Predicting Retail Sales Success: A Comparative Analysis of Machine Learning, Neural Networks, and Time Series Models

Sai Sharan Paspunuri

Final Thesis Report

July 2024

TABLE OF CONTENTS

DEDICATION

This thesis is dedicated to my loving parents and my wonderful wife. Mom and Dad, your unwavering support and belief in me have been my greatest source of strength and motivation. Your encouragement has been a constant reminder to persevere, even in the face of challenges.

To my wife, your patience, understanding, and love have been my rock throughout this journey. Your endless support and sacrifices have made it possible for me to reach this milestone. Thank you for always being there for me, cheering me on, and believing in my dreams.

This work is a tribute to your faith in me and a heartfelt thank you for being my pillars of strength.

ACKNOWLEDGEMENTS

I am deeply grateful to my supervisor, Abraham, whose guidance and support have been invaluable throughout this journey. Your insights and encouragement have been crucial to the completion of this thesis, and I cannot thank you enough for your patience and dedication.

I also want to extend my heartfelt thanks to Liverpool John Moore's University for providing a fantastic environment for learning and growth. The resources and opportunities available here have significantly contributed to my academic and personal development.

To my parents, thank you for your unwavering support and belief in me. Your constant encouragement has been a source of strength and motivation, helping me to persevere through the toughest times. To my dear wife, your love, patience, and understanding have been my rock. Your support has been essential in helping me stay focused and committed to this journey. Thank you for always being there for me and for believing in my dreams.

This thesis reflects the collective support and encouragement from all these incredible individuals. Thank you for being a part of this journey and for making this achievement possible.

ABSTRACT

This study aims to assess the efficacy of diverse forecasting methods in predicting retail sales, with a focus on guiding retail businesses in selecting optimal methods tailored to specific business scenarios. Utilizing Walmart's sales data, the investigation delves into factors like store types, sizes, departments, and holidays. A comparative analysis is performed to compare the performance of advanced deep learning models with traditional machine learning and time series models. The findings illuminate instances where alternative models surpass traditional counterparts in both accuracy and efficiency. By showcasing the superiority of certain models, this research equips retail businesses with valuable insights for enhancing forecasting practices, ultimately contributing to informed decision-making and improved operational strategies in the retail sector.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARIMA..... Auto Regressive Integrated Moving Average

ANN..... Artificial Neural Networks

ANCDE..... Attentive Neural Controlled Differential Equations.

CNN ..... Convolution neural network

CPI..... Consumer Price Index

DL..... Deep Learning

DT..... Decision Tree

EDA..... Exploratory Data Analysis

GBRT..... Gradient Boosted Regression Tree

KNN..... K-Nearest Neighbors' Algorithm

LSTM..... Long Term – Short Term Memory

MAE..... Mean Absolute Error

MAPE..... Mean Absolute Percentage Error

NCDE..... Neural Controlled Differential Equation

NN..... Neural Networks

RF..... Random Forest

RNN..... Recurrent Neural Network

RMSE..... Root Mean Squared Error

SARIMAX.....Seasonal Autoregressive Integrated MA with Exogenous Regressor.

SVM..... Support-vector machines

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The background introduction highlights the critical role of retail sales forecasting in the industry, emphasizing its impact on decision-making, production, and supply chain management. It acknowledges the challenges posed by factors such as trends, seasonality, and the complexity of the retail market. The study positions itself within the existing landscape of forecasting methods, including time series analysis, statistical models, and the emerging field of deep learning. The background underscores the need for accurate forecasting to enhance business strategy, reduce operational costs, and improve customer satisfaction.

Dividing retail sales forecasting into time series analysis and forecasting, further categorized into statistical and deep learning methods. This provides a structured framework for evaluating and comparing different forecasting approaches. The comparative analysis, challenges, and conclusion sections contribute to a comprehensive understanding of the research landscape and its implications for the retail industry.

This research emphasizes the growing complexity of forecasting retail sales due to factors such as a high number of products, shorter product lifecycles, and aggressive marketing campaigns. Retail sales promotions, including various types, are identified as a confounding factor. The study positions itself within the existing literature by acknowledging the prevalence of univariate methods and the emergence of ML techniques in retail sales forecasting. The focus on post-promotional periods is identified as a gap in the current research landscape, motivating the study to address this aspect comprehensively.

The global pandemic significantly impacted the economy, particularly the retail and leisure industries. The study focuses on Walmart, a major retail player, to analyze the aftermath and provide a forecasting model for weekly sales. Exploratory Data Analysis (EDA) reveals insights into store types, sizes, and departments, contributing to better understanding and model accuracy.

Traditional time series methods like ARIMA and Smoothing face challenges in handling non-linear patterns common in real-world sales series. The research also advocates for the use of

advanced forecasting methods, such as Neural Networks and Machine Learning, citing their ability to capture nonlinearity effectively. The background also highlights the dataset's origin from a Kaggle and emphasizes the absence of additional features that could contribute to explaining store sales patterns.

Time Series Analysis in Business:
Time series analysis is highlighted as a valuable tool for extracting meaningful statistics and properties from data in business environments. Emphasis on the crucial role of time series forecasting models in determining future sales and aiding in business management.

Time Series Analysis Components:
Introduction of the four major components of time series data: Level, Trend, Seasonality, and Noise. Recognition of the significance of understanding these components for accurate predictions and decision-making.

The study is grounded in the evolution of supply chain management systems, emphasizing the role of technology, particularly artificial intelligence, in enhancing forecasting accuracy. Traditional demand planning methods are contrasted with the hybrid CNN-LSTM model, highlighting the need for more effective forecasting techniques in the retail industry. The research also reinforces the choice of the hybrid model based on the effectiveness of CNNs and LSTMs in time-series forecasting.

This research also involves a review of the evolution from "shallow" neural networks to deep learning. The paper discusses the mathematical background of neural networks, starting from single layer perceptron to deep neural networks. It highlights the challenges associated with optimization, the choice of activation functions, and the need for preprocessing in deep learning. The study emphasizes the potential advantages of deep neural networks in handling raw data without extensive manual feature engineering.

This comprehensive analysis provides the context for the subsequent exploration of the proposed deep-embedded network architecture and its application in the three business analytics case studies. The research contributes to the ongoing efforts to enhance predictive modeling using deep learning techniques in operational research and business analytics.

1.2 Problem Statement

Traditional time series methods struggle to capture the non-linear patterns exhibited in real-world sales series, such as those in retail. The research addresses the limitations of conventional methods like ARIMA and Smoothing in handling nonlinearity and explores advanced forecasting methods to enhance accuracy. The study aims to provide effective forecasting models for sales time series, considering the challenges posed by the absence of additional explanatory features.

The post-pandemic era poses challenges for the retail industry, demanding accurate sales forecasts. The study addresses this by building models to predict Walmart's sales. The goal is to assist management teams in proactive decision-making, saving costs, and avoiding errors in staff scheduling.

Retail sales forecasting is challenged by the complex dynamics of promotions, particularly in predicting the impact during and after promotional periods. Existing methods often rely on judgmental approaches, and there is a need to explore the effectiveness of other forecasting techniques. The study identifies the post-promotional period as a critical phase that requires specific attention for accurate inventory planning.

The conventional machine learning models, including linear models and tree-based approaches, may not fully capture the complexity of business analytics tasks, leading to suboptimal predictive performance. The research addresses this limitation by exploring and evaluating the application of deep learning models, and intense neural networks, in various business scenarios.

We will be building & comparing forecasting models using the time series, ML, and Deep neural networks and optimizing them. We will be comparing and evaluating the different metrics. We will be also suggesting the model to be used based on the business application or the business problem to get efficient results.

1.3 Aim and Objectives

The primary aim of this thesis is to evaluate and select the most suitable model for accurately predicting retail sales, considering the complexities of the retail environment and leveraging a comprehensive comparison of machine learning models, neural networks, and time series

models. This research will assess the effectiveness of these models, identify their strengths and limitations, and provide guidance for retailers in choosing the most appropriate approach.

The objectives are as follows:

1. **Conduct a Comprehensive Literature Review:** Review existing studies on retail sales forecasting to establish a foundational understanding of the techniques, identify research gaps, and define the scope of this research.
2. **Model Development and Evaluation:** Construct, implement, and evaluate the performance of a range of forecasting models, including machine learning, neural networks, and time series techniques, using a real-world retail dataset.
3. **Data Exploration and Feature Engineering:** Conduct thorough exploratory data analysis and identify key variables influencing sales, optimizing the dataset for improved model accuracy.
4. **Comparative Analysis and Insights:** Compare the performance of different models using relevant metrics, identifying their strengths and limitations, and highlighting the factors that influence model accuracy in a retail context.
5. **Recommendations for Retail Forecasting:** Based on the evaluation results, provide recommendations for retailers on selecting the most effective models for their specific needs, considering factors like data availability, model complexity, and desired accuracy.

1.4 Research Questions

1. How do you optimize the data features based on the business problem?
2. Which time series models provide accurate predictions?
3. What is the time frame for considering historical data and making future predictions?
4. Does the size of the training dataset impact results?
5. Which algorithm, ARIMA or LSTM, performs a more accurate prediction of time series data?
6. Which forecasting model performs better for supermarket sales prediction among the time series forecasting models?
7. How do different time-series forecasting models handle variations, trends, and outliers in supermarket sales data?

8. How does the hybrid CNN-LSTM model perform in forecasting demand compared to traditional methods?
9. What are the optimal parameters for the Deep Neural Network, and how do they impact the forecasting results?
10. What are the key parameters affecting the performance of the time series models, and how do they influence the forecast accuracy?

1.5 Scope of the Study

This research aims to study retail sales forecasting and its importance in making decisions, managing inventory, and supply chain operations in the retail industry. The study will evaluate various forecasting methods, including time series analysis, statistical techniques, and deep learning models. It will examine how these methods can enhance business strategy, minimize operational costs, and improve customer satisfaction.

The research focuses on forecasting Walmart's sales by analyzing historical data. It explores various factors such as store types, sizes, departments, and holidays. The dataset includes information on store types, item sizes, dates of sale, temperature, fuel prices, holidays, etc.

1.6 Significance of the Study

Sales forecasting plays a crucial role in various industries, particularly in retail, for efficient inventory planning and business decision-making.

Accurate sales forecasting benefits businesses by improving liquidity and reducing operating costs. Controlling merchandise stock aids in warehouse management and customer satisfaction. The study explores different model techniques, providing insights into effective algorithms for sales prediction, and contributing to efficient inventory management.

The comparison between different models provides insights into their performance and limitations in the absence of additional features. By comparing and analyzing various methods, including traditional statistical approaches and modern deep learning techniques, the research aims to contribute valuable insights for improving the accuracy and efficiency of retail sales predictions.

The research highlights the variety of forecasting models available and the importance of choosing the best method for specific uses.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This literature review provides a comprehensive examination of the existing body of knowledge related to retail store sales prediction. By exploring the historical development, applications, and various approaches to predicting retail sales, this chapter aims to contextualize the research within the broader academic discourse and identify gaps that the study aims to address.

The review begins with the history of predicting retail store sales, tracing the evolution of methodologies from traditional statistical models to advanced machine learning and deep learning techniques. Understanding this historical progression is crucial for appreciating the advancements and limitations that have shaped current practices.

Next, the focus shifts to the practical applications of retail sales prediction. Accurate sales forecasting is integral to various business operations, including inventory management, supply chain optimization, and strategic planning. This section discusses how businesses leverage predictive analytics to enhance their decision-making processes and improve overall efficiency.

Different types of retail store sales predictions are then categorized, such as short-term versus long-term forecasts and aggregate versus individual item-level predictions. Each type of prediction has unique requirements and challenges, necessitating tailored approaches and methodologies.

The inherent challenges in predicting retail store sales are addressed, exploring factors such as seasonality, promotional events, market volatility, and consumer behavior. This section examines these challenges and the ways researchers and practitioners have attempted to overcome them.

A core part of the literature review examines the various techniques used for retail store sales prediction. From traditional time series analysis and statistical models to cutting-edge machine learning and neural networks, this section provides a detailed overview of the methodologies

employed in the field. Comparative analysis of these techniques' sheds light on their relative strengths and weaknesses.

The review also identifies the determinants of prediction success, emphasizing the importance of data quality, feature selection, model accuracy, and interpretability. Understanding these factors is essential for developing robust and reliable forecasting models.

The review then highlights related research publications, summarizing key findings from seminal and recent studies. This aims to situate the current research within the broader academic context and identify areas where further investigation is needed.

The discussion synthesizes the insights gained from the review, critically evaluating the state of research in retail sales prediction. This section highlights unresolved issues, emerging trends, and potential future directions for research.

By thoroughly examining the existing literature, this chapter aims to build a solid foundation for the current research, ensuring it is grounded in established knowledge while addressing significant gaps and challenges in the field.

2.2 Applications of Sales Prediction in Retail

Accurate sales forecasting is essential for retail businesses, as it provides insights that inform critical decision-making across various aspects of operations. This section explores how sales predictions impact retail businesses, focusing on key areas relevant to the thesis.

Operational Efficiency: Precise forecasts allow retailers to manage inventory effectively, reducing stockouts and overstocking, leading to lower storage costs and higher customer satisfaction. Accurate forecasting also optimizes capital usage for inventory replenishment, enhancing cash flow. Staff scheduling can be optimized to meet peak demand while minimizing labor costs.

Strategic Decision-Making: Sales forecasts enable businesses to make informed decisions about expansion strategies, promotional planning, and pricing strategies. By predicting demand, retailers can identify growth opportunities, maximize sales and profitability through targeted promotions, and adjust prices strategically to optimize revenue.

Supply Chain Optimization: Accurate sales forecasts are vital for manufacturers to align production with anticipated demand, ensuring efficient resource allocation. Forecasts also streamline transportation logistics by optimizing delivery schedules and minimizing costs.

E-commerce Security: While not directly related to forecasting, detecting fraudulent activity in e-commerce is essential for protecting revenue and maintaining customer trust. This involves identifying unusual patterns in customer behavior.

These applications highlight the critical role of accurate sales forecasting in driving success in the retail industry, from optimizing daily operations to making strategic business decisions.

2.3 Types of Retail Sales Prediction: A Taxonomy

This section provides a taxonomy of different approaches used for retail sales prediction, which are often categorized into two main branches:

2.3.1 Traditional Statistical Methods

Traditional time series analysis methods leverage historical sales data to identify patterns and trends for predicting future sales. These methods, often applied in retail to forecast product demand, assume that past patterns will continue into the future.

Exponential smoothing simplifies data by calculating weighted averages of past observations, making it suitable for forecasting products with stable demand, such as basic groceries. It requires careful selection of the smoothing parameter (alpha) to balance recent trends with historical patterns.

Time series decomposition breaks down data into its components (trend, seasonality, noise) for individual forecasting, useful for products with strong seasonal patterns (e.g., holiday merchandise). It requires selecting appropriate decomposition parameters, which can be computationally intensive.

ARIMA (Autoregressive Integrated Moving Average) models capture non-stationary patterns, useful for products with trends and seasonality (e.g., seasonal clothing). It involves determining the order of autoregressive, integrated, and moving average components.

Holt-Winters Seasonal Method is a specialized technique for capturing both trend and seasonality, making it particularly useful for forecasting products with strong seasonal fluctuations.

SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors) extends the ARIMA model by allowing the inclusion of external factors like weather or promotions. This model offers the potential to capture the influence of these factors on sales but requires careful selection of relevant variables and model parameters.

While computationally efficient, traditional time series methods might be less accurate than machine learning models for highly non-linear relationships or when dealing with complex retail environments.

2.3.2 Machine Learning and Deep Learning Methods

Machine learning and deep learning models offer a powerful approach for predicting retail sales, particularly for large and complex datasets. These techniques excel at identifying patterns and relationships that might be missed by traditional statistical methods.

Supervised Learning forms the basis for many of these models, where the algorithm learns from labeled data to associate input features with known sales values.

Regression Models are used to predict continuous values like sales volume:

Linear Regression:

- Algorithm: Fits a line to the data that minimizes the sum of squared errors.
- Strengths: Simple, easy to interpret, computationally efficient, and can be effective when the relationship between the variables is linear.
- Retail Applications: Suitable for forecasting sales of products with a relatively simple linear relationship with influencing factors, such as basic products with consistent demand.

Support Vector Machine (SVM):

- Algorithm: Aims to find the optimal hyperplane that separates data points into different classes or predicts a continuous value by maximizing the margin between the hyperplane and the data points.
- Strengths: Can effectively handle non-linear data, robust to outliers, and often performs well with smaller datasets.
- Retail Applications: Suitable for forecasting sales of products with complex relationships with influencing factors, such as seasonal products with fluctuating demand.

Gradient Boosted Regression Trees (GBRT):

- Algorithm: Sequentially builds multiple decision trees, each tree correcting the errors made by the previous trees. It uses gradient descent to minimize the overall error.
- Strengths: Very accurate, can handle high-dimensional data, robust to outliers and missing values, and often outperforms other methods in sales forecasting.
- Retail Applications: Widely used for forecasting sales of a variety of retail products, especially when dealing with complex relationships between variables and large datasets.

Random Forest (RF):

- Algorithm: Builds multiple decision trees on different random subsets of the data and then averages their predictions to make a final prediction.
- Strengths: High accuracy, robust to outliers, handles high-dimensional data well, and is less prone to overfitting compared to single decision trees.
- Retail Applications: Widely used for forecasting sales of a variety of retail products, particularly when dealing with complex relationships between variables and large datasets.

Deep Learning utilizes neural networks with multiple layers to learn intricate patterns and relationships from data.

Long Short-Term Memory (LSTM):

- Algorithm: LSTM uses internal gates to control the flow of information through the network, helping to address the vanishing gradient problem. It has a memory cell that stores information over longer periods.
- Strengths: Can learn long-term dependencies in time series, robust to noise and missing data, and often outperforms other RNNs for time series forecasting.
- Retail Applications: Excellent choice for forecasting sales of products with long-term seasonal patterns, complex cyclical demand, or those with historical trends, for example, fashion trends, holiday merchandise, or products with seasonal fluctuations.

Convolutional Neural Networks (CNNs):

- Algorithm: CNNs use convolutional filters to extract features from the data, learning patterns that might be missed by other models.
- Strengths: Can learn from local patterns in sequential data, effective for detecting trends and seasonality, and can handle large datasets efficiently.
- Retail Applications: Suitable for forecasting sales of products with shorter-term patterns or seasonality, for example, products with short shelf lives, promotional items, or those with rapid shifts in demand.

Attentive Neural Controlled Differential Equations (ANCDEs):

- Algorithm: It uses attention mechanisms to weigh different parts of the input data based on their relevance to the prediction task. It integrates attention into the framework of controlled differential equations to handle irregular time series.
- Strengths: Can learn from complex non-linear patterns, handles irregular time series well, and may be more interpretable than other deep learning models.
- Retail Applications: Promising for forecasting sales with unpredictable patterns or those affected by external factors, for example, sales affected by unexpected events like holidays or weather changes.

These machine learning and deep learning techniques offer a powerful arsenal for addressing the complex challenges of retail sales forecasting, providing potential for improved accuracy and more robust predictions.

2.4 Challenges in Retail Sales Forecasting

Retail sales forecasting presents a unique set of challenges due to the complexity of retail environments and the constantly evolving nature of customer behavior. These challenges can significantly impact the accuracy of forecasting models and necessitate the use of sophisticated techniques to overcome them. Here are some of the key challenges highlighted in my research.

2.4.1 Data-Related Challenges

Data Availability and Completeness

- **Limited Data Access:** Not all retailers have access to extensive and comprehensive supply chain data, making it challenging to build robust forecasting models. This is especially true for smaller businesses, as they might not have the resources or infrastructure to collect and manage large amounts of data.
- **Missing Data:** Even when data is available, it often suffers from missing values, either due to data entry errors, technical glitches, or incomplete records. Handling missing data effectively is crucial for building accurate forecasting models.
- **Data Silos:** Data is often spread across different systems and departments within a retail organization, leading to fragmented information and difficulty in integrating all relevant data for forecasting.

Data Quality

- **Outliers:** Outliers, or extreme values in the data, can significantly distort model training and prediction. Identifying and handling outliers appropriately is essential for accurate forecasting.
- **Inconsistent Data Formats:** Data might be collected and stored in different formats or with different units of measurement, making it difficult to combine and analyze effectively.
- **Data Noise:** Real-world data is often noisy, containing random fluctuations or errors that can obscure underlying patterns.

Data Imbalance

- **Uneven Distribution:** Sales data often exhibits imbalances, with some products selling significantly more than others. This can make it challenging to train models that accurately predict the sales of less popular items, as the model might be biased towards the majority class.
- **Long Tail Products:** Retailers often have a long tail of products that sell infrequently. This poses a challenge for forecasting, as there might be limited historical data available for these products.

Incomplete Information

- **Hidden Factors:** Retailers often face situations where they lack complete information about factors influencing sales, such as competitor actions, changing consumer preferences, or emerging trends. For example, a new competitor entering the market might impact sales but not be immediately captured in the data.
- **Unforeseen Events:** Unexpected events like natural disasters, economic crises, or political changes can significantly disrupt sales patterns and make accurate predictions challenging.
- **Limited Customer Data:** Retailers might have limited access to customer-level data, such as demographics, purchase history, or online behavior. This can restrict the ability to develop more personalized forecasting models.

2.4.2 Model Complexity and Limitations

Nonlinearity

- **Complex Relationships:** Retail sales often exhibit non-linear relationships, meaning that simple linear models like linear regression might not be able to capture the complex patterns in sales data. For example, demand for a product might not increase linearly with price reductions; it could plateau or even decrease at a certain point.
- **Seasonality and Cyclical Patterns:** Retail sales are often affected by seasonal cycles (e.g., holiday shopping, weather changes) or cyclical patterns, which introduce non-linearity.

- **Promotional Effects:** Sales promotions can have non-linear effects on demand, with stockpiling effects and post-promotional dips that are not easily captured by linear models.

Long-Term Dependencies

- **Trend Identification:** Accurately forecasting sales over long periods requires models that can effectively capture long-term trends and seasonal patterns, especially for products with a long lifecycle or those affected by long-term economic cycles.
- **Limited Memory:** Traditional statistical models might struggle to capture these long-term relationships, while some machine learning models, like LSTMs, are specifically designed to handle long-term dependencies in sequential data.

External Factors

- **Economic Fluctuations:** Economic factors like inflation, unemployment, and consumer confidence can significantly impact overall retail sales. These fluctuations might not be directly reflected in sales data, making it challenging to account for them in forecasting models.
- **Competitor Activity:** Competitor actions, like price changes, new product launches, or marketing campaigns, can significantly affect sales patterns, making it essential to incorporate competitor data into forecasting models.

2.4.3 Model Selection and Interpretation Challenges

Conflicting Results

- **Model Suitability:** Retailers often face the challenge of choosing the most appropriate forecasting model from a range of options, as different models can yield varying and even conflicting results. This can be especially challenging when dealing with complex data or when the underlying relationships between variables are not fully understood.
- **Data-Specific Model Choice:** A model that performs well for one product or category might not be optimal for another. Selecting the best model often requires experimentation and careful consideration of data characteristics.

Model Interpretability

- **Black Box Models:** For businesses to trust and act on model predictions, it is crucial to understand how the models arrive at their forecasts. This is particularly challenging for complex models like deep neural networks, where the decision-making process can be opaque.
- **Transparency:** A lack of transparency in how models make predictions can hinder adoption and decision-making. Businesses need to understand the model's logic to trust its outputs and make informed decisions.

2.4.4 Overcoming the Challenges

To address these challenges, researchers have explored several strategies:

- **Advanced Models:** Deep learning models like LSTM, CNN, and hybrid architectures are increasingly used to capture complex patterns and handle large datasets.
- **Feature Engineering:** Creating new features from existing data (e.g., incorporating seasonality, promotional periods, or economic indicators) can enhance model accuracy.
- **Ensemble Methods:** Combining multiple models (e.g., using different deep learning architectures or combining statistical and machine learning models) can improve model performance and robustness.
- **Data Preprocessing:** Techniques like data cleaning, outlier removal, and feature scaling are crucial for preparing data for model training.
- **Domain Expertise:** Incorporating domain expertise from retail professionals can be invaluable in selecting the right models, interpreting results, and identifying key factors that might not be readily apparent in the data.

2.5 Techniques used for Retail Sales Prediction: A Methodology Overview

This section provides a comprehensive review of various methodologies employed for retail sales prediction, spanning both traditional statistical methods and cutting-edge machine learning and deep learning approaches. These techniques are categorized into two main branches as mentioned below.

2.5.1 Traditional Statistical Methods

Traditional time series analysis methods rely on statistical principles and historical sales data to forecast future sales, assuming that past patterns will continue. They are often suitable for scenarios with relatively stable data patterns and a moderate amount of data.

Time Series Analysis:

- Exponential Smoothing: This method calculates a weighted average of past observations, giving more weight to recent data.
- Model Building: The smoothing parameter (alpha) controls the balance between recent and historical data.
- Key Parameters: Alpha (α), Beta (β) for trend smoothing, and Gamma (γ) for seasonal smoothing.

ARIMA (Autoregressive Integrated Moving Average):

- Model Building: Combines autoregressive, integrated, and moving average components.
- Key Parameters: p (autoregressive order), d (differencing order), q (moving average order), and seasonal parameters (P, D, Q).

SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors):

- Model Building: Extends ARIMA to incorporate exogenous variables that influence sales.
- Key Parameters: p, d, q, P, D, Q from ARIMA and the selection and transformation of relevant exogenous variables.

These statistical models are computationally efficient and can be useful for capturing trends and seasonality in retail sales. However, they might be less accurate than machine learning models when dealing with highly non-linear patterns or when external factors are heavily influencing sales.

2.5.2 Machine Learning and Deep Learning Methods

Machine learning and deep learning models utilize powerful algorithms to learn complex patterns within data, potentially offering greater accuracy and flexibility, especially when handling large datasets.

Supervised Learning: This category involves training models on labeled data, where the model learns to associate input features with a known output.

Regression Models: These models predict a continuous value, like sales volume.

Linear Regression:

- Model Building: Fits a line to the data that minimizes the sum of squared errors.
- Important Parameters:
    - Coefficients (β): Represent the strength of the linear relationship between each independent variable and the dependent variable.
    - Intercept (α): Represents the value of the dependent variable when all independent variables are equal to zero.

Support Vector Machine (SVM):

- Model Building: Aims to find the optimal hyperplane that separates data points into different classes or predicts a continuous value by maximizing the margin between the hyperplane and the data points.
- Important Parameters:
    - Kernel Function: Determines how the data is transformed into a higher-dimensional space.
    - Regularization Parameter (C): Controls the balance between maximizing the margin and minimizing the classification error.
    - Gamma (γ): For non-linear kernels, controls the influence of individual data points.

Gradient Boosted Regression Trees (GBRT):

- Model Building: Sequentially builds multiple decision trees, each tree correcting the errors made by the previous trees. It uses gradient descent to minimize the overall error.
- Important Parameters:
    - Learning Rate: Controls the step size for each tree.
    - Number of Trees: Determines the total number of trees.
    - Maximum Depth: Limits the depth of each decision tree, helping to prevent overfitting.
    - Subsampling Ratio: Controls the proportion of data used for training each individual tree.

Random Forest (RF):

- Model Building: It builds multiple decision trees on different random subsets of the data and then averages their predictions to make a final prediction.
- Important Parameters:
    - Number of Trees: Determines the number of decision trees.
    - Maximum Depth: Limits the depth of each decision tree.
    - Minimum Samples per Leaf: Specifies the minimum number of data points required to create a leaf node in a decision tree.

Bootstrap Aggregation (Bagging): Randomly samples data with replacement to create multiple training sets for individual trees.

Deep Learning: These methods leverage complex artificial neural networks with many layers to learn intricate patterns and relationships from data. They are often preferred for large datasets and tasks requiring sophisticated pattern recognition.

Recurrent Neural Networks (RNNs):

- Model Building: Involves defining the number of hidden layers, the number of neurons per layer, and the activation functions.
- Important Parameters:
    - Hidden Layers: The number of layers in the network.
    - Neurons per Layer: Determines the number of neurons in each hidden layer.

- Activation Functions: Controls the non-linearity of the network.

Long Short-Term Memory (LSTM):

- Model Building: Involves defining the number of LSTM layers, the number of neurons per layer, and the activation functions.
- Important Parameters:
  - LSTM Layers: The number of LSTM layers.
  - Neurons per Layer: Determines the number of neurons in each LSTM layer.
  - Activation Functions: Controls the non-linearity of the network.

Convolutional Neural Networks (CNNs):

- Model Building: Involves defining the number of convolutional layers, the size of the convolutional filters (kernels), the pooling layers, and the fully connected layers.
- Important Parameters:
  - Convolutional Filters: Specifies the number and size of convolutional filters.
  - Pooling Layers: Used to reduce the dimensionality of the data.
  - Fully Connected Layers: Connect the output of the convolutional layers to the prediction layer.

Attention-Based Methods: Attentive Neural Controlled Differential Equations (ANCDEs):

- Model Building: Involves defining the number of layers, the number of neurons, the activation functions, and the attention mechanisms used.
- Important Parameters:
  - Attention Mechanism: The choice of attention mechanism.
  - Differential Equation: The type of differential equation used to model the temporal dynamics of the data.

2.5.3 Data Preprocessing

Data Cleaning: This involves removing irrelevant or inaccurate data points from the dataset. This might include addressing missing values, outliers, and duplicate records.

Feature Selection: Selecting the most relevant features (independent variables) from the data can significantly improve model performance. Feature selection involves identifying variables that strongly correlate with the target variable (sales).

Feature Engineering: This involves creating new features from existing data, such as:

Lagged Features: Creating features that represent past values of the target variable (e.g., sales from previous weeks or months).

Seasonal Features: Adding features that encode seasonal patterns, like holiday dates or months of the year.

External Features: Incorporating external data sources, such as economic indicators, weather data, or competitor activity.

2.5.4 Model Evaluation: Ensuring Accuracy and Reliability

K-Fold Cross-Validation: This technique involves dividing the data into multiple folds, using some folds for training and others for testing. This process is repeated multiple times with different folds, ensuring a more robust evaluation of the model's performance.

Holdout Validation: This method involves splitting the data into a training set and a separate testing set. The model is trained on the training set and then evaluated on the unseen testing set.

Error Metrics: Common error metrics used to assess the accuracy of forecasting models include:

- Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values.
- Root Mean Squared Error (RMSE): Calculates the square root of the average of squared errors.
- Mean Absolute Percentage Error (MAPE): Measures the average percentage error between predicted and actual values.

2.6 Determinants of Prediction Success in Retail Sales Forecasting

This section delves into the critical factors that influence the accuracy of retail sales forecasting models. Achieving reliable predictions requires careful consideration of both data-centric and model-centric aspects.

Data Quality: Accurate forecasting relies on high-quality data that is clean, relevant, representative, and up to date. Missing values, inconsistencies, outliers, and outdated data can significantly compromise model performance.

Data Availability: Access to sufficient and diverse data is crucial, particularly for complex models like deep learning networks. Including data sources such as customer demographics, competitor activity, economic indicators, and weather data can improve model accuracy.

Model Selection: Choosing the right algorithm and model architecture based on the characteristics of the sales data and the desired forecasting accuracy is essential. Simple models might suffice for stable data, while complex models are needed for highly dynamic or non-linear data patterns.

Feature Engineering: Careful feature selection and transformation, including the inclusion of relevant variables (e.g., product category, price, promotions, seasonality) and the appropriate handling of categorical data, can significantly improve model effectiveness.

Model Optimization: Tuning hyperparameters within the chosen model architecture and employing robust evaluation techniques like k-fold cross-validation are essential for optimizing model performance and ensuring that the model generalizes well to new data.

By carefully addressing these data-centric and model-centric factors, researchers and practitioners can significantly enhance the accuracy and reliability of retail sales forecasting models.

2.7 Research Gaps and Opportunities in Retail Sales Forecasting

While the research discussed in this review has advanced the field of sales prediction, there remain significant gaps and opportunities for further investigation, particularly in the context of retail sales forecasting. Here are some key areas where more research is needed, including your thesis topic as a compelling research opportunity:

2.7.1 Handling Complexities of Retail Environments

- **Impact of Promotions:**
    - Post-Promotional Period Forecasting: Most research focuses on the promotional period itself but understanding the post-promotion effect on sales (e.g., stockpiling, dips) is crucial for effective inventory management.
    - Promotional Strategy Optimization: Research needs to explore how to use forecasting to optimize promotional strategies, considering product selection, pricing, and timing to maximize sales and minimize negative post-promotion effects.
- **Seasonal and Cyclical Patterns:**
    - Beyond Simple Seasonality: Many models handle simple seasonality, but retail faces complex cyclical patterns (e.g., fashion trends) that require advanced models to capture effectively.
    - Integrating External Factors: Research needs to examine how to integrate external factors, like weather data or economic indicators, to improve forecasting accuracy, especially for products impacted by seasonal fluctuations.
- **Data Limitations:**
    - Limited Historical Data: Forecasting products with short lifecycles or those introduced recently can be challenging due to the limited availability of historical sales data. New methods need to be developed for these scenarios.
    - Data Scarcity for Smaller Retailers: Smaller retailers might not have access to large amounts of data, making it difficult to train advanced models. Research into data augmentation techniques and more data-efficient models is needed.

2.7.2 Leveraging Deep Learning for Retail

- **Beyond Univariate Forecasting:**
    - Multivariate Modeling: Univariate time series forecasting is prevalent, but incorporating multiple variables (e.g., pricing, promotions, economic indicators, competitor activity) into multivariate models can significantly improve accuracy.

- Hybrid Deep Learning Architectures: Research into hybrid models that combine the strengths of different deep learning architectures, such as CNNs and LSTMs, is promising for capturing both local and global patterns in sales data.
- Explainability and Interpretability:
- Black Box Models: Deep learning models often lack transparency, making it difficult for retailers to understand how the models make predictions. Developing methods to interpret and explain deep learning models is crucial for building trust and enabling informed decision-making.

- **Real-time Forecasting:**
  - Dynamic Updates: Research on real-time forecasting systems that can dynamically update predictions as new data becomes available is critical for quickly adapting to changing market conditions and ensuring timely decisions.

2.7.3 Enhancing Sales Prediction Accuracy: A Research Opportunity

Evaluating and Selecting Optimal Models: Currently, retailers face the challenge of choosing the best model from a variety of options, often lacking objective comparisons and guidance. This research will contribute significantly to this field by providing a rigorous evaluation of various models and establishing guidelines for selecting the most effective approach based on specific retail contexts.

2.7.4 Integrating Other Considerations

- Model Ensemble Methods: Investigating the effectiveness of ensemble methods, which combine predictions from multiple models to improve accuracy and robustness, can lead to more reliable forecasts.
- Dynamic Pricing: Exploring the relationship between dynamic pricing strategies and sales forecasting is essential for retailers looking to optimize pricing based on real-time demand.

2.8 Discussions

This section summarizes the key findings from the research reviewed and outlines promising future research directions, particularly those that are most relevant to our research topic.

Key Findings from the Research

Deep Learning's Potential: Several studies demonstrate the potential of deep learning models, including LSTM, CNN, and hybrid architectures, to achieve higher accuracy in sales forecasting compared to traditional statistical methods.

Importance of Feature Engineering: Across various research, incorporating relevant features like store size, product category, promotional periods, and economic indicators significantly improves model performance, especially for complex sales patterns.

Benefits of Ensemble Methods: Combining predictions from multiple models (e.g., different deep learning models or a mix of statistical and machine learning approaches) can enhance robustness and accuracy.

Challenges of Post-Promotional Forecasting: The post-promotional period often sees a decline in sales, and models need to be able to capture this effect accurately. Research indicates that incorporating "promotional period" as a feature can significantly improve predictions during this time.

Handling Time Series Data: The use of specific time series methods like ARIMA and LSTM has proven effective for capturing long-term dependencies and cyclical patterns in sales data.

Future Research Directions

Improving Model Interpretability: Deep learning models can be "black boxes," making it difficult to understand why they make certain predictions. Future research should focus on developing techniques to improve the interpretability and explainability of deep learning models, enabling better trust and informed decision-making.

Advanced Feature Engineering for Retail: Exploring more sophisticated feature engineering techniques to capture complex relationships between sales and influencing factors (e.g.,

seasonality, promotions, competitor activity) is crucial. This could involve incorporating data from diverse sources, like customer demographics, social media trends, or economic indicators.

Developing Real-Time Forecasting Systems: Researching real-time forecasting systems that can dynamically update predictions as new data becomes available is essential for retailers to react quickly to changing market conditions and optimize their operations.

Addressing Data Imbalance: Exploring techniques for handling data imbalances, such as oversampling, under sampling, or using cost-sensitive learning algorithms, can significantly improve the accuracy of predictions for less popular or infrequent products.

2.9 Summary of Research Findings

This section provides a concise overview of the key findings from the research reviewed. The research highlights the growing role of machine learning and deep learning in enhancing the accuracy of sales forecasting, particularly in retail:

Machine Learning in Retail: Multiple papers demonstrate the effectiveness of machine learning models, including Random Forest, Gradient Boosted Regression Trees (GBRT), and Support Vector Machines (SVM), for forecasting retail sales. These models offer advantages over traditional statistical methods in handling complex data patterns and achieving higher accuracy, particularly when dealing with large datasets.

Deep Learning for Retail: Deep learning models, such as LSTM, CNN, and hybrid architectures, have proven effective in capturing complex temporal relationships in sales data, handling seasonal variations, and improving forecasting accuracy, particularly for long-term predictions. The use of attention mechanisms further enhances these models by focusing on the most relevant information within the data.

Importance of Data Preprocessing: Many research papers underscore the importance of data preprocessing, including data cleaning, feature engineering, and handling missing values, to ensure the quality and relevance of data used for model training.

Research Gap: However, several studies highlight the need for more comprehensive research on comparing and evaluating different models (statistical, machine learning, and deep learning) in real-world retail contexts.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter lays out the methodological foundation for this study, which seeks to enhance sales prediction accuracy for a chain of retail stores by comparing and evaluating the performance of machine learning models, neural networks, and time series models. The primary goal is to establish a comprehensive framework for selecting the optimal model for retail sales forecasting, considering diverse data characteristics and specific retail challenges.

We will employ a structured research approach that includes these key elements:

Research Methodology: This section outlines the research design, data analysis techniques, and model evaluation methods used to ensure rigor and credibility.

Data Selection: We will describe the data sources used for this study, highlighting their relevance to retail sales and the features they contain.

Data Pre-processing & Transformation: This section explains how we address data quality issues, such as missing values and outliers, to prepare the data for model training and evaluation.

Interactive Visual Analytics: We will discuss how interactive visual analytics techniques help us to gain insights from the data, understand patterns, and explore model performance.

Interpretation/Evaluation: This section defines the evaluation criteria and methods used to assess model accuracy, including metrics and statistical tests to compare different models objectively.

Below is the Research Pipeline for our comparative analysis of model's performance for retail sales prediction.

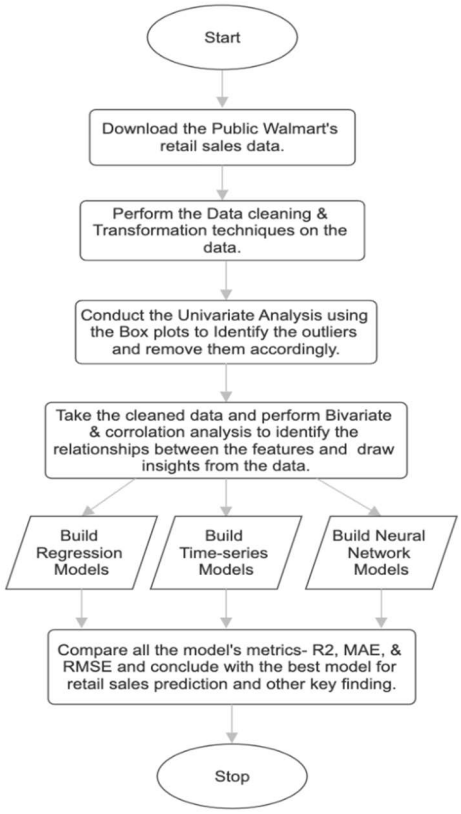


Figure 1: Research Pipeline for comparative analysis of retail sales prediction.

Summary: The final section provides a concise overview of the methodologies employed in this chapter, emphasizing their contribution to the overall research objectives.

3.2 Research Methodologies

The research design follows a comparative analysis approach. This involves building and evaluating multiple sales prediction models from different categories. Using EDA, we will be also drawing a few inferences.

3.2.1 Regression Models:

- **Linear Regression:** A basic model that establishes a linear relationship between independent variables and a dependent variable.
- **LASSO Regression:** A linear regression technique that performs regularization to reduce the number of variables used in the model.

- **Elastic Net:** Combines L1 and L2 regularization to balance bias and variance.
- **RIDGE Regression:** A linear regression technique that uses L2 regularization to prevent overfitting.
- **KNN Regression:** A non-parametric method that predicts a value based on the average of the "k" nearest neighbors in the training data.
- **SVR Regression (Support Vector Regression):** A non-linear regression technique that uses support vectors to build a model for prediction.
- **XG Boost Regression:** A powerful ensemble method that combines multiple decision trees using gradient boosting.
- **Random Forest Regression:** An ensemble method that combines multiple decision trees to improve prediction accuracy.
- **LG Boost Regression:** A gradient boosting algorithm known for its speed and accuracy.
- AdaBoost Regression: An ensemble method that sequentially builds models, each focusing on correcting errors made by previous models.

The choice of regression models in this study was driven by their ability to handle the complexities of retail sales data and address common challenges in forecasting:

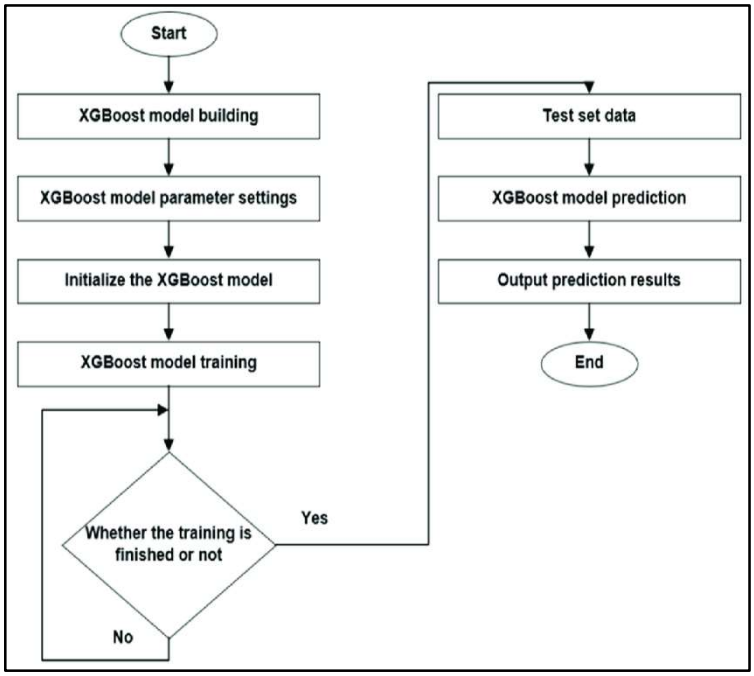


Figure 2: XG Boost model prediction flowchart

Linear Regression: This simple model served as a baseline to understand the basic linear relationships within the data. It provided a starting point for comparison with more complex models.

Regularized Linear Models (Lasso, ElasticNet, Ridge): These models were included to address overfitting and reduce the impact of irrelevant features, which can be common in high-dimensional datasets like retail sales data.

Non-linear Models (SVM, KNN): We included models like SVM and KNN to investigate non-linear relationships within the data, as retail sales are often influenced by complex factors that are not easily captured by linear models.

Ensemble Methods (XGBoost, Random Forest, LightGBM, AdaBoost): These powerful ensemble methods were chosen because they often achieve higher accuracy by combining multiple decision trees. They are particularly well-suited for handling large, complex datasets and reducing overfitting.

This selection strategy aimed to explore a range of model types, from simple linear models to more sophisticated ensemble methods, to identify those best suited for accurately predicting retail sales and handling the specific challenges of this type of data.

3.2.2 Time Series Models

ARIMA (Autoregressive Integrated Moving Average): A traditional statistical model that captures trends and seasonality in time series data.

SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors): Extends ARIMA to incorporate external factors that influence sales.

Auto ARIMA: A method that automatically identifies the best ARIMA parameters (p, d, q) for a given time series.

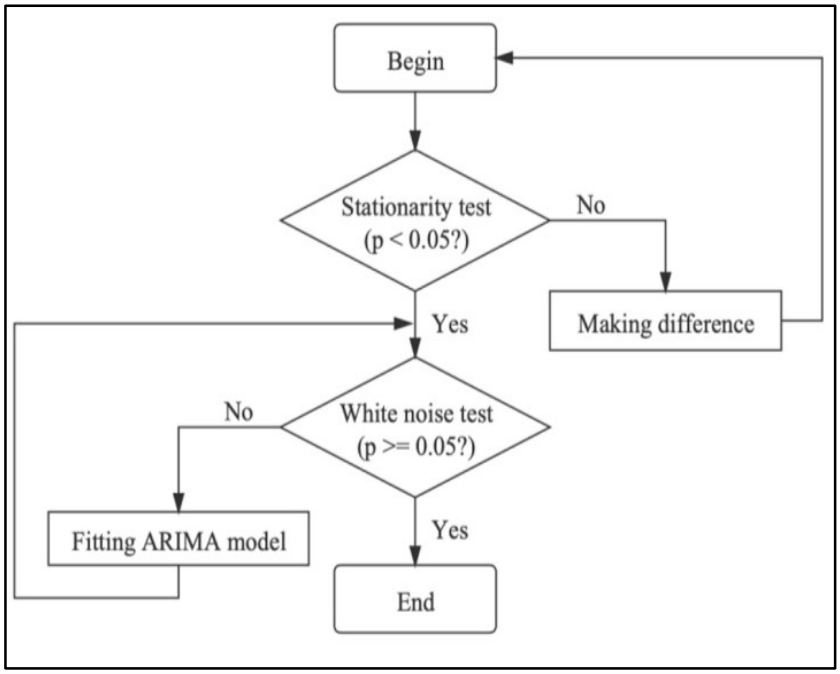


Figure 3: Flowchart of the ARIMA Model.

The selection of ARIMA, SARIMAX, and Auto ARIMA was driven by their ability to handle the specific characteristics of time series data prevalent in retail sales forecasting:

ARIMA: This widely used model captures trends and seasonality, providing a baseline for comparison.

SARIMAX: By allowing for the inclusion of external factors, SARIMAX addresses the need to incorporate variables like promotions and economic indicators that influence retail sales.

Auto ARIMA: Automating parameter selection through Auto ARIMA streamlined the process and potentially identified more optimal settings for the ARIMA model.

These models were chosen to explore both traditional statistical methods and more advanced approaches capable of incorporating external factors, offering a comprehensive evaluation of time series techniques for retail sales forecasting.

3.2.3 Deep Learning Models

Artificial Neural Network (ANN): A basic neural network architecture with multiple layers.

Tuned ANN: An ANN that has been optimized for the specific task through hyperparameter tuning.

LSTM Bidirectional: A variant of LSTM that processes data in both forward and backward directions, potentially capturing more complex patterns in time series data.

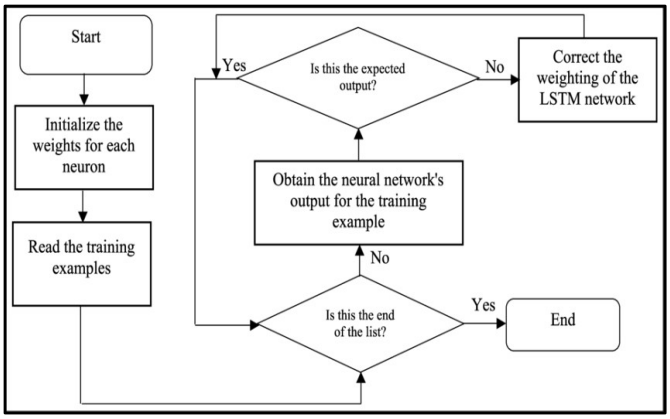


Figure 4: Flowchart of the LSTM Model.

The chosen neural network models were selected to explore the potential of deep learning for capturing complex patterns and temporal dependencies within retail sales data:

ANN (Artificial Neural Network): This basic architecture served as a foundation, providing a baseline for comparison with more sophisticated models. It helped demonstrate the potential of neural networks for time series forecasting.

Tuned ANN: By hyperparameter tuning the ANN, we sought to optimize its performance for this specific dataset, potentially improving its accuracy and addressing challenges like overfitting.

LSTM Bidirectional: Bidirectional LSTM was chosen to explore its ability to capture more intricate temporal dependencies in sales data by processing the data in both forward and backward directions. This approach is particularly valuable for time series with complex patterns.

This selection aimed to assess the effectiveness of different neural network architectures, from basic to more sophisticated models, to identify those that best suited for handling the complex dynamics of retail sales data.

3.3 Data Selection

We carefully selected the datasets that is representative of retail sales data and contain features relevant to our research objectives. For this research, I have selected Walmart's Retail Sales data. The datasets will include information such as:

Historical Sales Data: This will form the foundation of our models, providing insights into past sales trends and patterns.

External Factors: We will consider external data sources like economic indicators, weather data, or competitor activity to improve model accuracy and capture wider context.

The Key features from the dataset are Promotions, CPI index of the location, Fuel price, Temperature, Unemployment, Dept, Store size & Type of the store. The data processing will be applied to these key features and train them on all the models.

3.4 Data Pre-processing & Transformation

Raw sales data often needs to be cleaned, pre-processed, and transformed before it can be used for model training. This involves:

Data Cleaning: Removing irrelevant data, handling missing values (e.g., imputation or deletion), and addressing outliers.

Feature Engineering: Creating new features from existing data to improve model performance, such as: date columns, rolling sales, and more.

- **External Features:** Adding economic indicators, weather information, or competitor activity data.
- **Data Transformation:** Scaling or normalizing features to improve model training and reduce bias.

3.5 Interactive Visual Analytics

We will utilize interactive Business analytics tool like power BI to explore the data, gain insights into patterns, and assess model performance. This will involve:

- **Data Visualization:** Using graphs, charts, and other visualizations to understand data relationships and identify key features.
- **Interactive Tools:** Employing interactive visualization tools that allow for dynamic exploration of data and model performance.

3.6 Model Evaluation and Selection

We will use a combination of metrics to evaluate the performance of different models:

- **Error Metrics:** We will measure prediction accuracy using metrics like MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R Squared.
- Root Mean Squared Error (RMSE) is a metric used to evaluate the accuracy of time series forecasting models. It measures the average difference between the predicted and actual values.

  Formula: $RMSE = \sqrt{[\Sigma(y_i - \hat{y}_i)^2 / n]}$
  Where: n is the number of data points
  y_i is the actual value of the i-th data point
  $\hat{y}_i$ is the predicted value of the i-th data point

- Mean Absolutes Error (MAE): MAE measures the average absolute difference between the predicted values and the actual values. It essentially calculates the average magnitude of the errors, regardless of their direction (positive or negative).

  Formula: $MAE = (1/n) * \Sigma|y_i - \hat{y}_i|$
  where: n is the number of data points
  y_i is the actual value of the Ith data point
  $\hat{y}_i$ is the predicted value of the Ith data point

- R-squared (R^2): R-squared (R^2) is a statistical measure that represents the proportion of the variance in the dependent variable (the variable you are trying to predict) that is explained by the independent variables (the predictor variables) in your model.

  Formula: $R^2 = 1 - (SSR / SST)$

where: SSR is the sum of squared residuals (the difference between the predicted values and the actual values) and SST is the total sum of squares (the difference between each actual value and the mean of all actual values).

- **Cross-Validation:** We will use k-fold cross-validation to assess the generalization ability of models and ensure robustness to different data splits.

3.7 Summary

This research aims to compare the accuracy of sales predictions for a chain of retail stores by evaluating various machine learning and deep learning models. We'll be building and testing a diverse range of models, including regression models like linear regression, SVM, random forest, and XGBoost; ensemble methods like bagging, boosting, and stacking; time series models like ARIMA, SARIMAX, and Auto ARIMA; and deep learning models like LSTM Bidirectional and various neural network architectures.

To ensure a robust analysis, we will carefully select datasets representative of retail sales data, incorporating historical sales, product information, and relevant external factors like economic indicators and employment data. We will then meticulously clean and transform the data to address any quality issues, including missing values, outliers, and inconsistencies. We'll use interactive visual analytics to explore data patterns and gain insights.

To evaluate model performance, we will utilize a combination of error metrics (MAE, RMSE, MAPE) and employ k-fold cross-validation to assess the generalization ability of models. Finally, a univariate analysis will be conducted to understand individual features and their relationships with sales.

This comprehensive methodology will enable us to objectively compare the models and ultimately identify the optimal model for predicting retail sales, considering both data characteristics and real-world retail challenges.

CHAPTER 4

ANALYSIS

4.1 Introduction

This chapter delves into the core analysis of this research, detailing the methodology and procedures used to investigate the effectiveness of various machine learning, neural networks, and time series models in a retail setting. The overarching aim is to evaluate and select the optimal model for predicting retail sales, considering the unique challenges and complexities of the retail environment.

This chapter outlines a structured approach to data analysis and model development, beginning with the preparation and transformation of the chosen dataset. We will then conduct a comprehensive examination of the data's characteristics through univariate analysis, outlier detection, and interactive visualization using Power BI. This detailed exploration of the dataset will provide insights into data patterns.

The subsequent sections delve into the implementation of different prediction techniques:

Regression & Ensemble Models: We will explore the performance of established machine learning regression models, including Support Vector Machine (SVM), Gradient Boosted Regression Trees (GBRT), and Random Forest, to predict sales based on the identified features.

Time Series Models: We will examine the applicability of traditional time series models like ARIMA and Prophet, focusing on their ability to leverage historical sales patterns and seasonal trends.

(ANN) & (LSTM) Networks: The chapter will investigate the potential of deep learning models, including ANN and LSTM, to capture complex non-linear relationships and long-term dependencies in sales data.

To optimize model performance, we will conduct experiments exploring:

Hyperparameter Tuning: We will systematically adjust model parameters to identify the configurations that achieve the highest accuracy for each model.

Feature Engineering: We will explore the effectiveness of creating new features from existing data to improve the model's ability to capture relevant patterns.

This chapter provides a detailed account of the data analysis and model building process, laying the foundation for the evaluation and comparison of different models in the subsequent chapters. It aims to contribute to a deeper understanding of the capabilities and limitations of different modeling approaches for retail sales forecasting.

4.2 Dataset Preparation & Transformation

The procedures undertaken to prepare and transform the Walmart retail sales dataset for use in model development. The dataset, obtained from Kaggle, spans three years (2010-2012) and consists of three primary components: store data, feature data, and sales data.

Initially, the data was comprised of three separate tables, each containing distinct information. To create a comprehensive dataset suitable for analysis, we performed a merge operation, joining the tables on the "Store" column as the primary key. This resulted in a single dataset encompassing all relevant information for sales forecasting.

The final dataset includes the following features:

- Store: Unique identifier for each retail store.
- Dept: Department within the store (e.g., clothing, electronics).
- Date: Date of the sales transaction.
- Weekly_Sales: Total weekly sales for the store and department.
- IsHoliday: Indicator variable denoting whether the week contains a holiday.
- Type: Type of store (e.g., "A", "B", "C").
- Size: Size of the store (in square feet).
- Temperature: Average temperature for the week.
- Fuel Price: Average fuel price for the week.
- MarkDown1-5: Promotional markdown values for various events (potential missing values).
- CPI: Consumer Price Index for the region.
- Unemployment: Unemployment rate for the region.

Data Cleaning and Transformation

Handling Missing Values: The "MarkDown" columns, representing promotional markdown values, contained missing values. These missing values were replaced with zeros, assuming that a lack of promotional activity translates to zero markdown.

Data Type Conversion: The "IsHoliday" and "Size" columns were converted to categorical and integer data types, respectively, to improve data representation and model compatibility.

Feature Engineering: New features were extracted from the "Date" column, including "Year", "Month", "Day of the Week", "Day", and "IsWeekend". This enriched the dataset with additional temporal information relevant for capturing seasonal patterns and day-of-week effects.

These data preparation and transformation steps were crucial for ensuring data consistency, accuracy, and suitability for building and evaluating forecasting models. This cleaned and enriched dataset provided a strong foundation for the analysis and model development process, ensuring that the models were trained on reliable and relevant data.

4.3 Understanding the Dataset Attributes

This section provides an analysis of the attributes within the Walmart retail sales dataset, revealing key characteristics that will inform model development and analysis. The dataset's structure and attributes are carefully examined to gain a comprehensive understanding of its potential for sales forecasting.

4.3.1 Categorical Variables

IsHoliday: This categorical variable represents whether a particular week includes a holiday. The dataset shows a balanced distribution of holiday and non-holiday weeks, indicating that there is sufficient data to analyze the impact of holidays on sales.

Type: This categorical variable categorizes stores into three types ("A", "B", "C"). While categories "A" and "B" exhibit a balanced distribution, category "C" appears slightly skewed, suggesting that these stores might have different sales patterns or characteristics that require further consideration during model building.

4.3.2 Temporal Variables

Date: The "Date" column, along with its derived features, provides a rich temporal dimension for analyzing sales patterns.

Day of Week and Day: As the sales data reflects weekly sales aggregated on the same day each week, the "Day of Week" and "Day" columns contain a single value, indicating that the dataset focuses on analyzing weekly sales patterns. These columns are therefore not considered for model training due to their limited variability.

4.3.3 Key Dataset Characteristics

Store and Department Diversity: The dataset encompasses data from 45 stores across 65 departments, providing a comprehensive view of sales patterns across various retail locations and product categories.

Promotional Activity: The inclusion of five promotional markdown features ("MarkDown1" to "MarkDown5") suggests the dataset can be used to understand the impact of promotions on sales.

Economic Indicators: The dataset includes economic indicators like the Consumer Price Index (CPI), Unemployment rate, and Fuel Price. These variables provide a broader context for understanding the economic factors that might influence sales trends.

This understanding of the dataset's attributes will help us for the subsequent analysis and model development. By identifying the nature and distribution of different variables, we can select appropriate models and features to capture the underlying relationships and dynamics that drive sales patterns within the retail environment.

4.4 Univariate Analysis and Outlier Detection

This section details the initial exploratory data analysis, focusing on univariate analysis and outlier detection, which are essential steps to prepare the dataset for further modeling and analysis.

Univariate Analysis:

To gain a preliminary understanding of the distribution and characteristics of individual numerical variables, we conducted a univariate analysis using boxplots. These boxplots provided a visual representation of the central tendency, spread, and potential outliers for each variable.

Outlier Identification and Treatment:

Univariate analysis revealed the presence of outliers in the "Weekly Sales", "MarkDown1-5", and "Unemployment" variables. Outliers represent data points that deviate significantly from the expected range of values, potentially skewing statistical analysis and model training.

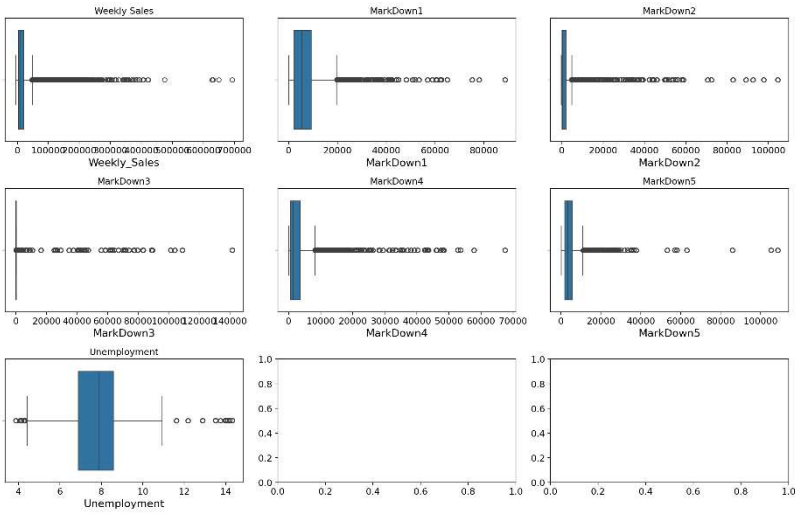


Figure 5: Before Outlier Treatment

To address these outliers, a robust outlier treatment strategy was implemented, focusing on retaining data points within a reasonable range while preserving the essential information within the dataset. We removed outliers that fell outside two standard deviations from the mean, capturing approximately 95% of the data distribution. This approach ensured that the remaining data accurately represented the typical sales patterns and economic conditions while minimizing the influence of extreme values.

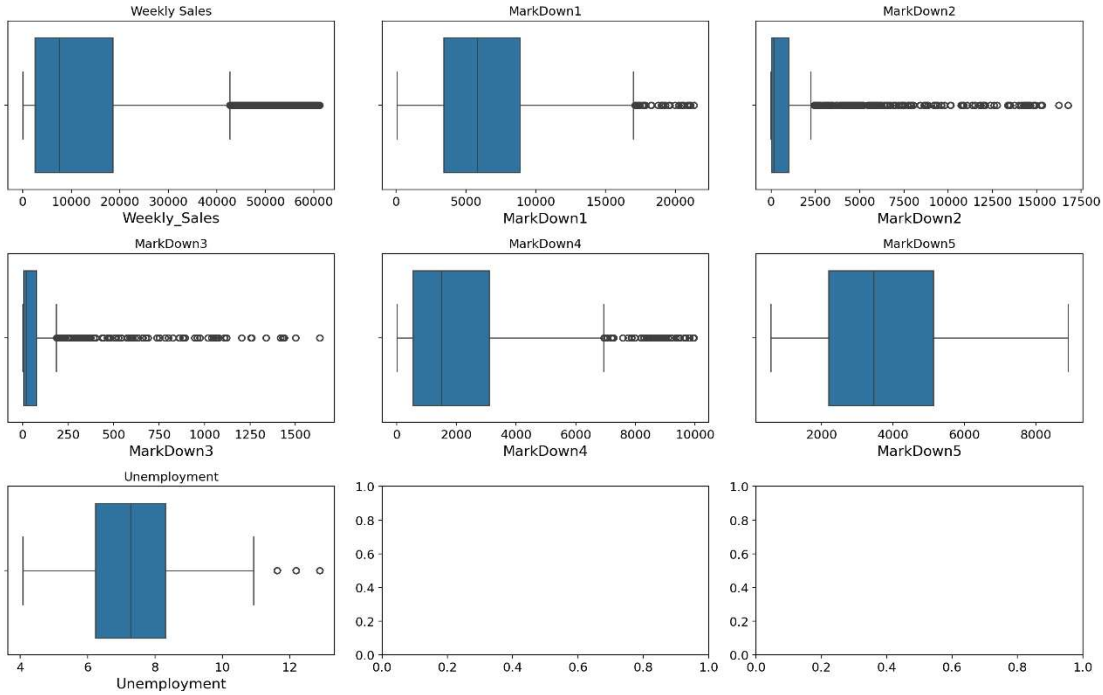


Figure 6: After Outlier Treatment

Data Preparation for Further Analysis:

This outlier treatment process resulted in a cleaner dataset, more suitable for subsequent analysis and model building. The removal of outliers enhanced the reliability and accuracy of both exploratory data visualization in Power BI and model training, leading to more robust insights and improved predictive power.

4.5 Unveiling Relationships: Data Visualization with Power BI

After diligently preparing and cleaning our dataset, we turned to the power of visualization to explore the relationships between different variables and gain a deeper understanding of the underlying dynamics driving sales. Power BI, a robust business intelligence tool, enabled us to create interactive and insightful visualizations.

Bivariate Analysis and Insights:

Our primary focus was on conducting bivariate analysis, investigating the relationships between pairs of variables. This involved creating various visualizations, including scatter plots, line charts, and bar charts, to uncover potential correlations and patterns. These visualizations helped us to:

Visualize Trends and Patterns: Visualizations facilitated the identification of trends and seasonality in the data. This knowledge informed the selection and design of appropriate forecasting models.

The Visualizations are as follows:

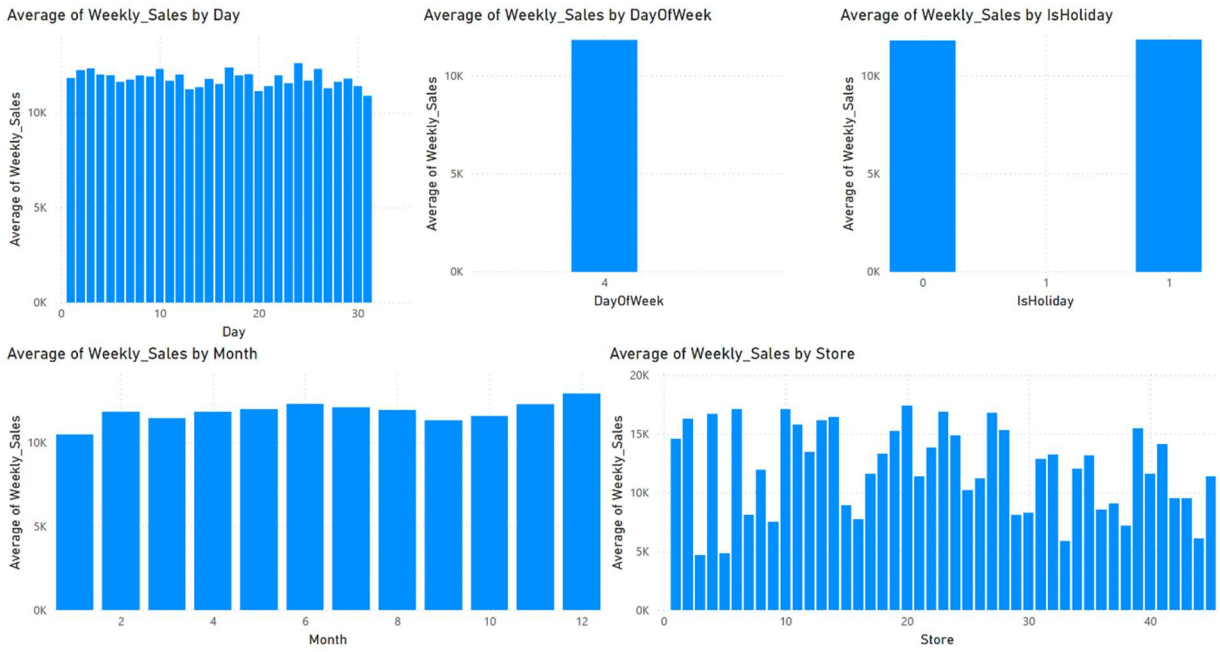


Figure 7: Bar plots of Day of Month, Holiday, Month, Store against Target variable – Weekly sales.

Insights:

- **Day of Month Impact:** Our analysis of the relationship between the day of the month and weekly sales revealed no discernible trend. This suggests that sales performance does not appear to be significantly influenced by the specific day of the month.
- **Holiday Effect:** The visualization clearly indicates that holidays do not have a significant impact on sales. This finding suggests that promotional periods and other marketing initiatives might be more influential in driving sales than holidays alone.
- **December Demand:** Sales data indicates a clear peak in sales during December, highlighting the strong impact of holiday shopping on overall retail performance.
- **Store-Specific Performance:** Our analysis revealed significant differences in sales performance across different stores. Stores 3, 5, 33, 38, and 44 consistently had the lowest sales, while stores 20 and 10 showed the highest sales. This suggests that store-

specific factors, such as location, store size, product mix, or local market conditions, significantly influence sales performance.

The Other Visualizations are as follows:

Below are the visuals of the CPI, Fuel Price, Temperature, & Unemployment against our target variable 'Weekly_Sales'.

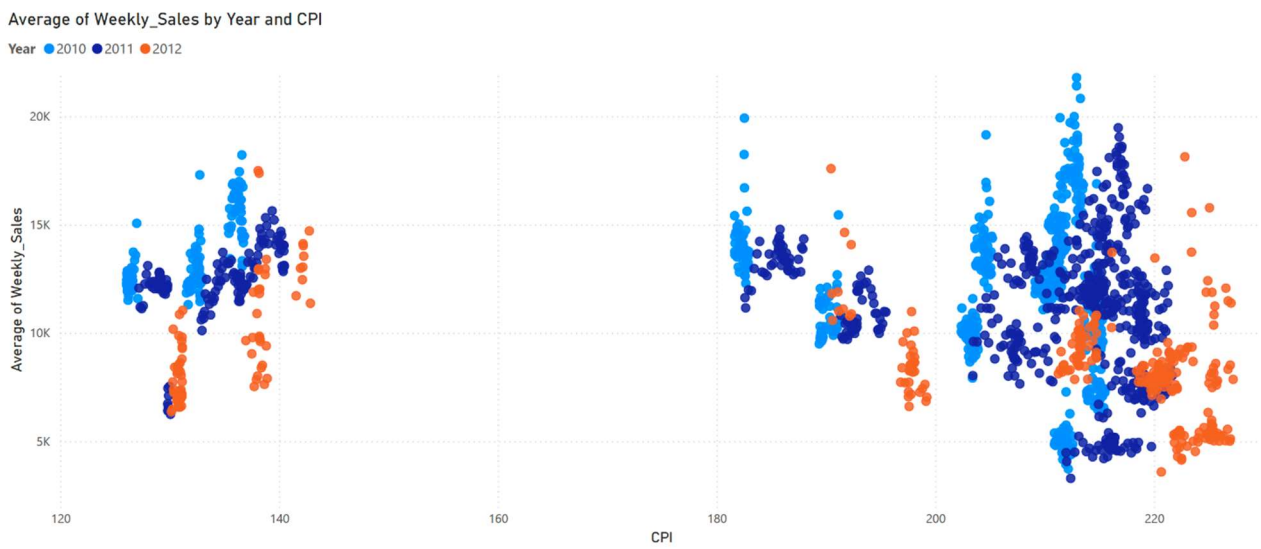


Figure 8: Scatter plot of CPI against Weekly sales categorized by year.

Insights:

- N**o Strong Correlation:** There is no apparent strong linear relationship between average weekly sales and CPI. The data points are scattered across the plot, suggesting that CPI might not be a significant factor influencing average weekly sales.
- **Year-wise Trends:** The data suggests that average weekly sales might be slightly higher in 2011 compared to 2010 and 2012. However, this trend is not consistent and requires further investigation.
- **Limited CPI Range:** The CPI values are clustered within a relatively narrow range (between 120 and 230). This might be limiting the ability to observe a clear relationship between CPI and sales.

Fuel Price Vs Weekly Sales

Figure 9: Scatter plot of Fuel Price against Weekly sales categorized by year.

Insights:

- **No Strong Correlation:** The data points are scattered across the plot, suggesting a lack of a strong linear relationship between average weekly sales and fuel price. There is no clear trend indicating that higher fuel prices lead to either higher or lower sales.
- **Similar Sales Distribution:** Across all three years, the average weekly sales seem to be distributed within a similar range, regardless of the fuel price fluctuation. This suggests that fuel price might not be a significant factor driving overall sales variations.
- **Potential Year-Wise Trends:** A closer look reveals that the average weekly sales for 2010 seem to be slightly higher than those for 2011 and 2012, particularly when the fuel price is around 3.0. However, this trend needs further investigation with additional analysis.

Temperature Vs Weekly Sales

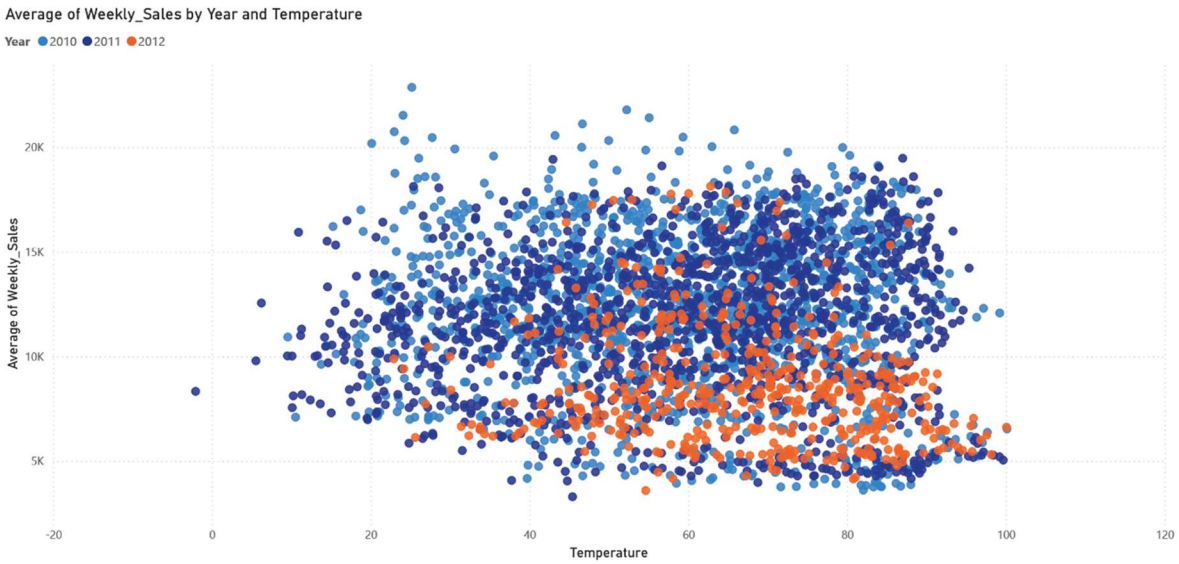


Figure 10: Scatter plot of Temperature against Weekly sales categorized by year.

Insights:

- No Strong Correlation: There doesn't seem to be a clear linear relationship between average weekly sales and temperature. The data points are scattered across the plot, suggesting that temperature might not be a significant factor directly driving sales fluctuations.
- Similar Sales Distribution: Across all three years, the average weekly sales seem to be distributed within a similar range, despite the varying temperatures. This reinforces the idea that temperature might not be a major driver of sales variations.
- Potential for Non-linear Relationship: The scatter plot shows a slight downward trend, indicating that sales may decrease slightly as the temperature increases. However, it is not a very strong trend and requires further investigation.

Unemployment Vs Weekly Sales

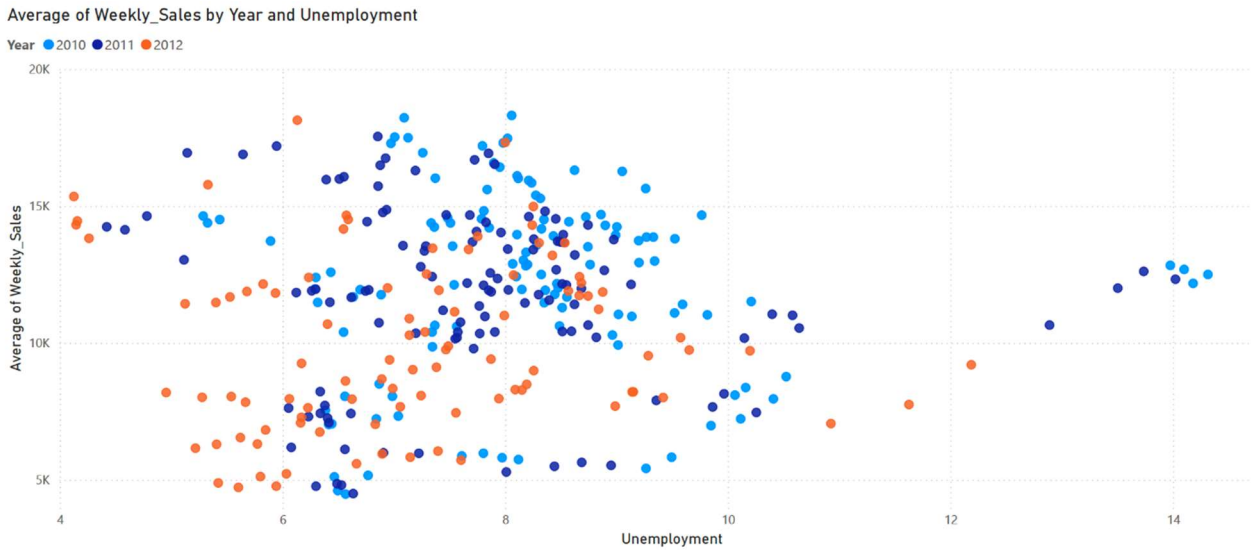


Figure 11: Scatter plot of Unemployment against Weekly sales categorized by year.

Insights:

- **No Clear Correlation:** There's no apparent linear relationship between average weekly sales and unemployment rate. The data points are spread out, suggesting that unemployment might not be a significant factor driving sales fluctuations.
- **Similar Sales Range:** Across all three years, the average weekly sales seem to fall within a comparable range, despite the varying unemployment rates. This reinforces the idea that unemployment might not be a major driver of sales variations.
- **Potential Year-Wise Trends:** A closer look suggests that the average weekly sales in 2010 and 2012 might be higher compared to 2011, particularly when the unemployment rate is between 6 and 8. However, this is not a consistent trend and requires further analysis.

Correlation Map:

To gain a comprehensive understanding of the interrelationships between all variables, we constructed a correlation map. This visualization graphically illustrated the strength and direction of correlations between each pair of variables. The correlation map served as a valuable tool for:

Supporting Model Selection: The correlation map assisted in identifying features with high correlation to Weekly_Sales, providing guidance for selecting relevant features for model training.

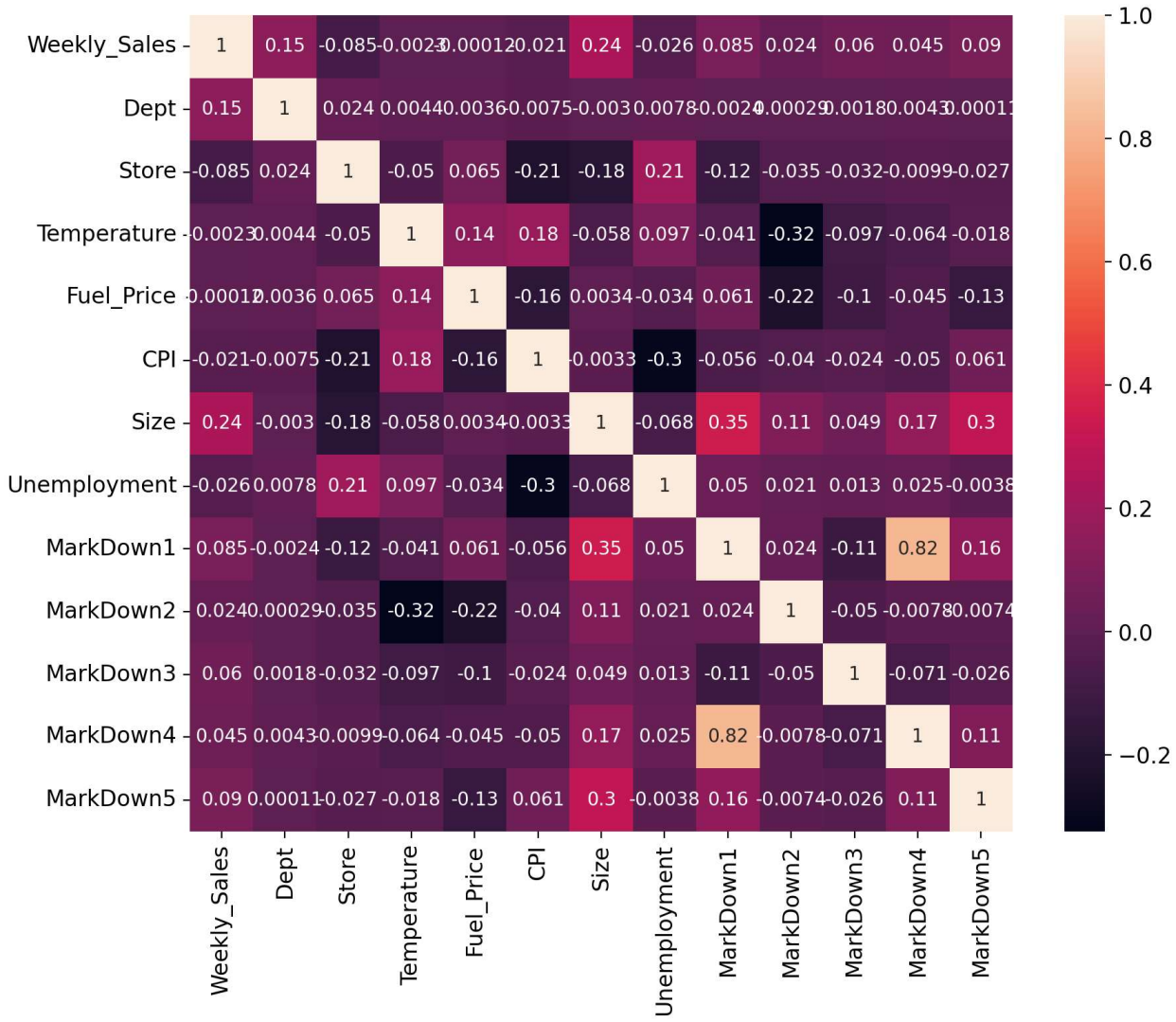


Figure 12: Correlation Map of numerical variables.

Insight: Our correlation map analysis revealed a surprising finding: there was no strong or statistically significant relationship between the target variable (Weekly_Sales) and most of the independent variables examined. While a few variables showed a very slight negative correlation, their influence was negligible. The highest correlation observed was with "Size," indicating a moderately weak positive correlation of 0.24. Based on this analysis, we can conclude that the selected independent variables do not appear to have a direct or inverse causal relationship with weekly sales.

This visualization-driven approach using Power BI proved highly effective in uncovering hidden patterns and relationships within the dataset, providing invaluable insights for model development and analysis.

4.6 Building Regression Models

With a cleaned and transformed dataset, we embarked on constructing and evaluating a diverse set of regression models to predict retail sales. Our goal was to identify the model architectures and hyperparameters that yielded the highest accuracy and best predictive performance for this specific dataset.

Model Selection and Hyperparameter Tuning:

We implemented a range of regression models, each with its unique strengths and capabilities, to capture different aspects of the sales data. For each model, we conducted hyperparameter tuning using GridSearchCV, systematically exploring a range of hyperparameter values to identify the optimal configuration that minimized model error and maximized performance. The best parameters found for each model are listed below:

- Lasso Regression: This model uses L1 regularization to reduce the impact of less important features, promoting sparsity.
    - Best Parameters: {'alpha': 10, 'max_iter': 1000}
- Elastic Net: Combining L1 and L2 regularization, this model offers a balance between feature selection and model stability.
    - Best Parameters: {'alpha': 100, 'l1_ratio': 3}
- Ridge Regression: Employing L2 regularization, this model prevents overfitting and enhances stability.
    - Best Parameters: {'alpha': 100}
- KNN Regression (K-Nearest Neighbors): This model uses a distance-based approach to predict sales by considering the sales of similar data points.
    - Best Parameters: {'metric': 'euclidean', 'n_neighbors': 10, 'weights': 'uniform'}
- SVR Regression (Support Vector Regression): This model uses a margin-based approach to find a function that best fits the data and minimizes errors.
    - Best Parameters: {'C': 100, 'epsilon': 0.01, 'kernel': 'rbf'}

- XGBoost Regression: This powerful gradient boosting algorithm iteratively constructs trees, each addressing the errors of the previous ones, to achieve high accuracy.
    - Best Parameters: {'colsample_bytree': 0.9, 'learning_rate': 0.1, 'max_depth': 9, 'n_estimators': 300, 'subsample': 0.9}
- Random Forest Regression: This ensemble model combines multiple decision trees to enhance predictive power and reduce variance.
    - Best Parameters: {'max_depth': None, 'max_features': 'auto', 'min_samples_split': 2, 'n_estimators': 200}
- LightGBM (Light Gradient Boosting Machine): A gradient boosting algorithm known for its speed and efficiency, particularly for large datasets.
    - Best Parameters: {'boosting_type': 'gbdt', 'objective': 'regression', 'metric': {'l2', 'mae'}, 'num_leaves': 31, 'learning_rate': 0.05, 'feature_fraction': 0.9}
- AdaBoost Regression: This model combines multiple weak learners (e.g., decision trees) in a sequential manner to create a more robust and accurate prediction.
    - Best Parameters: {'base_estimator__max_depth': 7, 'learning_rate': 0.1, 'n_estimators': 50}

Model Evaluation:

The performance of these regression models was assessed using three common metrics: R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). The evaluation results, discussed below and in subsequent sections, provided a basis for comparing the effectiveness of different regression models and informed the selection of the most promising models for further analysis.

Table 1: Comparison of regression models evaluation metrics

Regression Model Performance			
Model Name	**R Squared**	**RMSE**	**MAE**
XGBoost Regression	0.95165	3014.199	1812.385
Random Forest Regression	0.94765	3137.518	1640.912
LGBoost Regression	0.82263	5842.45	4014
AdaBoost Regression	0.7286	7226.53	5264.93
KNN Regression	0.17608	12442.98	8806.89

LASSO Regression	0.06354	13265.54	9994.91
Elastic Net	0.06354	13265.54	9994.91
RIDGE Regression	0.06352	13265.73	3187.56
SVM Regression	0.00959	13642.32	8587.45

4.7 Building Time Series Models

This section details the development and evaluation of time series models designed to capture the temporal patterns and trends present in the data. Time series models are particularly well-suited for forecasting data that exhibits sequential dependencies and recurring patterns over time.

Model Selection and Parameter Identification:

We implemented 3 time series models, each with specific characteristics to handle different aspects of time series data. We conducted stationarity checks (using the Augmented Dickey-Fuller test) and examined the autocorrelation (ACF) and partial autocorrelation (PACF) plots to identify the optimal model parameters for each model.

Here's a breakdown of the models we built:

- ARIMA (Autoregressive Integrated Moving Average): ARIMA is a widely used statistical model that incorporates autoregressive (AR), integrated (I), and moving average (MA) components to capture trends, seasonality, and non-stationary patterns.
    - Model Parameters: The ARIMA model's parameters (p, d, q) represent the order of the autoregressive, integrated, and moving average components. The Augmented Dickey-Fuller test confirmed the stationarity of our data, indicating that differencing (d) was not required. Analysis of the ACF and PACF plots suggested a model order of (1,0,1), indicating a strong dependence on the previous time step and the previous forecast error.

- Auto ARIMA: To automate the process of finding the optimal ARIMA model parameters, we employed the Auto ARIMA algorithm. This approach explores various combinations of parameters (p, d, q) within a defined range and selects the model with

the lowest Akaike Information Criterion (AIC) score, indicating the best balance between model complexity and prediction accuracy.

    - Best Parameters: The Auto ARIMA algorithm determined that the optimal model order was ARIMA (1,1,1) (0,0,0), implying a single autoregressive term, a single differencing term, and a single moving average term, with no seasonal components.

- SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors): SARIMAX is an extension of ARIMA that allows the incorporation of exogenous variables, providing a richer model capable of considering external factors that might influence sales patterns.
    - Model Parameters: Based on the stationarity check and ACF/PACF analysis, we implemented a SARIMAX model with an order of (1,0,1), indicating a strong dependence on the previous time step and the previous forecast error. A seasonal component with a period of 52 was included, reflecting the weekly nature of the data over three years.

Model Evaluation:

The performance of these regression models was assessed using three common metrics: R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). The evaluation results, discussed below and in subsequent sections, provided a basis for comparing the effectiveness of different regression models and informed the selection of the most promising models for further analysis.

Table 2: Comparison of time-series models evaluation metrics

Time Series Model Performance			
Model Name	**R Squared**	**RMSE**	**MAE**
ARIMA	0.000282	13942.26	10789.3
SARIMAX	-0.0007	13945.18	10958.075
Auto ARIMA	-0.00076	13945.18	10958.07

4.8 Building Neural Network Models

This section delves into the construction and evaluation of neural network models, leveraging the power of deep learning to capture complex relationships and patterns within the data. Our focus is on identifying the most effective neural network architecture for predicting sales trends, considering the specific characteristics of the dataset.

Model Selection and Architecture:

We implemented three neural network models, choosing architectures that are particularly well-suited for time series data:

- Artificial Neural Networks (ANN): This model uses multiple layers of interconnected nodes (neurons) to learn complex patterns from data.
    - Architecture: The chosen ANN model comprised four dense layers with varying numbers of nodes (ranging from 256 to 32). A dropout layer (with a rate of 0.2) was included to prevent overfitting. Activation functions included ReLU (Rectified Linear Unit) for hidden layers and a linear activation function for the output layer.
- Tuned ANN: To optimize the ANN model's performance, we employed the Keras Tuner library for hyperparameter optimization. This process systematically explored various configurations of the ANN architecture, including the number of layers, activation functions, dropout values, and optimization algorithms. The Keras Tuner identified the optimal configuration, maximizing model accuracy.
- Bidirectional LSTM: This model uses a bidirectional LSTM architecture, processing the time series data in both forward and backward directions to capture both past and future context. This allows for a more comprehensive understanding of the data's temporal dependencies.
    - Architecture: This model featured four dense layers, a dropout rate of 0.2, and activation functions (ReLU and linear). It employed the Adam optimizer and used the "return_sequences" parameter in the LSTM layer to preserve information about the sequence.

Model Evaluation:

The performance of these regression models was assessed using three common metrics: R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). The evaluation results, discussed below and in subsequent sections, provided a basis for comparing the effectiveness of different regression models and informed the selection of the most promising models for further analysis.

Table 3: Comparison of Neural Network models evaluation metrics

Neural Network Model Performance			
Model Name	**R Squared**	**RMSE**	**MAE**
ANN	0.36308	10940.14	7548.15
Tuned ANN	0.29192	11535.09	7951.03
LSTM Bidirectional	-0.35194	16129.86	10250.99

4.9 Optimizing Model Performance: Exploring Feature Impact and Ensemble Methods

This section focuses on a series of experiments designed to further enhance the performance of our most promising sales forecasting model, XGBoost, by investigating the impact of different feature combinations and exploring the potential benefits of model ensembles. The goal is to identify the best model configuration for achieving the highest accuracy and robustness in predicting retail sales.

Initial Model Selection:

Our initial model evaluation revealed that the XGBoost Regressor and Random Forest Regressor models consistently achieved the highest R-squared values, indicating strong performance in capturing the relationships within the sales data. Given XGBoost's superior R-squared score, it was chosen as the primary model for further optimization experiments.

Experiment Design and Execution:

To investigate the impact of specific features on model performance, we conducted a series of controlled experiments:

- **Experiment 1:** Removing Promotional Columns: We evaluated XGBoost's performance without using the "MarkDown" columns (representing promotional discounts). This experiment sought to understand the influence of promotional activities on sales predictions.
- **Experiment 2:** Removing Store Size and Type: We tested the model's performance excluding the "Size" and "Type" columns to determine their impact on the model's accuracy.
- **Experiment 3:** Using Only Promotional Columns: We trained the model solely on the promotional features ("MarkDown1-5") to assess their predictive power.
- **Experiment 4:** Combining Promotional Columns: We combined all five promotional columns into a single feature, aiming to capture a more comprehensive view of promotional activity and its impact on sales.

We also performed Experiment 4 by combining other variable combinations but it did not yield any notable results.

Model Performance Evaluation:

The performance of these regression models was assessed using three common metrics: R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). The evaluation results, discussed below and in subsequent sections, provided a basis for comparing the effectiveness of different regression models and informed the selection of the most promising models for further analysis.

Table 4 Comparison of XG Boost model's Experiments evaluation metrics.

Experiments on XG Boost Model			
Experiments	**R Squared**	**RMSE**	**MAE**
XG Boost Model Experiment 1	0.95165	3014.199	1812.385
XG Boost Model Experiment 2	0.93661	3451.309	2176.194
XG Boost Model Experiment 3	0.00699	13664.701	10408.86
XG Boost Model Experiment 4	0.95235	2992.263	1760.227

Below are the plots of RMSE of the training dataset & validation Dataset of all four experiments.

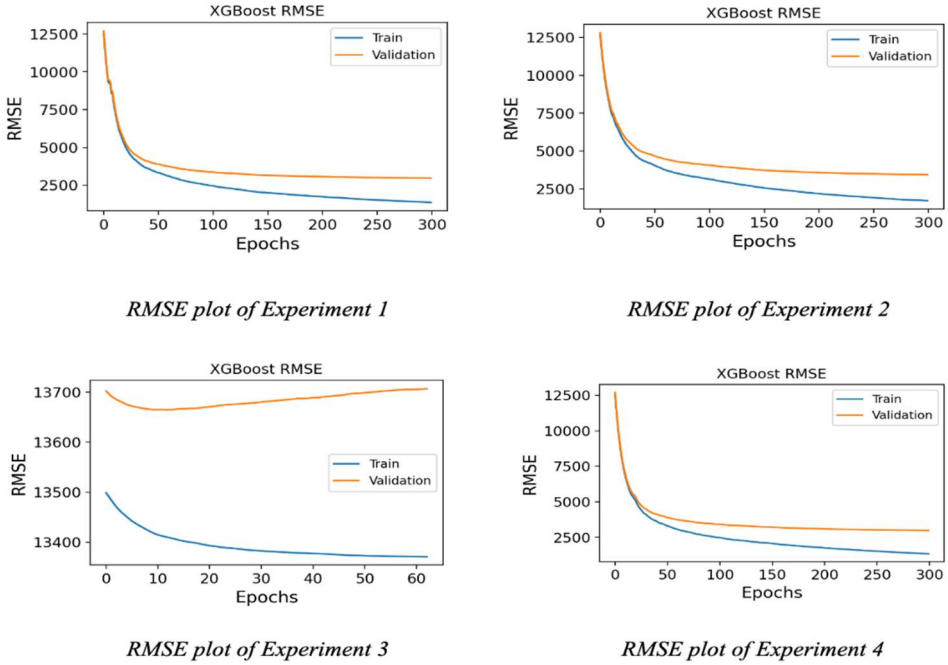


Figure 13: RMSE plots of training data Vs validation data.

4.10 Summary

This chapter details the analysis and model development process for investigating sales prediction accuracy in a retail setting. The dataset, sourced from Kaggle and comprising Walmart retail sales data, was cleaned, transformed, and explored using univariate analysis, outlier detection, and interactive visualization with Power BI.

A range of forecasting models were implemented, including machine learning models (Support Vector Machine, Linear Regression, Gradient Boosted Regression Trees, Random Forest), time series models (ARIMA, Prophet), and deep learning models (ANN, LSTM). Hyperparameter tuning was applied to optimize model performance.

To assess the impact of different feature sets, experiments were conducted, exploring various combinations of promotional columns, store size and type, and a combined promotional feature. The performance of each model was evaluated using metrics like R-squared, MAE, and RMSE.

The chapter provides a detailed account of the data analysis and model building process, laying the groundwork for the subsequent chapters that focus on comparing the performance of these models in a retail environment.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Introduction

This chapter delves into the heart of our research, presenting a comprehensive evaluation of the various forecasting models developed and tested in the previous chapter. We will analyze the performance of each model, comparing their metrics.

Our analysis will focus on drawing key insights from the model evaluation results, addressing the following:

Model Performance Comparison: We will present a detailed comparison of the models' performance across different evaluation metrics, highlighting the strengths and limitations of each approach.

Model Suitability: We will discuss the relative suitability of each model for retail sales forecasting, considering the specific challenges and requirements of a retail environment.

Addressing Error Metrics: This chapter will delve into a deeper exploration of why certain models yielded high MAE and RMSE values, seeking explanations for these deviations and identifying potential areas for further optimization.

Through this comprehensive evaluation, we aim to identify the most effective model for retail

sales forecasting, providing valuable insights and recommendations for practitioners in the retail industry.

5.2 Evaluation: Comparison of Model Metrics and Insights

This section presents a comparative analysis of the performance of various forecasting models, evaluating their ability to predict retail sales using the metrics: R-squared, Root Mean

Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics provide valuable insights into the strengths and limitations of each model type.

Regression Models:

Gradient Boosting Machines (XGBoost, LightGBM): These models consistently outperformed other regression models, achieving significantly higher R-squared values and lower RMSE and MAE. This suggests that gradient boosting techniques are particularly effective for capturing complex relationships within the retail sales data.

Random Forest Regression: This model also demonstrated strong performance, indicating its ability to handle complex data patterns and reduce overfitting.

Other Regression Models (Lasso, ElasticNet, Ridge, KNN, SVR): These models exhibited lower R-squared values and higher error metrics compared to gradient boosting and random forest. The simpler linear models (Lasso, ElasticNet, Ridge) were less effective in capturing the non-linear relationships present in the sales data. KNN and SVR, while effective in some cases, did not perform as well as the more sophisticated boosting and ensemble methods.

Time Series Models:

ARIMA and Seasonal ARIMA (SARIMAX): These traditional statistical models struggled to capture the complexities within the retail sales data, resulting in very low R-squared values and high error metrics. This indicates that these models may not be the most suitable for predicting retail sales when considering factors like promotional activity and external influences.

Auto ARIMA: While attempting to optimize model parameters, Auto ARIMA did not improve the performance of the ARIMA model, suggesting that this automated approach might not be sufficient for finding the best parameters for this dataset.

Neural Network Models:

LSTM Bidirectional: This deep learning model, despite its potential to handle temporal patterns, yielded the lowest R-squared value and the highest error metrics. This highlights the challenges in training deep learning models effectively and finding the optimal configurations for this dataset.

ANN and Tuned ANN: These models showed moderate performance, demonstrating that basic ANN architectures may not be sufficient to achieve the highest accuracy when dealing with complex retail sales data.

Below is the comparison table of the model performances.

Table 5 Comparison of all the model's evaluation metrics.

Model Performance			
Model Name	**R Squared**	**RMSE**	**MAE**
XG Boost Regression	0.95165	3014.199	1812.385
Random Forest Regression	0.94765	3137.518	1640.912
LG Boost Regression	0.82263	5842.45	4014
AdaBoost Regression	0.7286	7226.53	5264.93
ANN	0.36308	10940.14	7548.15
Tuned ANN	0.29192	11535.09	7951.03
KNN Regression	0.17608	12442.98	8806.89
LASSO Regression	0.06354	13265.54	9994.91
Elastic Net	0.06354	13265.54	9994.91
RIDGE Regression	0.06352	13265.73	3187.56
SVR Regression	0.00959	13642.32	8587.45
ARIMA	0.000282	13942.26	10789.3
SARIMAX	-0.0007	13945.18	10958.075
Auto ARIMA	-0.00076	13945.18	10958.07
LSTM Bidirectional	-0.35194	16129.86	10250.99

This comparative analysis of different models highlights the strengths and limitations of various forecasting approaches for retail sales. These findings will inform the next stage of our research, focusing on exploring strategies to further improve model performance and identify the most promising models for predicting retail sales.

5.2.1 Inferences from the Experiments on XG Boost Model

Here are some inferences based on the conducted experiments, focusing on the performance of the XG Boost model:

Experiment 1: Model Performance without Promotional Columns:

Significant Drop in Accuracy: The R-squared value decreased from 0.95165 to 0.93661, suggesting that promotional columns contribute significantly to the model's accuracy. This indicates that promotional activities play a crucial role in predicting sales patterns.

Experiment 2: Model Performance without Size & Type Columns:

Moderate Impact: The R-squared value dropped to 0.93661, which is only slightly lower than Experiment 1. This suggests that while store size and type are important, their influence is less significant compared to promotional information.

Experiment 3: Model Performance with Promotional Columns Only:

Dramatic Decrease in Accuracy: The R-squared value plummeted to 0.00699, showing a significant decrease in model performance when only promotional columns are used. This indicates that promotional columns alone are not sufficient to predict sales accurately. Other factors are crucial for model accuracy.

Experiment 4: Model Performance with Combined Promotional Columns:

High Accuracy Retained: The R-squared value (0.95235) is nearly identical to the initial XGBoost model (0.95165) which indicates that combining all promotional columns into a single feature doesn't significantly improve the model's accuracy.

Table 4 Comparision of XG Boost model's Experiments evaluation metrics

Experiments on XG Boost Model			
Experiments	**R Squared**	**RMSE**	**MAE**
XG Boost Model Experiment 1	0.95165	3014.199	1812.385
XG Boost Model Experiment 2	0.93661	3451.309	2176.194
XG Boost Model Experiment 3	0.00699	13664.701	10408.86
XG Boost Model Experiment 4	0.95235	2992.263	1760.227

Key Insights:

- Importance of Promotional Data: Promotional columns ("MarkDown" features) play a vital role in accurately predicting sales using the XGBoost model.
- Complementary Data Sources: The model relies on a combination of features, not solely on promotional information, to achieve optimal performance.
- Further Research: While combining all promotional columns into a single feature didn't enhance accuracy, further research could explore methods to better capture the individual and collective effects of promotions on sales.

5.3 Discussion on the Results: Analyzing Model Performance and Limitations

While the evaluation of various models revealed that Gradient Boosted Regression Trees (XGBoost and LightGBM) and Random Forest Regression demonstrated superior performance with high R-squared values, it is crucial to address the consistently high Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) values observed across all models. These high error metrics highlight the inherent challenges of accurately predicting retail sales, pointing to potential limitations within the dataset and the models themselves.

Dataset Limitations:

Several factors related to the dataset might be contributing to the high error. A few possibilities are mentioned below.

Feature Limitations: The dataset may lack crucial features that strongly correlate with sales. For instance, the absence of competitor data, pricing information, and marketing campaign details could hinder model accuracy.

Limited Data Variety: The dataset, while spanning three years, might not adequately capture the full range of promotional activities, store types, or seasonal patterns, limiting the models' ability to generalize effectively.

Data Bias: The data might be biased toward specific store types or regions, making it challenging for models to accurately predict sales for underrepresented groups.

Temporal Data Challenges: The complexity of temporal patterns in retail sales might not be fully captured by the dataset. The nuances of seasonality, trends, and holiday effects might not be adequately represented.

Data Noise: Real-world data often contains inherent noise, which can obscure underlying patterns and contribute to model errors.

Model Complexity and Overfitting:

While complex models offer the potential for higher accuracy, they can also be susceptible to overfitting the training data. Overfitting occurs when the model learns the specific nuances of the training set but fails to generalize well to new data. This can lead to high error values when the model encounters unseen data.

5.4 Summary

This chapter comprehensively evaluated the performance of various forecasting models – including machine learning (XGBoost, Random Forest), time series (ARIMA, SARIMAX), and deep learning (ANN, LSTM) – against the Walmart retail sales dataset. The results revealed that gradient boosting models (XGBoost and LightGBM) consistently achieved the highest accuracy, while traditional time series models and deep learning approaches faced challenges in capturing the complex dynamics of retail sales.

The analysis highlighted the importance of data quality, the need for relevant features, and the ongoing challenge of effectively modeling non-linear relationships and seasonal trends within retail data. While the findings revealed the strengths of some models, they also underscored the need for further research to address the limitations of current forecasting techniques and enhance model accuracy in retail environments.

While this research has shed light on the potential of various models for retail sales forecasting, further investigation is needed to address the challenges of complexity, interpretability, and real-world implementation. Future work should focus on developing hybrid architectures, enhancing model explainability, and exploring data augmentation techniques to improve prediction accuracy. Moreover, researching methods for integrating these models seamlessly into existing retail operations will be crucial for practical application and driving meaningful business impact.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This chapter draws together the key findings and conclusions of this research, examining the effectiveness of various machine learning, neural network, and time series models for predicting retail sales. The study aimed to enhance sales prediction accuracy for a chain of retail stores by evaluating and comparing the performance of these models, considering the unique challenges of the retail environment.

The research involved a comprehensive analysis of the Walmart retail sales dataset, including data cleaning, transformation, and exploratory data analysis. This was followed by the development and evaluation of several prediction models:

- Regression Models: We investigated the performance of linear regression, Support Vector Machines (SVM), Gradient Boosted Regression Trees (GBRT), and Random Forest models.
- Time Series Models: Traditional statistical models like ARIMA and Prophet were implemented and compared.
- Neural Networks: Deep learning models such as ANN and LSTM were explored.

We conducted controlled experiments to assess the impact of various feature combinations and hyperparameter tuning on model accuracy. The analysis revealed the strengths and limitations of different model types, highlighting the importance of data quality, feature selection, and model optimization in achieving accurate retail sales predictions.

This chapter will build upon these findings, drawing conclusions about the most effective models and identifying key areas for future research.

6.2 Discussion and Conclusion

Based on the analysis and evaluation conducted in the previous chapters, gradient boosting models, particularly XGBoost and LightGBM, emerged as the most promising approaches for retail sales forecasting. These models demonstrated superior accuracy, as indicated by higher

R-squared values and lower error metrics, compared to other regression models, time series models, and neural networks. This suggests that these advanced machine learning techniques are particularly effective in capturing the complex relationships and non-linear patterns present in retail sales data.

However, it is important to acknowledge the limitations inherent in the dataset and their potential impact on model performance. The consistently high RMSE and MAE values observed across various models suggest that the dataset might be incomplete or biased, hindering the models' ability to achieve perfect accuracy. Specifically, the limited availability of features, such as customer demographics, product categories, competitor information, and detailed promotional data, could be contributing to the error metrics.

Furthermore, the dataset's focus on weekly sales data might not fully capture the nuances of daily or seasonal variations, particularly in the context of evolving consumer behavior and market dynamics.

Despite these limitations, the research highlights the potential of gradient boosting models for retail sales forecasting.

6.3 Contribution to knowledge

This research contributes to the field of retail sales forecasting by providing a comprehensive evaluation and comparison of various machine learning, neural network, and time series models in a real-world retail setting. The study's key contributions include:

- **Comparative Evaluation of Forecasting Models:** This research offers a rigorous evaluation of different model types, including gradient boosting, random forest, time series, and deep learning approaches, using a large and relevant retail dataset. The comparative analysis provides insights into the strengths and limitations of each model type, guiding practitioners in choosing the most appropriate model based on specific needs.
- **Identifying Key Determinants of Success:** The research highlights the critical importance of data quality, feature selection, and model optimization for achieving accurate sales predictions. The findings emphasize that while advanced modeling techniques offer significant potential, data availability and quality remain critical for successful forecasting.

- **Exploring the Role of Promotional Activity:** The research demonstrates the substantial influence of promotional activity on sales patterns. This finding underscores the importance of incorporating promotional data into forecasting models to improve accuracy.

By conducting this comprehensive analysis, this research offers valuable insights and contributes to a deeper understanding of the effectiveness and limitations of different models for forecasting retail sales. The findings provide practical recommendations for retailers seeking to optimize their forecasting practices and improve business outcomes.

6.4 Future Recommendation

This research provides a solid foundation for continued exploration and advancements in retail sales forecasting. Key areas for future research include:

- Hybrid Architectures: Investigating hybrid architectures that combine different deep learning models, such as NCDEs, ANCDEs, and other attention-based models, holds significant potential for capturing complex patterns and achieving greater accuracy.
- Model Interpretability and Explainability: Developing techniques to enhance the interpretability and explainability of these complex models is crucial. This could involve using visualization tools, creating model summaries, or incorporating techniques like feature attribution methods.
- Data Augmentation and Feature Engineering: Exploring methods for data augmentation (e.g., synthetic data generation, feature transformations) and more sophisticated feature engineering techniques can further enhance model performance, particularly when dealing with limited data or complex relationships.
- Multi-Step Forecasting: While the research focused on single-step forecasting, exploring multi-step forecasting models that predict sales over multiple future time periods can provide more comprehensive insights and support long-term planning.
- Integration with Business Processes: Investigating how these advanced forecasting models can be seamlessly integrated into existing retail operations and decision-making processes is crucial for practical implementation and business impact.

These future research directions offer promising avenues for enhancing retail sales forecasting, providing retailers with more accurate predictions and valuable insights to optimize their operations and improve customer experiences.

REFERENCES

1. Wanchoo, K. (2019). Retail Demand Forecasting: a Comparison between Deep Neural Network and Gradient Boosting Method for Univariate Time Series. 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, pp. 1-5. doi: 10.1109/I2CT45611.2019.9033651.

2. IEEE Xplore. (n.d.). Retail Sales Forecast Based on Machine Learning Methods. [online] Available at: https://ieeexplore.ieee.org/document/10129443 [Accessed 6 December 2023].

3. Wang, F. and S, A.J. (2022). Using Regression Algorithms to Forecast Merchandise Sales in the Presence of Independent Variables. [online] IEEE Xplore. doi: https://doi.org/10.1109/ICCSIE56462.2022.00030.

4. IEEE Xplore. (n.d.). Comparing Statistical and Machine Learning Methods for Sales Forecasting During the Post-promotional Period. [online] Available at: https://ieeexplore.ieee.org/document/9672954 [Accessed 6 December 2023].

5. IEEE Xplore. (n.d.). A Comprehensive Analysis of Retail Sales Forecasting Using Machine Learning and Deep Learning Methods. [online] Available at: https://ieeexplore.ieee.org/document/10245887 [Accessed 6 December 2023].

6. Earnest, A., Chen, M.I., Ng, D. and Sin, L.Y. (2005). Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. BMC Health Services Research, 5(1). doi:https://doi.org/10.1186/1472-6963-5-36.

7. Kraus, M., Feuerriegel, S. and Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. European Journal of Operational Research, 281(3), pp. 628–641. doi:https://doi.org/10.1016/j.ejor.2019.09.018.

8. Sheo Yon Jhin, Shin, H., Hong, S., Park, S. and Park, N. (2023). Attentive neural controlled differential equations for time-series classification and forecasting. Knowledge and Information Systems. doi: https://doi.org/10.1007/s10115-023-01977-5

9. IEEE Xplore. (n.d.). Demand Forecasting in Supply Chain Management using CNN-LSTM Hybrid Model. [online] Available at: https://ieeexplore.ieee.org/document/10307665 [Accessed 6 December 2023].

10. IEEE Xplore. (n.d.). A Comparison of Prediction Algorithms in Food Sales with Different K-Folds Cross-Validation. [online] Available at: https://ieeexplore.ieee.org/document/10331578 [Accessed 6 December 2023].

11. IEEE Xplore. (n.d.). Correlation Recurrent Units: A Novel Neural Architecture for Improving the Predictive Performance of Time-Series Data. [online] Available at: https://ieeexplore.ieee.org/document/10264112 [Accessed 6 December 2023].

12. IEEE Xplore. (n.d.). Deep Learning Based Behavior Anomaly Detection within the Context of Electronic Commerce. [online] Available at: https://ieeexplore.ieee.org/document/10297230 [Accessed 6 December 2023].

13. Siami-Namini, S., Tavakoli, N. and Siami Namin, A. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). doi:https://doi.org/10.1109/icmla.2018.00227.

14. IEEE Xplore. (n.d.). Forecasting the Stability of COVID-19 Vaccine Companies Stock Market using LSTM and Time-series Models. [online] Available at: https://ieeexplore.ieee.org/abstract/document/10139884 [Accessed 6 December 2023].

15. Google Scholar. (2021). [online] Available at: https://scholar.google.com/scholar_lookup?title=Exploration+of+Time-Series+Models+on+Time+Series+Data&author=Kulakou [Accessed 6 December 2023].

16. Kulakou, S., Ragab, N., Midoglu, C., Boeker, M., Johansen, D., Riegler, M.A. and Halvorsen, P. (2022). Exploration of Different Time Series Models for Soccer Athlete Performance Prediction. Engineering Proceedings, 18(1), p.37. doi:https://doi.org/10.3390/engproc2022018037.

APPENDIX A: RESEARCH PROPOSAL

ENHANCING SALES PREDICTION ACCURACY FOR A CHAIN OF RETAIL STORE SALES: EVALUATING AND SELECTING THE OPTIMAL MODEL BY COMPARING MACHINE LEARNING MODELS, NEURAL NETWORKS, AND TIME SERIES MODELS.

SAI SHARAN PASPUNURI

Research Proposal

NOVEMBER 2023

Index -

Abstract

This research proposal aims to assess the efficacy of diverse forecasting methods in predicting retail sales, with a focus on guiding retail businesses in selecting optimal methods tailored to specific business scenarios. Utilizing Walmart's sales data, the investigation delves into factors like store types, sizes, departments, and holidays. A comparative analysis is performed to compare the performance of advanced deep learning models with traditional machine learning and time series models.

The findings illuminate instances where alternative models surpass traditional counterparts in both accuracy and efficiency. By showcasing the superiority of certain models, this research equips retail businesses with valuable insights for enhancing forecasting practices, ultimately contributing to informed decision-making and improved operational strategies in the retail sector.

1. Background

The background introduction highlights the critical role of retail sales forecasting in the industry, emphasizing its impact on decision-making, production, and supply chain management. It acknowledges the challenges posed by factors such as trends, seasonality, and the complexity of the retail market. The study positions itself within the existing landscape of forecasting methods, including time series analysis, statistical models, and the emerging field of deep learning. The background underscores the need for accurate forecasting to enhance business strategy, reduce operational costs, and improve customer satisfaction.

Dividing retail sales forecasting into time series analysis and forecasting, further categorized into statistical and deep learning methods. This provides a structured framework for evaluating and comparing different forecasting approaches. The comparative analysis, challenges, and conclusion sections contribute to a comprehensive understanding of the research landscape and its implications for the retail industry.

This research emphasizes the growing complexity of forecasting retail sales due to factors such as a high number of products, shorter product lifecycles, and aggressive marketing campaigns. Retail sales promotions, including various types, are identified as a confounding factor. The study positions itself within the existing literature by acknowledging the prevalence of univariate methods and the emergence of ML techniques in retail sales forecasting. The focus on post-promotional periods is identified as a gap in the current research landscape, motivating the study to address this aspect comprehensively.

The global pandemic significantly impacted the economy, particularly the retail and leisure industries. The study focuses on Walmart, a major retail player, to analyze the aftermath and

provide a forecasting model for weekly sales. Exploratory Data Analysis (EDA) reveals insights into store types, sizes, and departments, contributing to better understanding and model accuracy.

Traditional time series methods like ARIMA and Smoothing face challenges in handling non-linear patterns common in real-world sales series. The research also advocates for the use of advanced forecasting methods, such as Neural Networks and Machine Learning, citing their ability to capture nonlinearity effectively. The background also highlights the dataset's origin from a Kaggle and emphasizes the absence of additional features that could contribute to explaining store sales patterns.

1.1 Time Series Analysis in Business:

Time series analysis is highlighted as a valuable tool for extracting meaningful statistics and properties from data in business environments. Emphasis on the crucial role of time series forecasting models in determining future sales and aiding in business management.

1.2 Time Series Analysis Components:

Introduction of the four major components of time series data: Level, Trend, Seasonality, and Noise. Recognition of the significance of understanding these components for accurate predictions and decision-making.

The study is grounded in the evolution of supply chain management systems, emphasizing the role of technology, particularly artificial intelligence, in enhancing forecasting accuracy. Traditional demand planning methods are contrasted with the hybrid CNN-LSTM model, highlighting the need for more effective forecasting techniques in the retail industry. The research also reinforces the choice of the hybrid model based on the effectiveness of CNNs and LSTMs in time-series forecasting.

This research also involves a review of the evolution from "shallow" neural networks to deep learning. The paper discusses the mathematical background of neural networks, starting from single-layer perceptrons to deep neural networks. It highlights the challenges associated with optimization, the choice of activation functions, and the need for preprocessing in deep learning. The study emphasizes the potential advantages of deep neural networks in handling raw data without extensive manual feature engineering.

This comprehensive analysis provides the context for the subsequent exploration of the proposed deep-embedded network architecture and its application in the three business

analytics case studies. The research contributes to the ongoing efforts to enhance predictive modeling using deep learning techniques in operational research and business analytics.

2. Problem Statement Or Related Work

Traditional time series methods struggle to capture the non-linear patterns exhibited in real-world sales series, such as those in retail. The research addresses the limitations of conventional methods like ARIMA and Smoothing in handling nonlinearity and explores advanced forecasting methods to enhance accuracy. The study aims to provide effective forecasting models for sales time series, considering the challenges posed by the absence of additional explanatory features.

The post-pandemic era poses challenges for the retail industry, demanding accurate sales forecasts. The study addresses this by building models to predict Walmart's sales. The goal is to assist management teams in proactive decision-making, saving costs, and avoiding errors in staff scheduling.

Retail sales forecasting is challenged by the complex dynamics of promotions, particularly in predicting the impact during and after promotional periods. Existing methods often rely on judgmental approaches, and there is a need to explore the effectiveness of other forecasting techniques. The study identifies the post-promotional period as a critical phase that requires specific attention for accurate inventory planning.

The conventional machine learning models, including linear models and tree-based approaches, may not fully capture the complexity of business analytics tasks, leading to suboptimal predictive performance. The research addresses this limitation by exploring and evaluating the application of deep learning models, and intense neural networks, in various business scenarios.

We will be building & comparing forecasting models using the time series, ML, and Deep neural networks and optimizing them. We will be comparing and evaluating the different metrics. We will be also suggesting the model to be used based on the business application or the business problem to get efficient results.

3. Research Questions

- How do you optimize the data features based on the business problem?
- Which time series models provide accurate predictions?
- What is the time frame for considering historical data and making future predictions?
- Does the size of the training dataset impact results?
- Which algorithm, ARIMA or LSTM, performs a more accurate prediction of time series data?
- Which forecasting model performs better for supermarket sales prediction among the time series forecasting models?

- How do different time-series forecasting models handle variations, trends, and outliers in supermarket sales data?
- How does the hybrid CNN-LSTM model perform in forecasting demand compared to traditional methods?
- What are the optimal parameters for the Deep Neural Network, and how do they impact the forecasting results?
- What are the key parameters affecting the performance of the time series models, and how do they influence the forecast accuracy?

4. Aims & Objectives

The main aim of the thesis is to build and compare the time series models, Machine learning models, and deep NN models to forecast retail sales.

We will be comparing all the metrics and evaluating the performance of the models. Using the results, we should be able to suggest the business case-based model preference.

The objectives are as follows:

1. Build Time series, Machine Learning, and Deep NN models using the retail sales data.
2. Consider the business variables like CPI, Temperature, Fuel, Store type, and more from the data while processing and building the model.
3. Evaluate the models using the key metrics. This will help us understand the usability of the model based on the business scenario.
4. Compare the metrics and efficiencies. Understand the advantages and disadvantages of each model.
5. Based on the advantages, suggest the business case-wise model suggestion.

5. Significance of the Study

Sales forecasting plays a crucial role in various industries, particularly in retail, for efficient inventory planning and business decision-making.

Accurate sales forecasting benefits businesses by improving liquidity and reducing operating costs. Controlling merchandise stock aids in warehouse management and customer satisfaction. The study explores different model techniques, providing insights into effective algorithms for sales prediction, and contributing to efficient inventory management.

The comparison between different models provides insights into their performance and limitations in the absence of additional features. By comparing and analyzing various methods,

including traditional statistical approaches and modern deep learning techniques, the research aims to contribute valuable insights for improving the accuracy and efficiency of retail sales predictions.

The research highlights the variety of forecasting models available and the importance of choosing the best method for specific uses.

6. Scope of the study

This research aims to study retail sales forecasting and its importance in making decisions, managing inventory, and supply chain operations in the retail industry. The study will evaluate various forecasting methods, including time series analysis, statistical techniques, and deep learning models. It will examine how these methods can enhance business strategy, minimize operational costs, and improve customer satisfaction.

The research focuses on forecasting Walmart's sales by analyzing historical data. It explores various factors such as store types, sizes, departments, and holidays. The dataset includes information on store types, item sizes, dates of sale, temperature, fuel prices, holidays, etc.

7. Research Methodology:

This research utilizes a mixed-method approach, combining both quantitative and qualitative methods to evaluate the performance of retail sales forecasting models. The research employs historical sales data from a U.S.-based retailer - Walmart across the USA. The dataset includes information on sales and various promotion types. The research methodology involves EDA, data preprocessing, baseline projection, and the application of univariate and machine learning (ML) models.

7.1 Data Preprocessing:

The dataset goes through a meticulous preprocessing process, which includes checking for any missing values. Various methods are used to handle the missing data such as replacing gaps with zeros, filling repeated values, removing gaps, and concatenating arrays. All the features, including time series and external data, are defined in detail, which improves the accuracy and depth of the analysis.

Here is the data that will be used for this research -
https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast

The Key features from the dataset are: Promotion type, CPI index of the location, Fuel price, Temperature, & Type of the store. The data processing will be applied to these key features and train them on all the models.

7.2 Stationarity Check:

The stationarity of the time series data is checked using the KPSS test. This test helps determine whether the statistical properties of the series remain consistent over time. If necessary, log transformation is applied to stabilize the series.

7.3 Train/Test Splitting:

To tackle the problem of predicting sales with incomplete data, the dataset is divided into two parts: a training set and a testing set. The training set constitutes 80% of the data, while the remaining 20% is used for testing. The testing and validation sets are selected based on recent instances to ensure the reliability and accuracy of the predictions.

7.4 Baseline Projection:

A Moving Average model with 4 periods estimates baseline demand using sales data from normal periods. Promotional and post-promotional periods are identified using a promotional calendar, which is crucial for subsequent analysis.

7.5 Univariate and ML Methods:

The analysis will make use of a selection of univariate models, including Exponential Smoothing (ETS), Exponential Smoothing with a time-invariant regressor parameter (ETSX), and Autoregressive Moving Average (ARIMA). Additionally, machine learning (ML) models will be employed. To assess the effectiveness of ML methods in automatically identifying post-promotional periods, improving forecast accuracy, and outperforming univariate methods, multiple hypotheses will be formulated.

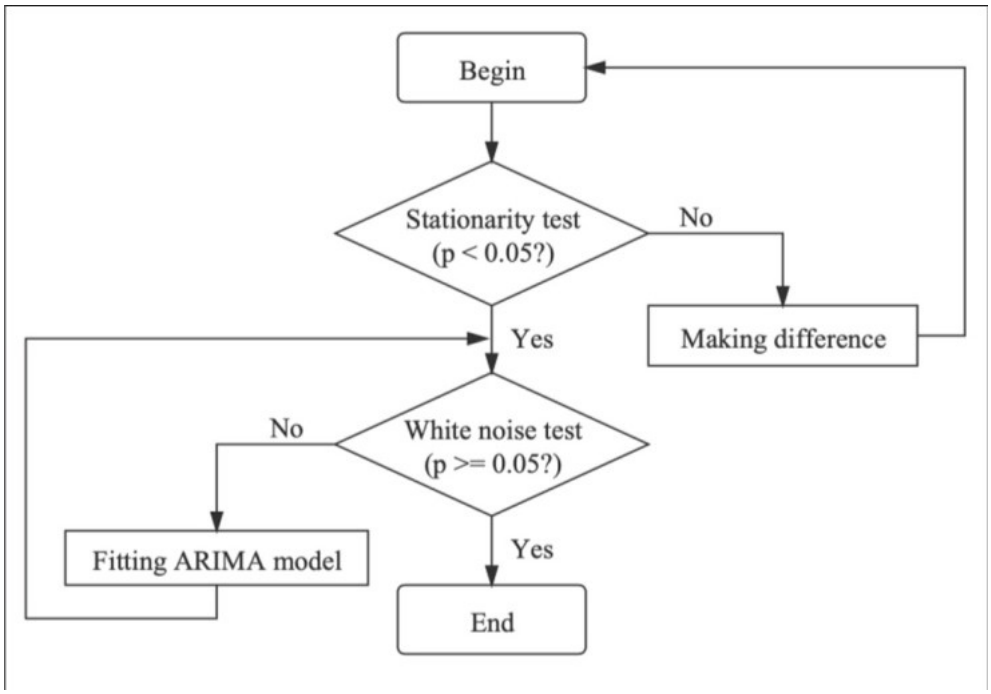


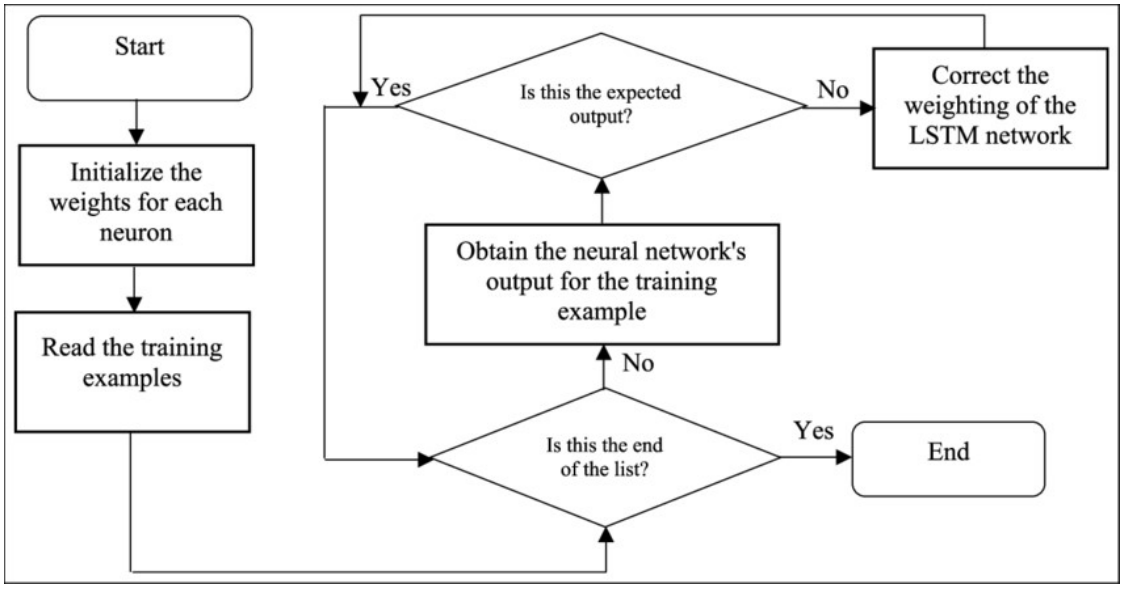
Figure 1: Flowchart of the ARIMA Model.

7.6 Neural Network Models:

Long Short-Term Memory (LSTM) is used as a sequential neural network to capture complex and nonlinear patterns in demand data. A 1D Convolutional Neural Network (CNN) is utilized for feature extraction in raw time-series data.

Figure 2: Flowchart of the LSTM Model.

CNN-LSTM



Model:

The hybrid model is an improved architecture that combines the strengths of CNN and LSTM. This model enhances efficiency, enables better feature extraction, allows for customization, and reduces noise and variability. CNNs reduce dimensionality, extract spatial features, and improve efficiency. On the other hand, LSTMs capture temporal correlations that enable accurate forecasting.

7.8 Evaluation Metrics:

To measure the performance of a model, we use standard metrics. These metrics vary depending on the type of prediction task (regression or classification) and include mean squared error (MSE), mean absolute error (MAE), Root Mean Square Error (RMSE) for the forecasts, Mean Absolute Percentage Error (MAPE) for overall accuracy, explained variance (R2), and area under the curve (AUC). By using these metrics, we can calculate the performance of the model and compare the results with other models.Mean Absolute Percentage Error shows us how much inaccuracy we should expect from our predictions on average.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i} * 100\right|$$

Root Mean Squared Error (RMSE) is a metric used to evaluate the accuracy of time series forecasting models. It measures the average difference between the predicted and actual values.

$$RMSE = \sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N}(x_i - m_i)^2}$$

For the Machine learning & the Deep neural networks models, we will be using the confusion metrics to evaluate the performance of the models. The metrics are as follows-

True Positive - TP; True Negative - TN; False Positive - FP; False Negative - FN;

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad \text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The research compares the performance of deep learning models, including deep neural networks (DNNs) and proposed deep-embedded network architecture, against traditional machine learning models and time series models to evaluate the model performances.

8. Requirement Resources:

For this thesis research, I will be using the following resources:

- Walmart's Sales Dataset - (Source: Kaggle)
- Google Collab Notebook Account
- Libraries like NumPy, Pandas, Sciket Learn, OpenCV, Tensorflow, Matplotlib Pyplot, Seaborn, Tensorflow, etc.,
- GPUs to process the Deep NN models. (Purchasing at least 100 computing GPU units)
- A working computer with a high-speed internet connection.

9. Research Plan

Gantt Chart - Thesis Submission Plan - https://docs.google.com/spreadsheets/d/140fZRzHCzBFihF-h7qA5cb9bse-J8s__hat8zIJrTIc/edit#gid=1692151692

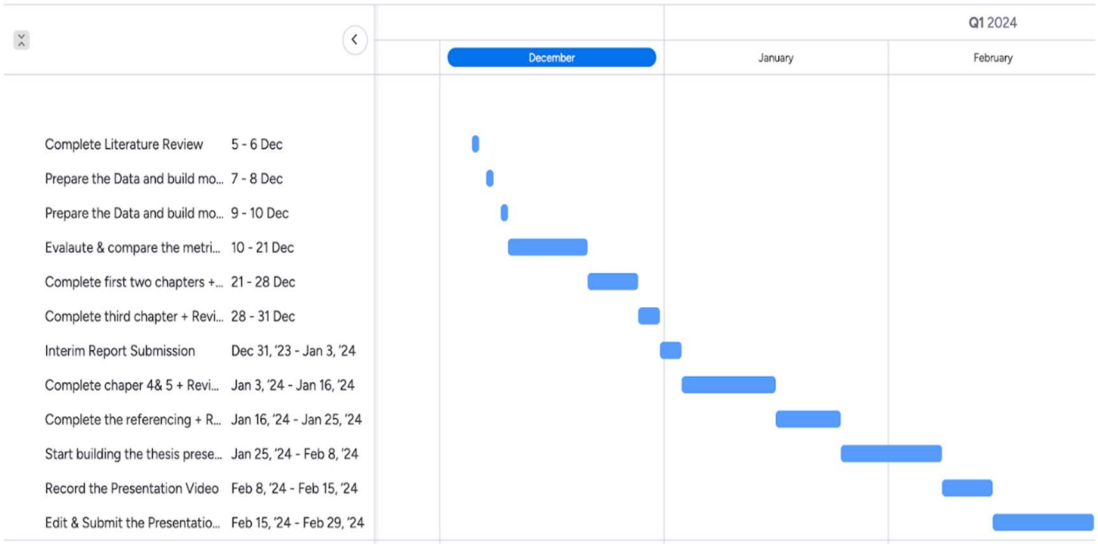


Figure 3: Gantt chart of the research plan.

References

1. K. Wanchoo, "Retail Demand Forecasting: a Comparison between Deep Neural Network and Gradient Boosting Method for Univariate Time Series," 2019 IEEE 5th

International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033651.

2. ieeexplore.ieee.org. (n.d.). Retail Sales Forecast Based on Machine Learning Methods | IEEE Conference Publication | IEEE Xplore. [online] Available at: https://ieeexplore.ieee.org/document/10129443 [Accessed 6 Dec. 2023].

3. Wang, F. and S, A.J. (2022). Using Regression Algorithms to Forecast Merchandise Sales in the Presence of Independent Variables. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICCSIE56462.2022.00030.

4. ieeexplore.ieee.org. (n.d.). Comparing Statistical and Machine Learning Methods for Sales Forecasting During the Post-promotional Period | IEEE Conference Publication | IEEE Xplore. [online] Available at: https://ieeexplore.ieee.org/document/9672954 [Accessed 6 Dec. 2023].

5. ieeexplore.ieee.org. (n.d.). A Comprehensive Analysis of Retail Sales Forecasting Using Machine Learning and Deep Learning Methods | IEEE Conference Publication | IEEE Xplore. [online] Available at: https://ieeexplore.ieee.org/document/10245887 [Accessed 6 Dec. 2023].

6. Earnest, A., Chen, M.I., Ng, D. and Sin, L.Y. (2005). Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. BMC Health Services Research, 5(1). doi:https://doi.org/10.1186/1472-6963-5-36.

7. Kraus, M., Feuerriegel, S. and Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. European Journal of Operational Research, 281(3), pp.628–641. doi:https://doi.org/10.1016/j.ejor.2019.09.018.

8. Sheo Yon Jhin, Shin, H., Hong, S., Park, S. and Park, N. (2023). Attentive neural controlled differential equations for time-series classification and forecasting. Knowledge and Information Systems. doi:https://doi.org/10.1007/s10115-023-01977-5.

9. ieeexplore.ieee.org. (n.d.). Demand Forecasting in Supply Chain Management using CNN-LSTM Hybrid Model | IEEE Conference Publication | IEEE Xplore. [online] Available at: https://ieeexplore.ieee.org/document/10307665 [Accessed 6 Dec. 2023].

10. ieeexplore.ieee.org. (n.d.). A Comparison of Prediction Algorithms in Food Sales with Different K-Folds Cross-Validation | IEEE Conference Publication | IEEE Xplore. [online] Available at: https://ieeexplore.ieee.org/document/10331578 [Accessed 6 Dec. 2023].

11. ieeexplore.ieee.org. (n.d.). Correlation Recurrent Units: A Novel Neural Architecture for Improving the Predictive Performance of Time-Series Data | IEEE Journals & Magazine | IEEE Xplore. [online] Available at: https://ieeexplore.ieee.org/document/10264112 [Accessed 6 Dec. 2023].

12. ieeexplore.ieee.org. (n.d.). Deep Learning Based Behavior Anomaly Detection within the Context of Electronic Commerce | IEEE Conference Publication | IEEE Xplore. [online] Available at: https://ieeexplore.ieee.org/document/10297230 [Accessed 6 Dec. 2023].

13. Siami-Namini, S., Tavakoli, N. and Siami Namin, A. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). doi:https://doi.org/10.1109/icmla.2018.00227.

14. ieeexplore.ieee.org. (n.d.). Forecasting the Stability of COVID-19 Vaccine Companies Stock Market using LSTM and Time-series Models | IEEE Conference Publication | IEEE Xplore. [online] Available at: https://ieeexplore.ieee.org/abstract/document/10139884 [Accessed 6 Dec. 2023].

15. Google.com. (2021). Google Scholar. [online] Available at: https://scholar.google.com/scholar_lookup?title=Exploration+of+Time-Series+Models+on+Time+Series+Data&author=Kulakou [Accessed 6 Dec. 2023].

16. Kulakou, S., Ragab, N., Midoglu, C., Boeker, M., Johansen, D., Riegler, M.A. and Halvorsen, P. (2022). Exploration of Different Time Series Models for Soccer Athlete Performance Prediction. Engineering Proceedings, [online] 18(1), p.37. doi:https://doi.org/10.3390/engproc2022018037.