research includes ARIMA models for time series forecasting and linear pro gramming for price determination and the sales forecasting is improved. Therefore, the model can be effectively used in real life decision making processes by the management of the organization. The findings of the study also offer possi bilities for further investigation of the integration of different optimization models with more intricate forecasting methods to fine tune the pricing and ordering decisions for greater customer satisfaction and operational performance in the case of retail distribution centers. The

model can also be applied to other contexts with similar ordering activities and goals. The contribution of this study is the integration of optimization models with forecasting techniques to solve complex problems in logistics and management and show the applicability of such approaches in real life settings [17]. In this paper, an advanced Deep Reinforcement Learning technique known as Deep De terministic Policy Gradient (DDPG) is applied for optimal bidding strategies in renewable energy markets. The study highlights the use of historical price data in decision making to minimize costs and enhance project cash flow. Although there are limitations of small dataset and volatile market prices, DPG is demonstrated to be effective in modelling complex environmental dynamics and optimizing bidding strategies for control structures. Some challenges include ensuring robust ness in different market scenarios. The approach holds promise for cost efficient project execution and decision making in a changing renewable energy market context [18]. Forecasting techniques are useful in sales prediction, banking, healthcare and stock market analysis. The effectiveness of time series models such as ARIMA, Logistic Exponential Models and Facebook Prophet in enhancing the accuracy of the prediction has been established by studies. It has been found that FB Prophet is better than the conventional models because it is capable of handling seasonality and missing data. Our study is a continuation of the above works where we apply ARIMA, SARIMA, XGBoost and FB Prophet for sales forecasting to proper business trend prediction and decision making [19]. As data grows exponentially, businesses are using machine learning algorithms for sales predictions and decisions. XG Boost is found as a powerful model in the literature and with the help of feature engineering boosts the forecasting accuracy. Research proves that XGBoost performs better than classical models, e.g., Logistic Regression and

Ridge Re gression, having lower RMSSE values. These findings are further supported by our research where we develop ARIMA, SARIMA, XGBoost and Prophet to improve sales forecasting, more accurate business trends and data insights [20]. Emerging Local energy markets (LEMs) are getting more focus because of increasing consumer's energy preference variation. The literature also identifies the challenge of heterogeneous energy valuation and recommends innovative market mechanisms. Studies suggest auction and merit order based approaches to deal with these complexities, while incorporating consumer preference models such as Borda count voting mechanism. Our research is informed by these insights and we use ARIMA, SARIMA, XGBoost and Prophet to improve sales forecasting, for more accurate demand forecasting and data driven deci sion making in a dynamic market context [21]. As markets for medium and long-term electricity have been expanding, the issue of market power concentration has become more apparent. Research points to the need of quantitative analysis and key evaluation indicators to ensure market stability and fair competition. Some studies suggest combination weighting methods for assessing market power in electricity retail sectors and therefore enhanced regulatory oversight. Our research is a continuation of the aforementioned methodologies which include ARIMA, SARIMA, XGBoost, and Prophet for sales forecasting, which enable better demand prediction and data driven decision making in a competitive business environment [22]. Digital marketing is used by businesses to harness social media and email marketing to increase customer in teraction and purchase willingness. Research shows that the relationships with the customers are better when they are stronger and that is why social media and email marketing are becoming essential in the current business world. Stud ies apply SEM-PLS modelling to examine the relationships between

marketing activities and consumer responses, thus supporting their ability to enhance sales and brand equity. Our research goes further on these findings by applying machine learning models (ARIMA, SARIMA, XGBoost and Prophet) for sales forecasting, which can help businesses to engage in data-driven marketing decisions to enhance revenue [23]. As Turkey's energy market has been growing rapidly, the power sector has been liberalised in order to fulfil the rising investment demands. Research highlights the introduction of electricity market laws that unbundles the market, has bilateral contracts and balancing mechanisms for the market to function efficiently. Studies emphasize the cost reflective tariffs and regulatory frameworks to stabilize the market structure. Our research goes one step further on top of these insights by using ARIMA, SARIMA, XGBoost and Prophet for sales forecast ing, to improve the demand forecasting and strategic decision making in dynamic markets [24] . Integration of the cross border electricity market provides the opportunities of power trade balancing and market efficiency. The impact of market design differences such as energy only and capacity based markets on interconnector capacity allocation is highlighted in research. The studies indicate that the price variations in the balancing market affect trade incentives but are restricted by risks and small imbalance volumes. These findings are further built upon in our research where we use ARIMA, SARIMA, XGBoost and Prophet to forecast sales, optimize market demand prediction and make strategic business plan ning decisions [25]. Customer segmentation is important for targeted marketing, but effectively clustering consumers in accordance with actual Customer Requirement Data (CRD) is difficult. Research introduces K-Means clustering based on Rubin Index (RI) to increase the accuracy of segmentation by optimizing the performance of the clusters. The effectiveness of this approach is validated

using IBM, Telco and Cell2Cell datasets where RI-K-Means is found to outperform traditional methods such as Multi Layer Perceptron with SMOTE by achieving 85.22% accuracy. These insights are then further built upon in our study where we employ K-Means, GMM, DBSCAN and ensemble clustering for customer segmentation, and optimize marketing strategies and business decision making [26]. As a result of the development of e-commerce, more and more companies apply machine learning clustering algorithms for customer demographic segmentation. The study includes K-Means, GMM, and BIRCH to cluster customers according to their gender, age, annual income, and spending score; the performance is evaluated using the Davies-Bouldin Score. The result of the study shows that K-Means is the best for structured data while other algorithms work well in some situations. Our study builds on these insights by applying K-Means, GMM, DBSCAN and ensemble clustering to improve customer segmentation, marketing strategies and targeted engagement [27]. Customer segmentation is widely used for differential management with the purpose of matching the offered services to the needs of different customers with the goal of increasing profit. Research applies DBSCAN and K-Means clustering for customer segmentation in electricity markets, which allows power exchanges to classify trading behaviors and market stability. The results of the study show that the segmentation performed using the real operational data is more accurate and reliable. These insights are then extended in our study where we implement K-Means, GMM, DBSCAN and ensemble clustering to enhance customer segmentation, marketing strategies and business intelligence [28]. Market styles are patterns of stock trends that can assist the investor in the formation of the optimal price forecast for maximum profit. Research reveals problems with the traditional market style discrimination based on the

scale values instead of the actual feature contributions. Several factor models and Gini index based feature selection criteria are suggested to enhance the accuracy of market trend identification and stock price forecasting. Our study is a continuation of the above mentioned ideas and in this paper, we apply machine learning models (ARIMA, SARIMA, XGBoost and Prophet) for sales forecasting which can help in improving trend analysis and data driven decision making in the financial and business organizations [29].An analysis of the wholesale market uses machine learning models to forecast the demand for products and therefore to improve the management of inventories. Research notes that conventional ML models are inadequate, and thus ensemble techniques are employed to enhance the forecasting precision. Some studies suggest SMOreg and Kstar algorithms, which improve SVM-based training called SMOreg for better demand forecasting. Our study follows these results by applying ARIMA, SARIMA, XGBoost, and Prophet models to forecast sales patterns to enable wholesale dealers to make prudent production and inventory decisions based on data analysis [30].Public Relations (PR) plays a very important role in communication and is often times criticized for being associated with hype. Research shows that using PR techniques improves interaction with people and understanding of messages which makes it useful for technical communicators. The literature review reveals that employing PR strategies increases communication effectiveness. Our study contributes to extending these insights by applying sentiment analysis and machine learning models (SVM, Naive Bayes, and Logistic Regression) to examine the public perception and brand communication, and recommend optimal PR practices for business effectiveness [31]. Interactions with business and media are an important factor in public relations and crisis management. Research highlights the following as key

strategic frameworks to improve corporate communication and reputation management: Attitude, Information dissemination, Media collaboration, and Staff management (AIMS). Studies stress the necessity of positive PR approaches to prevent conflicts and influence the perception of the public. Our study goes one step further to these insights, using NLP-based sentiment analysis and machine learning models (SVM, Naive Bayes and Logistic Regression) to analyze public sentiment and improve the PR strategies for effective crisis management [32]. Russia: the reorganization of the higher education system is aimed at the improvement of the educational quality, digital revolution and the integration with industry. The role of PR and advertising studies in aligning curricula with the industry demand: research. Reviews show that monitoring industry trends is a good way of ensuring that graduates of PR programs are skilled enough in the area. Our study contributes to the above by applying sentiment analysis and machine learning models (SVM, Naive Bayes, Logistic Regression) on the analysis of public perception and optimization of PR strategies in the higher education and corporate communication [33]. Distance learning is a novel approach to teaching that has received a significant amount of attention especially in light of health concerns. The literature also examines the impact of new media technology on learning outcomes of students, contrasted with conventional learning. Some research focuses on graduate satisfaction in Public Relations and Advertising programs, evaluating the efficacy of remote education. Our study contributes to these insights by using NLP-based sentiment analysis and machine learning models (SVM, Naive Bayes and Logistic Regression) to assess public perception and PR strategies in education and corporate communication [34].AI is changing the practice of public relations by doing menial work and letting professionals work on the strategic aspects of the

field. I say this because tools driven by artificial intelligence are available to help with content generation, target market analysis, and prediction of trends in the media which will help to increase the efficiency of communication. The literature reviewed in this paper shows that large language models are useful in the PR workflows to develop specific messages and improve the engagement tactics. Our study adds value to these findings by applying NLP-based sentiment analysis and machine learning models (SVM, Naive Bayes, and Logistic Regression) to understand the public perception and help in the PR decision-making process with the help of AI based insights [35]. Text classification is a challenge of achieving high accuracy with efficient computing. Many mod els fail in handling the high dimensional data. The reviewed literature reveals that neural networks, Word2Vec, and Cosine Similarity are effective methods of reducing the dimension of data without compromising on the classification accuracy. The studies show that there is a great reduction ratio of 36% 45% in all the datasets with no reduction in efficiency, which makes the text analysis more efficient. The purpose of this study is to continue from these findings and to use NLP-based sentiment analysis and machine learning models (SVM, Na¨ ıve Bayes, and Logistic Regression) for improving public relations practices through the use of AI-based text classification and sentiment analysis [36]. News that is controversial draws a lot of the public's attention, but understanding the opposite point of view is not easy. Research: The Disputant Relation-Based Method as introduced to classify news issues by determining key participants or actors in the debate (disputants) to simplify the process. Studies used unsupervised SVM classification to segment news while the most recent work has investigated both supervised and unsupervised SVM approaches for better accuracy. Our study is based on these findings and for news classification

and public perception analysis in the field of public relations we use NLP-based sentiment analysis and machine learning models (SVM, Naive Bayes and Logistic Regression) to improve the process [37].This paper focuses on sentiment analysis in natural language processing (NLP), which deals with the problem of identifying emotions in text using both lexicon-based and machine learning approaches. A comparison with VADER (a rule-based lexicon method) is made with Logistic Regression, and it is found that Logistic Regression has a higher accuracy of classifying sentiment in airline-related tweets at 79%. Studies show that machine learning models are better at capturing complex sentiment structures. Our study is a continuation of these insights; we employ public relation strategies optimization with the help of NLP-based sentiment analysis and machine learning models (SVM, Naive Bayes, and Logistic Regression) to examine the public perception [38].AI is changing the way we use strategic communication and Public Relations (PR) by improving cam paign management and crisis response. Research shows that AI improves message personalization and media monitoring positively. Studies also reveal that crisis management plays the role of a mediator that increases the effect of the use of AI on the PR. Our study contributes to the above by applying sentiment analysis using NLP and machine learning models (SVM, Naive Bayes, and Logistic Regression) to analyze public perception and recommend PR strategies based on AI-generated insights [39].It is essential that technologists, scientists, and industry leaders always get clear messaging from public relations (PR) programs. Research shows that poorly executed PR strategies can cause organizations to fail to articulate their vision, fail to position themselves as thought leaders, and fail to influence the public discourse. Studies highlight the function of PR in narrative shaping and creating informed debate. Our study, which is based on senti-

ment analysis of unstructured text using natural language processing (NLP) and machine learning models (SVM, Naive Bayes, and Logistic Regression) to analyze the public perception, and to recommend improved PR strategies for higher audience interaction is built on these insights.

Chapter 4

DESIGN AND METHODOLOGY

This study examines how machine learning and deep learning approaches might improve strategic decision-making in critical fields like human resources (HR), sales, marketing, and public relations (PR). The datasets used in this study were obtained from Kaggle to assure diversity, reliability, and relevance for each business function. An organized analytical framework, depicted in Figure 1, directs this study's approach to integrating several business operations. Organizations can use data-driven insights across several domains to systematically review and refine their strategy, resulting in increased efficiency and effectiveness. The interrelated nature of various business operations ensures that insights from one area help to improve another, which enables ongoing optimization.

4.1 BUSINESS OPERATIONS FRAMEWORK

HR analytics is essential for analyzing staff patterns, which impact overall productivity and operational stability. We began by conducting Exploratory Data Analysis (EDA) to identify patterns in employee performance, retention, and overall work force efficiency. To prepare the data for machine learning models, categorical variables were transformed using feature engineering techniques. Different models, including Random Forest, XGBoost, Support Vector Machines, K-Nearest Neigh bors, Logistic Regression, and LightGBM, were trained and evaluated. Model perfor-

mance was evaluated using confusion matrices, accuracy, precision, recall, and F1-scores to deter mine the most suitable Model. HR analytics data enables businesses to enhance their hiring strategies, boost employee engagement, and improve workforce planning. Understanding trends in personnel attrition and performance is crucial for operational stability, cost control, and long-term corporate success. Sales analytics is crucial for a company's success as it has a direct impact on financial stability and expansion. So, to understand the factors that influence sales we have first conducted EDA to analyze sales data to assess seasonal trends, consumer demand, and external economic issues. The Augmented Dickey-Fuller test is used to assess stationarity and determine the requirement for differentiating approaches. ARIMA, SARIMA, XGBoost, and Prophet are time series forecasting models used to predict sales trends and demand fluctuations. Forecasting models are evaluated using perfor mance metrics, including MAPE, MSE, and R^2. In addition, model comparisons are performed using test-train accuracy to choose the best-performing model for robust forecasting. Sales insights are directly relevant to marketing analytics, which analyzes consumer behavior to optimize engagement efforts. To improve feature engineering, datasets undergo prepro cessing using techniques such as Principal Component Analysis, Autoencoders, and Uniform Manifold Approximation and Projection (UMAP). Customer segmentation is then performed based on purchasing patterns, utilizing clustering methods like K-Means, Gaussian Mixture Model, DBSCAN, and Louvain. The quality of segmentation is evaluated through various metrics, including the Silhouette Score, Variance Ratio Cri terion (VRC), Dip Tests, and Davies-Bouldin Index (DB Index). These insights help businesses create targeted market ing strategies, such as tailored promotions, strategic pricing adjustments,

and improved customer engagement. PR analytics, which is important for maintaining brand reputation, employs natural language processing (NLP) to assess brand perception and sentiment. To standardize cus tomer evaluations, the dataset is preprocessed using stopword removal, tokenization, stemming/lemmatization, and TF-IDF vectorization. To analyze public impression, sentiment clas sification algorithms identify reviews as positive or negative, using both machine learning and deep learning methodologies. Performance indicators, including precision, recall, F1-score, confusion matrices, and ROC-AUC scores, are used to assess model efficacy. Sentiment insights enable firms to better al ter their messaging strategy, improve customer service, and manage brand reputation. This study takes a comprehensive approach to business optimization by merging information from HR, Sales, Market ing, and Public Relations. Workforce improvement and sales planning drive marketing methods, whereas public relations analytics ensure brand consistency and consumer satisfaction. This integrated architecture enables businesses to boost effi ciency, raise profits, and make data-driven strategic decisions.

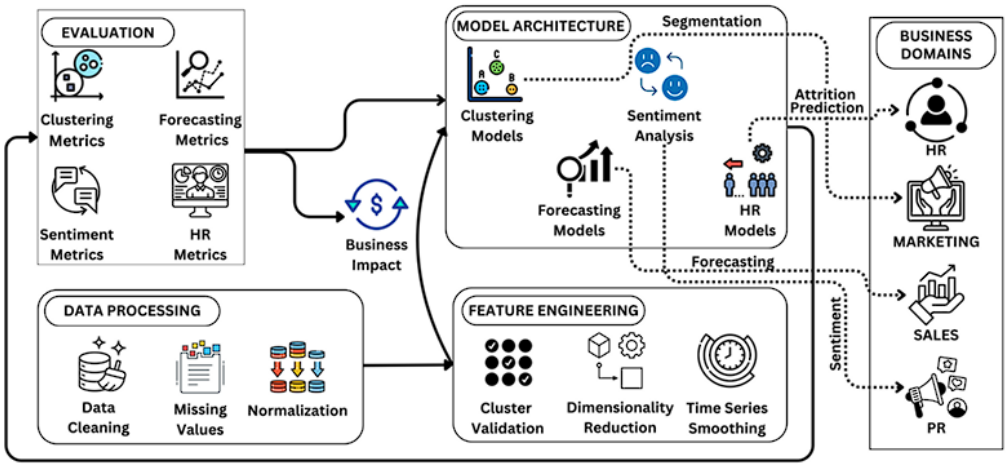


Figure 4.1: Workflow of the Proposed Methodology

4.2 DATA COLLECTION AND DATASET OVERVIEW

This study is based on four datasets that reflect diverse operational areas: human resources, sales, marketing, and public relations. Each dataset has structured attributes relevant to its domain, allowing for complete business analysis and predictive modeling.

1. Human Resources Dataset: The HR dataset obtained from Kaggle [40] consists of 1,470 employee records with 35 features, including numerical and categorical variables. Key features include age, attrition, business travel, daily rate, job role, job satisfaction, performance rating, work-life balance, and years at the company. The major purpose of analyzing this dataset is to gather knowledge about employee attrition, job satisfaction, and workforce retention.

2. Sales Dataset: The Sales dataset, also collected through Kaggle [41], which has 6,435 weekly sales observations with variables such as Store, Date, Weekly Sales, Holiday Flag, Temperature, Fuel Price, Consumer Price Index (CPI), and Unemployment. This dataset is used in time series forecasting to estimate sales patterns and evaluate the impact of economic factors on business performance.

3. Marketing Dataset: The Marketing dataset, also collected from Kaggle [42], contains 8,950 client records with 18 attributes, which include both numerical and category information. Key features are Customer ID, Balance, Purchases, Credit Limit, Payments, and Tenure. The major goal of this dataset is to better marketing and financial strategies by segmenting customers and assessing credit risks.

4. Public Relations (PR) Dataset: The PR dataset sourced from Kaggle [43] includes 20,000 customer reviews with three attributes: label, title, and

review. This dataset is used for sentiment analysis, which evaluates brand impressions and public sentiment.

4.3 DATA PREPROCESSING AND FEATURE ENGINEERING

1. **Data Cleaning:** Handling missing values was an important step in guaranteeing data consistency. Missing categorical variables in the HR dataset were imputed using mode, and numerical values were filled with median imputation. The Sales dataset did not include any missing values. Mean imputation was used to address missing values in MINIMUM_PAYMENTS and CREDIT_LIMIT from the Marketing dataset. In the PR dataset, column name mistakes were fixed, and rows with missing values were eliminated. Redundant elements were removed to increase model efficiency.

2. **Data Transformation:** The HR dataset utilized label encoding for binary categorical variables and one-hot encoding for multi-class categorical variables. Min-max scaling was used for feature normalization. Weekly Sales data in the Sales dataset was aggregated by month to allow for trend analysis. The PR dataset was TF-IDF vectorized to provide textual data for NLP-based sentiment analysis.

3. **Outlier Detection and Removal:** Outlier detection approaches such as the Interquartile Range (IQR) method and Z-score analysis were used to find and exclude extreme data.To improve model reliability, outliers in salaries and customer purchases were thoroughly filtered.

4. **Feature Engineering:** New features have been created to improve prediction performance. In the HR dataset, a promotion rate ratio was included. The Sales dataset was augmented with time-based

features. The Marketing dataset included a Credit Utilization Ratio calculation. The PR dataset used advanced NLP techniques to boost sentiment classification accuracy. This comprehensive methodology offers an organized, data-driven approach to company optimization, with machine learning and deep learning techniques used to give actionable insights for decision-making and strategic planning.

4.4 VALIDATION TECHNIQUES

1. **ADF Test:** Confirming Data Suitability for Forecasting: Stationarity tests were done to validate that appropriate trans formations have stabilized the dataset, ensuring that ARIMA and SARIMA models can produce reliable forecasts. ADF test produced the results which we can see below, ADF Test for Monthly Sales:

ADF Statistic:	-6.164473786749467
p-value:	$7.044293727364576 \times 10^{-8}$
Critical Values:	
1%	-3.661428725118324
5%	-2.960525341210433
10%	-2.6193188033298647

The ADF test shows that the monthly sales data is stationary, with an ADF statistic of-6.164 and a p-value of 7.04e-08, rejecting the null hypothesis of non-stationarity. Because the data's mean and variance are stable, no additional transfor mations are required, allowing forecasting models such as ARIMA, SARIMA, and Prophet to be applied directly. Figure 2 shows the patterns and variations in the time series data. The blue line helps to display the original dataset, which shows fluctuations across time, and the red line indicates the 6-month rolling

mean, which smoothes out short-term fluctuations and displays long-term patterns. The green line represents the 6-month rolling standard deviation, indicating variations in variability over time. A constant rolling mean and standard deviation reveal stationarity, which is required for accurate modeling and forecasting in time series analyses.

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum \delta_i \Delta Y_{t-i} + \epsilon_t$$

where:

- Y_t is the time series,
- α is the intercept term,
- β is the trend coefficient,
- γ is the coefficient of the lag term,
- δ_i are autoregressive terms, and
- ϵ_t is the error term.

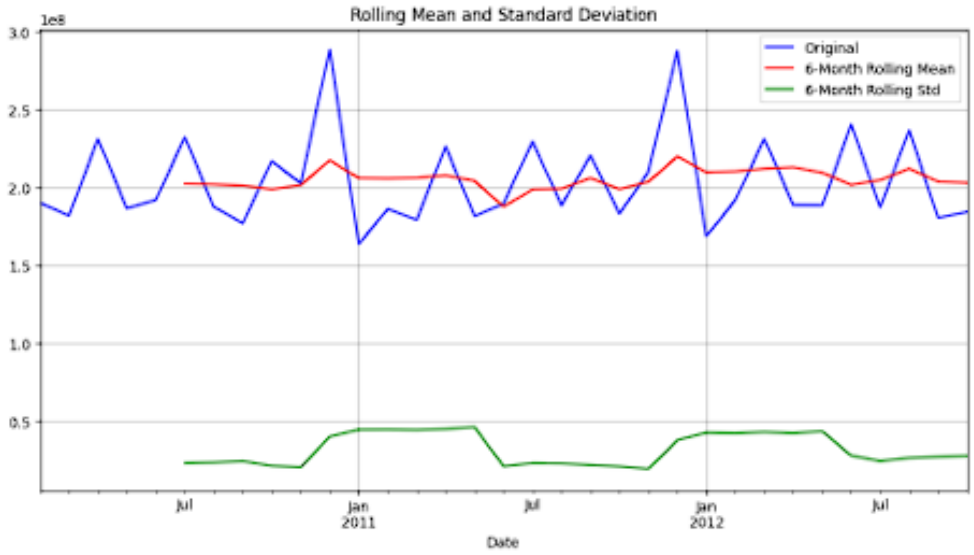


Figure 4.2: Rolling Mean and Standard Deviation

2. **Dip Test for Unimodality:** A statistical test determines if data are of unimodal (one mode) distribution or multimodal (several peaks, many clusters). Hartigan's Dip Test calculates the deviation from unimodal distribution and gives a p-value to set up significance. The project performed an unimodality test (run Kolmogorov–Smirnov and Anderson–Darling tests on feature distribution) to determine whether customer data contained more than one inherent group. A p-value less than 0.05 showed that data was not of one standard curve, so it was multimodal. This showed that the dataset had separate segments and not a single group, and thus clustering was appropriate.
   The **Dip Test statistic** is computed as:

$$D_n = \sup |F_n(X) - F(X)|$$

   where:

   - $F_n(X)$ is the empirical cumulative distribution function (ECDF).
   - $F(X)$ is the theoretical unimodal distribution.

   If the *p-value* $<$ **0.05**, it suggests that the dataset is **multimodal**, meaning that it contains **multiple clusters**.

3. **Silhouette Score:** The silhouette coefficient measures how well each data point lies within its cluster relative to others. It ranges from–1 to +1, with higher values meaning the point is much closer to points in its cluster than points in other clusters; this provides a gauge of cluster cohesion and separation from different clusters. An average silhouette near 1 indicates well-separated, compact clusters, whereas values near zero or negative indicate overlapping or ill-defined clusters. In

this project, silhouette scores were calculated for different clustering solutions to assess quality. For example, the average silhouette for K=2 clusters was slightly higher than for K=7 clusters; this implies that two broad clusters had marginally better separation than seven smaller clusters. However, the difference was minor, and both environments possessed medium silhouette values (much lower than the ideal of ¿0.5). So, silhouette analysis alone was not enough– although K=2 was better by a little, it was not significant enough to deem the clustering "good." We used this information in an effort to understand that other values were needed in order to choose the best value of K. The Silhouette Score is computed as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where:

- $a(i)$ is the **average intra-cluster distance**.
- $b(i)$ is the **average inter-cluster distance** to the closest cluster.

A silhouette score:

- **Close to 1** $\rightarrow$ Indicates **well-separated clusters**.
- **Near zero** $\rightarrow$ Suggests **overlapping clusters**.
- **Negative values** $\rightarrow$ Indicate **misclassification**.

4. **Davies–Bouldin Index:** The Davies–Bouldin Index (DBI) evaluates clustering by examining the average sim ilarity between each cluster and its most similar counter part. It computes, for each cluster, a ratio of the cluster's internal dispersion to the distance between that cluster and the nearest other cluster, then averages these ratios across all

clusters. A lower DBI indicates better cluster separation. Unlike silhouette (where higher is better), smaller values are superior for DBI. In the project, DBI was calculated using *davies_bouldin_score* for various cluster counts. The results were precise: K=7 clusters yielded a much lower DB index than K=2. The DBI for 7 clusters was significantly smaller, indicating more compact and far-apart clusters. This significant drop in DBI at K=7 exhibited a defined clustering pattern. Thus, the DBI metric proved that seven clusters pro vide a finer segmentation than two clusters, thereby definitively settling the value of K in the final model. It is calculated as:

$$DBI = \frac{1}{N} \sum \max \left(\frac{\sigma_i + \sigma_j}{d(i,j)} \right)$$

N represents the number of clusters, i and j are cluster indices, σ_i is the average intra-cluster distance of cluster i, σ_j is the average intra-cluster distance of cluster j, and $d(i,j)$ is the between-cluster distance. The lower DBI means the clusters are adequately separated and well-defined.

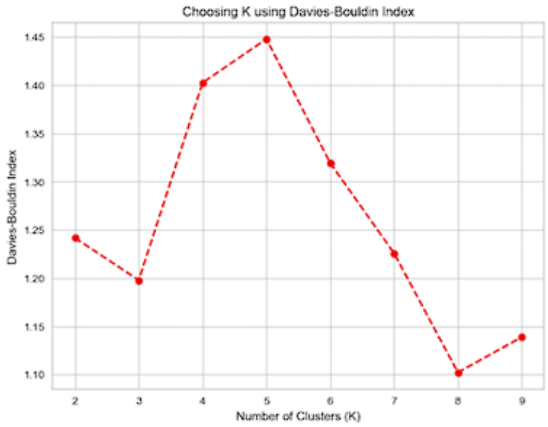


Figure 4.3: Choosing K using Davies-Bouldin Index

Chapter 5

IMPLEMENTATION

5.1 MACHINE LEARNING MODELS FOR EMPLOYEE ATTRITION PREDICTION:

5.1.1 Logistic Regression: Baseline Model for Interpretability:

We chose Logistic Regression (LR) as our baseline classifier because it is interpretable and necessary for HR settings where decision-making must be explainable. LR models the probability of attrition through a sigmoid function of a linear function of input features. The coefficients tell HR professionals how job satisfaction and work-life balance, among other factors, impact attrition. LR, however, does not support non-linear relationships. The probability of an employee leaving (attrition) is given by:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Where β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables, and e represents Euler's number.

5.1.2 Random Forest: An Ensemble Approach to Prevent Overfitting:

Random Forest (RF) was chosen to solve the issue of overfitting a single decision tree by training many trees on subsets of bootstrapped data, averaging their prediction, and voting for better generalization. RF revealed

that important variables pointing towards attrition were monthly compensa tion, total years employed, and years with existing managers. RF was more accurate than LR but much more computation ally demanding in terms of power, hence less efficient for real time computing. The final prediction is obtained as follows:

$$Prediction = \frac{1}{N}\sum_{i=1}^{N} T_i(X)$$

Where N represents the number of trees and $T_i(X)$ is the prediction from the i^{th} decision tree, the Gini impurity criterion used for tree splitting is given by:

$$Gini = 1 - \sum_{i=1}^{c} p_i^2$$

Where p_i denotes the probability of class i in a given node.

5.1.3 XGBoost:Gradient Boosting for Better Performance:

XGBoost improves predictability by gradient boosting, incrementally reducing classification error rates. It is based on the assumption that each tree contributes to learning from the previous trees.The support of L1 and L2 regularizations prevents overfitting,making it suitable for high-dimensional data. It provides a stronger recall and F1Score than other models and is especially ideal for HR analysis when false negatives are to be avoided.Nevertheless, its high complexity demands proper hyperpa-rameter tuning, i.e., learning rate, depth,and number of estimators.XGBoost minimizes the loss function through gradient boosting, expressed as:

$$L(\Theta) = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

$l(y_i, \hat{y}_i)$ represents the loss function, and $\Omega(f_k)$ is the regularization term.

5.1.4 LightGBM: For High-Big Data:

LightGBM, a gradient-boosting big data framework, compared employee attrition prediction.LightGBM has a tree grown leaf-wise that aims to boost training speed without precision loss. Its feature importance analysis established job involvement, relationship satisfaction,and working overtime as large-scale predictors for the risk of attrition.Being computationally light, it qualifies as an eligible candidate to use in the real-time application of HRdashboards.LightGBMutilizeshistogram-basedlearning to enhance efficiency. Its objective function is given by:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \lambda \sum_{j=1}^{J} \|w_j\|^2$$

Where λ is the regularization parameter and w_j represents leaf weights in decision trees.

5.1.5 Support Vector Machine: Maximizing Decision Boundaries:

Support Vector Machine (SVM) classified employees on optimized decision boundaries.SVM mapped input features in to higher-dimensional space us interradial basis function (RBF) kernel to attempt dichotomization between non attrition and attrition instances.Even though SVM produced similar accuracy, it took more computation and thus was ineffective with big data,particularly categorical variables,for one-hot encoding.SVM comes up with an optimal hyperplane that gives the maximum margin between two classes, that is, w·x+b=0.Wherew is a weight vector, x is the feature vector,

and b is the bias. The optimization function for classification is:

$$\min\left(\frac{1}{2}\|w\|^2\right) \quad \text{subject to} \quad y_i(w \cdot x_i + b) \geq 1$$

Where y_i denotes the class label (+1 or −1).

5.1.6 K-Nearest Neighbors:Instance-Based Classification:

K-Nearest Neighbors (KNN) classified employees based on the most prevalent class out of their closest neighbors.KNN performed well in the identification of localized patterns of turn over in particular departments or tenure groups.Euclidean distance measure and ideal value of K at 5 had to be selected for performance, though KNN did not perform on data with high dimensions.Thus, the use of feature selection techniques is suggested. The distance between two points is calculated using the Eucliden distance formula:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

where X and Y are data points, and n represents the number of features.

5.1.7 Naive Bayes:A Probabilistic Model for Speed and Simplicity:

Naive Bayes(NB) was also used in this study as a light weight categorical HR data classifier.Because it assumes feature independence,it computed posterior probabilities computationally efficiently in real-time prediction.Although notas robust as ensemble methods,its speed and tolerance to missing data made it a handy utility for exploratory analysis. For continuous features, Gaussian Naive Bayes assumes a normal distribution, formulated

as:

$$P(x_i \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{-(x_i - \mu_k)^2}{2\sigma_k^2}\right)$$

where x_i is the feature value, C_k represents the class, μ_k is the mean of the feature for class k, and σ_k^2 is the variance of the feature for class k.

5.2 SALES FORECASTING MODELS FOR BUSINESS INSIGHTS:

Before applying forecasting models,we conducted preliminary checks and preprocessing to ensure the data met model assumptions.We aggregated the sales data to a consistent temporal frequency and confirmed its stationary with an Augmented Dickey-Fuller test,which yielded a test statistic of The result is -6.164 and a p-value of 7.04×10^{-8}. allowing us to reject the null hypothesis of non-stationarity; this indicated a stable mean and variance, so no further differencing was needed. We then tried out four forecasting models– ARIMA, SARIMA, XGBoost, and Prophet– training each on historical data and testing their out-of-sample performance on a hold-out test set. The motivation, implementation, and performance trade-offs for each model are discussed in the following subsections.

5.2.1 ARIMA (Auto Regressive Integrated Moving Aver age):

ARIMA is a classic time series forecasting model that represents future values in terms of past values and forecast errors. We have selected ARIMA as a baseline model since it is good at short-term forecasting of stationary data. It is a basic but stable representation of the series and is very common in demand forecasting, such as inventory planning and large retailers' sales forecasts. Its appeal is that it can model auto correlations regardless of the given predictors, and therefore, it's a good place to begin generating sales

predictions.

- **Implementation Methodology:** Upon verification of stationarity using the ADF test, we established a differencing order d=0 for the ARIMA model. Following the Box-Jenkins methodology, we checked the ACF and Partial ACF plots to select the auto regressive order p and moving-average order q. Small orders (0–3) were our interest to prevent overfitting, focusing on the most parsimonious model that preserved the correlation structure, resulting in an ARIMA(p,0,q) model. Our findings indicated that a low-order order was adequate for the sales figures since higher-order orders had little effect on the AIC or validation fit.

- **Hyper parameters and Impact:** The most critical ARIMA hyper parameters are (p,d,q), where d=0 because the series is stationary. The choice of p and q is essential: too low can result in underfitting, and too high can result in overfitting. We optimized these parameters on the training set using AIC and cross-validation to identify weekly sales patterns with out unnecessary complexity. The ARIMA model was around 89.78% accurate on the test set and 91.3% when trained. It picked up the underlying trends and short-range dependencies well with good generalization without overfitting too much.

- **Efficiency and Interpret ability:** ARIMA models are lightweight and quick to train, allowing efficient fitting to our aggregated sales series in seconds. Even with larger datasets, ARIMA's estimation method remains effective for moderate values of p and q, though more extensive data may require optimization. The model is distinguishable, and coefficients reflect prior sales impact on subsequent performance in AR and MAterms. Diagnostic plots affirmed model fit as revealing no

residual autocorrelation. Disadvantages of ARIMA include linearity assumptions and failing to fit seasonality or exter nal regressors apart from being added on (e.g., ARIMAX). ARIMA is susceptible to abrupt structural change as well. While restricted, ARIMA is a decent baseline for making sales predictions with good interpret ability and accuracy.

5.2.2 SARIMA (Seasonal ARIMA):

SARIMA (Seasonal ARIMA) is a variation of the ARIMA model with the addition of seasonal terms to deal with periodic patterns in data, such as annual cycles in retail store sales. With the expected yearly seasonality of Walmart sales (higher sales during holiday sea sons and midyear low points), using SARIMA makes it easier to capture such phenomena. It contains extra auto regressive and moving average terms at seasonal lags (12 months or 52 weeks) and a seasonal differencing term where necessary. It has effectively forecast seasonality in firms such as airlines and retail. SARIMA was thus selected to see whether adding seasonality would enhance our predictions.

- **Implementation Methodology:** In order to construct the SARIMA model, we had to specify the frequency of seasonality and orders (P,D,Q) and non-seasonal parameters (p, d,q). We assigned the seasonal frequency (s) 52 weeks as our data is weekly and is over a period of over one year. Due to high stationarity, normal differencing (d = 0) and seasonal differencing (D = 0) were unnecessary. We tried small values of P and Q using seasonal ACF/PACF plots with large spikes at lag 52, which indicated the application of seasonal AR(1) or MA(1) models. We tried various combinations such as SARIMA(p,0,q)×(P,0,Q)s=52 and checked them with AIC and validation performance without over-

fitting. We selected a simple seasonal model that balanced goodness-of-fit and simplicity, ensuring it captured repeating yearly effects well without overfitting to noise since seasonality.

- **Performance and Trade-offs:** The SARIMA model's test ing performance was similar to that of non-seasonal ARIMA on our data, achieving a test set accuracy of 89.07%, which fell short of ARIMA's 89.78%. Interestingly, SARIMA's train ing accuracy was lower at 84.4%, indicating that adding seasonality terms may have sacrificed training fit but im proved generalization. This suggests that seasonality was weak or redundant in the low-dimensional ARIMA model. The Holiday-influenced sales peaks could be non-occasion-based annual trends or partially driven by the Holiday Flag, which SARIMA could not effectively exploit. Overall, SARIMA did not contribute as much as ARIMA, supporting the fact that increased complexity pays off only when there is a clear seasonal signal in the data.

- **Computational Efficiency:** Adding seasonality terms that were burdensome and computer-intensive made SARIMA. However, we only trained SARIMA for seconds with the size of our dataset. Fitting grows linearly with observation counts. Hence, our 3-year-length weekly series was in ratio. It could give more extended series or multiple cycles to cause the rise in computation times and convergent problems, but it wasn't in our situation.

- **Interpretability:** SARIMA retains much of the inter pretability of ARIMA, where predictions can be explained by seasonality. For instance, it can state, "This week's sales will be like those from 52 weeks ago," using an AR(1) seasonal term. SARIMA's seasonality is not as

easily interpretable as Prophet's explicit terms, but it remains an open linear model. Our sample found a small annual effect—sufficient to justify adding the seasonal term but not strong enough to significantly impact forecasts for a non-seasonal model over time. SARIMA is effective for testing structured seasonality, performing well when seasonality is present but requiring diligent complexity control. If additional seasonal cycles were significant, we would expect SARIMA to outperform a non-seasonal ARIMA. However, in our case, the performance boost was negligible, reflecting the trade-off between increased explanatory power and higher parameter tuning complexity.

5.2.3 XGBoost Regression:

XGBoost is an ensemble tree model using gradient boosting for high-precision prediction on structured data. Relative to ARIMA/SARIMA, which only forecasts historical time series values, XGBoost learns many features and can model non-linear relationships. XGBoost lets us use other variables from the dataset, such as holiday dummies and economic measures, capturing interaction effects of the subtle sort, e.g., the effect of temperature on sales. This model's ability to process large datasets and provide high accuracy has made it trendy in companies such as Netflix, Uber, and Airbnb for forecasting. XGBoost will do much better if these additional regressors affect sales in ways time based models may not be able to.

- **Implementation Methodology:** We defined the prediction problem for XGBoost as a regression problem by constructing feature vectors for every time point with features such as Hol iday Flag, Temperature, Fuel Price, CPI, and Unemployment. Temporal features or lagged sales were also considered since XGBoost does not have in-built tem-

poral knowledge. We used only past data for every training sample to predict the sales for the subsequent period. The data were split chronologically, including the training set of the first period and the test set of the second period, to maintain causal structure. Given the data we had, we already had an XGBRegressor with the right hyperparameters chosen so that we wouldn't overfit. We began with 50 trees, a depth of 3, and a learn ing rate of around 0.15 for rapid fitting. We used L1 and L2 regularization as well as 80% subsampling per tree to help with generalization. Although cross-validation tuning of hyperparameters would be desirable, manual tuning by experimentation was sufficient for the best bias-variance trade off. Our XGBoost model was designed to detect non-linear effects but not overfit.

- **Performance and Results:** The XGBoost model was good on the training set with an accuracy of 90.35% but performed slightly less than 86.22% on the test, reflecting some over fitting; this was to be anticipated in light of the low number of observations and the complexity of the model. The rigid time series structure can do damage to poor lagged sales tree models. 86% test accuracy is tolerable, though. XGBoost made good use of exogenous features, capturing holiday weeks are related to higher sales and high gas prices hurt sales. Feature importance analysis should confirm that Holiday Flag is a good predictor and behaves as expected.

- **Computational Efficiency:** Speed and scalability are strong suits of XGBoost. Model training is milliseconds to several seconds, even on thousands of training instances, since it is C++-optimized and parallelizable. Even on millions of data instances, XGBoost handles big data efficiently by splitting the job across threads or clusters. Meanwhile,

models such as ARIMA/SARIMA or Prophet would be inappropriate for managing long series across multiple stores. Still, XGBoost can include "store" as an extra feature in a global model. The performance makes XGBoost better for prominent data usage than for one-series models.

- **Interpretability:** Ensemble techniques such as XGBoost lose some interpretability compared to models such as ARIMA or Prophet, which have interpretable parameters capturing the components of a time series. The XGBoost prediction is derived from numerous decision trees, so the explanations are not as straightforward. The feature importance score or SHAP values may reveal leading factors on sales, e.g., Hol iday Flag and CPI. However, we do not have an explicit equation for the prediction, so XGBoost is a "black box" com pared to ARIMA/SARIMA's transparency. Although XGBoost produced results by incorporating outside modules, it never dominated the base time-series models in our testing since it may have been overwhelmed by the brief but consistent character of the dataset and thus favored more straightforward methods.

5.2.4 Prophet (Facebook/Meta Prophet):

Prophet is an open-source time series forecasting library created by Face book's data science team with a focus on business time series usability. It is a model that includes a piecewise linear or lo gistic growth trend model augmented with seasonal cycles and holiday terms and applies Bayesian estimation for flexibility and uncertainty. We selected Prophet for its ability to handle seasonal patterns, holiday impacts, and irregularities in sales data. Its design minimizes tuning and adapts well to real-world retail

challenges, making it a reliable choice, as evidenced by its use at companies like Facebook, Google, and LinkedIn.

- **Implementation Methodology:** To use Prophet, preparing the data in the shape of a two-column data frame of ds (date) and y (value) is necessary. We had weekly sales aggregating data with dates being the time grid. Prophet has default yearly seasonality modeling, which fits our weekly data, and it also can model holiday effects by defining a list of essential holiday dates like Super Bowl weekend, Labor Day, Thanksgiving, and Christmas week. As with other models, this holiday inclusion enables Prophet to pick up its impact on sales without requiring manual feature engineering. The hyperparameters of Prophet's trend flexibility are gov erned by possible changepoints, a prior scale, and the Fourier order for seasonals. We used the default options: it detected the model's trend changepoints in the first 80% of the data, employed *"a standard"* Fourier series order (e.g., 10 for annual seasonality), and defaulted on holiday assumptions. Manual tuning was not required unless we noticed a lousy fit, e.g., raising the change point before a more flexible trend. While Prophet can oversmooth extreme spikes, this generally avoids overfitting. For example, consistent holiday season sales peaks permitted effective learning by Prophet. Upon fitting, we explored the decomposition of the forecast into trend and annual seasonality, which validated reasonable patterns.

- **Performance and Results:** Prophet worked best among the models we tested, with 91.58% accuracy on the test set, a hair above the rest in the high 80s to low 90s. Its accuracy on training was even higher at 97.7%, which indicates some overfitting to the training set; this did not

harm its test accuracy, though. Prophet also captured actual patterns in the time series, including holidays and seasonality, which enabled it to be generalized. For example, it accurately predicted the Thanksgiving holiday season shopping frenzy and the resulting post-holiday slowdown. It was better than models such as ARIMA and SARIMA, which had no direct holiday data. Prophet's success demonstrates the benefit of combining domain knowledge in predictions.

- **Computational Efficiency:** Training Prophet on our data was not too time-consuming. Stan, which it employs for optimization and sampling, is slower than ARIMA's analytical approach, but computation is only a matter of seconds for our data; this is because of the problem size and Prophet's default mode of computing a Maximum A Posteriori solution rather than large amounts of MCMC sampling. For long datasets, such as daily data for decades, we can observe longer runtimes or memory problems, but techniques such as changepoint reduction or parallel processing can prevent this. Overall, Prophet is scalable in single-series forecasting and provides automated modeling with development time saved, although not as quickly as XGBoost on large datasets.

- **Interpretability:** The Prophet model is very interpretable and thus accessible to non-statisticians. Once the model has been fit, trends, year-over-year seasonality, and holiday im pacts can be examined. Our trend component in our analysis revealed small year-over-year increases in sales between 2010 and 2012. Our seasonality by year revealed spikes in late November and December and a minor spike in early spring, which aligns with normal retailing behavior. The holiday com ponent revealed higher sales on holidays, which helps in in ven-

tory allocation decisions. In contrast to ARIMA/SARIMA, where there isn't a well-defined distinction of trend and sea sonality, and XGBoost, where additional analysis is necessary for effects, Prophet provides an easy decomposition and thus is a choice in business forecasting.

5.3 CLUSTERING TECHNIQUES FOR IMPROVED SEGMENTATION:

5.3.1 Gaussian Mixture Models (GMM) for Probabilistic Clustering:

Gaussian Mixture Models (GMM) provide a soft clustering response to K-Means as customers are distributed across multiple clusters with probabilities, not hard separa tions. Iteratively optimized assignment to clusters occurs in GMMusing the Expectation-Maximization algorithm by max imizing likelihood via updating covariances and means. This proves useful to incorporate overlapping customer movement and transition tendencies in marketing databases. Experimental outcomes indicate that GMM on raw data yielded a lower silhouette value (0.038) and Variance Ratio Criterion (543.10), while employing GMM on encoded data saw a considerable boost in these values (silhouette: 0.104, VRC: 1210.52), indicating that feature encoding increases the separation of clusters. GMM is still computationally demanding (6.62s for encoded data) relative to K-Means (2.23s) and thus better suited to applications that demand probabilistic partitioning over rigid categorical assignation.

$$p(x_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

Where $p(x_i)$ represents the probability of data point x_i belonging to a mixture of K Gaussian distributions, weighted by mixing coefficients π_k, with each Gaussian characterized by mean μ_k and covariance Σ_k.This probabilistic assignment allows for a more flexible and robust customer segmentation approach in marketing applications.

5.3.2 Deep Learning-Based Clustering: Autoencoders + K Means:

Autoencoders improve clustering by projecting high dimensional financial data into low-dimensional compressed latent space, improving cluster structure. A deep learning based approach allows K-Means to create denser, more well separated clusters, thereby improving segmentation accuracy. Variance Ratio Criterion (VRC) was enhanced to 1431.78, exhibiting significantly better cluster separateness compared to raw feature clustering. Additionally, the computation time of 0.4238s guarantees efficiency and the approach is ideally suited for large-scale customer segmentation where precision and scalability are paramount.

5.3.3 Non-Linear Feature Extraction: UMAP + K-Means:

UMAPwas applied to map high-dimensional finance informa tion into 2D space without sacrificing intricate relationships. Non-linear mapping improved cluster separation over the limitations of PCA. K-Means (K=7) was applied subsequently, benefiting from UMAP's ordered space for easier customer segments. Segmentation was further improved to a VRC Score of 29,066.21, a much higher number than PCA-based clustering. Computational efficiency remained adequate (1.12s running time), being effective on large marketing tasks. UMAP + K-Means produced crisper, more interpretable customer segments.

5.3.4 Density-Based Clustering: DBSCAN & HDBSCAN:

DBSCAN and HDBSCAN were used to pick out abnormal customer behavior and clusters that could not be linearly separated. DBSCAN doesn't require pre-definition of cluster amounts, building up clusters using density thresholds (ε, *min_samples*). HDBSCAN builds further on this with automatic optimization of the clusters and is, hence, more flexible. Experiments demonstrate that DBSCAN was able to spot the dense sub-populations but struggled with the high-dimensional space (Silhouette Score: -0.218, VRC: 187.35). DBSCAN processed more segmentation with poorer separation of clusters (Silhouette Score: -0.295, VRC: 101.98). Though they consume more computational time (DBSCAN: 1.15s, HDBSCAN: 2.24s), the models are extremely proficient at anomaly detection and hence become useful for selecting high-risk customers or fraud transactions. The **DBSCAN clustering mechanism** is governed by the concept of **density reachability**, where a point p is classified as a **core point** if it has at least *min_samples* points within an ε-radius neighborhood:

$$N_\varepsilon(p) = \{q \in D \mid d(p,q) \leq \varepsilon\}$$

Where $d(p,q)$ represents the Euclidean distance (or any other chosen metric) between points. For **HDBSCAN**, the **mutual reachability distance** extends this concept by modifying the core distance to dynamically adjust clusters:

$$d_{mrd}(a,b) = \max\{d(a,b), \text{core\_dist}(a), \text{core\_dist}(b)\}$$

5.3.5 Contrastive Learning + K-Means for Clustering:

To enhance customer segmentation, we employed contrastive learning by training a deep contrastive encoder that trans formed high-dimensional financial data into a low-dimensional latent space. The encoder is comprised of fully connected layers that reduce the feature dimensionality successively, capturing beneficial patterns in an encoded representation. The trained encoder projected the dataset into a low-dimensional space where clustering is more effective with better separation of customer behaviors. Following encoding, K-Means cluster ing was utilized on the obtained feature space. This resulted in significantly better compactness and readability of clusters, as indicated from the Variance Ratio Criterion (VRC) value of 8998.71. The model was efficient enough to classify cus tomers into discrete groups effectively, and thus it was worth considering for customized marketing strategies. Computation time took only 0.45s, and therefore it was realistic for actual world financial purposes. Contrastive learning ensures that similar data points are pulled together while dissimilar ones are pushed apart. The loss function used:

$$L = \sum_{i,j} \left(y_{ij} d_{ij}^2 + (1 - y_{ij}) \max(0, m - d_{ij})^2 \right)$$

Where:

- d_{ij} is the Euclidean distance between data points i and j,
- $y_{ij} = 1$ if i and j are similar, otherwise $y_{ij} = 0$,
- m is a margin that enforces separation between clusters.

5.4 PUBLIC RELATIONS: SENTIMENT TAGGING APPROACHES:

5.4.1 Logistic Regression:

Logistic Regression is a linear classifier model chosen because it is easy to interpret, simple, and effective with high-dimensional text data. Logistic Regression estimates the probability that a text document falls into a sentiment class (negative/positive) using a logistic function of a linear combination of features. Multinomial logistic regression was used for multi-class sentiment classification and handled multiple sentiment classes simultaneously. Prepro cessed features (TF-IDF vectors) are input into the classifier, which learns weights via gradient descent to minimize the log-likelihood of incorrect class labels. We utilized L2 regularization to prevent overfitting. Logistic Regression provides efficiency and transparency into feature importance, where we can determine terms that control sentiment prediction, which aids in explaining results to non-expert stakeholders.

- **Implementation and Performance:** After preprocessing, we applied a Logistic Regression model to the TF-IDF feature matrix. The model is fast-converging, thus computationally inexpensive. The model was precise and achieved around 83.8% in Amazon review sentiment classification, which was equal to one of the top-performing model (SVM). It accurately labeled most positive and negative instances, with accuracy and recall rates above 0.80 and a good F1 score. Public relations-wise, Logistic Regression provides accurate predic tions and easy-to-interpret outputs since its coefficients are keywords that affect sentiment. Its probabilistic outputs from its model also assist organizations in prioritizing dealing with negative sentiment. Its speed, ease, and accuracy balance make Logistic Regression apt for sentiment classifi-

cation.

5.4.2 Naive Bayes:

We used Naive Bayes because we felt it was fast, lightweight, and best suited for sentiment analysis from text on resource-poor machines. We utilized the Multinomial Naive Bayes variant because it is best designed to represent features of word frequencies. It follows from applying Bayes' theorem when features are assumed to be independent under the class label; this notwithstanding, it is a good performer with text and can deal with high dimensional ity nicely. We depicted the preprocessed text as TF-IDF vectors and applied Laplace smoothing to zero-probability situations. Training is efficient with rapid word counts and probability computation, and therefore, it is easy to update quickly on arriving data for real-time sentiment monitoring.

- **Implementation and Performance:** After training, the Naive Bayes classifier generates a probabilistic sentiment model, predicting sentiment by selecting the class with the highest posterior probability. Although it got approximately 82.48% in the sentiment analysis on Amazon review, it lagged behind more sophisticated models such as Logistic Regression and SVM. Naive Bayes performed well for some classes but showed balanced recall across sentiment classes, i.e., a conser vative strategy for capturing negative sentiment. It applies class prior and conditional independence and is bound to perform poorly at subtlety, i.e., sarcasm. Yet, its performance and fewer tuning requirements make it worth it, particularly for quick bulk social media data analysis. PR practitioners can use Naive Bayes as an early warning system for real-time sentiment tracking before using more so-

phisticated models. Naive Bayes sacrifices lower accuracy in exchange for breathtaking speed and gives a quality reference point in tracking sentiment.

5.4.3 Support Vector Machine (SVM):

We applied a linear kernel with a one-vs-rest strategy for multi-class sentiment **(positive/negative/neutral)** and trained the SVM with TF-IDF feature vectors. The SVM's regularization power enables it to generalize and not overfit. We optimized the regularization parameter C to balance maximizing margin and misclassifications. Platt scaling was used to measure confidence levels more interpretably for probability estimates of predictions.

- **Implementation and Performance:** Training the SVM on our data set produced one of the highest accuracies of all attempted models, competing with Logistic Regression, which averaged approximately 84.30% accuracy. The SVM demon strated intense precision and recall, especially for the dominant classes, and effectively managed the high-dimensional text features. While SVM and Logistic Regression showed similar performance due to both being linear models suitable for the sentiment data, SVM was notably more computationally intensive, requiring longer training times and significantly more memory as the dataset increased. After training, SVM is not unreasonable regarding prediction time and can be used for real-time inference. Its capability to make high precision sentiment predictions can be of great value to public relations, as it can rapidly detect negative sentiments and take action quickly on complaints or bad publicity.SVM is a black-box model and thus lacks the interpretability of logistic regression or decision trees. Hence, it can make it uncertain which variables contribute

to its predictions, which can be mitigated through feature weights analysis or interpretation tools. Despite these, the performance of SVM is priceless to sentiment analysis, primarily where precision is of great importance in PR strategy.

5.4.4 Random Forest:

To further diversify the model, we used the Random Forest with 100 decision trees, each trained on bootstrap training data samples and random feature subsets. Since it can also be used in multi-class classification, we used it in three-class sentiment classification from TF-IDF feature vectors of preprocessed text. In addition, we talked about feature importance scores to identify keywords significant in sentiment classification, providing interpretability by indicating positive or negative sentiment words.

- **Implementation and Performance:** Due to the numerous trees, the random forest model is computationally more expensive than single models like Logistic Regression, Naive Bayes, and SVM, with higher training time, though the process can be parallelized. In our experiments, Random Forest attained an accuracy of approximately 77.83%, which is lower than the best performance (84% for SVM/LR); this is consistent with previous work that showed that although Random Forests are very powerful, linear models are better with high-dimensional sparse text data. Most importantly, Random Forest showed weaker recall for the minority sentiment class (recall of 0.64), which is lower than Naive Bayes (0.82). It was minimally less precise for a few classes, showing overfitting. In general, it had a fantastic power-interpretability tradeoff with fewer outliers to contort it and solid collective conclusions. Its performance evaluation is of exceptional help

to public relations since it identifies words that generate sentiment, and PR professionals would be able to instantly view what's being thanked or criticized about their products. Not the most precise of performances, Random Forest's ensemble approach provides confidence in the forecasts' precision. Although computation ally costly, it can supplement linear models by identifying non-linear relationships and providing insightful results on the importance of features.

5.4.5 BiLSTM (Bidirectional Long Short-Term Memory) Model:

The BiLSTM model was applied for sentiment tag ging, considering past and future text dependencies. In contrast to regular LSTMs, BiLSTMs process sequences in both direc tions, enabling a richer contextual understanding of sentiment; this is beneficial in capturing subtle sentiment expression. The BiLSTM model comprises multiple layers, starting with an embedding layer of pre-trained GloVe word vectors. A spatial dropout layer with a dropout rate of 0.3 was used to manage overfitting. The bidirectional LSTM layer has 256 hidden units and a recurrent dropout of 0.2. Batch normalization is used to achieve stable training, after which GlobalMaxPooling1D is used to find the essential features. Two dense layers, 128 and 64 neurons with ReLU activations, and the last sigmoid activation layer are used for binary senti ment classification. It was optimized with an Adam optimizer, and the learning rate was 0.0003. The loss function employed was binary cross-entropy. Preprocessing included tokenization, removal of stopwords, and padding the sequence to 50 words. The training was done for 10 epochs, and a batch size 128 was employed. The model employed early stopping to avoid overfitting.

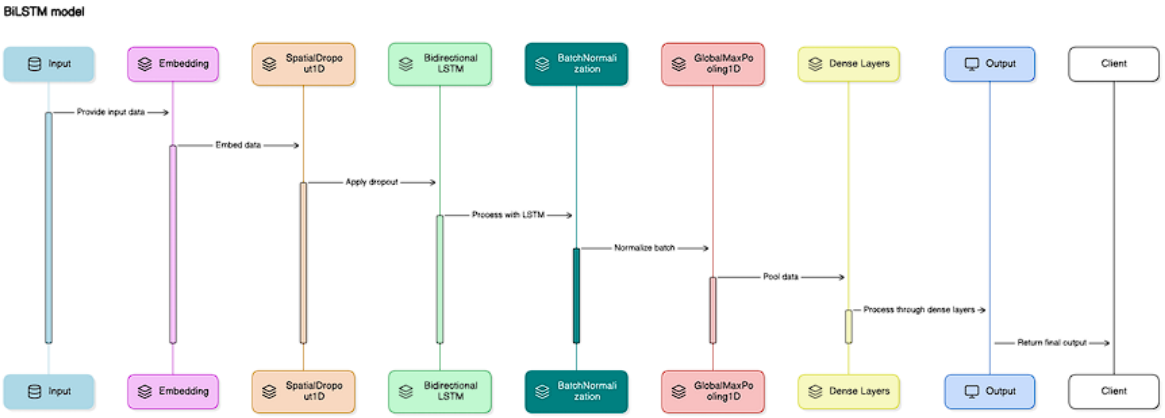


Figure 5.1: BiLSTM model performance over epochs

Performance measurement employed accuracy, precision, recall, and F1-score metrics. The model resulted in training, validation, and test accuracy of 83.0%, 83.0%, and 85.0%, respectively. The macro F1-score for the BiLSTM model was 0.85, with a strong balanced performance among sentiment classes. The confusion matrix had a high recall of positive sentiment at 0.88 and misclassifications of a higher percentage of negative sentiment. The accuracy-over-epochs plot showed consistent learning without overfitting, and the loss-over epochs curve showed good convergence. Overall, precision and recall values were more significant than 0.80, which ensured robust classification according to sentiment type. BiLSTM excels over Naive Bayes and Logistic Regression models based on traditional learning in sentiment analysis through its capacity to identify patterns dependent on the context. Being computationally expensive, it makes the clas sification accurate and more resilient, making it appropriate for sentiment tracking in real-time. Its capability to discover long-term dependency is helpful for public relations, brand sentiment tracking, and the detection of crises. BiLSTM's bidirectionality offers equitable sentiment classification rich with information about the consumers' sentiments.

Chapter 6

HARDWARE/ SOFTWARE TOOLS USED

Category	Tool/Specification	Description
Hardware	Processor: Intel Core i7/i9, AMD Ryzen 7/9	High-performance CPU for faster computations and model training.
	GPU: NVIDIA RTX 3080/3090, Tesla A100	Used for deep learning and AI model training acceleration.
	RAM: 16GB – 64GB DDR4/DDR5	Required for handling large datasets efficiently.
	Storage: SSD (512GB – 2TB)	Fast storage for quick data access and processing.
	Cloud Computing: AWS, Google Cloud, Azure	Used for scalable computing and AI model deployment.
Software	Programming Language: Python	Core language used for data science, machine learning, and NLP.
	Libraries: NumPy, Pandas, Matplotlib, Seaborn	Data handling, analysis, and visualization libraries.
	ML Frameworks: Scikit-learn, TensorFlow, PyTorch	Used for training classification and deep learning models.
	NLP Libraries: NLTK, SpaCy, RASA, Transformers (Hugging Face)	Used for text preprocessing, chatbot development, and NLP tasks.
	Visualization Tools: Tableau, Power BI, Plotly	Used for creating interactive dashboards and reports.
	Cloud Services: Google Colab, AWS S3, Azure ML	Used for cloud-based AI training and model deployment.

Table 6.1: Hardware and Software Tools Used

Chapter 7

RESULTS & DISCUSSION

7.1 HR DOMAIN

The performance summary table gives a comparison of seven different machine learning models in terms of four essential metrics: accuracy, precision, recall, and F1-score. The following is a more detailed description:

- Accuracy indicates the level of overall correctness dis played by each model. Models like Logistic Regression, SVM, and LightGBM achieved an accuracy of 90i.e., they correctly classified 90accuracy indicates good overall performance.

- Precision: Precision is the number of positive predictions that turned out to be true. In our table, SVM performs the best with a precision of 0.87, which means if it predicts an employee to attrite, then it is highly likely to be true. This is critical in HR applications where false positives will lead to unnecessary interventions.

- Recall: Recall is an estimate of how well one can perform to identify all the true cases of attrition. Naive Bayes is best in identifying most of the employees who are actually at risk of attrition with the highest recall of 0.60. But at the cost of lowering precision because high recall may lead to more false positives.

- F1 Score: F1-score is the harmonic mean of precision and recall, hence finding a balance between the two. Both Logistic Regression and

LightGBM have comparatively higher F1-scores, which indicates that they achieve a good balance in correctly classifying attrition cases while minimizing false positives.

Briefly, SVM exhibits extremely high accuracy performance, whereas Naive Bayes is better in recall. LightGBM and Logistic Regression exhibit relatively balanced performance across all the metrics of evaluation. These differences render the choice of model depending on the emphasis on reducing false positives, where the choice of SVM would be best, or on maximizing the identification of true attrition cases, where the choice of Naive Bayes would be best. This close examination assists HR to pick the best model for effectively enhancing employee retention.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	**0.90**	0.68	0.40	0.50
Random Forest	0.87	0.50	0.06	0.11
SVM (Support Vector Machine)	**0.90**	**0.87**	0.27	0.41
KNN (K-Nearest Neighbors)	0.89	0.77	0.21	0.33
XGBoost	0.89	0.72	0.27	0.39
Naive Bayes	0.67	0.22	**0.60**	0.32
LightGBM	**0.90**	0.73	0.33	0.46

Table 7.1: EMPLOYEE ATTRITION MODEL PERFORMANCE COMPARISON

7.1.1 Accuracy and Precision Analysis:

Accuracy is the overall accuracy of the prediction while precision is the model's ability to accurately identify actual attrites without identifying non attrites as attrites. Based on the evaluation metric, Logistic Regression, Support Vector Machine (SVM), and LightGBM had the best accuracy of 90%

which shows that they are good at making predictions. At the same time, Na¨ıve Bayes had the worst accuracy of 67% which shows that it is not well positioned to handle complex features. SVM had the best precision of 87%, LightGBM had 73%, and KNN had 77%, which means that they can decrease the number of false positives. On the other hand, Naive Bayes precision is 22% which is a clear indication of the model's inability to distinguish between attrition and non attrition cases.

7.1.2 Recall and F1 Score Analysis:

Recall is the accuracy of the model in identifying the true cases of attrition and F1-score combines precision and recall into a single value. The model with the highest measure of recall was the Naive Bayes model at 60% although it had a high measure of false positives. For instance, Random Forest had a very low measure of recall at 6% which means that it is not efficient in identifying the actual cases of attrition. The F1-score that aims to balance between precision and recall was achieved by Logistic Regression at 50%, LightGBM at 46% and SVM at 41%, which shows that these models are accurate in their classification. Random Forest had the worst F1-score of 11% which shows that it has low precision and recall.

7.1.3 Confusion Matrices for Employee Attrition Prediction Models:

The confusion matrices can help evaluate the performance of the employee attrition prediction models in detail. Logistic Regression, SVM and LightGBM were good; they correctly classified most of the samples as leaving the organization and had a good recall rate for the samples that left the organization.

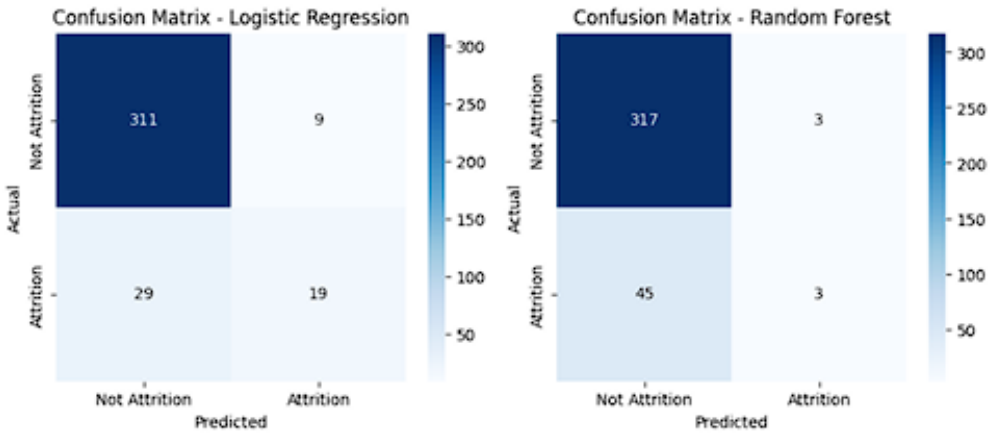


Figure 7.1: Confusion Matrices for Logistic Regression and Random Forest

SVM, however, performed very well in classification with low false positive rate and was able to identify the attrition cases well. Random Forest and KNN had a poor recall; they often mislabeled leaving employees as non-attributing employees.

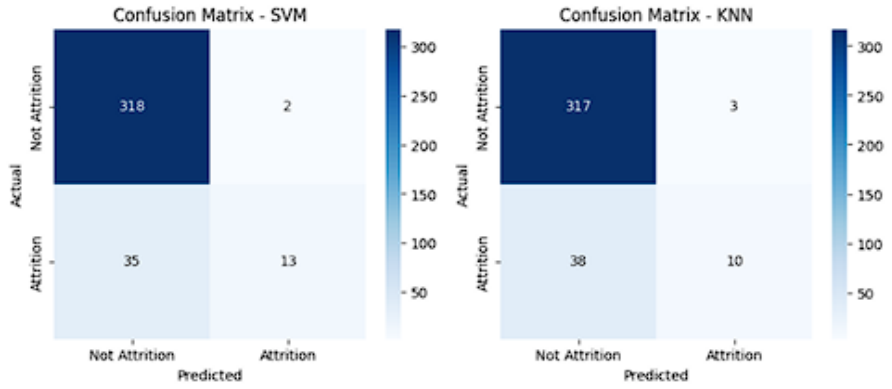


Figure 7.2: Confusion Matrices for SVM and KNN

This leads to a model that is quite protective in its predictions of non-attrition and does not alert the user to possible employee turnover. XGBoost had a moderate performance; it identified the attrition cases correctly and

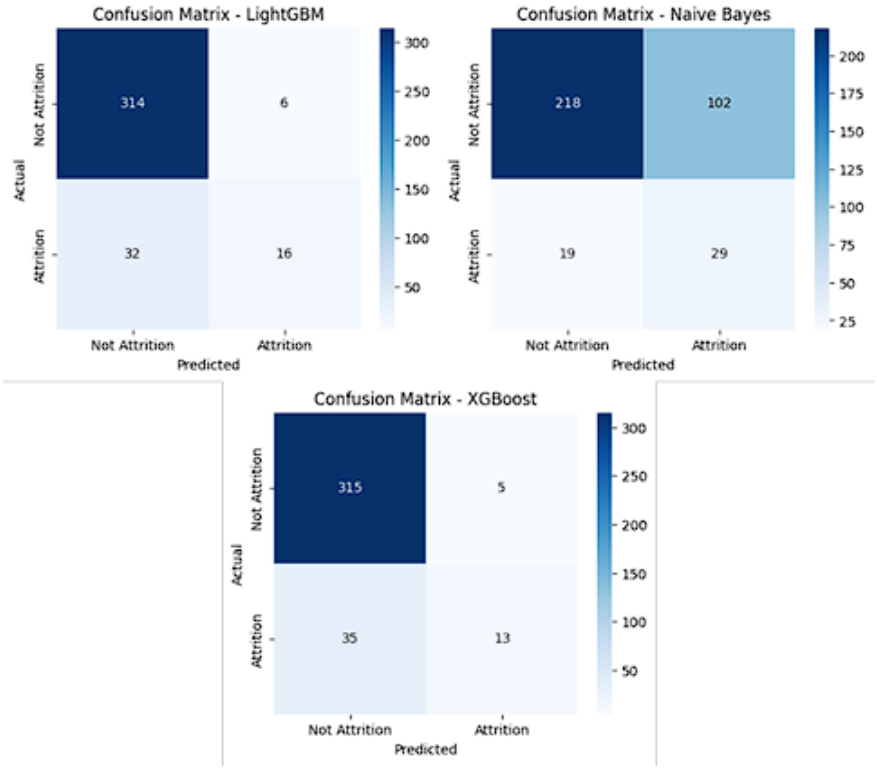


Figure 7.3: Confusion Matrices for LightGBM, Naive Bayes, and XGBoost

had a moderate false positive rate. However, it also misclassified 35 cases of attrition as non-attrition, which affects the recall. Naive Bayes was the worst performer; it misclassified 102 non-attrition cases as attrition, which shows that its prior probabilities are not suitable for the data set.

7.2 SALES DOMAIN

The analysis of weekly sales data over time *(Fig. 7.4)* demonstrates substantial seasonal trends, with periodic spikes most likely due to promotions, holidays, and economic situations. According to statistics, consumer behavior follows predictable annual cycles, allowing businesses to effectively manage stock levels and personnel planning in response to demand

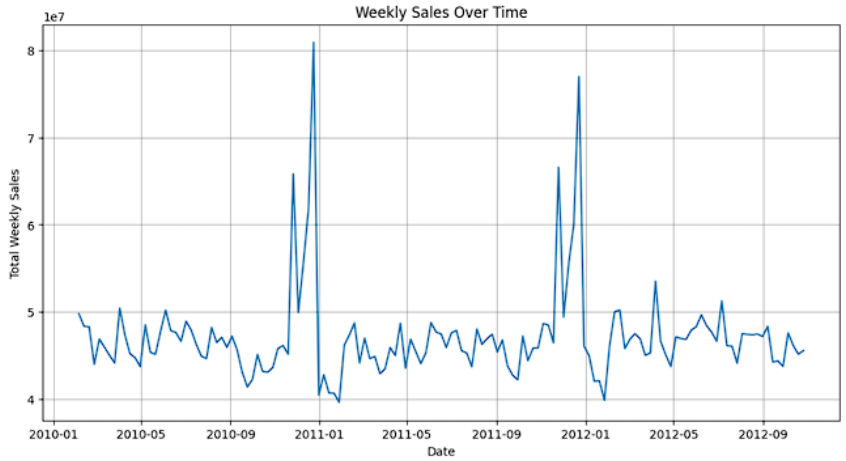


Figure 7.4: Confusion Matrices for LightGBM, Naive Bayes, and XGBoost

swings. Notably, significant gains in sales coincide with key shopping occasions, showing that targeted promotional actions may improve revenue during peak periods. The box plot in *(Fig. 7.5)* shows significant variation in weekly sales between holiday and non-holiday weeks. While holiday weeks have slightly higher median sales, there are more variations, reflecting inconsistent

Figure 7.5: Confusion Matrices for LightGBM, Naive Bayes, and XGBoost

higher median sales, there are more variations, reflecting inconsis-

tent consumer purchasing tendencies. Notable outliers in both holiday and non-holiday periods indicate that big sales increases can occur outside the traditional holiday seasons, most likely due to promotions or external reasons. These findings underscore the need of including both seasonal and non-seasonal effects in sales forecasting in order to enable dynamic marketing strategies and efficient inventory management. The correlation between temperature and sales in *(Fig. 7.6)* reveals a non-linear relationship in which extreme cold and hot conditions have a negative impact on in-store sales. This recommends that businesses should run seasonal promotions, such as discounts on weather-appropriate products and in creased internet buying incentives during severe weather

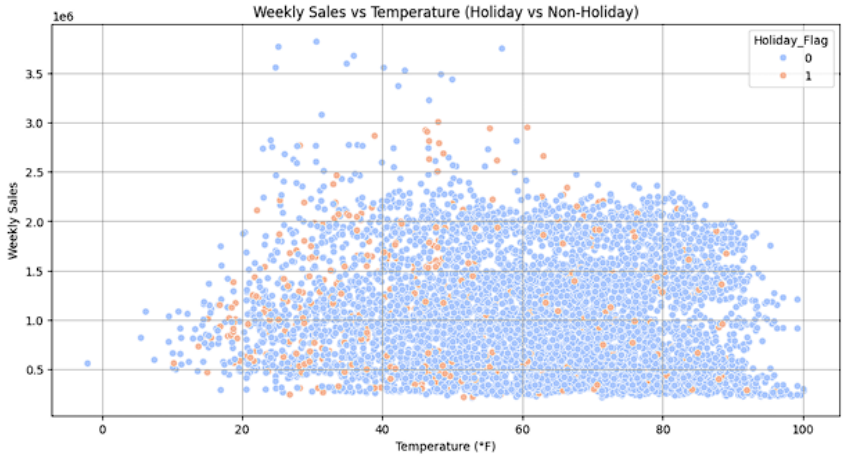
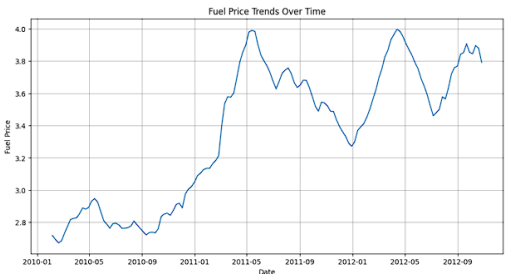
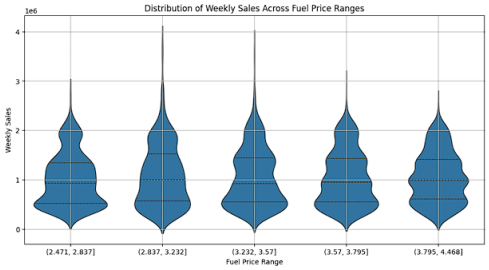


Figure 7.6: Confusion Matrices for LightGBM, Naive Bayes, and XGBoost

Fuel prices have a substantial impact on consumer pur chasing behavior. *(Fig. 7.7(a) and 7.8(b))* show that rising gasoline costs marginally reduce overall sales volume, notably in non essential retail groups. Consumers are likely to alter their buying habits to match rising transportation costs, emphasizing the significance of changing pricing structures and providing fuel-based incentives to maintain demand.



(a) Fuel Price Trends Over Time



(b) Distribution of Weekly Sales Across Fuel Price Ranges

Figure 7.7: Relationship Between Fuel Prices and Weekly Sales

(Fig. 7.8) shows the relationship between CPI and sales, demonstrating that inflationary pressures do not directly lower consumer spending. Instead, fluctuations in sales appear to be more influenced by external factors such as seasonal demand and macroeconomic events. This emphasizes the importance for firms to include economic data into their long-term pricing and inventory strategy rather than depending exclusively on inflation trends.

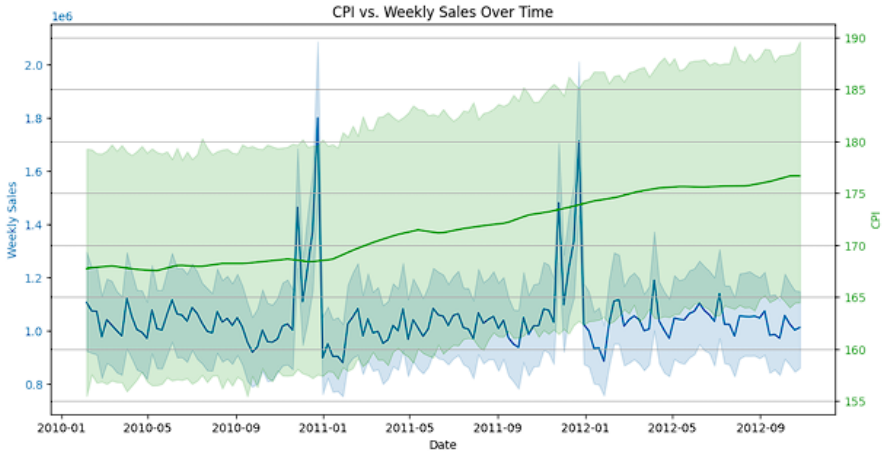


Figure 7.8: CPI vs. Weekly Sales Over Time

Feature correlation analysis. *(Fig. 7.9)* demonstrates that variables such as holiday schedules and fuel costs have the strongest link with sales, emphasizing their usefulness in predictive modeling.

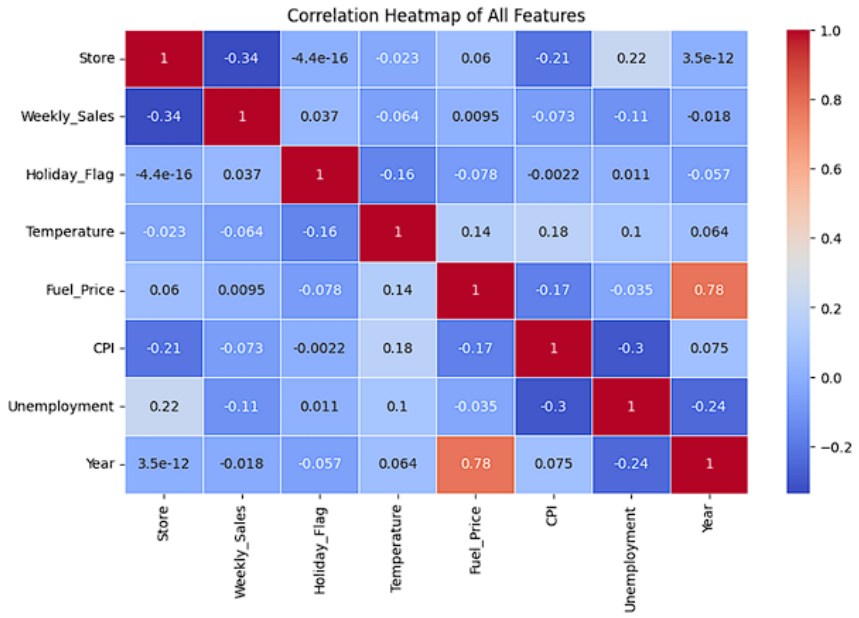


Figure 7.9: Correlation Heatmap of All Features

After confirming stationarity with the Augmented Dickey Fuller (ADF) test, which revealed that no additional trans formations were required, we used ARIMA, SARIMA, XG Boost, and Prophet models to forecast sales trends and assess their predictive performance. The model comparison shows different levels of effectiveness. ARIMA *(Fig. 7.10)* functioned well because the sales data was already stationary, requiring no significant adjustments. It reflected overall trends well, but suffered with unexpected increases induced by holidays or external factors such as fuel prices and CPI. Because it depended solely on historical values, it struggled to react to unanticipated changes. While it was useful for consistent trend predictions, it lacked flexibility in dealing with external forces.

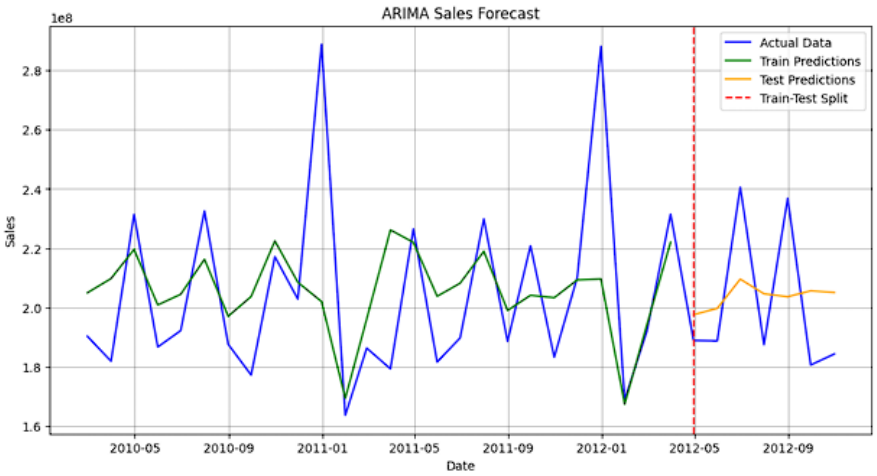


Figure 7.10: ARIMA Sales Forecast

SARIMA in *(Fig. 7.11)*, which extends ARIMA by include seasonal adjustments, did not show a substantial improvement. This was due to the data's weak and inconsistent seasonal trends, with many changes caused by external variables rather than a set seasonal trend. Because SARIMA does not automatically account for external variables, it performs similarly to ARIMA.

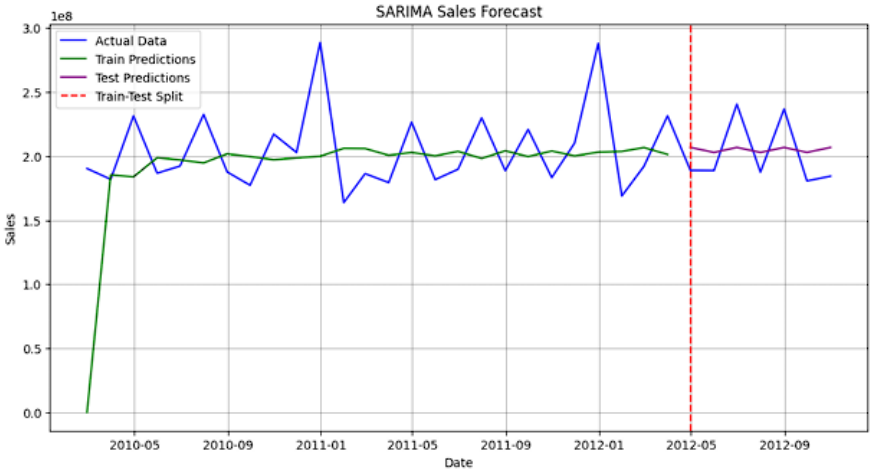


Figure 7.11: SARIMA Sales Forecast

XGBoost in *(Fig. 7.12)* was the least effective of the models, most likely because it struggled to capture time-dependent correlations in the data.

While it is effective for general predictive modeling, its inability to handle sequential dependencies makes it unsuitable for this form of time series forecasting. Unlike ARIMA and SARIMA, which directly predict trends and seasonality, XGBoost necessitated more feature engineering, which may not have been optimized for this dataset.

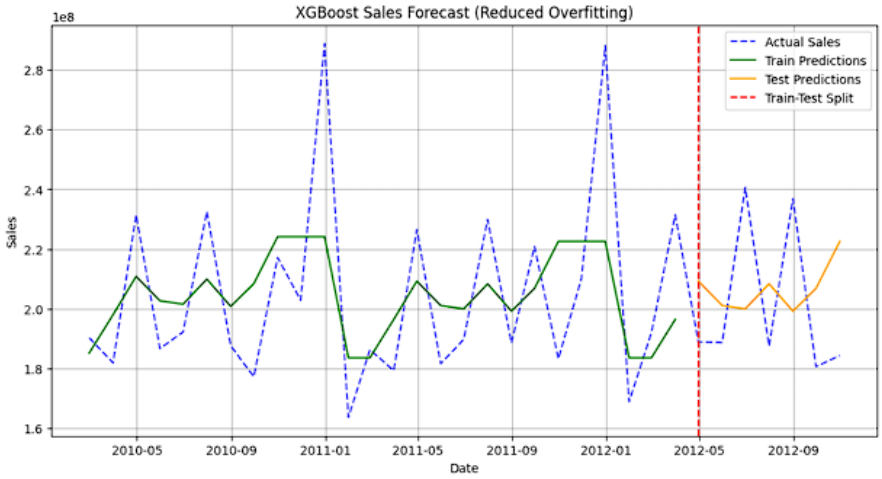


Figure 7.12: XGBoost Sales Forecast (Reduced Overfitting)

As seen in *(Fig. 7.13)*, Prophet performed the best, most likely due to its greater flexibility in dealing with trend shifts and missing data. It adapted flexibly to large changes and performed better in long-term predictions. However, it tended to smooth out sharp spikes in sales, making it less successful at capturing unexpected seasonal surges. Prophet's capacity to absorb external parameters led to its impressive performance, making it more versatile than ARIMA and SARIMA.

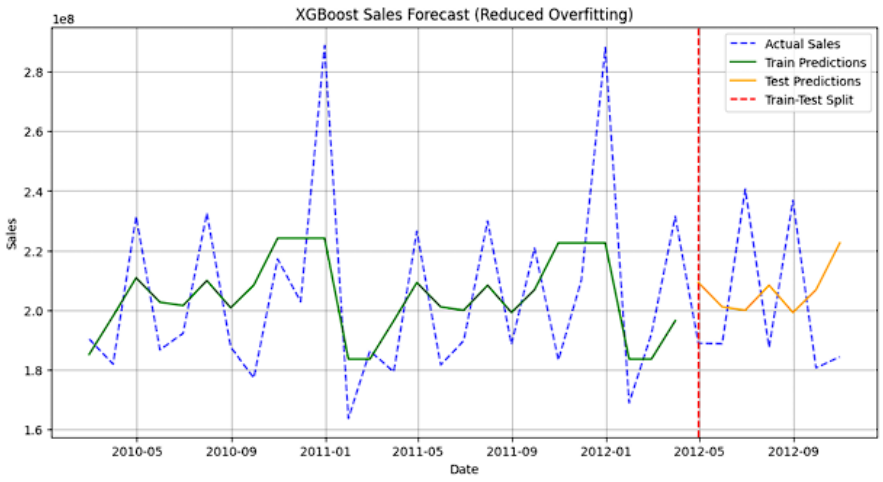


Figure 7.13: Prophet Sales Forecast

Finally, based on *(Fig. 7.14)*, we can conclude that Prophet was the best-performing model, providing the most flexibility and adaptability. ARIMA and SARIMA performed similarly, with ARIMA significantly ahead due to the data's steady nature. Despite its competence in general predictive modeling, XGBoost was the least effective at capturing time-series relationships, making it the worst performance on this dataset.

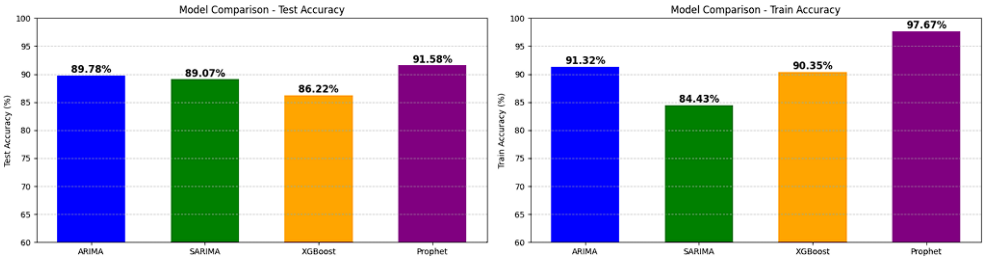


Figure 7.14: Model Comparison- Train and Test Accuracy

These findings can improve business operations in a variety of ways. Retailers can use holiday-based sales analytics *(Fig. 7.5)* to optimize inventory levels and marketing efforts. Our approach also emphasizes the need of clustering in marketing to concentrate on certain customers. Supply chain managers, on the other hand, can predict demand fluctuations based on

fuel price patterns *(Figs. .7(a) and 7.8(b))*, enabling cost-effective logistics planning. Financial planners can use CPI trends (Fig: 18) to align pricing strategies with economic condi tions, reducing the danger of inflationary pressures. Weather based marketing tactics *(Fig. 7.6)* can assist organizations in proactively adapting to seasonal fluctuations in consumer behavior. Businesses that incorporate predictive insights into their operational plans can increase revenue, improve customer happiness, and maintain a competitive advantage in a constantly evolving market.

7.3 MARKETING DOMAIN

The spending and credit behavior correlations are illustrated on the heatmap in *(Fig. 7.15)*. Between individual and total purchases, the correlation is very high (0.92), which means that large purchases aid significantly in total expenditure. There is also a high correlation (0.68) between purchase frequency and total spending, meaning that frequent purchases, though lower in ranking than large purchases, do contribute to accumulating total expenditure. From the behavior used in credit within the limits analyzed, the correlation does not show a stronger positive relation with higher balances (0.36). The correlation with balance is negative for full payment (-0.32) in such a way that those making full payments have less outstanding credit. There is also a strong correlation between the volume of transactions and the frequency of cash advances (0.80), while lower consumers with cash limits tend to rely more on cash advances (-0.18). Customers already using the services show more stable credit use (-0.32) and a greater likelihood of paying off the balances (-0.16). This improves the development of credit risk models and methods of customer segmentation to further control risk exposure and tailor financial services accordingly.

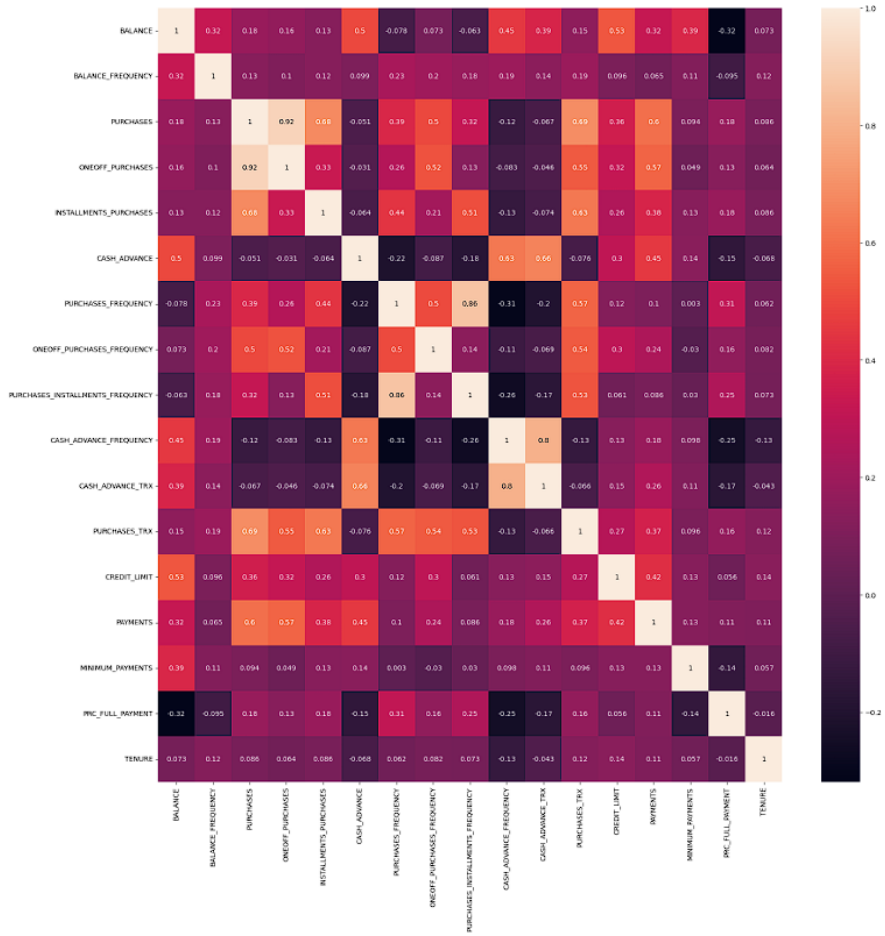


Figure 7.15: Correlation Heatmap

The scatter plots illustrated in *(Fig. 7.16)* show the clustering of customers according to their financial attributes and their purchasing behavior. While one-off purchases display a good linear relationship with total purchases, it seems that high-value transactions very much drive overall spending. There appears to be a positive correlation between installment purchases and total spending; that is, those who tend to make structured payments spend more. The clusters display a heterogeneous spending pattern. High-spending customers are expected to have high credit limits, but this does not imply that high-limit clients will always have high purchase activity; thus, they represent different behavior types. The association of cash advances with purchases is weak.

Model	Train Accuracy	Test Accuracy	Train R^2 Score	Test R^2 Score	Test MAE	Test MSE	Test RMSE
ARIMA	91.32%	89.78%	0.2015	0.1026	20,969,899.48	515,735,357,671,174.00	22,709,807.52
SARIMA	84.43%	89.07%	-1.5232	0.0534	22,304,352.19	543,982,255,153,634.38	23,323,427.17
XGBoost	90.35%	86.22%	0.3307	-0.5448	28,033,022.90	887,797,876,340,724.62	29,795,937.25
Prophet	97.67%	91.58%	0.9465	0.1006	17,213,653.18	516,880,443,514,011.62	22,735,004.81

Table 7.2: Performance Comparison of Different Models in Sales Forecasting

So, cash withdrawers cannot spend more in retail. The frequency distributions of different purchases reveal that mid-range spenders made several purchases with frequent repetition, whereas most high spenders make massive purchases that seldom occur. Whole payment behavior seems scattered throughout clusters, so different financial strategies must be utilized.

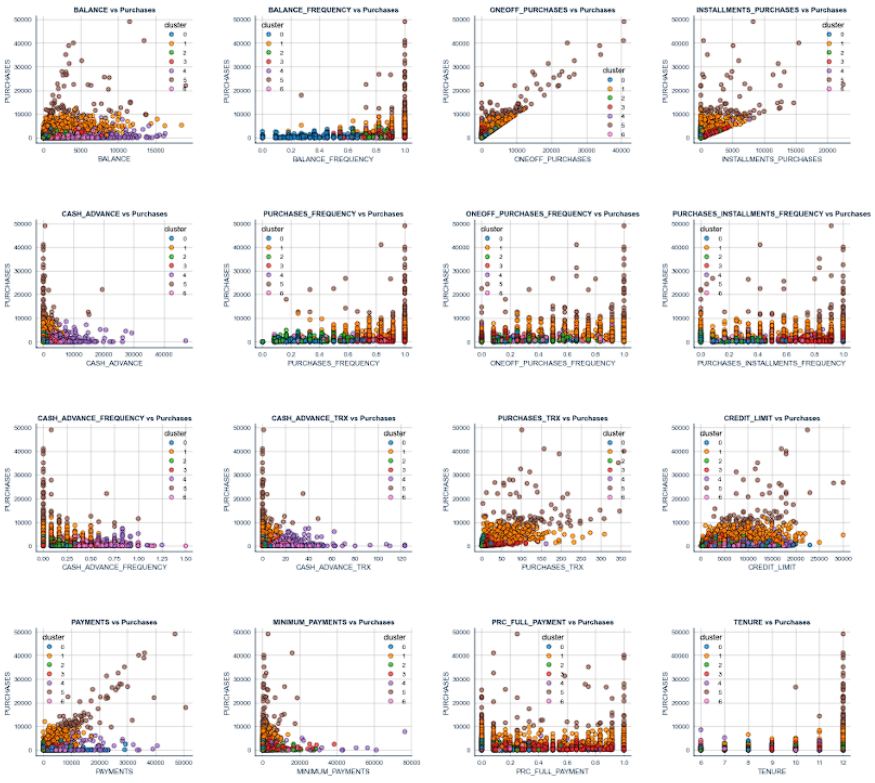


Figure 7.16: Clustered Scatter Plot Matrix- Purchase Behavior Analysis

Illustrated in *(Fig. 7.17)* is the distribution of financial behaviors across a total of 7 clusters. The 5th cluster displays largely inflated values of bal-

ance, one-off purchases, installment pur chases, payments, as well as credit limits, meaning they are high-value customers with very high volumes of transactions. The 4th cluster shows moderate-to-high values across cash advances and their frequency of transactions. Frequency dis tributions suggested that clusters 1, 3, and 5 exhibited similar patterns of frequency for balance, purchases, and one-off purchases, respectively. All the clusters mostly have the same tenure (around 8-12 units), while cluster 0 has the least values for the most metrics. The 2nd and the 6th clusters show very low financial activities in general but keep the relationship of accounts. The proportion of customers who pay in full varies the most between customers in the 1st and the 5th clusters, which highlights different repayment behaviors for those segments that are otherwise similar.

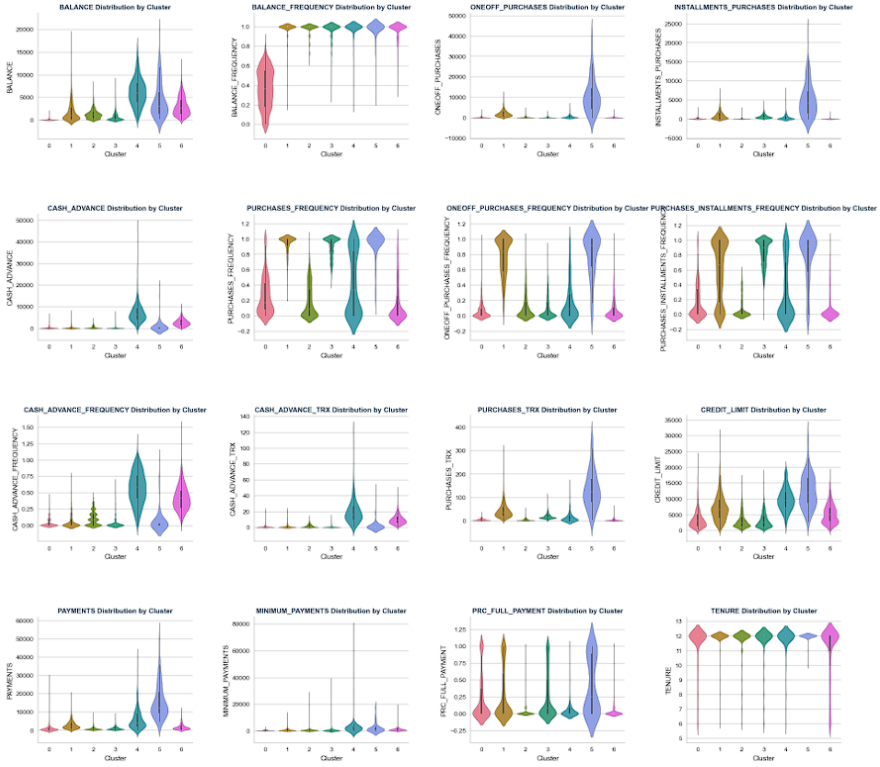


Figure 7.17: Violin Plot Matrix- Distribution of Financial Behaviors by Cluster

(Fig. 7.18) compares four clustering techniques on credit card cus-

tomer data for original and dimensionally-reduced formats. The original data plots (top row) reveal compressed distri butions with poor separation of clusters for both K-Means and GMM.

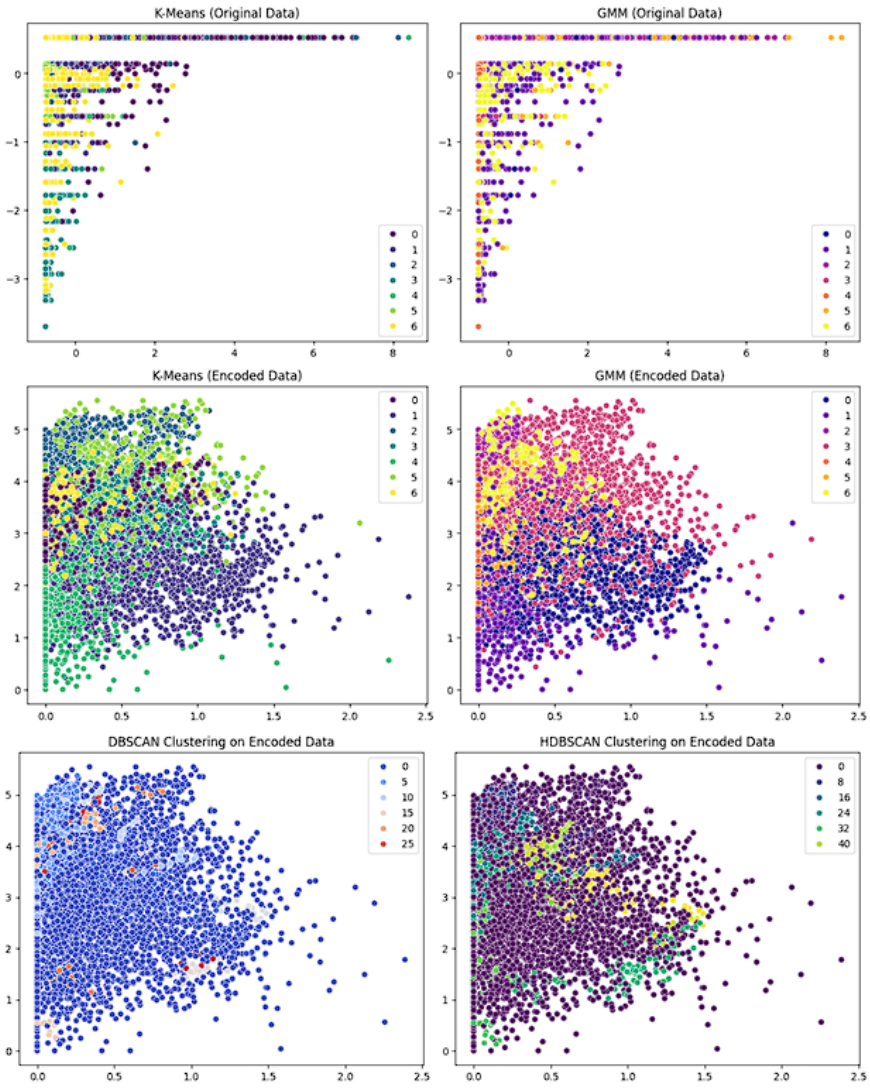


Figure 7.18: Comparison of Clustering Methods on Original and Encoded Data

After dimensional reduction (middle row), K Means produces clear boundaries between seven customer segments (clusters 0-6), particularly fixing cluster 0 at higher x-values. GMM's encoded implementation shows distinct clus ter 3 (pink) and cluster 6 (yellow) populations but with signifi-

cant overlap among other segments. DBSCAN (bottom left) predominantly identifies a single main cluster (blue) with minimal differentiation and sparse outlier points (orange/red). HDBSCAN (bottom right) creates a more granular segmenta tion with clusters showing density-based groupings that cap ture underlying financial behavior patterns. The visualization demonstrates that encoding dramatically improves clustering performance, with K-Means on encoded data yielding the most practical customer segmentation solution for targeted marketing applications.

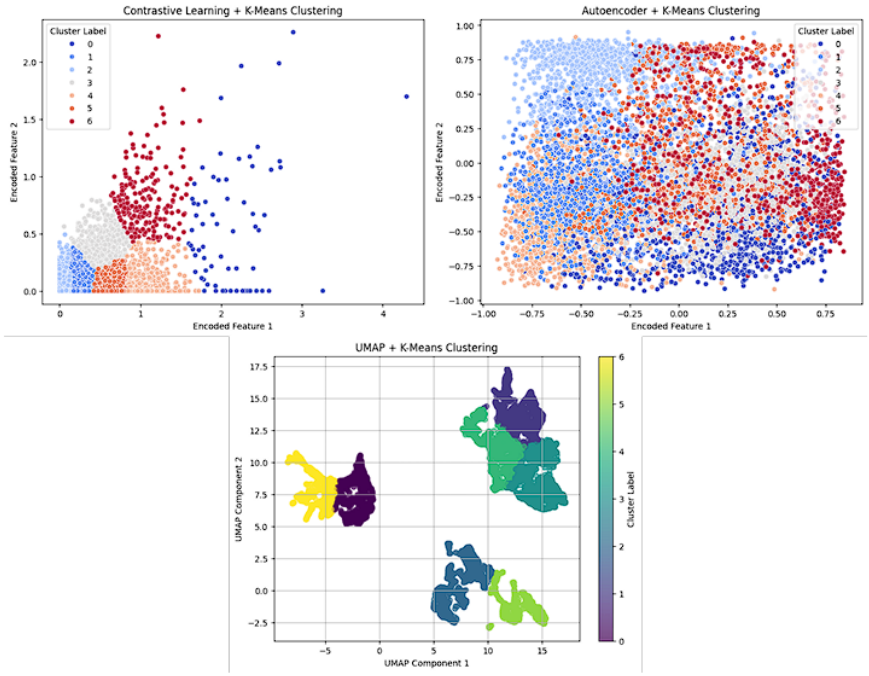


Figure 7.19: Comparison of Contrastive Learning, Autoencoder, and UMAP with K-Means Clustering

(Fig. 7.19) contrasts three dimensionality reduction techniques with K-Means clustering for the purpose of credit card customer segmentation. Contrastive Learning (top left) stands out by being able to create distinguishable clusters 0 (blue), 6 (dark red), and all the other ones; this also indicates success in picking out the major differences in customer financial behaviors. The second one, the Autoencoder method (top right), produced

a normalized distribution from-1.0 to 1.0 on both axes, although the overlap of clusters was very high for clusters 1 (light blue), 4 (pink), indicating lower feature extraction completeness regarding segmentation purposes. UMAP is able to offer far more separation due to the non-linear embedding space, where almost no clusters overlap, with the largest number of customers forming little islands for cluster 0 (purple) and cluster 3 (teal). Such visualization proves that UMAP is, in fact, the best among financial customer segmentation techniques, combining good dimensionality reduction with clear boundaries between clusters, enabling targeted marketing strategies.

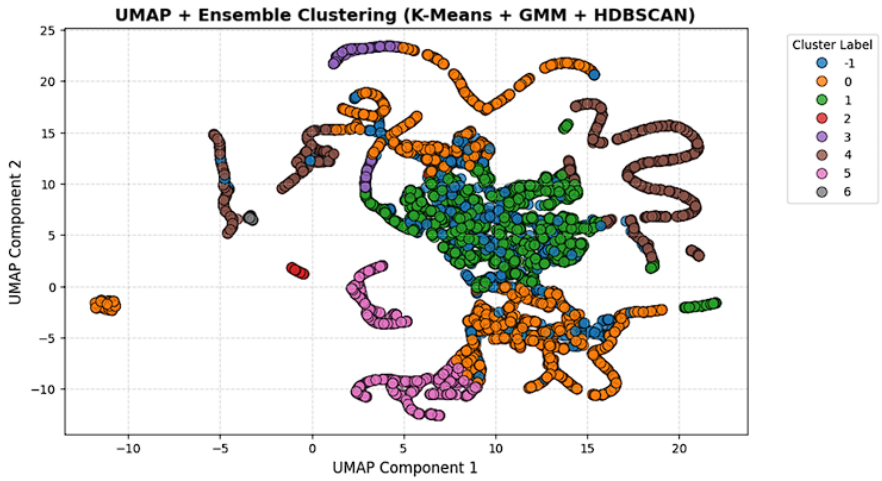


Figure 7.20: UMAP Visualization of Ensemble Clustering using K-Means, GMM, and HDBSCAN

(Fig. 7.20) presents the ensemble clustering results of K Means, GMM, and HDBSCAN algorithms glued together by UMAP dimensionality reduction. From this visualization, seven separate clusters with unique spatial patterns can be observed: cluster-1 (blue) clearly puts scattered points in the center; cluster 0 (orange) has formed elongated branches leaving the center and going outwards with quite an isolated group at (-10,-2); cluster 1 (green) consists of mostly a clustered core with some extensions; clusters

2 (purple) and 5 (pink) exhibit curved forms at the lower part of the plot; cluster 4 (brown) produces dramatic snake-like features that wrap around other clusters; cluster 6 (gray) appears small and resembles a curved segment in the upper region. The complex, non-linear borders between these clusters show that the ensemble approach has captured a variety of different data distributions, which would have been difficult to capture by any single clustering algorithm.

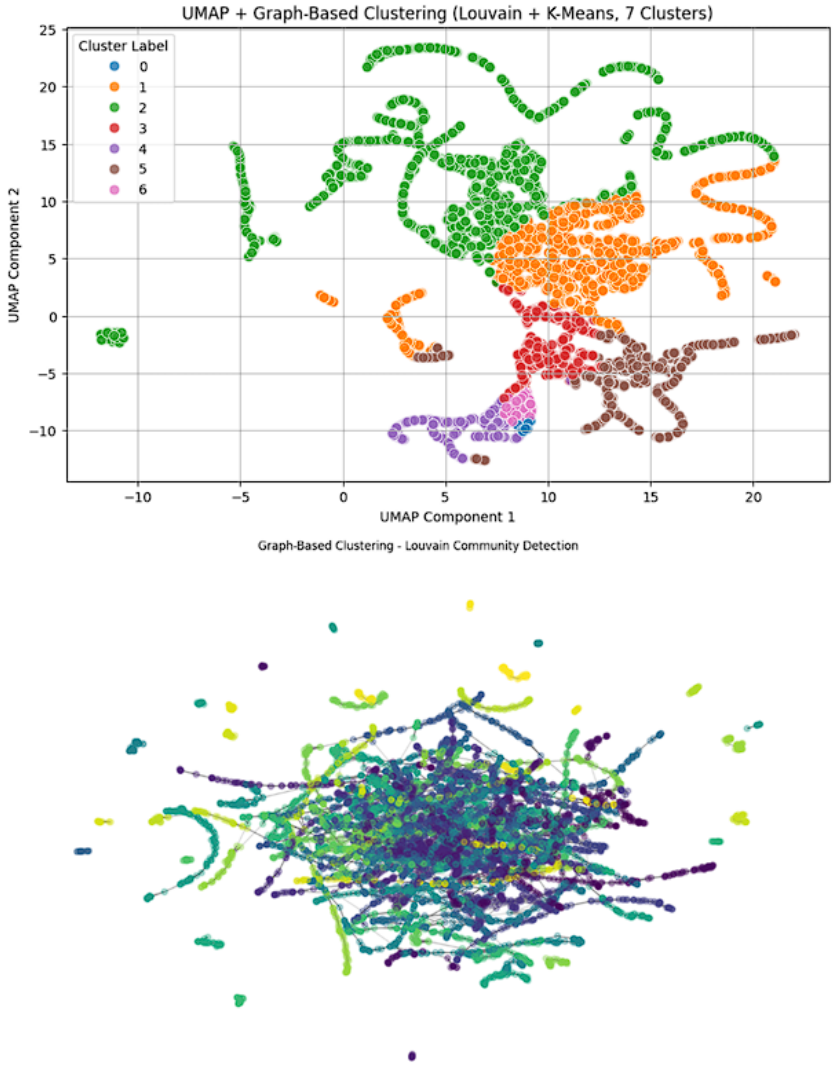


Figure 7.21: UMAP and Graph-Based Clustering using Louvain Community Detection and K-Means

(Fig 7.21) describes the findings of a hybrid graph clustering scheme

combining Louvain community detection with K Means (k=7) used on high-dimensional data. In the upper sec tion, the UMAP projection shows seven clusters, which may be visually differentiated based on topology: cluster 2 (green) builds branches-like structures in the upper quadrants; cluster 1 (orange) forms an aggregate mass from which branches diverge; clusters 3 (red), 5 (brown) in addition to 4 (purple) create well-defined curved manifolds lower down; cluster 0 (blue) not so significant appears only at critical junction points; and cluster 6 (pink), appears as isolated. The lower section visualizes the underlying network structure used by the Louvain algorithm to determine community modularity based on certain patterns of connectivity among nodes, where the density of edges corresponds to cluster assignation. It further shows how the algorithm managed to successfully preserve the local and global topological referents, keeping the non-convex structures intact, which were invisible to standard distance clustering techniques. The patterns suggest the efficacy of identifying innate data manifolds in environments of varying density without forced spherical or convex restrictions on the geometry of clusters.

(Fig. 7.22) shows distinctive behavioral traits of the seven segments of customers in terms of feature-cluster correlations. Cluster 0 is **"Cash-Advance Heavy Users"** and has a strong positive correlation in *CASH_ADVANCE_TRX (4.61)*, *PURCHASES (4.43)*, and *CASH_ADVANCE (4.16)*, depicting the customers who are heavy users of cash advances with high purchase frequency.

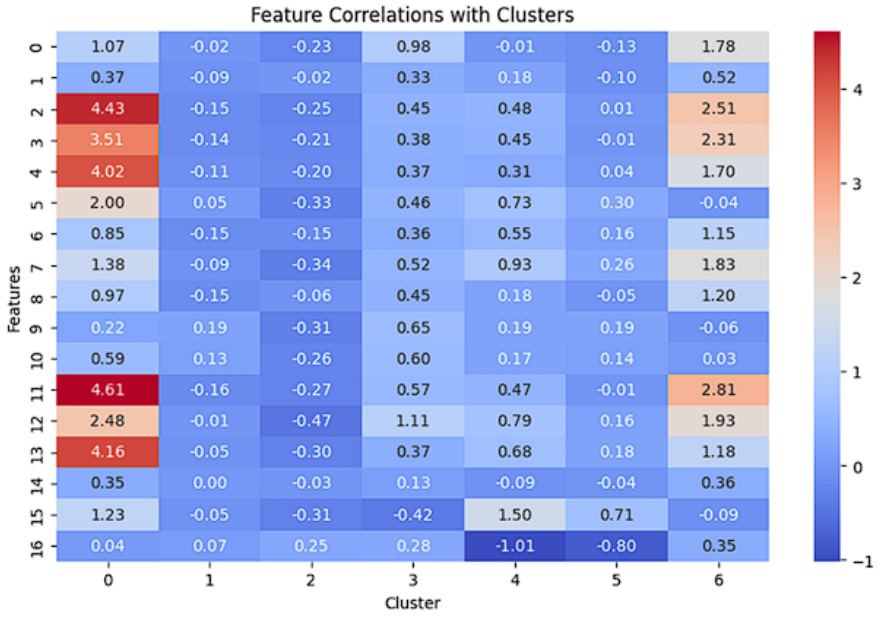


Figure 7.22: UMAP and Graph-Based Clustering using Louvain Community Detection and K-Means

Cluster 1, "Minimal Engagement", has low values of correlation on all the features, depicting the customers with minimum frequency in the credit card usage pattern. Cluster 2, **"Low Usage Accounts"**, has low to very low negative correlatedness, particularly with PURCHASES_TRX (-0.47), with low or inactivity accounts customers. Cluster 3, **"Stable Revolvers"**, has medium amounts of positive correlation between the majority of the attributes, particularly with *BALANCE_FREQUENCY (1.11)*, typified with balanced frequency but stable usage pattern customers. Table 3 illustrates the performance of various clustering algorithms in establishing that banks are able to obtain good insight into customers. The best is UMAP + Graph-Based Clustering at a Variation Ratio Criterion (VRC) of 9766.45 with perfect separation. This allows banks to differentiate between finer differences among customers, such as a young professional who spends using credit cards over holidays and a family that receives maximum rewards by purchasing items for groceries. In addition, the Ensemble

Method	Variance Ratio Criterion (VRC)	Computation Time (s)
K-Means (Original)	1365.864485	4.917399
GMM (Original)	543.108033	13.402929
K-Means (Encoded)	2258.523106	2.226011
GMM (Encoded)	1210.519953	6.617846
DBSCAN (Encoded)	187.346604	1.150230
HDBSCAN (Encoded)	101.982328	2.243511
UMAP + Clustering	29066.21	1.1212
Autoencoder + K-Means	1431.78	0.4238
Contrastive Learning + K-Means	8998.713569	0.449833
K-Means (Ensemble)	9110.559996	0.125614
GMM (Ensemble)	7676.627491	0.126673
HDBSCAN (Ensemble)	28044.207847	0.246344
Ensemble (K-Means + GMM + HDBSCAN)	435.064799	0.498631
UMAP + Graph-Based Clustering (Louvain + K-Means, 7 Clusters)	8998.713569	3.437985
Graph-Based Clustering - Louvain Community Detection	9766.455415	1.683769

Table 7.3: Performance Comparison of Different Clustering Models

methodology, through its integration of K-Means, Gaussian Mixture Models (GMM), and HDBSCAN, unites various analytical perspectives. Similarly, top-shelf banks utilize a broad array of financial know how to address sophisticated client needs.

7.4 PUBLIC RELATIONS DOMAIN

(Fig. 7.23) gives a comparison of accuracies for sentiment classification models. BiLSTM is the top-performing model among all models with 85.00% accuracy, followed by SVM

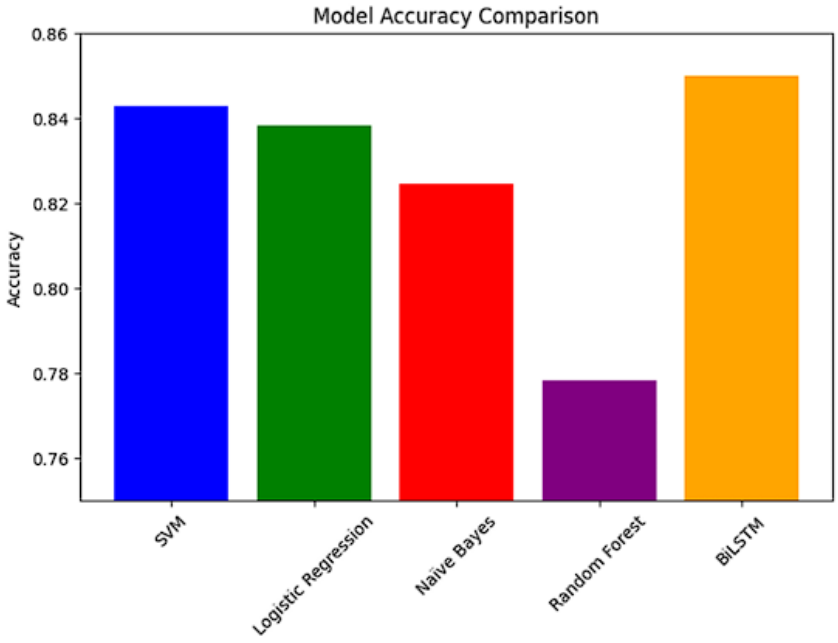


Figure 7.23: Comparison of sentiment classification model accuracies

(84.30%) and Logistic Regression (83.83%).While Naive Bayes is computationally efficient, its accuracy stands at 82.48%. Random Forest performs the worst with accuracy 77.83%,demonst rating that ensemble learning does not make much difference here in the text classification task under high dimensions.The results show the better performance of deep learning methods (BiLSTM) on sentiment analysis as well as better performance of conventional machine learning algorithms such as SVM and Logistic Regression.

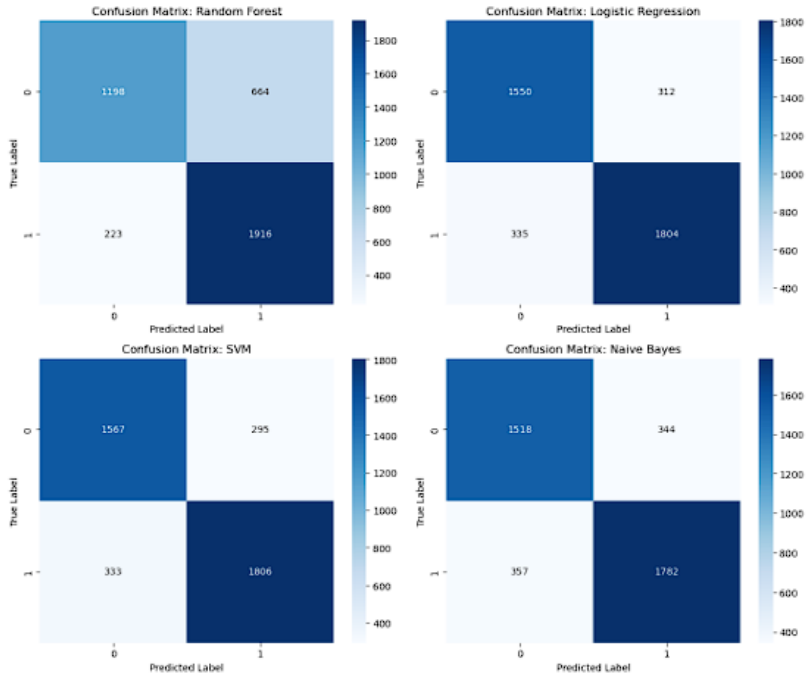


Figure 7.24: Confusion matrices for different sentiment classification models

(Fig. 7.24) illustrates the confusion matrices for the four machine learning algorithms: Random Forest, Logistic Regression, SVM, and Naive Bayes. The matrices show each algorithm's performance for correctly classifying sentiment labels on the diagonal in terms of true positive and true negative values.SVM and Logistic Regression have good classification accuracy and misclassify fewer examples compared to Random Forest and Naive Bayes.The Naive Bayes model exhibits greater false positive and false negative, indicating direction bias in the case of frequent sentiment classes.These results emphasize that the choice of model is relevant in sentiment classification in public relations analysis. *(Fig. 7.25)* is a plot of false positives and false negatives in sentiment classification models. One can quite clearly observe from the figure that Random Forest has the highest false negative rate by incorrectly classifying more than 650 instances and failing to flag them as correct positive sentiments. The others, Logistic Regression,

SVM, and Naïve Bayes, have a less skewed proportion of misclassifications. SVM and Naïve Bayes exhibit a more balanced ratio of false positives to false negatives, which reflects a more stable classification result. This observation points out that the emphasis should be placed on reducing false negatives in public relations sentiment analysis because leaving out negative sentiment can lead to delayed response in crisis management.

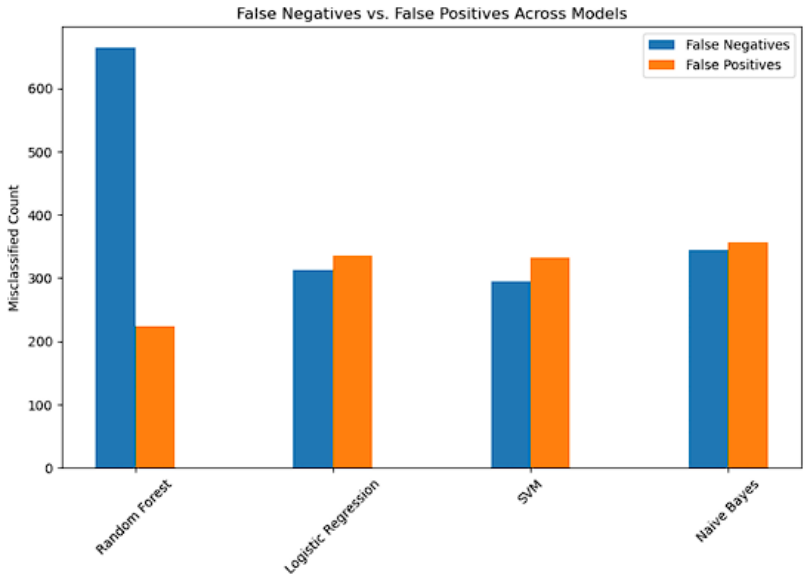


Figure 7.25: False Negatives vs. False Positives across different models

Fig. 7.26 gives the word cloud visuals for misclassified reviews for various models of sentiment classification, namely Naïve Bayes, SVM, Logistic Regression, and Random Forest. The most common words that get misclassified across all the models are *"book," "one," "time," "good," and "real."* They are used mostly in neutral or unclear contexts and thus hard for the models to classify. The frequency of the presence of words such as *"great"* and *"better"* in misclassified instances reveals the weakness of issues in sentiment polarity analysis, particularly of the mixed sentiments in consumer reviews. The above graph helps in outlining the boundary

of the importance of contextual knowledge of sentiment in public relations application as fuzzy words will inevitably lead to misinterpretations of consumers' sentiment.

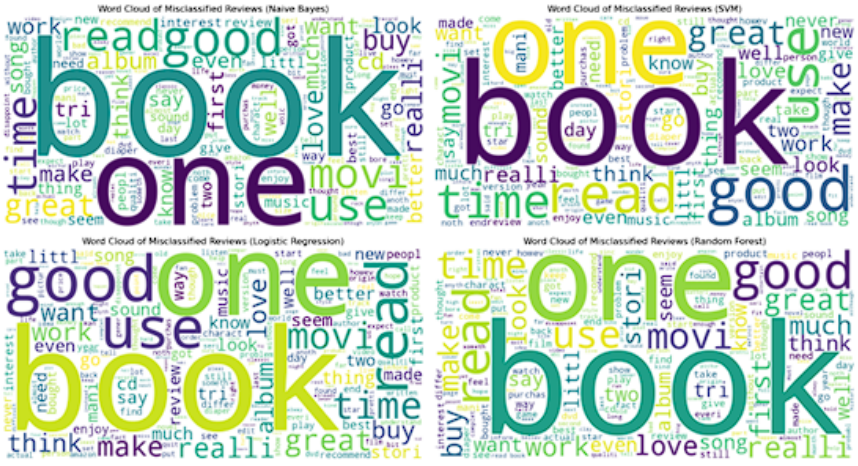


Figure 7.26: Word cloud visualization of misclassified reviews for different sentiment classification models

(Fig. 7.27) presents confusion matrices for the BiLSTM sentiment classification model. The left matrix presents the model's performance on the test set, and the right matrix presents the validation data results. The BiLSTM model possesses a high capability to classify, with high true positives and true negatives. There are not many positive examples that are classified as negative, which may result in a slight underestimation of positive sentiment.

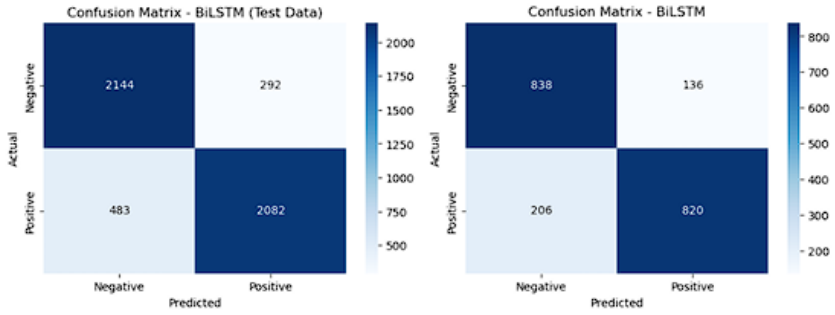


Figure 7.27: Confusion matrices for the BiLSTM model

However, overall classification accuracy remains accurate validating the effectiveness of deep learning models for sentiment analysis. *(Fig. 7.28)* shows the BiLSTM model performance across training epochs. The left plot shows the accuracy trend, where training accuracy increases monotonically and validation accuracy stabilizes at 82-85%, indicating good generalization of the model. The right plot shows the loss curves, where the training loss and validation loss decrease monotonically as epochs increase.

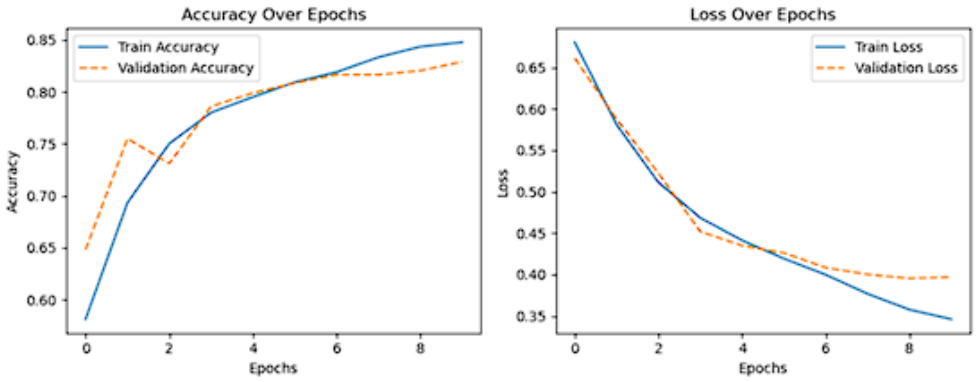


Figure 7.28: BiLSTM model performance over epochs

However, a narrow margin in posterior periods reflects less overfitting, implying that regularization or dropout adjustments can help improve. The results validate the effectiveness of BiLSTM in sentiment analysis with robust convergence in learning. *Table 7.3* shows a sentiment classification models compared in terms of performance, where there are strong strengths and drawbacks. BiLSTM has the highest accuracy (85.00%), yet the actual strength is in sequence learning and contextuality. In contrast with SVM and Logistic Regression using TF-IDF-based feature engineering, BiLSTM reads text in two directions, thus is able to catch fine sentiment changes as well as word dependencies. This makes it possible for highly proportionate recall over sentiment classes,

Model	Accuracy	Class	Precision	Recall	F1-Score
BiLSTM	85.00%	Negative	0.82	0.88	0.85
		Positive	0.88	0.81	0.84
		Macro avg	0.85	0.85	0.85
		Weighted avg	0.85	0.85	0.84
SVM	84.30%	Negative	0.82	0.84	0.83
		Positive	0.86	0.84	0.85
		Macro avg	0.84	0.84	0.84
		Weighted avg	0.84	0.84	0.84
Logistic Regression	83.83%	Negative	0.82	0.83	0.83
		Positive	0.85	0.84	0.85
		Macro avg	0.84	0.84	0.84
		Weighted avg	0.84	0.84	0.84
Naive Bayes	82.48%	Negative	0.81	0.82	0.81
		Positive	0.84	0.83	0.84
		Macro avg	0.82	0.82	0.82
		Weighted avg	0.82	0.82	0.82
Random Forest	77.83%	Negative	0.84	0.64	0.73
		Positive	0.74	0.90	0.81
		Macro avg	0.79	0.77	0.77
		Weighted avg	0.79	0.78	0.77

Table 7.4: Sentiment Analysis Model Performance Comparison

In contrast to models such as Random Forest with very low recall of 0.64 for negative sentiment classification. Even during the learning phase itself, BiLSTM is regular in generalization (83.00% accuracy), i.e., does not overfit even though it is having a more deeper structure. This makes it very well adapted to sarcasm, negations, and complex language structures, which are most useful for sentiment analysis to be used in PR. Even with the power of deep learning, SVM and Logistic Regression are still contenders at 84.30% and 83.83% respectively. Their power is that they use TF-IDF word representation, which is suitable at picking strong words. Sentiment classification is mostly linearly separable pattern, and an easy hyperplane (SVM) or

logistic function can effectively split positive and negative sentiment. SVM is better than Logistic Regression in the sense that it can be optimized for the margin between classes and thus is not highly sensitive to small variations. Both are bad with negation and multi-word expressions (e.g., "I was expecting a good book, but it was terrible" might be mislabeled because there are positive words like "good"). Alternatively, if computational cost is an issue, SVM or Logistic Regression would suffice instead of deep learning as long as sentiment phrases are not overly complex. Naïve Bayes at 82.48% accuracy is typical in demystifying its fault in dealing with word interdependencies. The algorithm assumes that the words are independent and are responsible for the sense, whereas in real text it's a false assumption (like for instance "not good" would be treated as a word, not individually as words). The outcome is highly accurate yet not highly recallful, Naïve Bayes never giving false positives yet at the cost of not being able to detect true negatives. It is a trade-off that makes it ideal for fast and light analysis, particularly in real-time surveillance where speed trumps accuracy. In widespread PR sentiment tracking, though, Naïve Bayes is not appropriate as it may not be able to detect significant negative sentiments. Of all the models, Random Forest does the worst with a mere 77.83% accuracy, and low negative recall sentiment (0.64) in particular. The main reason behind this is its poor ability to cope with high-dimensional sparse text data. Since Random Forest functions on the notion of using multiple decision trees trained on randomly chosen features, it has no way of understanding the complex word relationships required for sentiment classification. The model will show bias towards predicting a greater number of reviews as positive, leading to higher recall for positive sentiment (0.90) but unable to detect negative sentiment. Random Forest is therefore not appropriate for sentiment classification in

PR usage, where detection of negative sentiment is vital in reputation and crisis management. Another crucial thing to consider in model evaluation in such models is macro vs. weighted average difference. Most studies attempt to quantify accuracy alone, while real impact comes from whether or not the model will handle class imbalance. Macro averages treat both classes (negative and positive) as equally relevant, whereas weighted averages take account of true world class distribution. Low weighted average F1-score (i.e., 0.77 when true distributions are used) verifies poor sentiment prediction. High macro average but low weighted average indicating that the model is biased towards the majority class and thus could not be utilized for PR crisis management. Short-lived, BiLSTM is the superior model in sentiment analysis for detection of fine-grained sentiment variation and sense-based dependency. Naïve Bayes and SVM and Logistic Regression can be used where speed in computation is given a higher preference, while Naïve Bayes excels in identification speed but shallower in detection of dependence. Random Forest is not good when handling sentience subtlety and hence the worse option for sentiment analysis concerning PR.

Chapter 8

CONCLUSION

This research has indicated the ways Machine Learning (ML)methods can enhance business operations, such as sales forecasting, employee retention, emotion detection, and customer segmentation.Through clever computer applications and wide testing routines, businesses are in a position to utilize data-driven decision-making to streamline their operations and become more effective in planning. From the findings, it was determined that Prophet works best for holiday and seasonality trends when applying sales forecasting. XGBoost is the best among all models to predict who is likely to leave to enable HR departments to plan accordingly. BiLSTM is best to detect when people are shifting in their thought process about something, which can assist in public sentiment and brand management. K-Means Clustering performs well in recognizing various types of customers so that marketing teams can focus more intensely. Integration with ML makes a huge difference across the various departments in the sense that innovation from one department might make others more efficient by resulting in a best-case business setup from be ginning to completion. Prophet-based sales predictions favor marketing departments to plan for better campaigns according to anticipated demand. K-Means customer segmentation makes advertising more effective by targeting the appropriate crowd. In HR, BiLSTM sentiment analysis of employees provides insights to organizations regarding employee morale as well as turnover prediction through XGBoost. This enables organizations to intervene to retain their employees before they

leave. Public opinion analysis further assists organizations in aligning the employer's reputation of their organizations with what the employees desire, and thereby, HR policies will attain a good company reputation. In addition, customer intelligence also enables firms to expand through the support of marketing campaigns through the optimization of loyalty programs and the development of products in response to market trends. With the use of BiLSTM in monitoring what individuals feel, firms can monitor the mood of the public. On the other hand, XGBoost in HR analytics provides insights regarding customers that can be used in talent acquisition and talent management. Having the ability to feel what customers prefer, how contented employees are, and what the market requires immediately makes companies more robust and resilient as a whole.

8.1 SCOPE OF FURTHER WORK

With further advancements in AI and data science, there are several avenues to further develop and enhance this project. Future development can be done to improve the methodologies, implement more sophisticated AI methods, and integrate real-time analytics for dynamic decision making.

8.1.1 Improving Model Performance and Generalization

- **Hyperparameter Tuning:** Subsequent versions may work on model hyperparameter tuning based on automated methods like Grid Search, Random Search, and Bayesian Optimization.
- **Ensemble Learning:** Using multiple models with methods like Stacking, Blending, or Voting Classifiers to enhance predictive performance.
- **Transfer Learning:** Utilizing pre-trained models and adapting them to

business-specific data to build high-performing models with minimal data.

- **Explainable AI (XAI):** Adding interpretability methods such as SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) to explain AI decisions.

8.1.2 Integration of Automation and Real-Time Analytics

- Enabling real-time dashboards that update in real-time with fresh data, facilitating immediate decision-making.
- Creating automated retraining pipelines that adjust models to changing patterns of data.
- Deploying chatbots and AI assistants that employ NLP to offer insights and suggestions to HR, Sales, Marketing, and PR teams.

8.1.3 Deployment as a Unified Business Intelligence Platform

- Building a centralized dashboard consolidating HR analytics, sales forecasting, customer sentiment analysis, and public relations tracking.
- Designing cloud-based APIs to provide easy access to insights across various departments.
- Designing cloud-based APIs to provide easy access to insights across various departments.

8.1.4 Future Business Applications and Expansion

- Expanding the project across other departments such as Finance, Operations, and Supply Chain Management.

- Utilizing deep learning methods such as Reinforcement Learning for adaptive decision-making in marketing and selling strategies.
- Investigating the application of synthetic data generation in mimicking business situations and training AI models in multi-environment setups.

This project provides a solid basis for AI-based business intelligence, but the potential for growth is enormous. By ongoing model optimization, adding new data sources, and applying leading-edge AI methods, this system can be developed into a complete enterprise-grade AI-fueled decision-support tool. Future developments will concentrate on accuracy, automation, and ease of use, making AI-based insights more actionable for businesses globally.

REFERENCES

[1] **F. Jin and L. Wang**, "Evaluation and analysis of strategic human resource management based on multi-mode fuzzy logic control al gorithm," 2020 5th International Conference on Mechanical, Con trol and Computer Engineering (ICMCCE), Harbin, China, 2020, pp.

[2] **F. Guan**, "Human resource management innovation based on the view of knowledge management," 2010 IEEE International Confer ence on Management of Innovation Technology, Singapore, 2010, pp.

[3] **Y. Liu**, "Analysis of Human Resource Management Mode and Its Selection Factors Based on Decision Tree Algorithm," 2020 International Conference on Advance in Ambient Computing and Intelligence (ICAACI), Ottawa, ON, Canada, 2020, pp.

[4] **B. Galli**, "HRM's importance throughout the organization: analyzed through the concepts and views portrayed in eli goldratt's the goal," in IEEE Engineering Management Review, vol. 45, no. 3, pp.

[5] **T. Mihova, K. Angelov and A. Ferdov**, "Challenges to HR Specialists in High-Tech Enterprises," 2019 II International Conference on High Technology for Sustainable Development (HiTech), Sofia, Bulgaria, 2019,

[6] **U. J. Supraveen, S. S. Ali, K. S. Babu, V. R. Rao and G. N. S. Bandi**, "HR Analytics- The Measurement of HR Processes using a Methodical Approach," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp.

[7] **S. R. Akiti, A. Bathini and S. K. Kanagala**, "Random Forest based Fake Job Detection," 2024 4th International Conference on Perva sive Computing and Social Networking (ICPCSN), Salem, India, 2024, pp.

[8] **K. Li, M. Zhu, X. Zhang and H. Xia**, "A Gradient-Enhanced Decision Tree and XGBoost-Based Human-Job Matching Model," 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Nanjing, China, 2024, pp.

[9] **R. Bathija, V. Bajaj, C. Megnani, J. Sawara and S. Mirchandani**, "Revolutionizing Recruitment: A Comparative Study Of KNN, Weighted KNN, and SVM- KNN for Resume Screening," 2023 8th International Conference on Communication and Electronics Systems pp.

[10] **L. Gaojun and L. Boxue**, "The Research on Combination Fore casting Model of the Automobile Sales Forecasting System," 2009 International Forum on Computer Science-Technology and Applications, Chongqing, China, 2009, pp.

[11] **A. Jain, P. Gupta, H. K. Saran, D. S. Parmar, J. P. Bhati and D. Rawat**, "Forecasting Future Sales Using Linear Regression Approach," 2024 International Conference on Cybernation and pp.

[12] **Y. Ali and S. Nakti**, "Sales Forecasting: A Comparison of Traditional and Modern Times-Series Forecasting Models on Sales Data with Seasonality," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp.

[13] **P. Ghosh, O. Samanta, T. Goto and S. Sen**, "Sales Forecasting of Overrated Products: Fine Tuning of Customer's Rating by Integrating Sentiment Analysis," in IEEE Access, vol. 12, pp.

[14] **P. Goodwin, K. Dyussekeneva and S. Meeran**, "The use of analogies in forecasting the annual sales of new electronics products," in IMA Journal of Management Mathematics, vol. 24, no. 4, pp.

[15] **C. Aguilar-Palacios, S. Mu˜ noz-Romero and J. L. Rojo-´ Alvarez**, "Forecasting Promotional Sales Within the Neighbourhood," in IEEE Access, vol. 7, pp.

[16] **Z. Dong and H. Hao**, "Vegetable Sales Forecasting and Pricing Strategy Planning Based on ARIMA Algorithm and Linear Program ming Model," 2024 4th International Symposium on Computer Tech nology and Information Science (ISCTIS), Xi'an, China, 2024, pp.

[17] **C. R. Bhat, B. Prabha, C. Donald, S. Sah, H. Patil and F. A**, "SARIMA Techniques for Predictive Resource Provisioning in Cloud Environments," 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), Chennai, India, 2023, pp.

[18] **B. Kumar Jha and S. Pande**, "Time Series Forecasting Model for Supermarket Sales using FB-Prophet," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp.

[19] **Y. Niu**, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering," 2020 International Conference on Big Data Artificial Intelligence Software Engineering (ICBASE), Bangkok, Thailand, 2020, pp.

[20] **B. Richter, E. Mengelkamp and C. Weinhardt**, "Vote for your en ergy: a market mechanism for local energy markets based on the consumers' preferences," 2019 16th International Conference on the European Energy Market (EEM), Ljubljana, Slovenia, 2019, pp.

[21] **W. Fan, X. Song and C. Li**, "Market power evaluation of power retail market based on combination weighting method," 2024 3rd International Conference on Energy, Power and Electrical Technology (ICEPET), Chengdu, China, 2024, pp.

[22] **A. Bismo, S. Putra and Melysa**, "Application of Digital Marketing (social media and email marketing) and its Impact on Customer Engagement in Purchase Intention: a case study at PT. Soltius Indonesia," 2019 International Conference on Information Management and Technology (ICIMTech), Jakarta/Bali, Indonesia, 2019, pp.

[23] **F. H. Erdogan, S. Cetinkaya and E. Dusmez Tek**, "Market liberalization process and market arrangements in Turkey," 2008 5th International Conference on the European Electricity Market, Lisboa, Portugal, 2008, pp.

[24] **O. Gore, P. Spodniak and S. Viljainen**, "Participation of interconnected capacity in balancing markets: Benefits and practical steps," 2016 13th International Conference on the European Energy Market (EEM), Porto, Portugal, 2016, pp.

[25] **H. Huang**, "Research on Customer Segmentation and Marketing Us ing Rubin Index Based K Means Clustering," 2024 International Conference on Data Science and Network Security (ICDSNS), Tip tur, India, 2024, pp.

[26] **N. D. Sugiharto, D. Elbert, J. Arnold, I. S. Edbert and D. Suhartono**, "Mall Customer Clustering Using Gaussian Mixture Model, K-Means, and BIRCH Algorithm," 2023 6th Interna tional Conference on Information and Communications Tech nology (ICOIACT), Yogyakarta, Indonesia, 2023, pp.

[27] **X. Wang et al.**, "Electricity Market Customer Segmentation Based on DBSCAN and k-Means : —A Case on Yunnan Electricity Market," 2020 Asia Energy and Electrical Engineering Symposium (AEEES), Chengdu, China, 2020, pp.

[28] **P. Wu and X. Li**, "Market Style Discrimination via Ensemble Learning," 2022 IEEE 13th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2022, pp.

[29] **R. Bajaj, G. Bathla, A. Gupta, Anurag and L. Pawar**, "Optimized Ensemble Model for Wholesale Market Prediction using Machine Learning," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp.

[30] **P. Martell**, "It's not hype, it's communication: using public relations techniques to improve your technical messages," International Conference on Professional Communication,Communication Across the Sea: North American and European Practices, Guildford, UK, 1990, pp.

[31] **Qiao Mei, Dou Zhijie and Wang Dongping**, "The study on crisis public relations strategies of enterprise," 2010 International Conference on Future Information Technology and Management Engineering, Changzhou, China, 2010, pp.

[32] **L. V. Sharakhina and A. G. Trubnikova**, "Digitalisation of Public Relations Practice as an Issue to Meet Employers' Needs," 2018 IEEE International Conference "Quality Management, Transport and Information Security, Information Technologies" (ITQMIS), St. Petersburg, Russia, 2018, pp.

[33] **H.-T. Ho**, "Success factors for Using New Communication Technologies, Google Meet, in College Student's Graduation Production. Take Kun-Sun University Public Relations and Advertising Department for Example," 2023 International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan), PingTung, Taiwan, 2023, pp.

[34] **T. Volari´c, Z. Tomi´c and H. Ljubi´c**, "Artificial Intelligence Tools for Public Relations Practitioners: An Overview," 2024 IEEE 28th International Conference on Intelligent Engineering Systems (INES), Gammarth, Tunisia, 2024, pp.

[35] **A. Ababneh and Y. Sanjalawe**, "New Text Classification Strategy Based on a Word Embedding and Noise-Words Removal," 2023 24th International Arab Conference on Information Technology (ACIT), Ajman, United Arab Emirates, 2023, pp.

[36] **P. P. Raut and N. N. Patil**, "Classification of controversial news article based on disputant relation by SVM classifier," 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 2015, pp.

[37] **P. Dhanalakshmi, G. A. Kumar, B. S. Satwik, K. Sreeranga, A. T. Sai and G. Jashwanth**, "Sentiment Analysis Using VADER and Logistic Regression Techniques," 2023 International Conference on Intelligent

Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India, 2023, pp.

[38] **R. Jeljeli et al**, "Analyzing the Effect of Artificial Intelligence on Message Personalization and Media Monitoring in the United Arab Emirates-Based Public Relations," 2024 International Conference on Intelligent Computing Communication, Networking and Services (ICCNS), Dubrovnik, Croatia, 2024, pp.

[39] **D. C. Welch and P. D. Hill**, "The role of PR for technology," 2006 IEEE/UT Engineering Management Conference, Austin, TX, USA, 2006, pp.

[40] **P. Subhash**, "IBM HR Analytics Attrition Dataset," Kaggle Dataset, 2019. Available at: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

[41] **V. Ram**, "Walmart Sales Dataset of 45 Stores," Kaggle Dataset, 2021. at: https://www.kaggle.com/datasets/varsharam/walmart-sales-dataset-of-45stores

[42] **G. Dutta**, "Credit Card Marketing," Kaggle Dataset, 2021. Available at: https://www.kaggle.com/datasets/gauravduttakiit/creditcard-marketing

[43] **Kritanjali Jain**, "Amazon Reviews Dataset," Kaggle, 2023. Available: https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews