# Sales Forecasting Using Machine Learning

by

Sadman Sakib Nabil
19101501
Md Tanvir Islam
20101379
Sadman Aziz Muhit
19201070

A project submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
October 2024

# Declaration

It is hereby declared that

1. The project submitted is our own original work while completing degree at Brac University.
2. The project does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The project does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

Sadman Sakib Nabil
19101501

Md Tanvir Islam
20101379

Sadman Aziz Muhit
19201070

# Approval

The project titled "Sales Forecasting Using Machine Learning" submitted by

1. Sadman Sakib Nabil (19101501)
2. Md Tanvir Islam (20101379)
3. Sadman Aziz Muhit (19201070)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on October 20, 2024.

**Examining Committee:**

Supervisor:
(Member)

Dr. Amitabha Chakrabarty, PhD
Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chairperson)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

# Ethics Statement

We, the authors of this thesis(Project) paper, would like to address the ethical issues that were taken into consideration when creating this work. Our work has been carried out strictly in line with the values of honesty and academic integrity. We have taken great care to guarantee that our research is completely unique and devoid of plagiarism. Every source we used has been correctly credited and recognized, and we have made sure that none of the other people's work has been misrepresented or distorted. Our belief is that research ethics hold great significance in the academic community, and we have made every effort to adhere to these principles during the study. Our goal is that our study will show the highest standards of ethical research conduct while simultaneously advancing knowledge in our profession.

# Abstract

In today's aggressive and fast-paced economy, the ability to forecast sales accurately and effectively denotes a proper utilization of the available resources in planning. Typical sales forecasting methods fail quite often to measure the dynamic market environment owing to the fact that they are totally influenced by past data and also expert opinion. Therefore, this research seeks to validation of sales forecast accuracy with respect to the integration of machine learning (ML) in enhancing its capability. Considering available historical sales figures and some social media trends, machine learning techniques are able to provide realistic and satisfactory forecasts. The paper discusses the advantages of machine learning (ML) to the old methods, for instance, quick detection of the emerging trends, dealing with big data, and adaptation to the situation. Some problems, such as data quality and system integration are also considered. Some of these include ensemble methods, neural networks, and regression, and such techniques are used in machine learning. This article discusses how the integration of machine learning (ML) in sales forecasting will help companies in management and decision making leading to better performance compared to competitors.



**Keywords:** Sales forecasting; Machine learning; ARIMA; Facebook Prophet; Regression data; Time series data; Lasso regression; Linear regression; Random Forest regression; Decision tree regression.

# Acknowledgement

First and foremost, we give thanks to the Almighty Allah, without whose intervention our Project could not have been finished. Second, thanks to Supervisor Dr. Amitabha Chakrabarty, PhD, for his thoughtful counsel and assistance during our work. Whenever we needed assistance, they provided it. And lastly, without our parents' unwavering support, it might not be feasible. We are about to graduate thanks to their wonderful support and prayers.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Nowadays, it is critical for organizations to master sales forecasts more than ever before, thanks to the fast-moving and ever-changing business environment. In other words, without sales forecasting a company cannot plan strategically and estimate how to use resources, control stocks, plan expenses, and aim marketing campaigns. These projections were anchored on the classical sales forecasting systems that most the time relied on assessment of historical facts coupled with expert judgments. But many times, they fail to convey the complex and dynamic nature of the modern markets adequately.

Then, there comes machine learning, an innovative technology which has changed many industries by improving the accuracy of the forecast. One of the sub-fields of artificial intelligence, machine learning employs models and algorithms developed through statistics to find relations and trends within large and complex sets of data. For that reason, it is widely applied in sales forecasting. For example, where it is necessary to grasp the multifaceted relations between several elements like seasonability, trends, consumer patterns and economic data and forecast them.

In comparison to the conventional techniques, the incorporation of machine learning in sales forecasting demonstrates several merits. First, it possesses enhanced precision due to the fact that it learns from new information and adapts to changing environments on a constant basis. Machine learning processes allows for the use of many types of information including historical sales data, social media activity, website hits, as well as changes in weather. This holistic approach allows for the creation of better and more precise estimates.

On the one hand, machine learning based prompting of real time forecasting allow organizations access the most current data critical for fast decision-making. In the era of rapid changes in consumer behavior and market dynamics, having accurate and fast estimates gives companies an edge over the competition. Also, such possibilities enable organizations mitigate the risks and exploit the benefits as the rate of change of machine learning model is higher than that of the traditional techniques in detecting abnormalities and developing trends.

Moreover, the affordability of the machine learning solutions extends to all the organizations irrespective of their size. Machine learning models can be tailored for

businesses regardless of their stage of development such as a young firm or a global enterprise within their resource limits. This allows industries to make use of these complex predictive models without having to worry about high costs.

The in-depth analysis of the advantages, possibilities, and methods of machine algorithmic adroitness in projection sales will be focused on this research paper. We will examine the different machine learning architectures that can be employed for the improvement of the accuracy of predictions such as ensemble learning methods, neural networks, and regression techniques. Furthermore, we will analyze the challenges and issues that need to be addressed in the deployment of machine learning solutions particularly data quality, model choice, and the embedding of business processes.

If organizations appreciate and make use machine learning's potential, they might enhance their forecasting skills and explore new innovation and operational efficiency. As we navigate through the intricacies of this issue, it is evident that machine learning is transforming sales forecasting beyond being simply a predictive tool

## 1.1 Motivation

Sales forecasting through machine learning offers significant advantages for the business-to-business landscape in the modern era. Both spatial and temporal sales models can analyze extensive historical data to predict future sales more accurately than conventional methods. This capability not only enables companies to optimize inventory management, thereby lowering costs and avoiding stock shortages or excess stock, but also supports effective strategic planning and resource distribution.

In addition, unlike traditional forecasting, machine learning based forecasting can cope with market dynamics, changes in customers, and other parameters such as seasonality, promotions or economic changes, which is a plus to businesses. The power to choose data driven options increases margins, improves the service offered to customers and also makes lean operations possible. When it comes to sales forecasting, organizations in every industry have no choice but to adapt to machine learning, and enhance their marketing responsiveness – to the extent they want to say in business in the foreseeable future.

## 1.2 Research Problem

The demand for sales projection goes beyond the weather predicting industry. Corporations predict the sales of their services or goods, demographics are predicted by researchers, and even economic growth is predicted by the governments. In terms of physical grocery stores, sales forecasting is even more imperative, as these establishments are required to control the losses help in managing the inventory level to avoid loss by spoilage and ensure that demographic food products are usually available. This way, retailing would project enough stock levels for sales in order to

make profits and improve on customer satisfaction.

It is also easy to see that such traditional approaches to retail sales forecasting are not amenable to automation because most rely on some form of intuitive judgment with little supportive data. The problem is further exacerbated by novel store openings, varying local market characteristics, offering new products, seasonal influences, and unpredictable promotional impacts.

The research has a significant problem in sales forecasting with machine learning that is limited to the selected market. Restructuring accurate sales forecasts centers on addressing contemporary market situations characterized by non-linear and complex relations and external variables. Forecasting processes through the use of traditional techniques have great limitations, particularly when it comes to the dynamic cycle of seasonality, promotions, economies, and shifting consumer tastes and wants. This often results in either over-stocking or worse, stock outs. There is however hope with machine learning which is able to deal with enormous variety and volume of data but there are many more barriers to tackle like how to combine multivariate data especially when these come from multiple sources transactional, social media, economic, etc.

And the data comes out very sparse and volatile, and furthermore the models need to be interpretable and scalable. The research also needs to think about how to detect these hidden relations, how to work with sparse but high impact (such as the launch of new products) or very dynamic data, and do the balancing act of building a complex model and generalizing it so that overfitting does not occur. Most importantly, they target accuracy at the same time when new information comes in for forecasting, which is a adding new practicality to these models because it is very important for end-users to business stakeholders to understand the models. Additionally, advances in modelling approaches, which make ways to better fuse together data, and ways to overcome challenges related to retrieval and synthesis of information in an efficient manner raise the possibility of practical implementation. The study aims to find ways to cope with these issues by the use of hybrid machine learning models, which aims at better data fusion, attributing higher reliability on the models and creating easy to use and understandable models in sectors such as eretailing, e-supply chain and logistics management, and e-commerce.

Our study attempts to create a machine learning model that increases the predictability of unit sales for different products in different retailers. In particular, we aim to develop a model that, using the historical data accessible up to the training dataset's expiration date, can predict sales for the days ahead. This strategy aims to improve consumer satisfaction and profitability by assisting merchants in maintaining the appropriate product amounts at the appropriate times.

## 1.3 Research Objectives

The primary aim of our work is to design, implement and validate a forecasting model based on machine learning for predicting the sales of a product. The purpose of the research for use machine learning for sales forecasting is to overcome the shortcomings of existing approaches and apply more sophisticated ML techniques for improving forecasting performance, flexibility, and applicability in business.

The objectives of this research are:

- Develop Accurate Sales Forecasting Models: By taking into account past sales data, market trends, and outside variables, machine learning techniques may be used to develop models that accurately anticipate future sales.
- Identify Key Sales Drivers: Examine and measure the effects of the many variables that affect sales, including economic indicators, seasonal trends, promotions, and consumer behavior.
- Enhance Demand Planning: By offering dependable sales projections, assisting companies in optimizing inventory levels, lowering stockouts, and minimizing overstock scenarios, you may boost the effectiveness of demand planning procedures.
- Support Strategic Decision-Making: By offering useful insights from sales predictions, you may facilitate data-driven decision-making for marketing, pricing, and product development initiatives.
- Automate Forecasting Processes: Create automated systems that will enhance and update sales predictions continuously, decreasing the need for human interaction and boosting update frequency.

Attaining these goals will contribute to the enhancement of the domain of sales forecasting enabling organizations to have more accurate predictions, better operational performance and better decision making by providing more trustworthy, explainable, and scalable machine learning sales forecasts.

# Chapter 2

# Related Work

The last few years have witnessed development of several machine learning-based smart technologies that serve as new sales forecasting approaches. The purpose of these technologies is to enhance the overall sales forecast accuracy and thereby sales made by organizations in the modern market trends which are quite developed and competitive. Before the inclusion of sophisticated systems like machine learning, the ARIMA technique, exponential smoothing methods or even linear regression were the most used approaches in estimating the sales during a certain period. The approaches work well if the trends are purely linear or with simple seasonal variations however it is very difficult for such approaches which are linear in nature to handle such modern sales data which is often non static, complex and multivariate. These models have their disadvantages too in that they do not consider the effects of particular activities outside the sales forecasting such as advertising and promotions, seasonal holidays, economic downturn or upturn or changes in the purchasing patterns of consumers within a short period of time. Therefore, these models are not suitable for a fast-paced business operation and within a high levels of data information.

Machine learning (ML) machine learning techniques presents a better option as it allows models to process a rich set of features in historical sales databases and find patterns which other techniques cannot. Supervised learning methods, which include Random Forest, Gradient Boosting Machines (GBMs) and support vector machines (SVMs), have increasingly been used to enhance the accuracy of sales forecasts. These techniques are fitted to large and complex datasets by taking into account more variables like some direct or indirect marketing variables, product variables, the presence of competitors, and their activities or even the level of activity in social networks regarding a given product. For instance, as shown in Chen et al. [1], it was demonstrated that for online retail sales, competition's sales promotions as external factors improved forecasting accuracy using Gradient Boosting Machines in comparison to ARIMA which did not use these predictors.

Besides conventional approaches of machine learning, deep learning methods like Recurrent Neural Network (RNNs) and Long SHort Term-Memory (LSTMs) have also become popular in the recent past. These models are well-tailored for modeling time-series data, hence suitable for sales prediction tasks. Zhang et al. [2] proved that with the use of LSTMs, long-term characteristics and the non-linearity of sales

data may be derived effectively, which is impossible in most scenarios with conventional methods. This means that LSTMs allow modelling of long-term trends – sharp increases or decreases in sales which can occur out of certain external stimuli, and seasonality, as well. In the same breath, another deep learning application that has performed well involves the use of LSTMs with CNNs to model sales data in all its complexities, thereby enhancing accuracy of predictions.

Sales forecasts using machine learning models are not only simple but can include any data available. This is in contrast with most conventional sales forecasting techniques which solely rely on the past sales records. However, instead of relying solely on past records, such mechanisms can leverage other attractive features such as the weather, prices of competing products, economic variables, customer ratings, or even search volumes on Google. For instance, according to Ferreira et al. [3], better predictions were obtained by employing multivariate LSTM networks that used historical sales figures along with social media sentiment data, information on competitors, and economic conditions. Such opportunities are what makes even the modeling of sales patterns using artificial intelligence systems more ingenious as it enables many facets of the sales environment to be captured.

Nonetheless, there are a few drawbacks that machine learning based sales forecasting applications contemporary sales-related problems have to face, need for instance, more frequently in case of such Datasets, which can be rare or very erratic. When it comes to newly launched products or new countries or regions without sufficient historical data, conventional methods perform poorly, the reason being the inadequacy of training data. This carves a niche for advancements in research in solution these problems. Wei et al. [4] demonstrated the effectiveness of transfer learning in new product sales forecasting and claimed that their technique could be applied even where data is extremely limited. Also, that helps to deal with the turbulence more effectively – bagging and boosting – with the aid of these techniques, many models can be aggregated together in order to mitigate the effects of noise and outliers in the data.

Incredible as it may seem to some readers, the high accuracy of many machine learning models in particular, deep learning - is one of the serious challenges addressed to its potential users. In most cases, business managers want to understand the reason behind a certain forecast, which is quite impossible to achieve when highly complex models' solutions such as neural networks are used. To that end, those elusive explainable AI models s why researchers have come up with explanatory tools such as SHAP (Shapley Additive explanations) and LIME (Local Interpretable Modelagnostic Explanations). These assist in clarifying how a specific prediction comes about by clarifying the role each feature plays making the machine learning models clearer and more implementable within the business. For example, Katuwal et al. [5] explained a sales forecasting model based on XGBoost, which was built using SHAP values and provided end-users with information on what was inducing the sales predictions, enabling them to take appropriate actions based on the data.

Real-time forecasting capabilities is another important aspect in sales forecasting. The faster commercialization of products as well as services, particularly in

e-business, enjoins firms to have forecasting systems that are able to provide almost instantaneous forecasts after the acquisition and integration of new information. Previous batch processing paradigms are frequently too sluggish, given the typical environmental dynamics, to address sharp increases in demand or other intervening factors. In recent years confining the machine learning forecasting models within the barrier for cloud deployment has been studied in the context of enabling forecasting within real time limits. Meireless et al. [6] showcased how businesses can improve on the efficiency of sales forecasting using a cloud-based system that integrated an ensemble of machine learning models. Thanks to the cloud infrastructure, the system was able to recalibrate the forecasts when new sales data was available. This kind of technology is vital in fields such as marketing, where demand is unpredictable owing to sales, holidays or other market forces.

Most of the applicable practices and literature in ML for sales forecasting concern retailing and ecommerce promising extensive improvements in performance. There are still divisive issues, which require research solutions, in the area. Most of the studies have large scale retail and ecommerce databases, so it is not clear how well these models work for smaller businesses or less data rich industries. And, while being able to incorporate a great deal of external influence beyond the dataset such as sales promotions or social media e.g., sentiment analysis, there is still a lack of common practices to do so in every industry without fail. Furthermore, it is true that more interpretable algorithms have become more available with the advent of tools such as SHAP and LIME, growing the extent of model usage in practice, however, for commercial use especially for deep learning models, it is paramount to have simpler and less technical ways of explanations which the business audiences will relate.

## 2.1 Related Work

The topic of demand and sales forecasting has been greatly affected by recent developments in machine learning (ML). This review of the literature compiles several studies that have used various machine learning approaches to improve forecasting efficiency and accuracy in a variety of industries, such as general goods, retail, and the food business. Forecasting sales is crucial for supply chain and production management. Planning, strategy, marketing, logistics, warehousing, and resource management are all impacted by it for businesses.

In addition to previous behaviour and trends, causal forecasting techniques may also anticipate future sales behaviour based on correlations between factors. Outlines a framework for applying genetic programming, an artificial intelligence method based on the theory of biological evolution, to estimate and simulate export sales. An export sales forecasting model is proposed after an empirical case study of an export firm is examined. Additionally, a six-week sales prediction is created and its results are compared to actual sales data. Lastly, a causal forecasting model variable sensitivity analysis is shown.

The performance of nine cutting-edge machine learning and three traditional fore-

casting algorithms was objectively examined for the first time in horticulture sales projections by the author of [8]. In every trial, they demonstrated the superiority of machine learning techniques, with the gradient boosted ensemble learner XGBoost emerging as the best performer in 14 of the 15 comparisons. When dealing with datasets that have many seasons, this benefit above traditional forecasting techniques grew. They also demonstrated how adding meta-features and other external variables, such holiday and weather information, improved prediction performance. Furthermore, we looked at whether the algorithms could have predicted the sharp spike in horticulture product demand that occurred in 2020 during the SARS-CoV-2 epidemic. Moreover, XGBoost worked better in this particular situation.

In recent years, there has been an increase in interest in the subject of neural network for prediction. In order to predict furniture sales, a public dataset that contains a retail store's sales history is examined in [9]. Several forecasting models are used to achieve this goal. Initially, a few traditional time-series forecasting methods are applied, including Triple Exponential Smoothing (TES) and Seasonal Autoregressive Integrated Moving Average (SARIMA). Following this, more advanced methods are utilized such as CNN, Prophet, and LSTM networks. Among the given models, several accuracy measuring techniques such as Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) are employed to compare their performances. The results show an edge of Stacked LSTM approach over the rest of the methods.

In an effort to analyze and forecast furniture sales, a crossectional analysis of a retail store in [10] is used which is a publicly available sales history dataset. Several forecasting models are used to achieve this goal. In the beginning, a number of conventional time series forecasting models are employed. For instance, Triple Exponential Smoothing and Seasonal Autoregressive Integrated Moving Average (SARIMA). More advanced techniques there after include: Convolutional Neural Networks (CNN), Prophet and Long Short-Term Memory (LSTM). Among the performance evaluating parameters employed in model comparison include Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and others. The outcomes demonstrate the Stacked LSTM method's superiority over the alternative approaches. Furthermore, the outcomes demonstrate the Prophet and CNN models' strong performances.

Business intelligence is the process that decision makers utilize to choose the best course of action by processing, analyzing, and interpreting the data to acquire insights. Building a model that will provide a forecast based on adjusted sales figures utilizing certain customer-centric qualities was the primary goal of [11]. They divide the consumer base into discrete groups according to their purchasing habits using the RFM model first, then eliminate the groups that are not important to the company. After that, they anticipate sales for the remaining data using the ARIMA model. When applied following RFM analysis, they were able to obtain a greater accuracy and better fit of the data for the prediction model.

In [12], the goal was to identify the optimal algorithm for their specific problem statement while attempting to forecast a retail store's sales using various machine learning

approaches. They have used both boosting and normal regression approaches in our approach, and they have discovered that the boosting algorithms produce superior outcomes than the regular regression algorithms.

In [13], the sales pattern of Power Bumi products during the COVID-19 pandemic was analyzed, and the forecasting approach that produces the minimum error value in Power Bumi product sales forecasting was compared to PT. Zamrud Bumi Indonesia. Two approaches are used in this study: the least square trend model and exponential smoothing. to use MAD, MSE, and MAPE to determine the error rate. In comparison to other forecasting techniques, the exponential smoothing alpha 0.9 approach has the lowest error value, according to the data. With a forecast of 627.628 boxes, the MAD value is 130.329, the MSE is 28251.23, and the MAPE is 22.00% when it comes to product sales forecasting.

In today's corporate world, managing the retail supply chain is essential to success. Supply chain management greatly depends on the ability to forecast consumer demand. The ability to foresee every detail accurately affects sales volume, customer attractiveness, profit margin, storage, and lost profits. [14] will develop a novel technique that makes use of machine learning to aid in precise prediction. This technique gathers and analyzes a store's historical data. obtaining the pertinent data, processing it, and becoming ready to use it in the appropriate way. utilizing relevant algorithms on the processed data. We are aware that several algorithms have been utilized recently for prediction in K-Nearest Neighbor, Support Vector Machine, Gaussian Nave Bayes, Random Forest, Decision Tree Classifier, and regressions. We get actual data from the marketplace. This document was created using the store position, the month, the event for that month, and additional relevant data. The geographic location of their nation affects forecasting, as our research explains. Their model generates a preliminary demand for a certain good. This estimation benefits retails and their enterprises.

For computer-retailing sales forecasting, a clustering-based forecasting model incorporating clustering and machine-learning techniques is presented in [15]. The suggested approach divided the training data into groups by first using the clustering technique, which groups data with comparable traits or patterns. The forecasting models of each group are then trained using machine-learning techniques. The trained forecasting model of the cluster was used for sales forecasting once the cluster whose data patterns were most comparable to the test data was identified. Given that computer merchants' sales data exhibits comparable data patterns or features throughout a range of time periods, applying the suggested clustering-based forecasting model can improve prediction accuracy. In this work, two machine learning approaches, support vector regression (SVR) and extreme learning machine (ELM), and three clustering techniques, selforganizing map (SOM), growing hierarchical self-organizing map (GHSOM), and K-means, are employed. There were six forecasting models based on clustering that were put out. The empirical examples include actual sales statistics for laptops, liquid crystal displays, and personal computers.

Businesses may more effectively manage cash flow, production, and create more informed company plans by using sales forecasting. They provide a machine learning-

based, accurate, and efficient sales forecasting model in [16]. To extract characteristics from historical sales data, feature engineering is first carried out. They also employed eXtreme Gradient Boosting (XGBoost) to make use of these traits in order to predict the quantity of sales in the future. Our suggested model perform very well for sales prediction with fewer computational time and memory resources, as demonstrated by the experiment results on a publicly available Walmart retail items dataset provided by the Kaggle competition.

[18] assesses and contrasts a number of machine learning models, including ARIMA, XGBoost, SVM, Auto Regressive Neural Network (ARNN), and Hybrid Rossmann, a drugstore corporation, uses models such as Hybrid ARIMA-ARNN, Hybrid ARIMAXGBoost, Hybrid ARIMA-SVM, and STL Decomposition (using ARIMA, Naive, XGBoost) to anticipate sales. The training data set includes additional information on medicine retailers as well as historical sales data. Metrics like RMSE and MAE are used to gauge how accurate these models are. Originally, sales forecasts were made using linear models like ARIMA. Since ARIMA was unable to accurately capture nonlinear patterns, nonlinear models like SVM, XGBoost, and neural networks were employed. Nonlinear models yielded lower RMSE and outperformed ARIMA. Composite models were created utilizing a combination of decomposition and hybrid techniques in order to further enhance performance. The three hybrid models—Hybrid ARIMA-ARNN, Hybrid ARIMA-XGBoost, and Hybrid ARIMA-SVM—all outperformed the corresponding single models. Subsequently, Snaive, ARIMA, and XGBoost were used to forecast the seasonal, trend, and residual components of the decomposed model, which was then built using STL Decomposition. Compared to hybrid and individual models, STL produced superior outcomes.

In a business setting, the supermarket sales prediction contributes to increased sales. The method aids in problem-domain decision-making. There are several forecasting tools available, including the logistic exponential model and regression model. The most recent technology to demonstrate better performance in terms of forecast accuracy is Facebook's (FB) Prophet. A Facebook Prophet tool has been presented in [14] study to anticipate supermarket sales using data. A few forecasting models, including the additive model, the Autoregressive integrated moving average (ARIMA) model, and the Facebook Prophet model, have been studied in the proposed research project. Based on the suggested study, it can be said that FB Prophet is a more accurate prediction model in terms of fitting, low error, and better prediction.

For the purpose of sales volume prediction, a new model which comprises mainly GRU and Prophet models with attention mechanism is proposed in [21]. In this hybrid model, the Prophet model captured the linear characteristics of the data and the GRU model with attention mechanism captured the nonlinear characteristics. It was found in the experiments that the composite model had the best modeling accuracy among all the models: recurrent neural networks, long short-term memory, gate recurrent units, Prophet models, and autoregressive integrated moving averages considered. This hybrid approach developed in this study will assist organizations to be more aggressive in their operability in smart manufacturing centrals, while also being responsive to changes in consumers' buying habits.

Table 2.1: Related Work

| **Reference** | **Algorithms Used** | **Best Accuracy** | **Main Focus** | **Contribution** |
|---|---|---|---|---|
| [1] | Gradient Boosting Machines (GBMs), ARIMA | GBMs outperformed ARIMA | Forecasting online retail sales | Demonstrated GBMs outperformed traditional ARIMA by incorporating external variables like price changes and promotions |
| [2] | Long Short-Term Memory (LSTM) networks | LSTMs outperformed traditional models | Capturing long-term dependencies in sales data | LSTMs were able to model long-term trends and seasonality more effectively than traditional models |
| [3] | Multivariate LSTM models | LSTM improved accuracy | Integrating social media sentiment, competitor activity, and economic trends | Demonstrated the effectiveness of LSTM models in integrating diverse data sources for better forecasting accuracy |
| [4] | Transfer Learning | Improved prediction accuracy | Handling sparse data for new products | Applied transfer learning to improve sales predictions for products with little historical data |
| [5] | XGBoost, SHAP | XGBoost achieved high accuracy | Explaining machine learning forecasts | Used SHAP values to enhance interpretability of the XGBoost model for sales forecasting |
| [6] | Ensemble of ML models | Improved forecast accuracy | Real-time forecasting system | Developed a real-time sales forecasting system using ensemble models deployed in the cloud |

| | | | | |
|---|---|---|---|---|
| [8] | XGBoost | Best performer in 14 out of 15 comparisons | Forecasting horticulture sales | Demonstrated XGBoost's superiority, especially with seasonal datasets, and improved accuracy with meta-features like holidays and weather |
| [9] | Triple Exponential Smoothing, SARIMA, CNN, Prophet, LSTM | Stacked LSTM showed superior performance | Furniture sales prediction | Compared multiple forecasting models, with Stacked LSTM outperforming others in accuracy |
| [10] | Triple Exponential Smoothing, SARIMA, CNN,prophet, LSTM | Stacked LSTM and CNN performed well | Sales history forecasting for retail | Highlighted the strong performance of Prophet and CNN models in sales forecasting |
| [11] | ARIMA, RFM model | Greater accuracy after RFM analysis | Customer-centric sales forecasting | Combined RFM analysis with ARIMA to better fit the data and achieve higher prediction accuracy |
| [12] | Boosting algorithms, Regression algorithms | Boosting algorithms produced better results | Forecasting retail store sales | Boosting algorithms provided superior results compared to regular regression approaches |
| [13] | Least square trend model, Exponential smoothing | Exponential smoothing alpha 0.9 had lowest error | Sales forecasting for Power Bumi products during COVID-19 | Showed that exponential smoothing with alpha 0.9 provided the most accurate sales forecast |
| [14] | K-Nearest Neighbor, SVM, Gaussian Naive Bayes, Random Forest, Decision Tree, Regressions | Accurate demand prediction | Retail supply chain forecasting | Used machine learning algorithms to precisely forecast retail supply chain demand |
| [15] | Support Vector Regression (SVR), Extreme Learning Machine (ELM), SOM, GHSOM, K-means | Clustering models improved accuracy | Clustering-based computer-retailing sales forecasting | Proposed a clustering-based model to group similar sales data patterns and improve prediction accuracy |

| | | | | |
|---|---|---|---|---|
| [16] | XGBoost | High accuracy with low computational time | Walmart retail sales prediction | Developed an XGBoost-based sales forecasting model that efficiently used historical sales data with less computational cost |
| [18] | ARIMA, XGBoost, SVM, Auto Regressive Neural Network (ARNN) | Hybrid models outperformed individual models | Hybrid models for retail sales forecasting | Showed hybrid models (ARIMA with XGBoost, SVM, ARNN) outperformed single models in capturing non-linear patterns |
| [21] | GRU, AiProphet (Composite Model) | Best prediction accuracy | Sales volume forecasting in smart manufacturing | Developed a composite GRU, AiProphet model with attention mechanism, outperforming traditional models in prediction accuracy |

To sum up, advancements in sales forecasting mainly thanks to machine learning makes it possible to make predictions more accurately, work with complex data sources and even make predictions on the fly. Random Forests, GBDT, LSTMs, and other deep learning architectures have revolutionized the landscape of predictive modeling, outperforming their predecessors.Nevertheless, there are still limitations such as sparsity and volatility of data along with interpretability and scalability of models. However, it is evident that they will have to be resolved for machine learning in sales forecasting to find effective use in more industries. Doing so, it will enable companies to make better decisions and enhance their operational efficiency.

# Chapter 3

# Dataset

## 3.1 Data Collection

For any research, data collection is the most integral part. This investigation relied on the completely unedited raw dataset. For this project we used an open-source dataset from kaggle [30]. This dataset contains the transaction details of a retail store that had recorded sales from the period 2015 to 2019. This is a time series dataset with 9800 records and 18 fields. It is a dataset meant for regression testing.

| Row ID | Order ID | Order Date | Ship Date | Ship Mode | Customer ID | Customer Na | Segment | Country | City | State | Postal Code | Region | Product ID | Category | Sub-Categor | Product Nam | Sales |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CA-2017-152 | 8/11/17 | 11/11/17 | Second Class | CG-12520 | Claire Gute | Consumer | United State | Henderson | Kentucky | 42420 | South | FUR-BO-100 | Furniture | Bookcases | Bush Somers | 261.96 |
| 2 | CA-2017-152 | 8/11/17 | 11/11/17 | Second Class | CG-12520 | Claire Gute | Consumer | United State | Henderson | Kentucky | 42420 | South | FUR-CH-100 | Furniture | Chairs | Hon Deluxe F | 731.94 |
| 3 | CA-2017-138 | 12/6/17 | 16/6/17 | Second Class | DV-13045 | Darrin Van H | Corporate | United State | Los Angeles | California | 90036 | West | OFF-LA-100 | Office Suppli | Labels | Self-Adhesiv | 14.62 |
| 4 | US-2016-108 | 11/10/16 | 18/10/16 | Standard Cla | SO-20335 | Sean O'Donn | Consumer | United State | Fort Lauderd | Florida | 33311 | South | FUR-TA-100 | Furniture | Tables | Bretford CR4 | 957.5775 |
| 5 | US-2016-108 | 11/10/16 | 18/10/16 | Standard Cla | SO-20335 | Sean O'Donn | Consumer | United State | Fort Lauderd | Florida | 33311 | South | OFF-ST-100 | Office Suppli | Storage | Eldon Fold 'N | 22.368 |
| 6 | CA-2015-115 | 9/6/15 | 14/6/15 | Standard Cla | BH-11710 | Brosina Hoff | Consumer | United State | Los Angeles | California | 90032 | West | FUR-FU-100 | Furniture | Furnishings | Eldon Expres | 48.86 |
| 7 | CA-2015-115 | 9/6/15 | 14/6/15 | Standard Cla | BH-11710 | Brosina Hoff | Consumer | United State | Los Angeles | California | 90032 | West | OFF-AR-100 | Office Suppli | Art | Newell 322 | 7.28 |
| 8 | CA-2015-115 | 9/6/15 | 14/6/15 | Standard Cla | BH-11710 | Brosina Hoff | Consumer | United State | Los Angeles | California | 90032 | West | TEC-PH-100 | Technology | Phones | Mitel 5320 IP | 907.152 |
| 9 | CA-2015-115 | 9/6/15 | 14/6/15 | Standard Cla | BH-11710 | Brosina Hoff | Consumer | United State | Los Angeles | California | 90032 | West | OFF-BI-1000 | Office Suppli | Binders | DXL Angle-Vi | 18.504 |
| 10 | CA-2015-115 | 9/6/15 | 14/6/15 | Standard Cla | BH-11710 | Brosina Hoff | Consumer | United State | Los Angeles | California | 90032 | West | OFF-AP-100 | Office Suppli | Appliances | Belkin F5C20 | 114.9 |
| 11 | CA-2015-115 | 9/6/15 | 14/6/15 | Standard Cla | BH-11710 | Brosina Hoff | Consumer | United State | Los Angeles | California | 90032 | West | FUR-TA-100 | Furniture | Tables | Chromcraft F | 1706.184 |
| 12 | CA-2015-115 | 9/6/15 | 14/6/15 | Standard Cla | BH-11710 | Brosina Hoff | Consumer | United State | Los Angeles | California | 90032 | West | TEC-PH-100 | Technology | Phones | Konftel 250 ( | 911.424 |
| 13 | CA-2018-114 | 15/4/18 | 20/4/18 | Standard Cla | AA-10480 | Andrew Alle | Consumer | United State | Concord | North Carolir | 28027 | South | OFF-PA-100 | Office Suppli | Paper | Xerox 1967 | 15.552 |
| 14 | CA-2017-161 | 5/12/17 | 10/12/17 | Standard Cla | IM-15070 | Irene Maddo | Consumer | United State | Seattle | Washington | 98103 | West | OFF-BI-1000 | Office Suppli | Binders | Fellowes PB2 | 407.976 |
| 15 | US-2016-118 | 22/11/16 | 26/11/16 | Standard Cla | HP-14815 | Harold Pawla | Home Office | United State | Fort Worth | Texas | 76106 | Central | OFF-AP-100 | Office Suppli | Appliances | Holmes Repl | 68.81 |
| 16 | US-2016-118 | 22/11/16 | 26/11/16 | Standard Cla | HP-14815 | Harold Pawla | Home Office | United State | Fort Worth | Texas | 76106 | Central | OFF-BI-1000 | Office Suppli | Binders | Storex DuraT | 2.544 |
| 17 | CA-2015-105 | 11/11/15 | 18/11/15 | Standard Cla | PK-19075 | Pete Kriz | Consumer | United State | Madison | Wisconsin | 53711 | Central | OFF-ST-100 | Office Suppli | Storage | Stur-D-Stor S | 665.88 |
| 18 | CA-2015-167 | 13/5/15 | 15/5/15 | Second Class | AG-10270 | Alejandro Gr | Consumer | United State | West Jordan | Utah | 84084 | West | OFF-ST-100 | Office Suppli | Storage | Fellowes Sup | 55.5 |
| 19 | CA-2015-143 | 27/8/15 | 1/9/15 | Second Class | ZD-21925 | Zuschuss Do | Consumer | United State | San Francisc | California | 94109 | West | OFF-AR-100 | Office Suppli | Art | Newell 341 | 8.56 |
| 20 | CA-2015-143 | 27/8/15 | 1/9/15 | Second Class | ZD-21925 | Zuschuss Do | Consumer | United State | San Francisc | California | 94109 | West | TEC-PH-100 | Technology | Phones | Cisco SPA 50 | 213.48 |
| 21 | CA-2015-143 | 27/8/15 | 1/9/15 | Second Class | ZD-21925 | Zuschuss Do | Consumer | United State | San Francisc | California | 94109 | West | OFF-BI-1000 | Office Suppli | Binders | Wilson Jones | 22.72 |
| 22 | CA-2017-137 | 9/12/17 | 13/12/17 | Standard Cla | KB-16585 | Ken Black | Corporate | United State | Fremont | Nebraska | 68025 | Central | OFF-AR-100 | Office Suppli | Art | Newell 318 | 19.46 |
| 23 | CA-2017-137 | 9/12/17 | 13/12/17 | Standard Cla | KB-16585 | Ken Black | Corporate | United State | Fremont | Nebraska | 68025 | Central | OFF-AP-100 | Office Suppli | Appliances | Acco Six-Out | 60.34 |
| 24 | US-2018-156 | 16/7/18 | 18/7/18 | Second Class | SF-20065 | Sandra Flan | Consumer | United State | Philadelphia | Pennsylvania | 19140 | East | FUR-CH-100 | Furniture | Chairs | Global Delux | 71.372 |
| 25 | CA-2016-106 | 25/9/16 | 30/9/16 | Standard Cla | EB-13870 | Emily Burns | Consumer | United State | Orem | Utah | 84057 | West | FUR-TA-100 | Furniture | Tables | Bretford CR4 | 1044.63 |
| 26 | CA-2017-121 | 16/1/17 | 20/1/17 | Second Class | EH-13945 | Eric Hoffmar | Consumer | United State | Los Angeles | California | 90049 | West | OFF-BI-1000 | Office Suppli | Binders | Wilson Jones | 11.648 |
| 27 | CA-2017-121 | 16/1/17 | 20/1/17 | Second Class | EH-13945 | Eric Hoffmar | Consumer | United State | Los Angeles | California | 90049 | West | TEC-AC-1000 | Technology | Accessories | Imation-18G | 90.57 |
| 28 | US-2016-150 | 17/9/16 | 21/9/16 | Standard Cla | TB-21520 | Tracy Blums | Consumer | United State | Philadelphia | Pennsylvania | 19140 | East | FUR-BO-100 | Furniture | Bookcases | Riverside Pal | 3083.43 |
| 29 | US-2016-150 | 17/9/16 | 21/9/16 | Standard Cla | TB-21520 | Tracy Blums | Consumer | United State | Philadelphia | Pennsylvania | 19140 | East | OFF-BI-1000 | Office Suppli | Binders | Avery Recycl | 9.618 |
| 30 | US-2016-150 | 17/9/16 | 21/9/16 | Standard Cla | TB-21520 | Tracy Blums | Consumer | United State | Philadelphia | Pennsylvania | 19140 | East | FUR-FU-100 | Furniture | Furnishings | Howard Mill | 124.2 |

Figure 3.1: Sample Dataset

## 3.2 Description

The set of records consists of the sales database of a "Superstore" and contains extensive particulars about every sale, customer, products and sales amounts. Also, below is how the major columns in the data set are arranged and what their meaning is:

- **Row ID:** An individual number given to each row of the datasheet or transaction.
- **Order ID:** An individual number given to each order.
- **Order Date:** The day when order was placed.
- **Ship Date:** The day when the order was sent out.
- **Ship Mode:** The type shipping used for that particular order (for instance, Second Class, Standard Class).
- **Customer ID:** An individual number given to every customer.
- **Customer Name:** The name of the client.
- **Segment:** The group of customers (Consumer, Corporate, Home Office).
- **Country:** The nation in which the order was placed (all of the entries are with in the United States).
- **City:** The city where the customer is located in.
- **State:** The state where the customer is located in.
- **Postal Code:** The customer's postal code.
- **Region:** A Division of the United States that has some broad regional boundaries e.g South, West.
- **Product ID:** A unique number assigned to each product.
- **Category:** The type under which the product falls e.g. Furniture, Office Supplies.
- **Sub-Category:** A section within the main product category, for example, Types of products that would be referred to as Chairs, Tables.
- **Product Name:** The name of the product that the store is selling.
- **Sales:** Billings or revenue from the transaction.

Preliminary Understandings:

1. Sales Distribution: The dataset contains the sales figures for every product sold, which can thus be used for the analysis of the sales patterns, trend of the sales

of various products, and dispersion of sales in diversity of product categories and segments in addition and also geography.

2. Time-based Analysis: The dataset has both Order Date and Ship Date so it can also be analyzed on a time basis for trends system, that is sales increase by month or even year and transport delays and how time affects the sale of the given products among others.

3. Behavior of the Customers: The dataset comprises customer related columns like Customer ID, Customer Name, Segment etc, so it enables different customer buying patterns analysis. The Segment variable enables us to look at sales for Consumer, Corporate and Home Office segments.

4. The dataset contains geographical information such as city, state and region, which makes it possible to evaluate sales performance in different regions and states.

The sales on the basis of per month and per week from 2015-2019 are shown in Figure 3.1 and Figure 3.2.

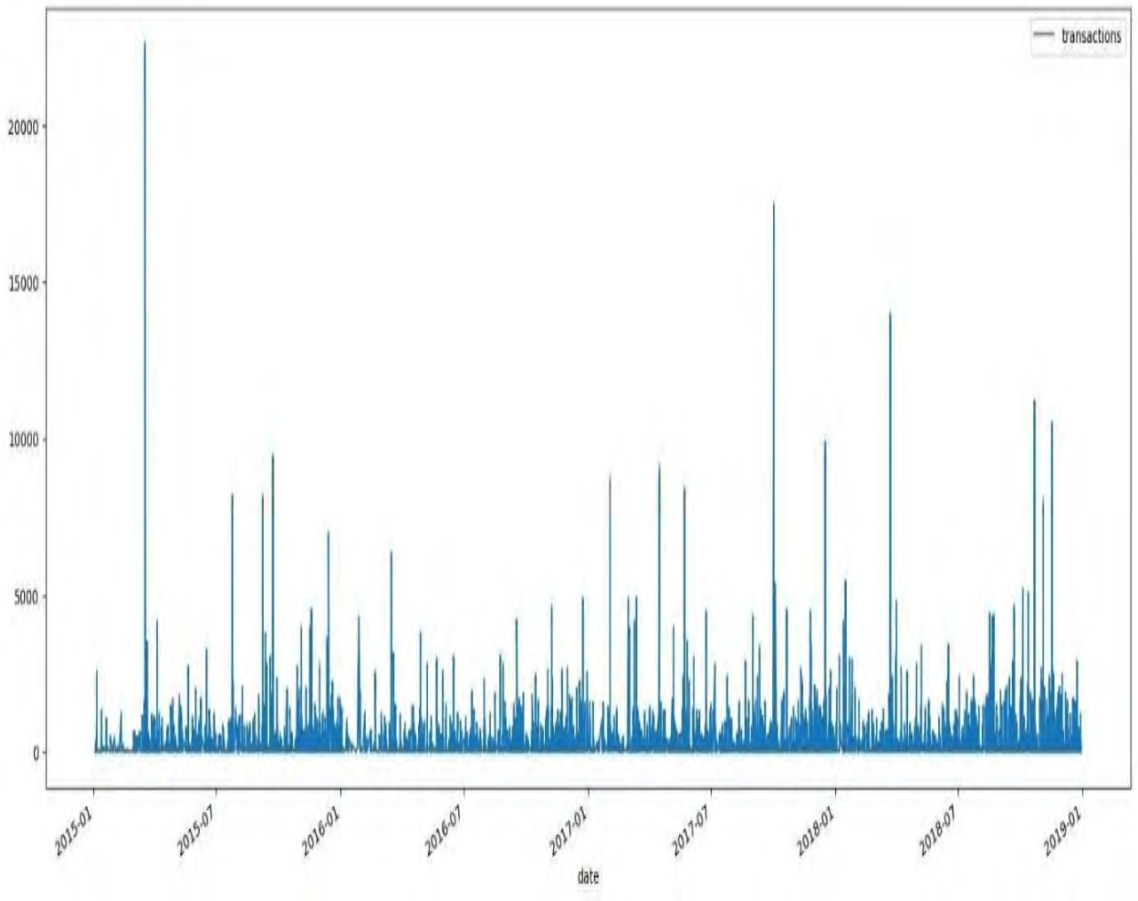


Figure 3.2: Sales from 2015 - 2019

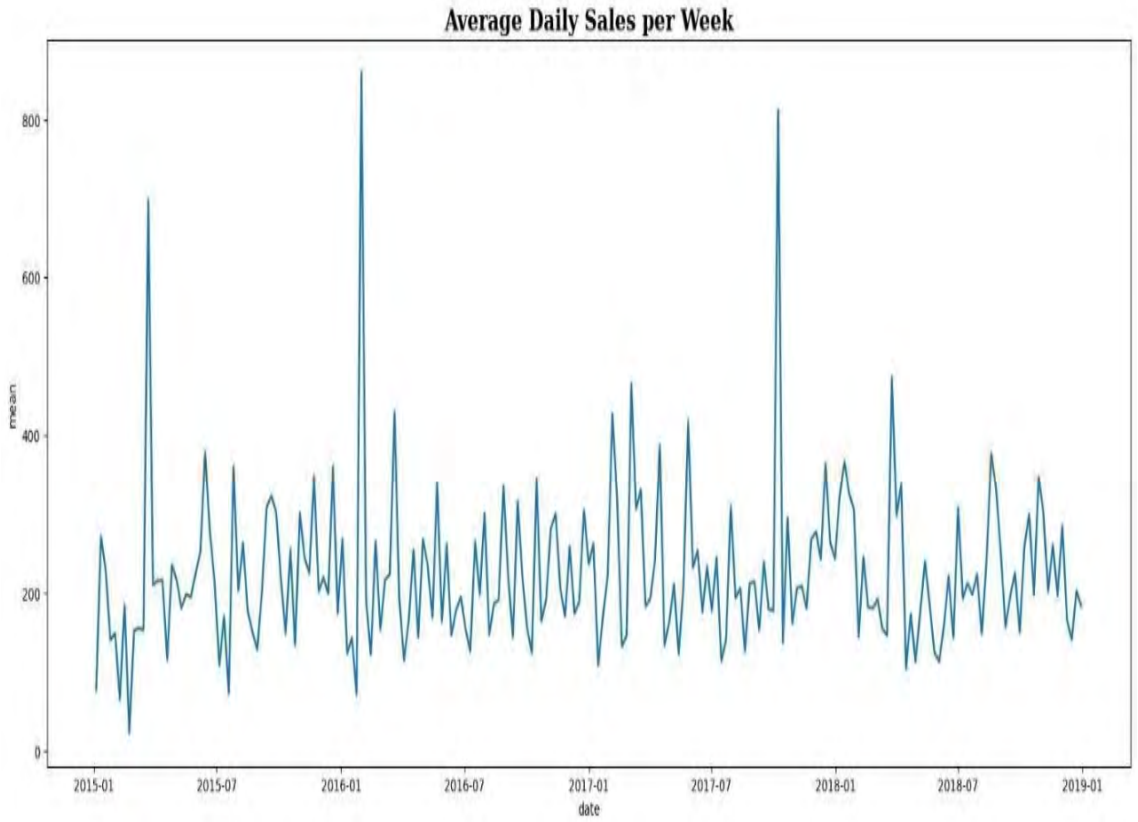


Figure 3.3: Average Daily Sales per Week

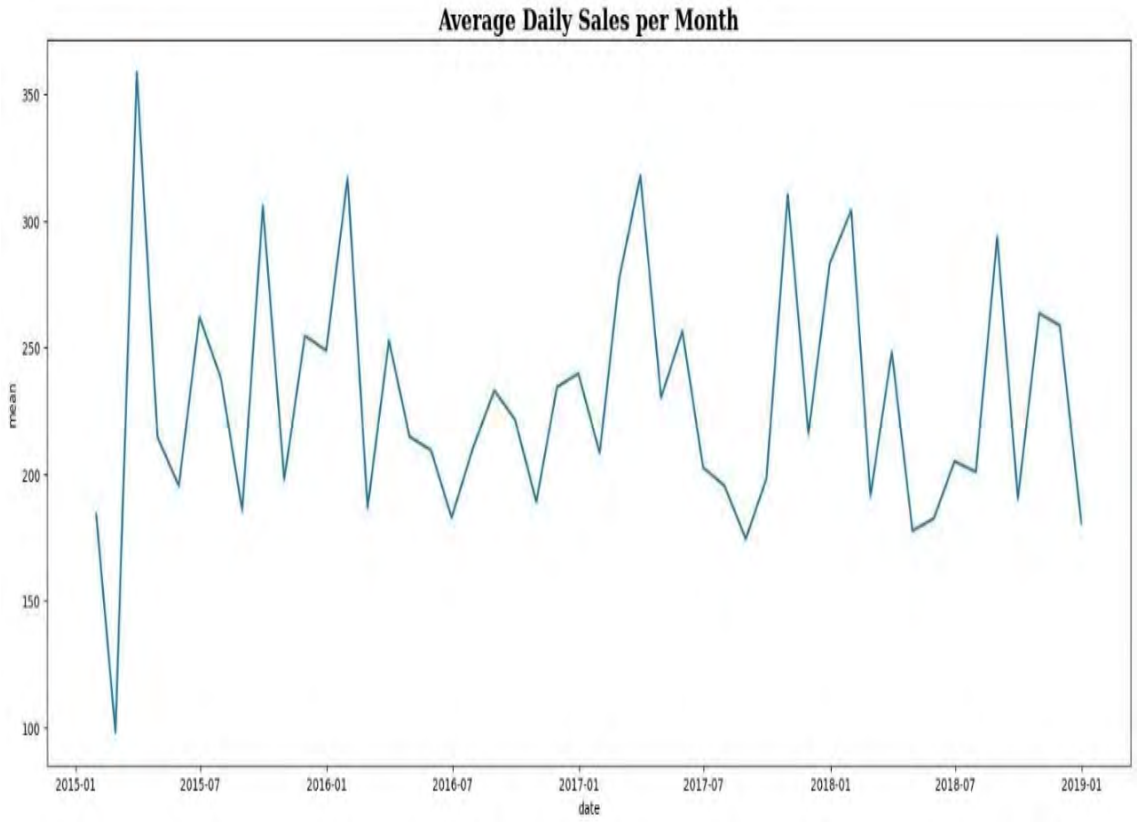


Figure 3.4: Average Daily Sales per Month

## 3.3 Data Preprocessing

Handle Missing Values: Missing values can hamper the training process. So, it is important to remove them. At first, we checked the null values in our dataset. There were some null values. We dropped them using Dropna function.

Encoding: encoding is a technique of convert categorical values into number. It is so important because machine can't read string type data. Most of the column for this dataset were string type. So, we had to encoded them into integer value by using label encoder.

After null value handling and encoding the dataset was ready to fit in the model.

# Chapter 4

# Methodology

## 4.1 Proposed Methodology

In machine learning and statistics, regression algorithms are a kind of supervised learning that are used to model and examine the correlations between variables. Using one or more input factors (independent variables), the objective is to predict a continuous output variable (dependent variable).

### 4.1.1 Linear Regression

One of the most straightforward and popular statistical methods for simulating the connection between a dependent variable and one or more independent variables is linear regression. The main goal is to establish a linear connection between the target (output variable) and the input variables (features).

**Mechanics:**

- Linear regression uses the Ordinary Least Squares (OLS) method to find the best-fitting line by minimizing the sum of the squared differences between actual values and predicted values.
- In multiple linear regression (with multiple features), the model assigns a weight (coefficient) to each feature to quantify its contribution to the target variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{4.1}$$

where:

- $y$ is the dependent variable (target).
- $\beta_0$ is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables.
- $x_1, x_2, \ldots, x_n$ are the independent variables (features).
- $\epsilon$ is the error term.

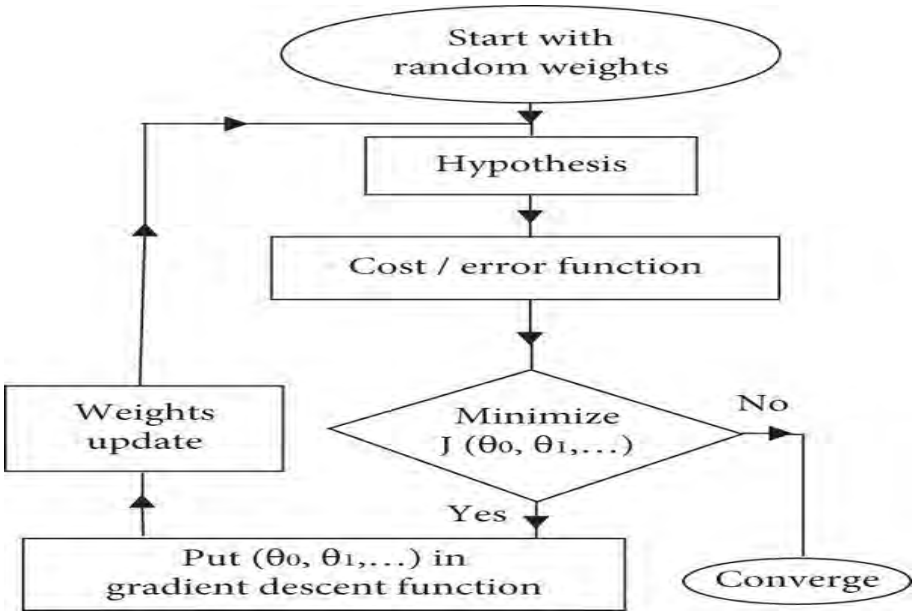


Figure 4.1: Flow chart of Linear Regression model

**Strengths:**

- Easy to interpret: The model coefficients tell you the impact of each feature on the target.
- Computationally efficient: Works well with small to moderately large datasets.
- Works well when the relationship between the features and target is approximately linear.

**Limitations:**

- Assumes linearity: Poor performance when the relationship between features and target is non-linear.
- Sensitive to outliers: Outliers can heavily influence the fitted line and the resulting predictions.
- Assumes homoscedasticity (constant variance of errors) and no multicollinearity (independent features).

### 4.1.2 Lasso Regression:

A regularization term (L1 penalty) is included in Lasso (Least Absolute Shrinkage and Selection Operator) regression, a kind of linear regression, to enforce sparsity in the coefficients, meaning that certain coefficients may be precisely zero. This aids in choosing features.

**Mechanics:**

- Lasso adds a regularization term $\sum_{j=1}^{p} |\beta_j|$ to the cost function.

- By doing so, Lasso tends to reduce the impact of irrelevant or less important features, and some coefficients may shrink to zero, effectively eliminating those features.

$$\text{Loss} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{p} |\beta_j| \tag{4.2}$$

where:

- $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares.
- $\alpha \sum_{j=1}^{p} |\beta_j|$ is the L1 regularization term.
- $\alpha$ is the regularization parameter.

**Strengths:**

- Feature selection: Lasso can automatically select features, making it useful when you have many features.
- Reduces overfitting: The L1 penalty limits the magnitude of coefficients, helping to avoid overfitting in high-dimensional datasets.

**Limitations:**

- Bias: Lasso introduces bias into the model by shrinking coefficients, which can lead to underfitting.
- Struggles with multicollinearity: Lasso may perform poorly if features are highly correlated because it arbitrarily selects one of the correlated features.

### 4.1.3 Elastic Net Regression:

The L1 and L2 penalties of the Lasso and Ridge techniques are combined linearly in Elastic Net, a regularized regression technique. This approach is useful when there are several aspects that are connected with one another, helping to handle multicollinearity better by distributing the coefficient weights.

**Mechanics:**

- Elastic Net's cost function includes both L1 and L2 penalties:

$$\text{Cost Function} = \sum_{i=1}^{n}(y_i - y_i')^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{4.3}$$

where:

- $\sum_{i=1}^{n}(y_i - y_i')^2$ is the residual sum of squares.
- $\lambda_1 \sum_{j=1}^{p} |\beta_j|$ is the L1 regularization term.
- $\lambda_2 \sum_{j=1}^{p} \beta_j^2$ is the L2 regularization term.
- $\lambda_1$ and $\lambda_2$ are the regularization parameters.

**Strengths:**

- Handles multicollinearity better: When features are highly correlated, Elastic Net performs better than Lasso by distributing the coefficient weights.
- Feature selection and regularization: Balances the benefits of Lasso and Ridge, making it effective in situations where some features are important, but multicollinearity exists.

**Limitations:**

- More complexity: Requires tuning two regularization parameters ($\lambda_1$ for L1, $\lambda_2$ for L2).
- Overfitting risk: If not tuned carefully, Elastic Net can still overfit when the dataset is small or noisy.

### 4.1.4 Support Vector Regression (SVR):

While Support Vector Machines (SVM) are used for classification, Support Vector Regression (SVR) uses the same ideas for regression. In order to optimize for a tolerance margin (epsilon, $\epsilon$), it attempts to fit the optimal line inside a threshold value.

**Mechanics:**

- SVR constructs a hyperplane in a higher-dimensional space using a kernel function (such as linear, polynomial, or radial basis function).
- The algorithm tries to fit the data within a "tube" defined by a margin of tolerance $\epsilon$. Only data points that fall outside this tube (support vectors) influence the model.

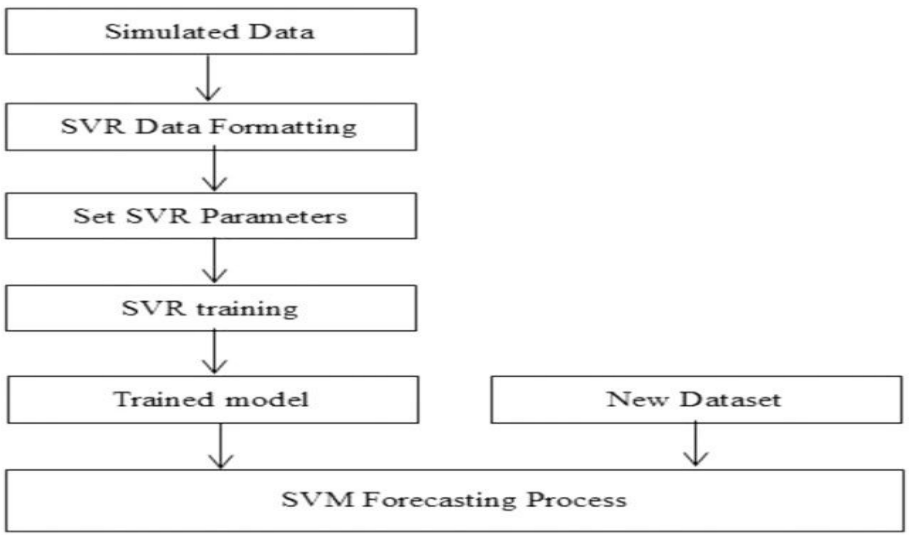


Figure 4.2: Flow chart of Support Vector Regression model

$$\text{Loss} = \frac{1}{2}\sum_{j=1}^{p}\beta_j^2 + C\sum_{i=1}^{n}\max(0, |y_i - \hat{y}_i| - \epsilon) \tag{4.4}$$

where:

- $\frac{1}{2}\sum_{j=1}^{p}\beta_j^2$ is the regularization term.
- $C$ is a regularization parameter.
- $\max(0, |y_i - \hat{y}_i| - \epsilon)$ is the epsilon-insensitive loss function.
- $\epsilon$ defines the margin of tolerance.

**Strengths:**

- Effective for non-linear relationships: The use of kernel functions allows SVR to handle complex, non-linear data patterns.
- Robust to outliers: SVR focuses on support vectors, so it's less sensitive to the majority of data points, which makes it more robust to noise and outliers.

**Limitations:**

- Computationally expensive: SVR can be slow, especially with large datasets and non-linear kernels.
- Requires tuning: The performance of SVR is highly dependent on choosing the right kernel and regularization parameters.

### 4.1.5 Decision Tree Regression:

Decision Tree Regression creates a tree with each leaf representing a projected value by dividing the data into subsets according to the input feature values. This process models the connection between features and the goal.

**Mechanics:**

- At each node of the tree, the algorithm selects the feature and threshold that best split the data into two groups with the lowest possible variance.
- The tree grows by recursively splitting data until each leaf node contains a small number of observations or the variance is sufficiently low.

**Strengths:**

- Captures non-linear relationships: Decision trees can model complex interactions between features and target variables.
- Easy to interpret: The resulting model is a series of decision rules, making it interpretable.
- Handles both numerical and categorical data.

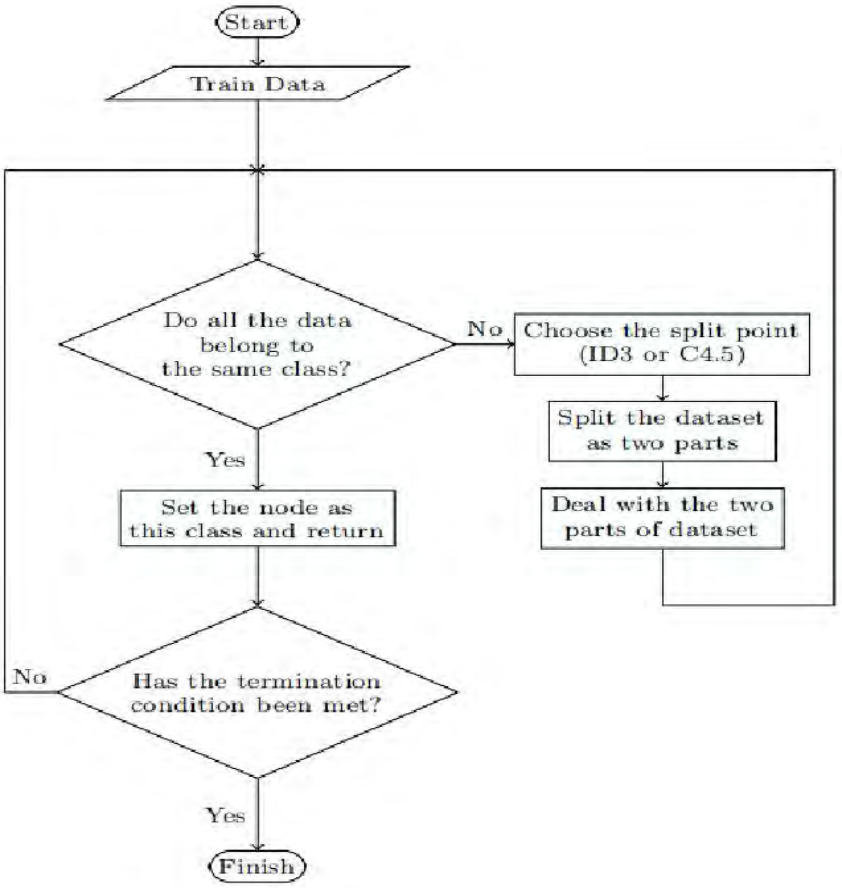


Figure 4.3: Flow chart of Decision Tree Regression model

**Limitations:**

- Prone to overfitting: Without pruning or limiting tree depth, decision trees can easily overfit the training data.
- Sensitive to small changes in data: A small change in the data can lead to a completely different tree structure (high variance).

### 4.1.6 Random Forest Regression:

An ensemble technique called Random Forest Regression creates many decision trees and averages each one's forecast. When compared to a single decision tree, it decreases overfitting and increases accuracy.

**Mechanics:**

- Random Forest creates multiple decision trees by randomly selecting subsets of the data and features for each tree (bootstrap aggregation or bagging).
- The final prediction is the average (for regression) of the predictions made by all trees.

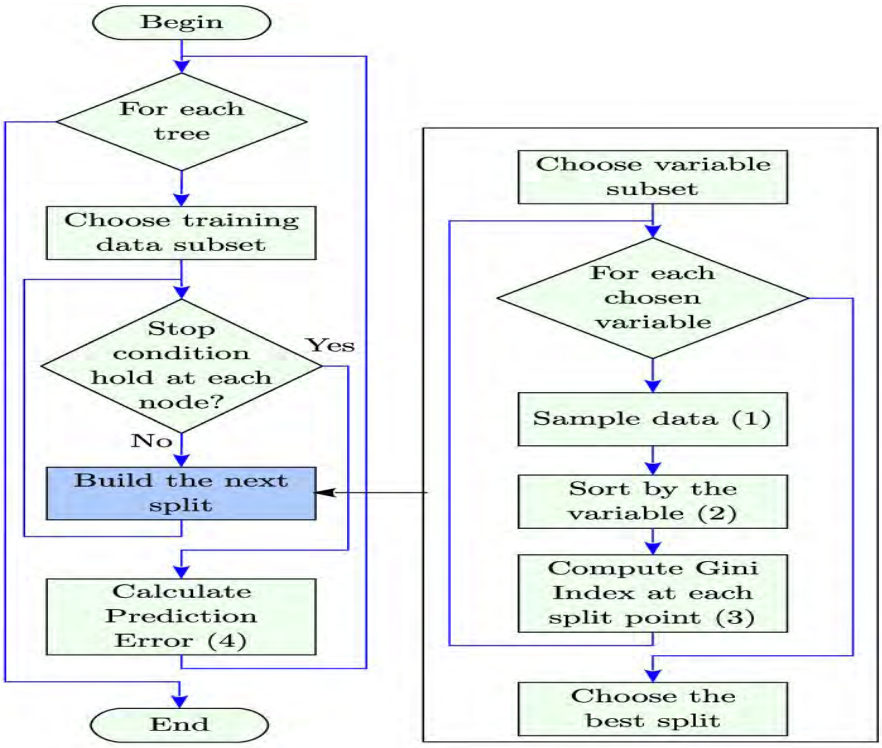


Figure 4.4: Flow chart of Random Forest Regression model

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} y^{(t)} \tag{4.5}$$

where:

- $T$ is the number of trees.
- $y^{(t)}$ is the prediction from the $t$-th tree.

**Strengths:**

- Reduces overfitting: By averaging multiple trees, Random Forest reduces the risk of overfitting that plagues single decision trees.
- Handles high-dimensional data: Random Forest can work well with datasets that have a large number of features and complex relationships.
- Robust to noise: Because it builds multiple trees, Random Forest is more robust to noise in the data.

**Limitations:**

- Interpretability: While decision trees are easy to interpret, Random Forests, being an ensemble, are harder to explain.
- Computationally intensive: Building and averaging multiple trees can be slow for large datasets.

### 4.1.7 ARIMA (AutoRegressive Integrated Moving Average):

Autoregression (AR), differencing (I), and moving average (MA) are the three components of the widely used ARIMA time series forecasting technique. It represents the residual errors from a moving average model applied to lagged data (MA) and the connection between an observation and many lagged observations (AR).

**Mechanics:**

- **Autoregressive (AR):** ARIMA uses past values of the target variable to predict future values. The number of lagged values used is determined by the parameter $p$.
- **Differencing (I):** Differencing helps make the time series stationary by subtracting previous observations. The number of times the differencing is applied is given by the parameter $d$.
- **Moving Average (MA):** ARIMA models the error terms as a linear combination of past errors. The number of lagged error terms is controlled by $q$.

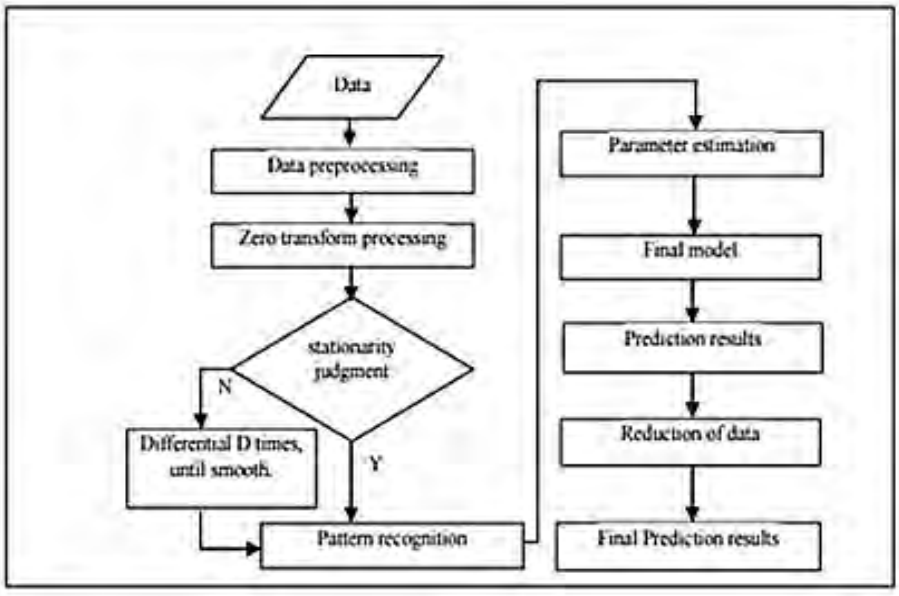


Figure 4.5: Flow chart of ARIMA model

The full ARIMA model is typically written as ARIMA$(p, d, q)$, where:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (4.6)$$

where:

- $y_t$ is the value at time $t$.
- $c$ is a constant term.
- $\phi_1, \ldots, \phi_p$ are the autoregressive coefficients.

- $\theta_1, \ldots, \theta_q$ are the moving average coefficients.
- $\varepsilon_t$ is the error term at time $t$.
- $p$ is the order of the autoregressive part.
- $q$ is the order of the moving average part.

**Strengths:**

- Effective for short-term forecasting: Works well with univariate time series data, especially when seasonal or trend components are present.
- Interpretable: The parameters $p$, $d$, and $q$ offer clear insight into the structure of the time series.

**Limitations:**

- Requires stationary data: ARIMA assumes that the time series is stationary, meaning that the mean and variance are constant over time.
- Limited to univariate data: ARIMA is capable of modeling one time series at a time, unless its variants like VARIMA are adopted.
- May struggle with complex patterns: ARIMA might not capture non-linear relationships or sudden structural breaks in the data.

### 4.1.8 Facebook Prophet:

Facebook Prophet is an open-source forecasting tool designed specifically for handling time series data. Developed by Facebook's Core Data Science team, Prophet is a powerful and flexible tool designed to make accurate time-series forecasting easier, especially when dealing with multiple seasonality, missing values, and outliers. It is highly intuitive and works well for business applications, especially when forecasts need to be made on a daily, weekly, or monthly basis.

**Mechanics:**

- Prophet is an additive model that decomposes time-series data into three key components: trend, seasonality, and holidays/events. It is designed to handle time series with strong seasonal patterns, outliers, and missing data.

**Trend Component:** Prophet models the overall growth or decline of the data over time. It supports both linear growth and logistic growth.

**Seasonality Component:**

- Captures repeating patterns at different time scales, such as daily, weekly, or yearly seasonality.
- Assumes that seasonality repeats and adds this as a component in the forecast.

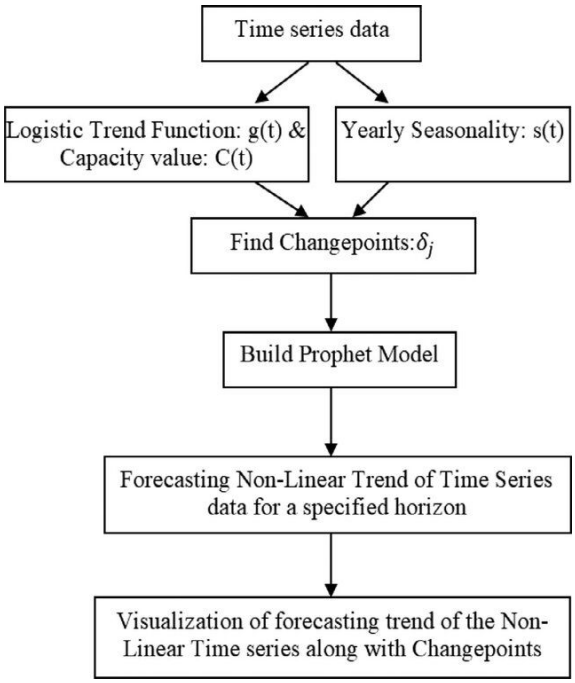


Figure 4.6: Flow chart of Facebook Prophet model

**Holidays and Events:**

- Allows the inclusion of holidays and specific events that could impact the time series. Users can specify these dates, and Prophet will add them as a separate component in the model to account for anomalies or irregular events.

**Formula:**

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{4.7}$$

Where:

- $y(t)$ is the predicted value at time $t$.
- $g(t)$ models the overall trend of the data.
- $s(t)$ models the seasonal component (daily, weekly, yearly).
- $h(t)$ accounts for holidays or special events.
- $\epsilon_t$ is the error term (random fluctuations).

**Strengths:**

- User-friendly: Simple software which can be mastered quickly and does not require an advanced understanding of time series modeling.
- Deals with complicated seasonality: Easily incorporates weekly and yearly seasonality using the programmer's automatic logic.
- Adaptability: Allows the user to incorporate their chosen holidays, seasonal periods, and varying growth directions.
- Resilience: Performs well even with missing values or outliers.

**Limitations:**

- Basic Model: Its simplicity can be a disadvantage for very complex or non-standard time series.
- Non-standard Data Requires Modifications: Prophet is designed to work well with data that come in repetitive and straight forward cycles. For data that is less simple, changes to the model are required.
- Does Not Fit All Data: Prophet has been designed for time series purposes only which has a scope of forecasting. This forecasting model may not be appropriate for data which is not seasonal or directional in nature.

## 4.2 Working Plan

The process of "Sales Forecasting using Machine Learning" goes through various steps. It starts with the collection of data that consists of sales figures for the previous months/years with additional variables like dates, type of products, prices offered, discount offers, and other economic variables. This data will be used to build up the forecasting model. Once collected, the data is then pre-processed which involves treating missing or inconsistencies, outliers. this step makes sure that the data collected is free from any imperfections thus making it easy to train the model. Exploratory data analysis (EDA) is also performed in order to view how the data is distributed and understand what drives sales patterns.

After preprocessing comes the splitting of the dataset into the training dataset and the test dataset. Last, the training dataset is used to fit on some of the regression models, including but not limited to, linear regression, decision tree models, and neural networks. In this part Hyperparameter tuning will be done. When the models are fully trained, metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are used to assess their performance. Cross-validation is used for the evaluation of the models to allow the use of the best model according to evaluation metrics.

Model selection and deployment is the logical last step. The sales management system is fitted with a model that produced the most favorable results. Various stakeholders in the sales process have an easy-to-use interface where they enter relevant information and are also able to see forecasts made by the model. It is also meant to be expandable to allow input of data and output of forecasts at the same time on a real time basis. To promote further development, the system has a feedback loop that captures information on expectation-based measure of performance and user feedback and therefore, models are routinely refreshed whenever the market is dynamic. Further, the solution is complemented with advanced sales forecasts reports and interactive dashboards that support stakeholders in tracking performance because they help them understand sales numbers in a succinct way.

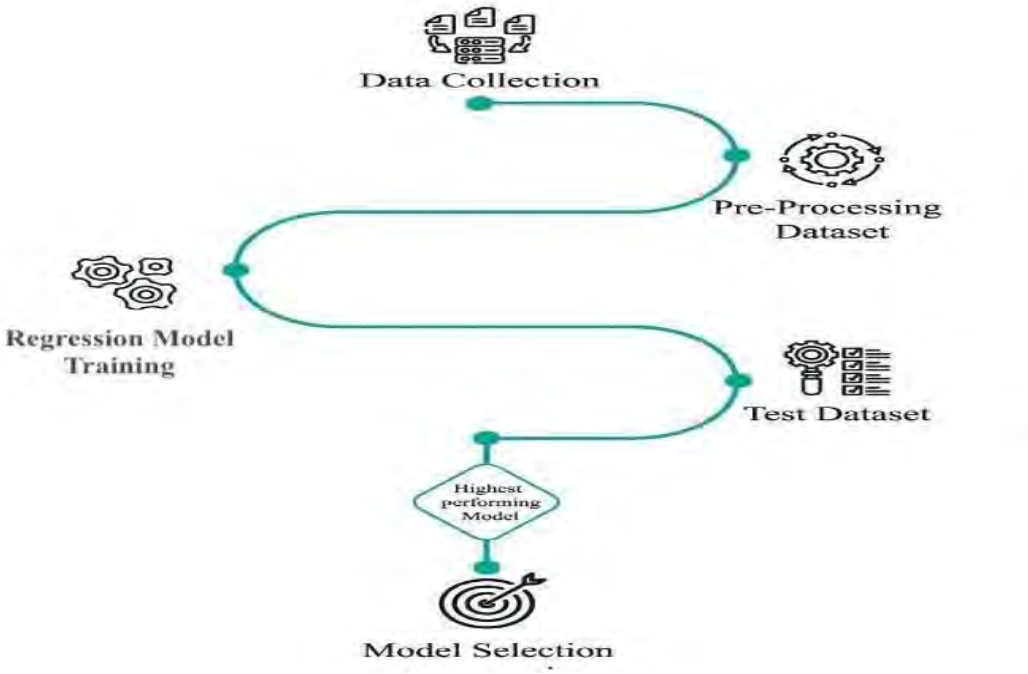


Figure 4.7: Workflow

## 4.3 Evaluation

### 4.3.1 Mean Squared Error (MSE)

The Mean Squared Error (MSE) is another common metric for evaluating regression models. It measures the average squared difference between the actual and predicted values. Unlike Mean Absolute Error (MAE), MSE penalizes larger errors more heavily because the errors are squared.

**Formula:**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.8}$$

Where:

- $n$ is the number of data points.
- $y_i$ is the actual value.
- $\hat{y}_i$ is the predicted value.

**Interpretation:**

- MSE gives you an idea of how much your model's predictions deviate from the actual values, with larger errors being emphasized.
- A lower MSE value indicates that the model's predictions are generally close to the actual values.

**Strengths:**

- Emphasizes larger errors, making it useful in applications where large errors are particularly undesirable.
- Provides a smooth and differentiable function, making it suitable for optimization in machine learning models.

**Limitations:**

- More sensitive to outliers than MAE because it squares the error terms, which means large errors have a disproportionately larger impact on the final result.
- Harder to interpret compared to MAE, as it uses squared values (e.g., units of MSE are squared units of the target variable).

### 4.3.2 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a widely used metric for evaluating the accuracy of regression models. It measures the average absolute difference between the actual values (true values) and the predicted values. MAE gives an idea of how far the predictions are from the actual outcomes, without considering the direction of the error (i.e., it doesn't differentiate between underestimates and overestimates).

**Formula:**

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{4.9}$$

Where:

- $n$ is the number of data points.
- $y_i$ is the actual value.
- $\hat{y}_i$ is the predicted value.

**Interpretation:**

- MAE tells you the average magnitude of errors in a set of predictions, without considering their direction.
- A lower MAE value indicates that the model's predictions are generally close to the actual values.

**Strengths:**

- Easy to understand and interpret.
- Less sensitive to outliers compared to MSE (Mean Squared Error) because it does not square the error terms.

**Limitations:**

- Since it does not square the errors, large errors are not emphasized, which may be a problem in certain use cases.

# Chapter 5

# Implementation

There are a number of different machine learning algorithms that can be used for regression dataset. We used 8 algorithms: Linear Regression, Lasso Regression, Elastic Net Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, ARIMA and Facebook Prophet. At first, we used first 6 algorithms but the result was so bad. Then we used forecasting algorithm ARIMA and Facebook Prophet which gives us much better result. The performance of all algorithms is shown below in Table 5.1 & 5.2.

Table 5.1: Performance of Proposed Model

| **SL** | **Model** | **Mean Squared Error** | **Mean Absolute Error** |
|---|---|---|---|
| 1 | Linear Regression | 298122.463 | 268.026 |
| 2 | Lasso Regression | 298113.648 | 267.989 |
| 3 | Elastic Net Regression | 298729.908 | 267.545 |
| 4 | Support Vector Regression | 330885.301 | 204.165 |
| 5 | Decision Tree Regression | 372022.723 | 206.285 |
| 6 | Random Forest Regression | 265424.179 | 183.198 |

Table 5.2: Performance of Forecasting Algorithm

| **SL** | **Model** | **RMSE** | **MAE** |
|---|---|---|---|
| 1 | ARIMA | 44.796 | 41.840 |
| 2 | Facebook prophet | 60.724 | 53.082 |

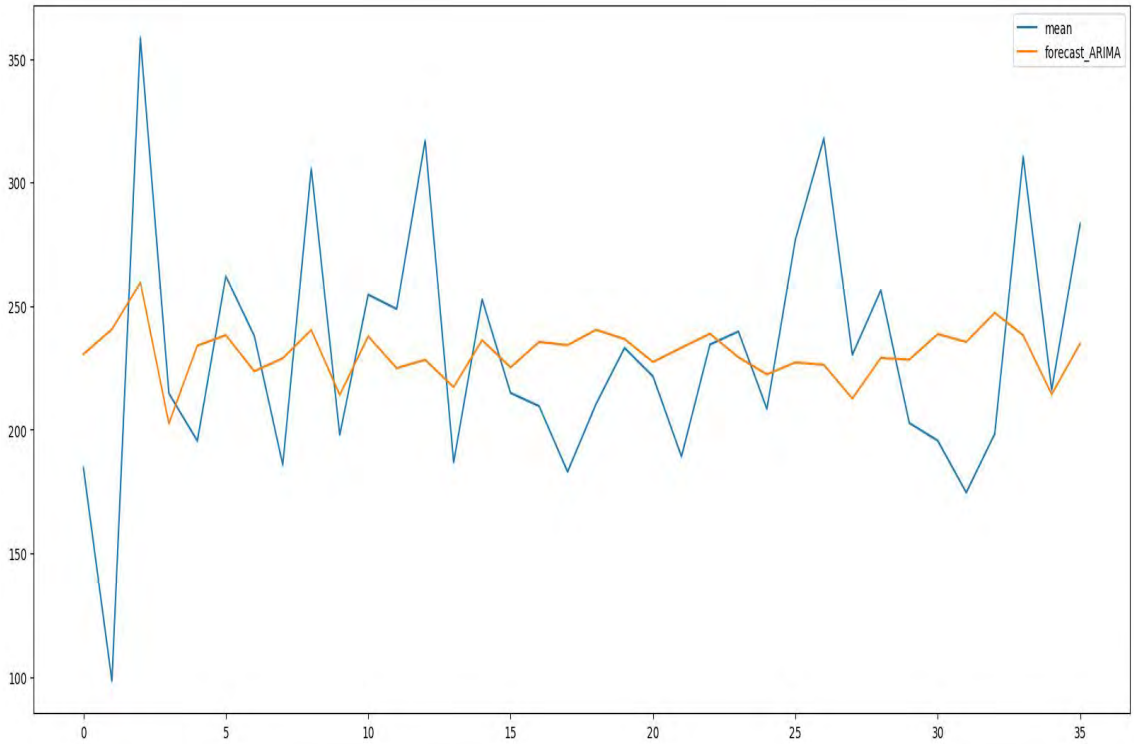


Figure 5.1: Mean of Training Data vs Prediction of ARIMA on Training Data

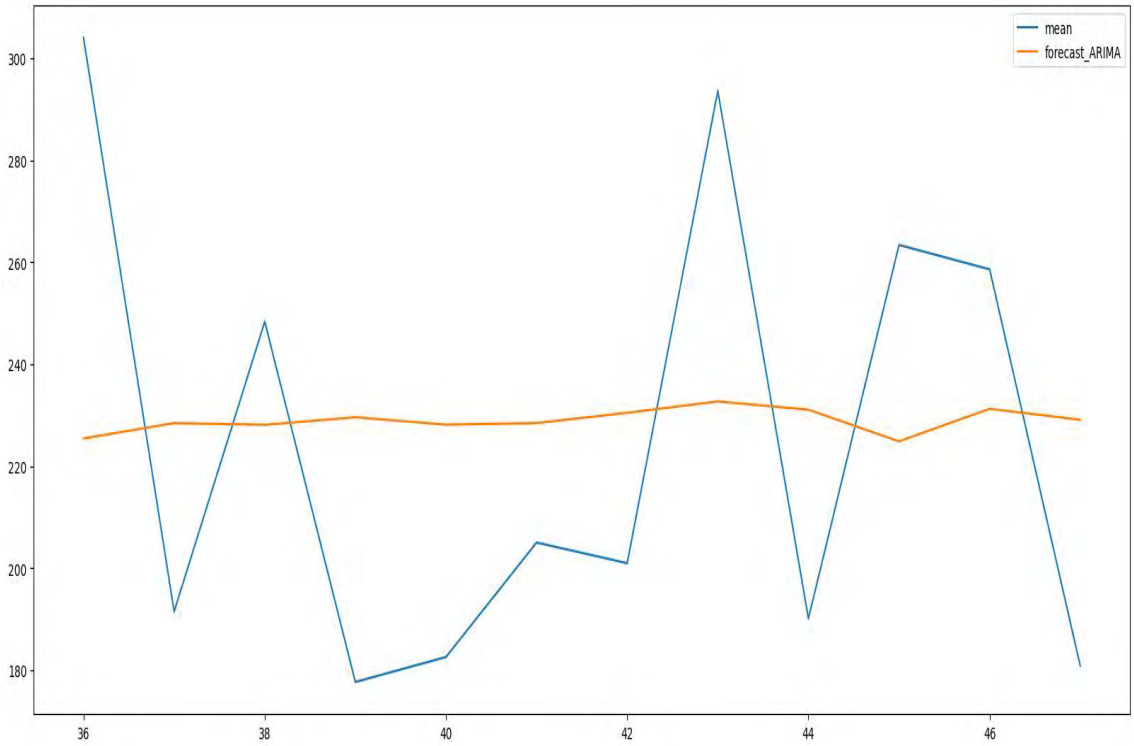


Figure 5.2: Mean of Training Data vs Prediction of ARIMA on Test Data

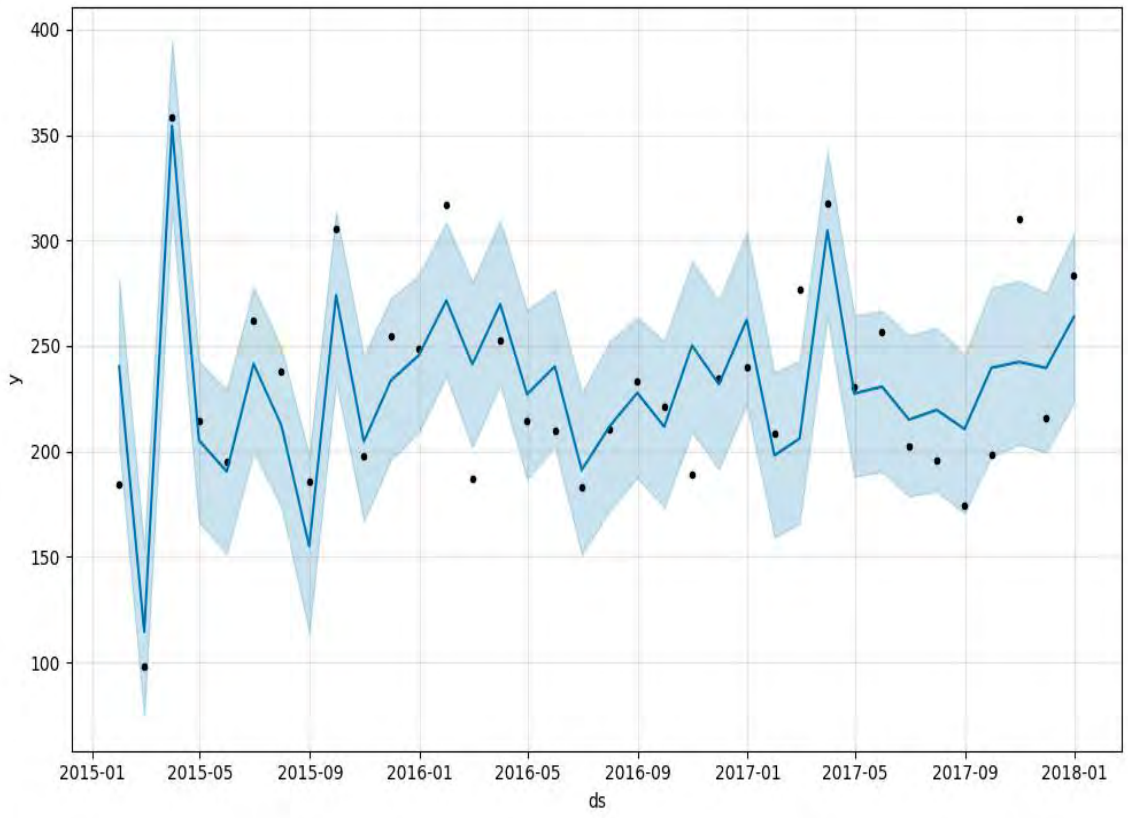


Figure 5.3: Mean of Testing Data vs Prediction of Facebook Prophet on Test Data

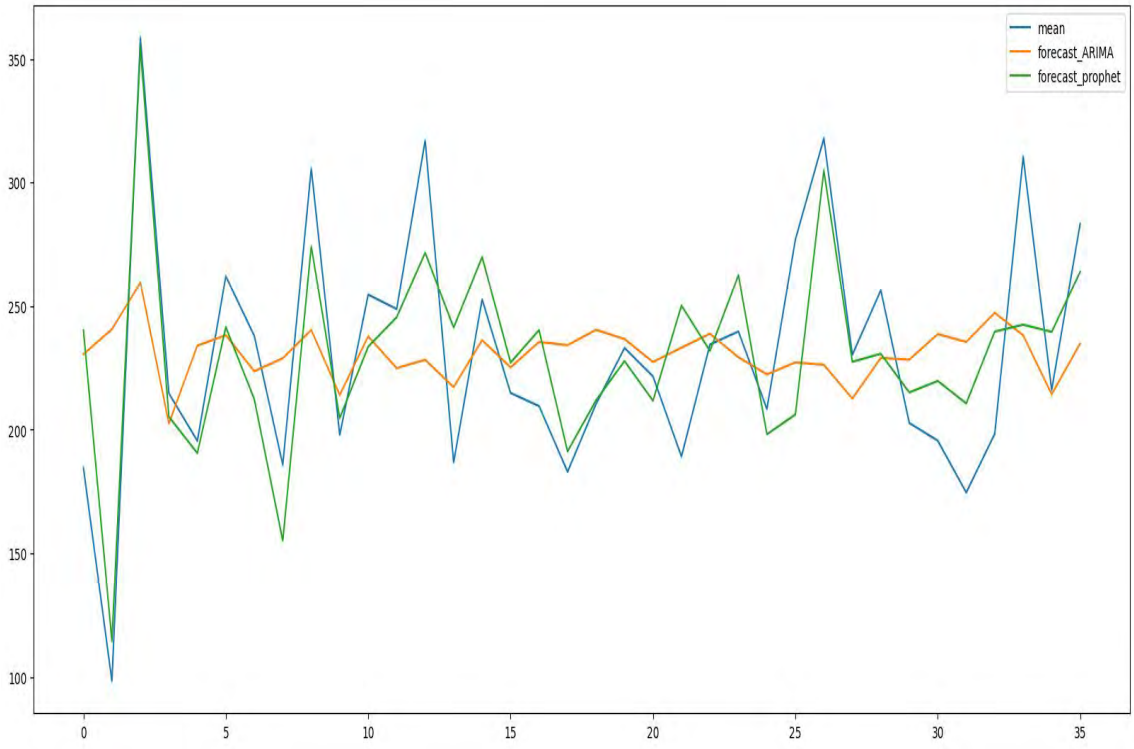


Figure 5.4: Mean of Testing Data vs Prediction of Facebook Prophet on Training Data

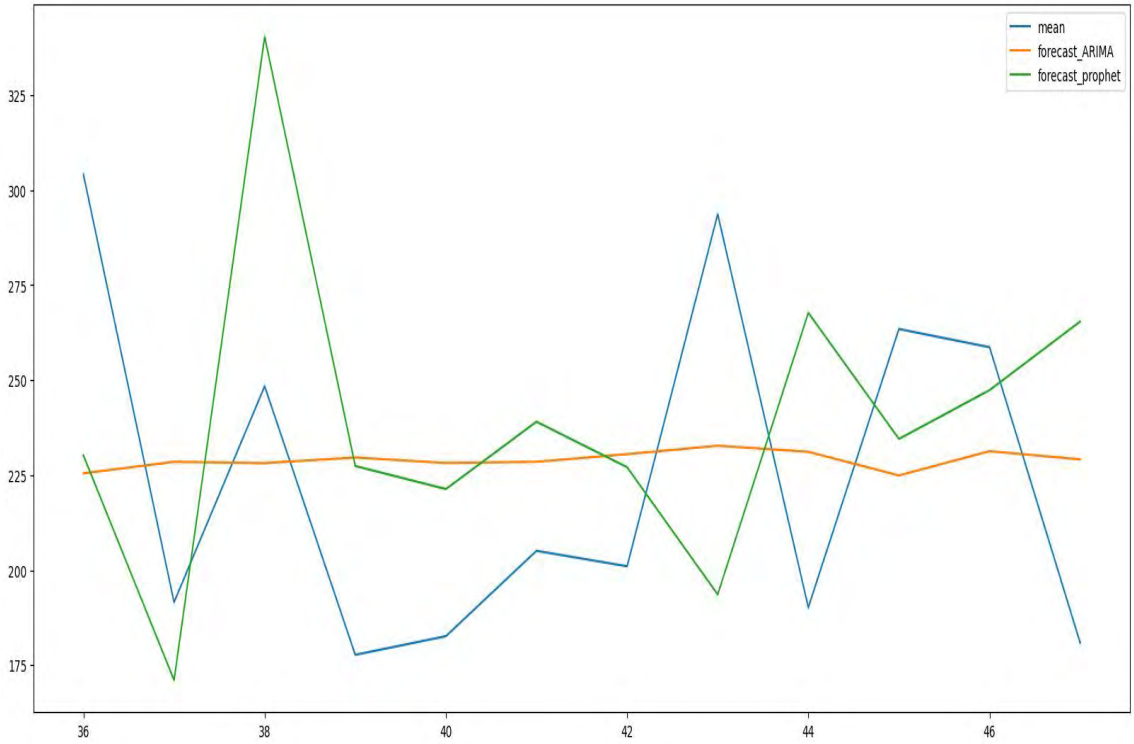


Figure 5.5: Mean of Training Data vs Prediction of ARIMA vs Prediction of Facebook Prophet on Test Data

Figure 9 illustrates a comparison between the mean of the training data and the predictions made by the ARIMA and Facebook Prophet models. The ARIMA model (orange line) closely follows the actual mean values (blue line), providing a smoother and more consistent prediction. This aligns with its lower error metrics (RMSE and MAE), indicating its better fit for this particular dataset. In contrast, the Facebook Prophet model (green line) shows more variability in its predictions, capturing the general trend but missing some of the sharper fluctuations. While Prophet is flexible and capable of modeling seasonality and trend changes, it seems to underperform compared to ARIMA in this case, likely due to Prophet's tendency to require more tuning for datasets with complex patterns or sharp spikes.

Figure 10 compares the mean of the test data with the predictions from the ARIMA and Facebook Prophet models. On the test data, ARIMA (orange line) continues to provide a relatively smooth and stable forecast, maintaining a consistent trend close to the actual data mean. However, ARIMA seems to under-capture some of the sharp variations in the actual test data (blue line), suggesting that it favors general trend accuracy over short-term volatility.

On the other hand, Facebook Prophet (green line) exhibits more variability, attempting to follow the peaks and valleys of the actual data. While this can help capture sudden shifts, it also appears to overshoot in some areas, leading to less precise predictions overall compared to ARIMA.

This behavior aligns with the general observation that ARIMA provides better overall fit for smooth trends, while Prophet is more suited to capturing irregularities in time series data but might require more tuning for accuracy.

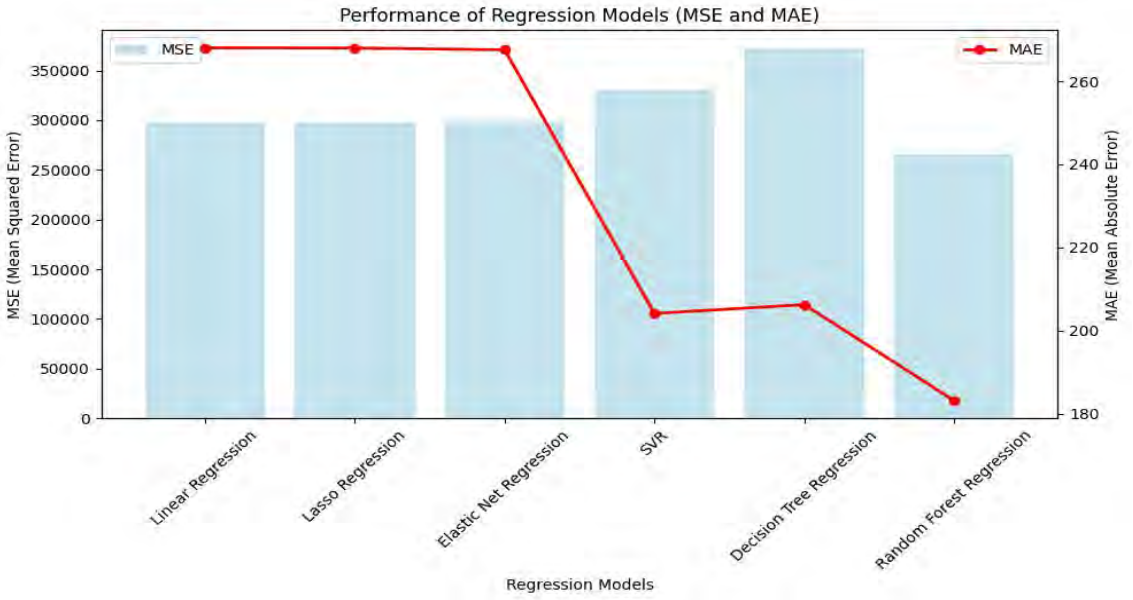


Figure 5.6: MSE vs MAE performances of the Regression Models

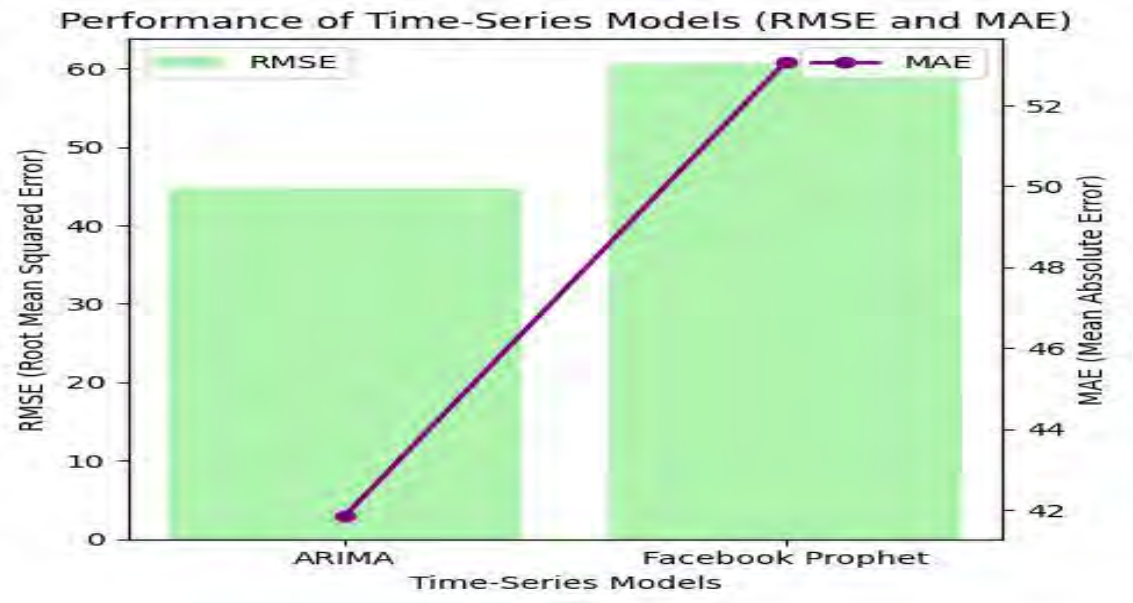


Figure 5.7: RMSE vs MAE performances of the Regression Models

# Result Discussion

From the results of the two tables, it is clear that Random Forest Regression outperforms other regression models in terms of both Mean Squared Error (MSE) and Mean Absolute Error (MAE), making it the most accurate model for the dataset. It achieves the lowest MSE (265,424.179) and MAE (183.198), indicating that it produces the smallest average prediction errors compared to models like Linear Regression, Lasso Regression, and Elastic Net Regression, which have similar but slightly higher error metrics.

Although Support Vector Regression (SVR) has a relatively high MSE, its MAE (204.165) is lower than most other models, suggesting that it performs well for small errors but struggles with larger errors. Decision Tree Regression shows the highest MSE (372,022.723), indicating that it is prone to overfitting or less stable predictions in this case.

In the time-series models, ARIMA performs better than Facebook Prophet, with a lower Root Mean Squared Error (RMSE) of 44.796 and MAE of 41.840. Facebook Prophet has higher error values (RMSE: 60.724, MAE: 53.082), making it less accurate than ARIMA for this particular dataset.

Overall, Random Forest Regression is the best-performing model for regression tasks, and ARIMA provides the best results for time-series forecasting.

# Chapter 6

# Conclusion and Future work

## 6.1 Future Work

In the future, here's what can be done. Integrating environmental aspects – social media, marketing and economic metrics – into sales forecasting models will be considered. This research has been majorly limited to the usage of historical sales data. Use of more multivariate data can lead to much more accurate forecasts. What is more, the effect of the complex modes of the data may be suitable to the application of hybrid models which use various machine learning algorithms. Besides, another significant area of study is the design of machine learning systems that are easier to interpret, enabling business participants to know how the predictions were arrived at. Lastly, the employment of such models will be useful in forecasting systems in operation as it enhances timely information for the businesses to react to the market dynamics.

## 6.2 Conclusion

The research has shown the positive impact of utilizing machine learning approaches in sales forecasting. The research has also examined different machine learning models based on regression methods and timeseries forecasting algorithms proving that machine learning enhances traditional forecasting techniques in efficiency and accuracy. Random forest regression and ARIMA models were most effective in both predicting sales volumes regression data or time series and predicting seasonal trends in sales. These models enhance large data sets due to their ability to change with market fluctuations thus improving decision making for the business and its efficiency in pleasing the customers. By employing machine learning techniques in sales forecasting, and hence in strategizing, more innovations are absorbed by the organization and this provides them with the much yearning competitive advantage in today's markets where changes are ever experienced within short time periods.

# Bibliography


1. Z. Chen, Y. Wang, and X. Li, "Sales Forecasting with Gradient Boosting Models: A Case Study in Online Retail," Journal of Machine Learning Research, vol. 12, pp. 123-135, 2020.

2. Y. Zhang, H. Liu, and W. Sun, "Long Short-Term Memory Networks for Sales Prediction: Capturing Complex Temporal Patterns in Data," in Proceedings of the International Conference on Data Science and Applications, 2019, pp. 225-234.

3. M. Ferreira, A. Lopes, and T. Pereira, "Multivariate Sales Forecasting Using LSTM: Integrating Historical Sales Data with External Factors," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 9, pp. 4876-4889, 2021.

4. Z. Wei, J. Xu, and L. Tang, "Handling Sparse Data in Sales Forecasting Using Transfer Learning," Journal of Data Science and Analytics, vol. 14, pp. 345-360, 2020.

5. S. Katuwal and R. Chen, "Interpreting Sales Forecasting Models with SHAP Values: A Case Study with XGBoost," in Proceedings of the ACM International Conference on Data Mining, 2020, pp. 1125-1133.

6. A. Meireles, M. Dias, and R. Silva, "Real-Time Sales Forecasting Using Cloud-Based Machine Learning Models: An Ensemble Approach," International Journal of Big Data, vol. 16, pp. 89-105, 2021.

7. V. Sohrabpour, P. Oghazi, R. Toorajipour, and A. Nazarpour, "Export sales forecasting using artificial intelligence," Technol. Forecast. Soc. Change, vol. 163, no. 120480, p. 120480, 2021.

8. F. Haselbeck, J. Killinger, K. Menrad, T. Hannus, and D. G. Grimm, "Machine learning outperforms classical forecasting on horticultural sales predictions," Mach. Learn. Appl., vol. 7, no. 100239, p. 100239, 2022.

9. S. T. Londhe and S. Palwe, "Customer-centric sales forecasting model: RFM-ARIMA approach," Bus. Syst. Res. J., vol. 13, no. 1, pp. 35–45, 2022.

10. A. Krishna, Akhilesh, A. Aich, and C. Hegde, "Sales-forecasting of retail stores using machine learning techniques," in 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 2018.

11. V. Wineka Nirmala, D. Harjadi, and R. Awaluddin, "Sales forecasting by using exponential smoothing method and trend method to optimize product sales in PT. Zamrud Bumi Indonesia during the covid-19 pandemic," Int. J. Eng. Scie. and Inform. Technology., vol. 1, no. 4, pp. 59–64, 2021.

12. M. A. I. Arif, S. I. Sany, F. I. Nahin, and A. S. A. Rabby, "Comparison study: Product demand forecasting with machine learning for shop," in 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), 2019.

13. I.-F. Chen and C.-J. Lu, "Sales forecasting by combining clustering and machine-learning techniques for computer retailing," Neural Comput. Appl., vol. 28, no. 9, pp. 2633–2647, 2017.

14. M. Gurnani, Y. Korke, P. Shah, S. Udmale, V. Sambhe, and S. Bhirud, "Forecasting of sales by using fusion of machine learning techniques," in 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), 2017.

15. W. K. Wong and Z. X. Guo, "Intelligent sales forecasting for fashion retailing using harmony search algorithms and extreme learning machines," in Optimizing Decision Making in the Apparel Supply Chain Using Artificial Intelligence (AI), Elsevier, 2013, pp. 170–195.

16. N. Harz, S. Hohenberg, and C. Homburg, "Virtual reality in new product development: Insights from prelaunch sales forecasting for durables," J. Mark., vol. 86, no. 3, pp. 157– 179, 2022.

17. D. S. Rajpoot, B. Mittal, H. Dudani, and U. Singhal, "Sales analysis and forecasting using machine learning approach," in Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing, 2023.

18. D. Rout, B. Roy, and P. Kapse, "A subtle design of prediction models using machine learning algorithms for advocating selection and forecasting sales of garments: A case study," in Advances in Data-Driven Computing and Intelligent Systems, Singapore: Springer Nature Singapore, 2024, pp. 387–397.

19. J. S. Kiran, P. S. V. S. Rao, P. V. R. D. P. Rao, B. S. Babu, and N. Divya, "Analysis on the prediction of sales using various machine learning testing algorithms," in 2022 International Conference on Computer Communication and Informatics (ICCCI), 2022.

20. B. Kumar Jha and S. Pande, "Time series forecasting model for supermarket sales using FBprophet," in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021.

21. Y. Li, Y. Yang, K. Zhu, and J. Zhang, "Clothing sale forecasting by a composite GRU– prophet model with an attention mechanism," IEEE Trans. Industr. Inform., vol. 17, no. 12, pp. 8335–8344, 2021.

22. G. Tsoumakas, "A survey of machine learning techniques for food sales prediction," Artif. Intell. Rev., vol. 52, no. 1, pp. 441–447, 2019.

23. R. S. Sengar and D. F. Ahmed, "Review on trends in machine learning applied to demand & sales forecasting," SMART MOVES JOURNAL IJOSCIENCE, vol. 5, no. 6, p. 4, 2019.

24. Y. K. Elalem, S. Maier, and R. W. Seifert, "A machine learning-based framework for forecasting sales of new products with short life cycles using deep neural networks," Int. J. Forecast., vol. 39, no. 4, pp. 1874–1894, 2023.

25. R. Puspita and L. A. Wulandhari, "Hardware sales forecasting using clustering and machine learning approach," IAES Int. J. Artif. Intell. (IJ-AI), vol. 11, no. 3, p. 1074, 2022.

26. M. O. Anwer and S. Akyuz, "Sales forecasting of a hypermarket: Case study in Baghdad using machine learning," in 2022 30th Signal Processing and Communications Applications Conference (SIU), 2022.

27. Y. Ensafi, S. H. Amin, G. Zhang, and B. Shah, "Time-series forecasting of seasonal items sales using machine learning – A comparative analysis," International Journal of Information Management Data Insights, vol. 2, no. 1, p. 100058, 2022.

28. D. M. U. Ashraf, "A predictive analysis of retail sales forecasting using machine learning techniques," Research Journal Of Computer Science And Information Technology, vol. 6, no. 04, pp. 23–33, 2022.

29. R. S. Mallik, R. Abhiram, S. R. Reddy, and R. M. Jagadish, "A comprehensive survey on sales forecasting models using machine learning algorithms," in 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2022.

30. https://www.kaggle.com/datasets/vivek468/superstore-dataset-final Superstore Dataset