Faculty of Computer
& Informatics Eng.
Departmant of Software
& Information.

كليـــة هندســـة الحـــاســـوب
والمعلومـــاتيـــــة.
قســـم هندســـة البـــرمجيـــات
ونظـم المعلومـــات.

Store’s Sales Forecasting

Senior Project (1) - Completed the requirements for obtaining a bachelor's degree in Informatics Engineering – Software Engineering and Information Systems

Prepared by

Muhammad Khaled Al-Hussein

Muhammad Kenan Al-Akkad

Muhammad Fakher Aqeel

Supervised by

Dr. Kadan Al-Joumaa

Eng. Aya Al-Aswad

Academic year

2023-2024

Faculty of Computer
& Informatics Eng.
Departmant of Software
& Information.

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

كليـــة هندســـة الحـــاســـوب
والمعلومـــاتيــــــة.
قســـم هندســـة البـــرمجيـــات
ونظم المعلومـــات.

التنبوء في مبيعات متجر

مشروع التخرج (1) - قدم إستكمالً لمتطلبات الحصول على درجة البكالوريوس في هندسة المعلوماتية - هندسة البرمجيات ونظم المعلومات

إعداد

محمد خالد الحسين

محمد كنان العقاد

محمد فاخر عقيل

إشراف

د.كادان جمعة

م.أية الأسود

السنة الدراسية

2024-2023

Supervisor Certification

I certify that the preparation of this project entitled "Store's Sales Forecasting", prepared by Muhammad Al-Hussein, Muhammad Kenan Al-Akkad, and Muhammad Fakher Aqeel, was made under my supervision at Department of Software and Information System Engineering / Faculty of Computer and Informatics Engineering in partial fulfillment of the requirements of the Degree of Bachelors of Software and Information System Engineering.

Name	**Name**	**Date**
Dr. Kadan Al-Joumaa	Eng. Aya Al-Aswad	5/11/2024
Signature	**Signature**	

Abstract

Sales forecasting is essential for retail management, helping businesses manage inventory, staffing, and marketing effectively. In today's data-driven world, advanced technologies like machine learning, deep learning, data science, and data analytics are replacing traditional methods for better accuracy and adaptability. This study focuses on using these technologies to build models for predicting store sales.

We explore different machine learning models, including time series analysis and regression, and discuss the steps for data preprocessing and feature engineering needed to make accurate predictions. The research also examines the impact of external factors like promotions, seasonality, and economic indicators on sales forecasts and how to incorporate these factors into the models.

Additionally, the study highlights the advantages of probabilistic forecasting, which provides a better understanding of the uncertainties in sales predictions. It emphasizes the importance of evaluating and selecting models carefully and the need for continuous refinement to keep up with market changes and consumer behavior.

Overall, this research demonstrates how machine learning can enhance sales forecasting in retail, giving businesses a competitive edge in a rapidly changing market.

الملخص

يعد توقعات المبيعات جزءًا أساسيًا من إدارة التجزئة، حيث تساعد الشركات على إدارة المخزون والموظفين واستراتيجيات التسويق بشكل فعال. في عالم اليوم المعتمد على البيانات، تحل التقنيات المتقدمة مثل التعلم الآلي، التعلم العميق، علم البيانات، وتحليل البيانات محل الطرق التقليدية لتحقيق دقة وتكيف أفضل. تركز هذه الدراسة على استخدام هذه التقنيات لبناء نماذج لتوقع مبيعات المتاجر.

نستكشف نماذج التعلم الآلي المختلفة، بما في ذلك تحليل السلاسل الزمنية والانحدار، ونناقش خطوات إعداد البيانات والهندسة المميزة اللازمة لتحقيق توقعات دقيقة. كما تبحث الدراسة في تأثير العوامل الخارجية مثل العروض الترويجية، والموسمية، والمؤشرات الاقتصادية على توقعات المبيعات وكيفية دمج هذه العوامل في النماذج.

بالإضافة إلى ذلك، تسلط الدراسة الضوء على فوائد التوقعات الاحتمالية، التي توفر فهمًا أفضل لعدم اليقين في توقعات المبيعات. كما تؤكد على أهمية تقييم النماذج واختيارها بعناية، والحاجة إلى تحسين مستمر لمواكبة التغيرات في السوق وسلوك المستهلك.

بشكل عام، توضح هذه الدراسة كيف يمكن أن يعزز التعلم الآلي توقعات المبيعات في قطاع التجزئة، مما يمنح الشركات ميزة تنافسية في سوق سريع التغير.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ARIMA	Autoregressive Integrated Moving Average
MAE	The average of the absolute differences between predicted and actual values. It gives a measure of the average magnitude of errors.
RMSE	Similar to MAE but emphasizes larger errors. It's the square root of the average of squared differences between predicted and actual values.
MAPE	Measures the percentage difference between predicted and actual values. It provides a relative error as a percentage.
R2	Measures the proportion of the variance in the dependent variable that is predictable from the independent variable.
Pros	Advantages
cons	Disadvantages
LR-ELM	Linear Regression Entry-Level Model
A-ELM	ARIMAR Entry-Level Model
ES-ELM	Exponential Smoothing Entry Level Model
LR-PM	Linear Regression Primary Model
A-PM	ARIMA Primary Model
ES-PM	Exponential Smoothing Primary Model
DeepAR	Probabilistic forecasting with autoregressive recurrent networks.

Chapter One

Introduction

Chapter One – Introduction

Forecasting is a process of making predictions or estimates about future events based on past and present data and analysis. It involves using various methods and techniques to anticipate future trends, outcomes, or developments in a specific area. Forecasting is commonly applied in various fields, including finance, economics, weather, sales, demand planning, and many others.

There are different methods of forecasting, and the choice of method depends on the nature of the data, the context of the prediction, and the available resources. For Example:

Time Series Analysis	Regression Analysis	Machine Learning Models
This involves analyzing historical data to identify patterns and trends over time. Techniques such as moving averages, exponential smoothing, and autoregressive integrated moving average (ARIMA) models are often used in time series analysis.	Regression models are used when there is a relationship between the variable to be predicted and one or more independent variables. It helps in understanding the strength and nature of the relationship.	Advanced techniques like machine learning algorithms, including decision trees, neural networks, and support vector machines, can be employed for forecasting tasks. These models are capable of handling complex patterns and relationships in data.

Table 1. 1. Forecasting Methods.

Qualitative Methods	Simulation and Scenario Analysis
In situations where historical data is limited or not available, qualitative methods involve expert judgment, market research, and subjective opinions to make predictions.	Simulation involves creating a model that represents a system and then running different scenarios to observe the outcomes. This is often used in complex systems where various factors interact.
The accuracy of forecasts can vary, and it's important to consider the limitations and uncertainties associated with any forecasting method. While forecasting provides valuable insights for decision-making, it's essential to regularly update predictions as new data becomes available and conditions change.	

Table 1. 2. Forecasting Methods.

In this chapter, we will be looking at:

1. **Introduction.**
2. **The Pros and Cons of Sales Forecasting.**
3. **The Aim of sales forecasting.**
4. **Sales forecasting models.**
5. **Report sections.**
6. **Summary.**

Introduction:

What is store sales forecasting? Store sales forecasting is a specific type of forecasting that focuses on predicting the future sales of a retail store. This process involves estimating the amount of revenue a store is likely to generate over a specific period, typically on a daily, weekly, monthly, or seasonal basis. Accurate sales forecasting is crucial for effective inventory management, staff scheduling, and overall business planning.

I. key aspects and methods involved in in-store sales forecasting:

Historical Data Analysis	**Time Series Analysis**	**Seasonal Adjustments**	**Promotion and Marketing Impact**
Retailers often analyze historical sales data to identify patterns, trends, and seasonality. This involves looking at sales figures for comparable periods in the past to understand how sales have behaved under similar conditions.	Time series forecasting techniques, such as moving averages, exponential smoothing, and ARIMA models, are commonly used in in-store sales forecasting. These methods take into account past sales data to project future sales.	Many retails businesses experience seasonality, with sales patterns influenced by factors like holidays, weather, or cultural events. Adjusting for these seasonal variations helps in making more accurate predictions.	Anticipating the impact of promotions, discounts, and marketing campaigns is essential. Store sales forecasting should consider how these factors influence customer behavior and purchasing patterns.
Market Trends and External Factors	**Inventory Levels**	**Customer Behavior Analysis**	**Technology and Data Analytics**
External factors, such as economic conditions, industry trends, and changes in consumer behavior, can impact store sales. Analyzing and incorporating these factors into the forecasting process enhances its accuracy.	Sales forecasting is closely linked to inventory management. Predicting sales accurately helps retailers maintain optimal inventory levels, reducing the risk of stockouts or overstock situations.	Understanding customer behavior, preferences, and purchasing patterns is crucial. This can involve analyzing data from loyalty programs, customer surveys, and other sources to gain insights into what drives sales.	increasingly used in in-store sales forecasting. These technologies can analyze large datasets quickly and identify complex patterns that may not be apparent through traditional methods.

Table 1. 3. Key Aspects & Forecasting Methods.

Accurate store sales forecasting helps retailers optimize their operations, improve customer satisfaction, and ultimately maximize profitability. It's an ongoing process that requires continuous refinement and adjustment based on actual sales data and changing market conditions.

The Pros and Cons of Store Sales Forecasting:

While store sales forecasting offers numerous benefits, it's important for businesses to be aware of its limitations and continually refine their methods based on real-time data and market feedback. The goal is to strike a balance between precision and adaptability in response to changing conditions.

Pros	Cons
Improved Inventory Management: Accurate sales forecasting enables retailers to maintain optimal inventory levels. This reduces the risk of stockouts, excess inventory, and associated costs.	**Inaccuracy and Uncertainty:** Forecasting is inherently uncertain, and inaccuracies can arise due to unforeseen events, changes in consumer behavior, or external factors.
Effective Staff Scheduling: With reliable sales predictions, retailers can schedule staff more efficiently, aligning workforce levels with expected customer traffic.	**Dependency on Historical Data:** Over-reliance on historical data may result in inaccurate predictions if significant changes occur in the market or if there are disruptions that deviate from historical patterns.
Cost Reduction: Efficient inventory management and staff scheduling lead to cost savings. Retailers can avoid unnecessary holding costs for excess inventory and optimize labor costs.	**Complexity of External Factors:** External factors such as economic conditions, political events, or unexpected market trends can be challenging to predict accurately, making it difficult to account for them in forecasts.
Enhanced Customer Satisfaction: Maintaining adequate stock levels ensures that customers can find the products they want when they visit the store, improving overall customer satisfaction.	**Data Quality Issues:** Poor data quality, incomplete data, or inaccurate data can negatively impact the accuracy of forecasts. It's crucial to ensure that the data used in forecasting is reliable.
Strategic Decision-Making: Sales forecasting provides valuable insights for strategic planning. It helps retailers make informed decisions about promotions, marketing campaigns, and expansion plans.	**Overemphasis on Short-Term Results:** Some forecasting methods may focus too much on short-term results, potentially overlooking long-term trends or shifts in consumer behavior.
Optimized Supply Chain: Retailers can work closely with suppliers to align production and delivery schedules with forecasted demand, creating a more streamlined and efficient supply chain.	**Resistance to Change:** Implementing forecasting processes may face resistance from employees or management who may be skeptical of the accuracy of predictions or resistant to adopting new methods.
Resource Allocation: Knowing when peak sales periods are likely to occur allows for better allocation of resources, both in terms of personnel and marketing efforts.	**Dynamic Market Conditions: R**apid changes in market conditions, technological advancements, or competitive landscapes can make it challenging for forecasts to keep pace with the evolving business environment.
Improved Cash Flow: By preventing overstock situations and reducing holding costs, accurate forecasting contributes to improved cash flow for the business.	**Model Complexity:** Advanced forecasting models, including machine learning algorithms, can be complex and require specialized expertise. Implementing and maintaining such models may pose challenges for some businesses.

Table 1. 4. Pros & Cons of Store Sales Forecasting.

The Aim of the Sales Forecasting Project:

The primary aim of this project is to develop a robust and accurate sales forecasting model. By doing so, we hope to mitigate the risks associated with inaccurate predictions and enable businesses to make informed decisions regarding their inventory management, production planning, and budget allocation.

Sales Forecasting Models:

Several forecasting models can be used for store sales forecasting, ranging from simple to more complex methods. The choice of a particular model depends on the nature of the data, the available resources, and the level of accuracy required. Here are some common store sales forecasting models:

I. **Moving Averages:**
   - **Description:** Moving averages involve calculating the average of a certain number of past data points to predict future sales.
   - **Application:** Simple Moving Average (SMA), Weighted Moving Average (WMA), and Exponential Moving Average (EMA) are commonly used.

II. **Time Series Analysis:**
   - **Description:** Time series analysis involves studying past sales data to identify patterns, trends, and seasonality. This forms the basis for predicting future sales.
   - **Application:** Autoregressive Integrated Moving Average (ARIMA) models are frequently used for time series forecasting.

III. **Regression Analysis:**
   - **Description:** Regression models establish relationships between the sales variable and other relevant independent variables, such as marketing spend, promotional activities, or economic indicators.
   - **Application:** Linear regression, multiple regression, and polynomial regression are examples.

IV. **Machine Learning Models:**
   - **Description:** Advanced machine learning algorithms can be employed for more complex and dynamic forecasting tasks. These models can handle non-linear relationships and intricate patterns in the data.
   - **Application:** Decision trees, Random Forests, Support Vector Machines, Neural Networks, and Gradient Boosting are examples of machine learning models used in sales forecasting.

V. **Neural Networks:**
   - **Description:** Neural networks, a subset of machine learning, simulate the human brain's ability to learn. They are capable of handling complex patterns and relationships in data.
   - **Application:** Feedforward neural networks and recurrent neural networks (RNNs) are commonly used in sales forecasting.

VI. **Seasonal Decomposition of Time Series (STL):**
- **Description:** STL decomposes time series data into three components: seasonal, trend, and remainder. This allows for a more granular analysis of patterns.
- **Application:** Particularly useful when dealing with data with strong seasonality.

VII. **ARIMA-X:**
- **Description:** Extends the ARIMA model by incorporating exogenous variables, such as external factors or marketing activities, to enhance forecasting accuracy.
- **Application:** Suitable when external variables significantly impact sales.

VIII. **Prophet:**
- **Description:** Developed by Facebook, Prophet is designed for forecasting with daily observations that display patterns on different time scales. It can handle missing data and outliers.
- **Application:** Particularly useful for retail sales forecasting due to its ability to handle holidays and special events.

IX. **Markov Chains:**
- **Description:** Markov chains model transitions between different states based on probabilities. It's useful when the future state depends only on the current state and not on the sequence of events that preceded it.
- **Application:** Suitable for predicting sales transitions between different product categories or states.

X. **Ensemble Methods:**
- **Description:** Ensemble methods combine the predictions of multiple models to improve overall accuracy and robustness.
- **Application:** Bagging (Bootstrap Aggregating), Boosting, and Stacking are examples of ensemble methods.

XI. **Deep AR Model:**
- **Description:** Deep AR is a deep learning RNN (recurrent neural network) sequence-to-sequence model specifically designed for time series forecasting. Developed by Amazon Web Services (AWS), it excels at handling the complexities inherent in time series data.
- **Application:** In retail settings, sales are influenced by various external factors (covariates) such as holidays, promotions, and store-specific events. By incorporating covariates into the forecasting model, Deep AR can better capture these influences and improve forecast accuracy. Research findings show that including time-, event-, and ID-related features significantly enhances forecast accuracy. All Deep AR models, with or without covariates, outperform the seasonal naïve benchmark, making Deep AR a powerful tool for store sales forecasting. Retailers can leverage this approach to make more accurate predictions and optimize their operations.

When implementing store sales forecasting models, it's crucial to continually evaluate and refine the models based on the latest data and business insights. The choice of a specific model depends on factors like the availability of data, the complexity of the sales patterns, and the business context.

Report Sections:

This report is divided into several Chapters. After this introduction, we will discuss the Project Topic & and a Reference Study in more detail. This will be followed by a chapter on the analytical Study and entry-level Models. We will then delve into the Practical Application. Finally, we will conclude the results of the project.

Summary:

The chapter presents a comprehensive view of sales forecasting, emphasizing its importance in business planning. It discusses the complexity of predicting sales, impacted by various factors, and aims to develop an accurate forecasting model to reduce risks from incorrect predictions. The report covers different models like Time Series Analysis, Regression Analysis, and Machine Learning, each with their strengths and weaknesses. It also addresses challenges in sales forecasting and proposes enhancements such as improving data quality and leveraging advanced analytics. The goal is to enable informed business decisions through robust sales forecasting.

Chapter Two

Project Topic & a Reference Study

Chapter Two – Project Topic & Reference Study.

Store sales forecasting is a critical process in the retail industry that involves predicting a retail store's future sales performance. This strategic undertaking relies on the analysis of historical sales data, market trends, and various influencing factors to anticipate consumer demand accurately. The primary goal is to optimize operational aspects such as inventory management, staff scheduling, and resource allocation. Different forecasting methods, including time series analysis, regression models, and advanced machine learning algorithms, are employed to generate precise predictions. By understanding past sales patterns, seasonal variations, and external influences like promotions or economic conditions, retailers can make informed decisions to enhance customer satisfaction, streamline supply chains, and ultimately maximize profitability. Store sales forecasting is an ongoing and dynamic process, requiring continuous refinement and adaptation to changing market conditions and consumer behavior.

In this chapter, we will be looking at:

1. **Introduction.**
2. **Reference Study.**
3. **Metrics exploration.**
4. **Understanding and Leveraging Data.**
5. **Used Tools.**
6. **Summary.**

Introduction:

In the realm of business analytics, the art of store sales forecasting plays a pivotal role in strategic decision-making for retail enterprises. This chapter serves as the foundation for a comprehensive exploration into the essential domain of store sales forecasting. The chapter begins by elucidating the critical significance of store sales forecasting in optimizing various facets of retail operations. It then delves into a substantial reference study that meticulously dissects three prominent forecasting models—linear regression, ARIMA, and triple exponential smoothing. These models are chosen for their proven efficacy in predicting store sales with varying degrees of complexity.

In addition to these traditional methods, the chapter introduces advanced approaches, including deep learning techniques, specifically focusing on the Deep AR model. Deep learning, with its capacity to handle large datasets and uncover intricate patterns, offers a powerful alternative to conventional models. The Deep AR model, leveraging recurrent neural networks, excels in capturing temporal dependencies and producing more accurate forecasts in complex, dynamic environments.

Reference study:

Embarking on the intricate landscape of sales forecasting methodologies, Time Series Analysis and Machine Learning emerge as powerful tools, each bringing its unique set of characteristics, advantages, and considerations to the table. Furthermore, the specific forecasting models—Linear Regression, ARIMA, and Holt-Winters Triple Exponential Smoothing—contribute distinctive strengths and limitations to the predictive analytics toolkit.

I. **Time Series Analysis:**

- **Definition:** Time Series Analysis is a method meticulously designed to model time-dependent data, focusing on forecasting future trends by scrutinizing critical patterns such as seasonality, trends, and cyclic behavior.
- **Pros:**

  i. **Effectiveness in Capturing Patterns:** Proficient in capturing and leveraging seasonal and historical patterns.
  ii. **Clear Trend Analyses:** Provides lucid trend analyses for better strategic planning.
  iii. **Suitability for Stable Markets:** Particularly effective in stable markets exhibiting consistent trends.

- **Cons:**

  i. **Assumption of Stability:** Assumes stability in historical patterns, posing challenges in rapidly changing markets.
  ii. **Sensitivity to Outliers:** May not adapt well to unexpected events and can be sensitive to outliers.
  iii. **Comparison:** Time Series Analysis excels in stable datasets with clear trends and seasonality. However, it may be less adaptable to sudden market changes compared to the flexibility offered by machine learning approaches.

- **Comparison:** Highly effective for stable datasets with clear trends and seasonality but less adaptable to sudden market changes or non-linear patterns compared to machine learning approaches.
- **Examples:** Predicting monthly sales figures based on historical data, and forecasting quarterly financial performance.

II. Machine Learning in Sales Forecasting:

- **Definition:** Machine Learning involves utilizing various algorithms to learn from and make predictions based on data, identifying complex patterns and relationships that traditional methods might miss.
- **Pros:**

  i. **Handling Complex Relationships:** Capable of handling non-linear and intricate relationships in data.
  ii. **Adaptation to New Trends:** Quick adaptation to new trends and incorporation of diverse data types and sources.
  iii. **Flexibility:** More flexible and powerful in handling dynamic, non-linear patterns.

- **Cons:**

  i. **Data Requirements:** Requires large amounts of data for effective model training.
  ii. **Complex Setup:** this can be complex to set up and understand.
  iii. **Data Dependency:** The accuracy of predictions heavily depends on the quality and quantity of the data.
  iv. **Comparison:** Machine learning stands out for its flexibility and adaptability to dynamic markets, especially suited for handling complex, non-linear patterns. However, it requires a substantial amount of data and computational resources compared to traditional Time Series Analysis.

- **Comparison:** More flexible and powerful in handling complex, non-linear patterns and adapting to new data, making it suitable for dynamic and rapidly changing markets. However, it requires more data and computational resources than traditional time series analysis.
- **Examples:** Using algorithms to predict future sales based on customer behavior, incorporating social media data for sales forecasts.

III. Deep Learning in Sales Forecasting:

- **Definition**: Deep learning, a subset of machine learning, involves neural networks with multiple layers that can model complex patterns in large datasets. In sales forecasting, deep learning models like Deep AR leverage recurrent neural networks to capture temporal dependencies and trends in sales data, offering a sophisticated alternative to conventional models.
- **Pros**:
  i. **Handling Complex Relationships**: Capable of managing non-linear and intricate relationships in data, providing more accurate predictions.

ii. **Adaptation to New Trends**: Quickly adapts to new trends and incorporates diverse data types and sources, making it highly responsive to changing market conditions.
iii. **Flexibility**: More flexible and powerful in handling dynamic, non-linear patterns compared to traditional forecasting methods.

- **Cons**:
  i. **Data Requirements**: Requires large amounts of data for effective model training, which can be a barrier for smaller datasets.
  ii. **Complex Setup**: Implementation can be complex and requires significant expertise and computational resources.
  iii. **Data Dependency**: The accuracy of predictions heavily depends on the quality and quantity of the data, making it sensitive to data inconsistencies.
- **Comparison**: Deep learning stands out for its flexibility and adaptability to dynamic markets, especially suited for handling complex, non-linear patterns. However, it requires a substantial amount of data and computational resources compared to traditional time series analysis, which can be more straightforward to implement and interpret.
- **Examples**: Utilizing algorithms to predict future sales based on customer behavior, incorporating social media data for more comprehensive sales forecasts, and leveraging recurrent neural networks in the Deep AR model for more precise temporal predictions.

IV. Sales Forecasting Models:

I. Linear Regression:

- **Definition:** Linear Regression is a statistical approach that models the relationship between a dependent variable and one or more independent variables.
- **Pros:**

  i. **Simplicity and Interpretability:** Simple and interpretable, providing a clear understanding of relationships.
  ii. **Transparent Results:** Offers transparent results, making it accessible for non-technical users.

- **Cons:**

  i. **Oversimplification:** May oversimplify complex trends and ignore seasonal or cyclic patterns.
  ii. **Limited Flexibility:** Less flexible in capturing intricate relationships compared to more complex models.
  iii. **Comparison:** Linear Regression, while straightforward, may not capture complex trends and patterns as effectively as more intricate models.

- **Comparison:** Straightforward to interpret but may lack sophistication in capturing intricate patterns compared to time series or machine learning methods.
- **Examples:** Predicting sales based on advertising spend, estimating revenue growth based on marketing efforts.

V. ARIMA (Autoregressive Integrated Moving Average):

- **Definition:** A complex model combining autoregression, differencing (to achieve stationarity), and a moving average. Effective for a wide range of time series data, especially non-seasonal patterns.
- **Pros:**

  i. **Versatility:** Effective for a wide range of time series data, especially non-seasonal patterns.
  ii. **Comprehensive Approach:** Incorporates autoregression, differencing, and a moving average for a holistic analysis.

- **Cons:**

  i. **Parameter Sensitivity:** Requires careful parameter selection, and may struggle with seasonal data.
  ii. **Complexity:** The complexity of the model may be challenging for some users.
  iii. **Comparison:** ARIMA is versatile, especially for non-seasonal patterns, but demands careful consideration of parameters and may not be ideal for datasets with strong seasonality.

- **Comparison:** This offers more flexibility for non-seasonal data compared to linear regression but may require more fine-tuning for seasonality.
- **Examples:** Forecasting monthly sales with a focus on non-seasonal trends, and predicting stock prices.

VI. Holt-Winters Triple Exponential Smoothing:

- **Definition:** Tailored for data with trends and seasonal patterns, applying three smoothing equations to capture level, trend, and seasonality.
- **Pros:**

  i. **Seasonal Handling:** Particularly strong in handling seasonality, making it suitable for data with trends and seasonality.
  ii. **Holistic Approach:** Applies three smoothing equations for comprehensive analysis.

- **Cons:**

  i. **Configuration Complexity:** Can be complex to configure, especially for users unfamiliar with the method.
  ii. **Effectiveness with Non-Seasonal Data:** Less effective when dealing with datasets lacking strong seasonality.
  iii. **Comparison:** Holt-Winters excels in capturing seasonal fluctuations but requires careful configuration and may not be as effective with non-seasonal data.

- **Comparison:** Excels in capturing seasonal fluctuations but may be less versatile in scenarios without clear seasonality.
- **Examples:** Predicting quarterly sales with distinct seasonality, forecasting holiday season retail demand.

VII. XGBoost Model:

- **Definition**: XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that uses a gradient boosting framework. It builds an ensemble of decision trees sequentially, where each tree corrects the errors of its predecessor, optimizing performance.
- **Pros**:

  i. **High Performance**: Known for its high accuracy and efficiency in handling large datasets.
  ii. **Feature Importance**: Provides insights into feature importance, which helps in understanding the contributing factors to sales.
  iii. **Scalability**: Can be scaled to large datasets and integrated into production systems.

- **Cons**:

  i. **Complexity**: Can be complex to tune and requires careful parameter optimization.
  ii. **Overfitting**: Prone to overfitting if not properly regularized.
  iii. **Computational Resources**: Requires significant computational power for training large models.

- **Comparison**: XGBoost is more robust and can handle a wide variety of data types and structures, making it suitable for complex forecasting tasks. However, it requires more effort in model tuning and computational resources compared to simpler models like linear regression.
- **Examples**: Predicting future sales by analyzing past sales data, customer demographics, and seasonal effects using an ensemble of decision trees.

VIII. Custom Deep Learning Neural Network:

- **Definition**: Custom Deep Learning Neural Networks involve designing and training neural networks with multiple layers tailored to specific forecasting needs. These networks can capture complex, non-linear relationships in sales data.
- **Pros**:

  i. **Customization**: Highly customizable to fit specific forecasting requirements and data characteristics.
  ii. **Complex Pattern Recognition**: Excels at recognizing complex patterns and interactions in data.
  iii. **Adaptability**: Can adapt to a wide range of forecasting scenarios and data types.

- **Cons**:

  i. **Development Time**: Requires significant time and expertise to design and train.
  ii. **Data Requirements**: Needs large amounts of high-quality data for effective training.
  iii. **Black Box Nature**: Can be difficult to interpret and understand the decision-making process.

- **Comparison**: Custom Deep Learning Neural Networks offer unparalleled flexibility and can model highly complex relationships in data. However, they require extensive computational resources and expertise in neural network design and training.
- **Examples**: Designing a network to predict sales by incorporating features like historical sales, promotions, holidays, and economic indicators.

IX. Deep AR Model:

- **Definition**: Deep AR is a deep learning model specifically designed for time series forecasting. It utilizes recurrent neural networks to capture temporal dependencies and trends in sequential data.
- **Pros**:

  i. **Temporal Dependencies**: Excellent at modeling temporal dependencies in sequential data.
  ii. **Accuracy**: Provides highly accurate forecasts by leveraging complex neural network architectures.
  iii. **Scalability**: Scalable to large datasets and can handle multiple time series simultaneously.

- **Cons**:

  i. **Data Intensive**: Requires large volumes of data for effective training.
  ii. **Computationally Expensive**: Demands significant computational resources for training and inference.
  iii. **Complexity**: Can be complex to implement and tune, requiring deep learning expertise.

- **Comparison**: Deep AR excels in capturing complex temporal patterns in time series data, making it ideal for scenarios with multiple, correlated time series. It is more powerful than traditional time series models but requires more data and computational resources.
- **Examples**: Using Deep AR to forecast daily sales across multiple stores, incorporating seasonality, trends, and holiday effects.

X. Research papers comparison table:

Title	Focus	Key Contributions	Methodologies	Applicability	Limitations
Deep AR: Probabilistic Forecasting with Autoregressive Recurrent Networks	Introduction and explanation of the Deep AR algorithm.	Establishes the framework for using recurrent neural networks for probabilistic forecasting.	Utilizes autoregressive recurrent networks trained on large, related time series datasets.	Broad applicability in any field requiring time series forecasting, such as retail, finance, and energy.	The paper primarily sets the stage without addressing specific operational challenges like handling sparse data or high-dimensional feature spaces.
Amazon Sage Maker Deep AR now supports missing values, categorical and time series features, and generalized frequencies	Enhancements to the Deep AR model to handle more complex data scenarios.	Introduces capabilities to handle missing values, utilize categorical data, and support generalized frequencies.	Extension of the original Deep AR model to be more robust and flexible.	Very applicable to real-world scenarios where data often has missing values and categorical features.	The paper might not provide comparative analysis with other models or discuss the computational overhead of these enhancements.
Comparison Analysis of Facebook's Prophet, Amazon's Deep AR+ and CNN-QR Algorithms for Successful Real-World	Comparative study of different forecasting models including Deep AR.	Evaluates the performance of Deep AR against other popular models, providing a benchmark for its effectiveness.	Empirical evaluation using real-world datasets to measure accuracy and performance.	Helps practitioners choose among different forecasting tools based on performance metrics.	May not delve deeply into why certain models perform better than others or under what specific conditions.

Sales Forecasting					
Robust Sales forecasting Using Deep Learning with Static and Dynamic Covariates	Enhancing deep AR models by incorporating a wide range of covariates.	Demonstrates how the inclusion of both static and dynamic covariates can improve forecasting accuracy.	Advanced feature engineering to integrate additional data into the forecasting model.	Useful for complex forecasting tasks where multiple influencing factors must be considered.	Could be computationally intensive; might require extensive data preprocessing.
Advanced Techniques in Time Series Forecasting with Deep AR	Exploring advanced techniques and methodologies to extend the Deep AR framework.	Likely introduces new methodologies or optimizations to enhance the Deep AR algorithm.	Could involve sophisticated model adjustments, tuning, or hybrid approaches.	Aimed at improving model performance and extending its use to more complex scenarios.	As an advanced topic, it may be less accessible to beginners and might require deep technical knowledge to implement.

Table 2. 1. Comparing Research Papers.

XI. Specific problems for each paper:

Title	**Deep AR: Probabilistic Forecasting with Autoregressive Recurrent Networks**	**Amazon Sage Maker Deep AR now supports missing values, categorical and time series features, and generalized frequencies**	**Comparison Analysis of Facebook's Prophet, Amazon's Deep AR+ and CNN-QR Algorithms for Successful Real-World Sales Forecasting**	**Robust Sales forecasting Using Deep Learning with Static and Dynamic Covariates**	**Advanced Techniques in Time Series Forecasting with Deep AR**
Specific Problem	As an introductory paper, it does not address the operational challenges of implementing the Deep AR model in real-world settings, such as dynamic environments where data patterns frequently change.	While it addresses missing data and categorical features, the paper might not cover the nuances of handling highly irregular time series or non-standard frequency data, which are common in	This comparative study might lack depth in explaining the underlying reasons for the performance differences between the models, such as the structural differences in the data that might favor	Integrating a wide range of covariates can introduce complexity in model training and tuning. The paper might not fully explore the trade-offs between model accuracy and interpretability when additional	Advanced techniques can often lead to increased model complexity, making the model harder to understand and explain. This can be problematic in industries where explain ability is crucial, such

		real-world applications.	one model over another.	covariates are included.	as finance and healthcare.

Table 2. 2. Research Papers Problems.

XII. Solution for specific problems for each paper:

Title	**Deep AR: Probabilistic Forecasting with Autoregressive Recurrent Networks**	**Amazon Sage Maker Deep AR now supports missing values, categorical and time series features, and generalized frequencies**	**Comparison Analysis of Facebook's Prophet, Amazon's Deep AR+ and CNN-QR Algorithms for Successful Real-World Sales Forecasting**	**Robust Sales forecasting Using Deep Learning with Static and Dynamic Covariates**	**Advanced Techniques in Time Series Forecasting with Deep AR**
Solution	Incorporate adaptive techniques that can adjust to changing data patterns dynamically. Methods like online learning or rolling re-training schedules can help the model stay relevant as underlying data distributions change.	Develop advanced imputation techniques tailored to the specific characteristics of time series data, such as using interpolation or regression models to estimate missing values based on observed temporal patterns.	Enhance the comparative analysis by including a deeper investigation into data characteristics that influence model performance. Techniques such as feature importance analysis and sensitivity analysis can provide insights into why certain models perform better under specific circumstances.	Implement dimensionality reduction techniques to manage the complexity introduced by a large number of covariates. Techniques such as Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) can reduce the number of input features to a manageable size while retaining most of the important information.	Focus on developing hybrid models that combine deep learning with simpler, more interpretable models. This can help maintain a balance between accuracy and explain ability. Techniques like model distillation, where a simpler model learns to replicate the performance of a more complex model, can also help in improving explain ability.

Table 2. 3. Solutions to Research Papers.

XIII. Deep AR Vs Previous Models:

i. Model Overview and Assumptions:

	Linear Regression	**ARIMA**	**Holt-Winters Triple Exponential Smoothing**	**Deep AR**
Model Overview and Assumptions	Assumes a linear relationship between the dependent variable and one or more independent variables. It doesn't inherently account for time-based dependencies unless time-related variables are explicitly included as features.	Combines autoregressive (AR) terms, differencing (I) to achieve stationarity, and moving average (MA) terms. It assumes a linear relationship in transformed stationary data and is best suited for data with no significant seasonal or non-linear patterns unless extended to SARIMA for seasonal effects.	Designed for time series data with trends and seasonal variations. It uses level, trend, and seasonal components, each updated exponentially over time, making it powerful for seasonal data.	A probabilistic, autoregressive model using recurrent neural networks, specifically designed for complex, nonlinear patterns that can vary over different segments of the dataset. It models the probability distribution of future values based on past values and covariates.

Table 2. 4. Comparing Deep AR with Old Models.

ii. Data Requirements and Preprocessing:

	Linear Regression	**ARIMA**	**Holt-Winters Triple Exponential Smoothing**	**Deep AR**
Data Requirements and Preprocessing	Requires careful feature selection and engineering to incorporate time dynamics. Sensitive to outliers and requires stationary data for reliable predictions.	The data must be made stationary, which often involves differencing and log transformations. Good at modeling autocorrelations but requires manual identification of model parameters (p, d, q).	Requires at least two full seasonal cycles for effective parameter estimation, making it less suitable for very short or non-seasonal series.	Handles large datasets with multiple correlated time series effectively. It can automatically learn from covariates and does not require stationarity. Deep AR is robust to missing values and the inclusion of additional covariates.

Table 2. 5. Comparing Data Requirements Based on Model.

iii. Complexity and Flexibility:

	Linear Regression	**ARIMA**	**Holt-Winters Triple Exponential Smoothing**	**Deep AR**
Complexity and Flexibility	Simple to implement and understand but lacks flexibility in handling non-linear relationships without transformation or adding interaction terms.	Can be complex to configure properly due to the need for initial stationarity analysis and parameter tuning (model order selection).	Relatively simple to implement but its parameters (alpha, beta, gamma) need tuning based on seasonal components. Less flexible for handling abrupt changes in trends or non-regular seasonal patterns.	Highly flexible and capable of modeling complex patterns but requires significant computational resources for training. The use of deep learning techniques allows it to adaptively learn from data characteristics.

Table 2. 6. Comparing models Strengths and Weaknesses.

iv. Typical Use Cases:

	Linear Regression	**ARIMA**	**Holt-Winters Triple Exponential Smoothing**	**Deep AR**
Typical Use Cases	Useful for forecasting when relationships between predictors and forecasts are expected to be linear and when the impact of individual factors needs to be understood.	Effective for non-seasonal time series data that can be made stationary. Widely used in economics and finance for its interpretability.	Best for data with stable seasonal patterns and trends, commonly used in retail and demand forecasting for seasonal products.	Ideal for complex datasets with multiple interacting time series, such as demand forecasting across multiple products/stores in retail, or in scenarios where probabilistic forecasts are needed.

Table 2. 7. Comparing models By Use.

v. Pros and Cons:

	Linear Regression	**ARIMA**	**Holt-Winters Triple Exponential Smoothing**	**Deep AR**

Pros	Simplicity, interpretability.	Effective at capturing relationships in lagged data.	Explicitly models seasonality and trend.	Handles complex, non-linear, and seasonal patterns, multiple series with covariates.
Cons	Poor handling of non-linearities, autocorrelation, and non-stationarity.	Does not handle seasonality or non-linearities well without modifications.	Assumes seasonality is fixed and regular, which may not hold in all cases.	Requires substantial computational power, more complex to tune and interpret compared to traditional methods.

Table 2. 8. Models Pros & Cons.

the choice of a sales forecasting methodology and model hinges on the specific characteristics of the data and the objectives of the business. Time Series Analysis and Machine Learning offer distinct advantages and considerations, and the selection of Linear Regression, ARIMA, or Holt-Winters depends on the nuances of the dataset at hand. The comprehensive overview provides valuable insights for businesses navigating the complexities of sales forecasting.

Metrics Exploration:

Evaluating the performance of forecasting models is essential to gauge their accuracy and reliability in predicting future outcomes. Metrics play a crucial role in quantifying the effectiveness of models across different dimensions. In this analysis, we delve into the key metrics employed to assess the performance of three distinct forecasting models: Linear Regression, ARIMA (Autoregressive Integrated Moving Average), and Holt-Winters. Each metric offers unique insights into the models' predictive capabilities, shedding light on their strengths and areas for improvement.

I. Linear Regression Metrics:

- **Mean Absolute Error (MAE):** The average of the absolute differences between predicted and actual values. It gives a measure of the average magnitude of errors.
- **Root Mean Squared Error (RMSE):** Similar to MAE but emphasizes larger errors. It's the square root of the average of squared differences between predicted and actual values.
- **Mean Absolute Percentage Error (MAPE):** Measures the percentage difference between predicted and actual values. It provides a relative error as a percentage.
- **R-squared (Accuracy):** Represents the proportion of the variance in the dependent variable that is predictable from the independent variable. R-squared ranges from 0 to 1, where 1 indicates perfect predictions.

II. ARIMA Metrics:

- **Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE):** Similar to linear regression, they measure the average magnitude of errors.
- **R-squared (R2):** Measures the proportion of the variance in the dependent variable that is predictable from the independent variable.
- **Adjusted R-squared (Adjusted R2):** Similar to R2 but adjusts for the number of predictors, providing a more accurate measure when there are multiple predictors.

III. Holt-Winters Metrics:

- **Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2):** Similar to linear regression, they measure the average magnitude of errors and the proportion of predictable variance.
- **Adjusted R-squared (Adjusted R2):** Adjusts R2 for the number of predictors, providing a more accurate measure in the presence of multiple predictors.

IV. XGBoost Metrics:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in predictions, providing a straightforward metric to understand forecast accuracy.
- **Root Mean Squared Error (RMSE)**: Highlights larger errors by squaring them, offering insight into the accuracy of predictions.
- **R-squared (R2)**: Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-squared (Adjusted R2)**: Adjusts the R2 value for the number of predictors in the model, providing a more accurate measure in models with multiple predictors.

V. Custom Deep Learning Neural Network Metrics:

- **Mean Absolute Error (MAE)**: Averages the absolute differences between predicted and actual values, offering a clear measure of prediction accuracy.
- **Root Mean Squared Error (RMSE)**: Penalizes larger errors more heavily, useful for understanding the impact of significant prediction errors.
- **Mean Absolute Percentage Error (MAPE)**: Provides the error as a percentage of the actual values, useful for understanding the error magnitude relative to the actual values.
- **Loss Function**: Often specific to the problem and model, such as mean squared error (MSE) for regression tasks, guiding the optimization during training.

VI. Deep AR Metrics:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of the forecasting errors, providing an intuitive measure of accuracy.
- **Root Mean Squared Error (RMSE)**: Emphasizes larger errors due to squaring the differences, useful for detecting significant prediction deviations.
- **Normalized Deviation (ND)**: Normalizes the absolute error by the sum of actual values, providing a scale-independent measure.
- **Quantile Loss**: Assesses the accuracy of probabilistic forecasts by measuring deviations at different quantiles, providing insights into the uncertainty of predictions.

These metrics help assess the accuracy and reliability of forecasting models, providing insights into how well the model's predictions align with actual data.

Understanding and Leveraging Data:

In the realm of store sales forecasting, the foundation lies in the effective understanding and utilization of data. Data, in its raw form, serves as the bedrock upon which accurate predictions and informed decision-making rest. In this section, we delve into the intricacies of what constitutes good data, how to choose an appropriate dataset, and the essential considerations when dealing with data for store sales forecasting. From dataset selection criteria to preprocessing operations and continuous monitoring, a comprehensive exploration awaits to guide businesses through the pivotal process of harnessing the power of data for optimal forecasting outcomes.

I. What is Data?

- Define data as raw facts and figures that can be processed to obtain meaningful information. In the context of store sales forecasting, data includes information about historical sales, customer behavior, promotions, economic indicators, and other relevant factors.

II. How to Choose a Dataset?

- Discuss the criteria for selecting a dataset. Consider factors such as the relevance of the data to your forecasting goals, the timeframe of the data, and the availability of key variables like sales figures, promotional data, and external factors.

III. Where to Find a Dataset?

- Explore potential sources for datasets. This could include internal data from the business, publicly available datasets from government agencies, industry-specific databases, or platforms like Kaggle that host a variety of datasets for different domains.

IV. What is a Good Dataset?

- Define a good dataset as one that is relevant, comprehensive, and representative of the factors influencing store sales. A good dataset should have a sufficient volume of data, minimal missing values, and include all relevant variables for accurate forecasting.

V. What the Dataset Should Look Like?

- Describe the ideal structure of a dataset for sales forecasting. It should typically have a timestamp for each data point, variables such as sales, promotional indicators, and any other relevant features. The dataset should be organized, with clear labels and consistent formatting.

VI. What the Dataset Should Have?

- Enumerate the essential variables a dataset for store sales forecasting should contain. This may include:
- Time-related variables (e.g., timestamps).
- Sales figures.
- Customer-related data.
- Information on promotions or marketing activities.
- External factors (e.g., economic indicators).

VII. How to Deal with the Dataset?

- Discuss the initial steps in handling a dataset for sales forecasting. This includes data cleaning to address missing values and outliers, exploring the distribution of variables, and checking for data consistency.

VIII. What Operations Should be Applied to the Dataset?

- Describe preprocessing operations such as normalization, scaling, or encoding categorical variables. Explain why these operations are necessary to ensure that the dataset is suitable for input into forecasting models.

IX. **Data Exploration Techniques:**

- Discuss exploratory data analysis (EDA) techniques to gain insights into the dataset. This involves visualizing trends, distributions, and correlations among variables to inform the modeling process.

X. **Handling Time Series Data:**

- If dealing with time series data, elaborate on techniques for handling temporal aspects. This includes identifying seasonality, trends, and any cyclical patterns in the sales data.

XI. **Quality Assurance and Validation:**

- Stress the importance of quality assurance in the dataset and how validation techniques, such as cross-validation, can be applied to ensure the reliability of the data for training and testing forecasting models.

XII. **Data Security and Privacy:**

- Touch on the importance of data security and privacy, especially if dealing with sensitive customer information. Emphasize compliance with relevant regulations and the ethical handling of data.

XIII. **Continuous Monitoring and Updates:**

- Highlight that the dataset is not static and should be continuously monitored and updated. New data points should be incorporated to improve the accuracy of forecasting models over time.

By addressing these aspects, we can handle and utilize data for store sales forecasting, from dataset selection to ongoing maintenance and improvement.

Used Tools:

The evolution of our sales forecasting models was empowered by a diverse array of tools, each playing a distinctive role in shaping the project's success:

I. **Python:**

- **Definition:** Python, a versatile programming language, served as the backbone for data analysis and model building.

- **Pros:**

  i. **Versatility:** Python's versatility allowed for comprehensive data analysis and efficient model development.
  ii. Rich Ecosystem: A rich ecosystem of libraries (e.g., NumPy, Pandas, Scikit-learn) enhanced functionality for statistical analysis and machine learning.

- **Cons:**

  i. **Learning Curve:** For those unfamiliar, there might be an initial learning curve associated with Python.

II. **Anaconda & Jupiter Notebook:**

- **Definition:** Anaconda, alongside Jupiter Notebook, provides an interactive environment for coding, data visualization, and model development.
- **Pros:**

  i. **Interactivity:** An interactive platform fosters real-time adjustments and exploration during the coding and modeling process.
  ii. **Ease of Sharing:** Jupiter Notebook facilitates easy sharing and collaboration

- **Cons:**

  i. **Resource Intensive:** Anaconda can be resource-intensive, requiring substantial disk space.

III. **Google:**

- **Definition:** Google tools were utilized for research and accessing cloud-based data and services.
- **Pros:**

  i. **Cloud Integration:** Seamless integration with cloud-based resources facilitated efficient data access and storage.
  ii. **Collaboration Tools:** Tools like Google Docs and Sheets provide collaborative platforms for team interaction.

- **Cons:**
- **Privacy Concerns:** Depending on the nature of data, there might be privacy concerns associated with cloud-based services.

IV. Excel:

- **Definition:** Excel played a pivotal role in preliminary data analysis and visualization.
- **Pros:**

  i. **User-Friendly:** Excel's user-friendly interface made it accessible for various team members.
  ii. **Quick Analysis:** Rapid data analysis and visualization capabilities for initial insights.

- **Cons:**

  i. **Limitations with Big Data:** Excel may have limitations when dealing with large datasets.

V. Power BI:

- **Definition:** Power BI contributed to creating dynamic dashboards and visualizations for interpreting forecasting results.
- **Pros:**

  i. **Visualization Capabilities:** Powerful visualization tools facilitated clear and interactive representation of forecasting outcomes.
  ii. **Integration with Other Tools:** Seamless integration with other Microsoft products enhanced overall efficiency.

- **Cons:**

  i. **Learning Curve:** Power BI's advanced features may have a learning curve for new users.

VI. Data Warehouse:

- **Definition:** A centralized data storage system for comprehensive analysis.
- **Pros:**

  i. **Centralization:** Centralized data storage streamlines data accessibility and management.
  ii. **Comprehensive Analysis:** Enables comprehensive analysis by consolidating data from various sources.

- **Cons:**

  i. **Implementation Complexity:** Setting up and maintaining a data warehouse can be complex.

VII. **SQL Server:**

- **Definition:** SQL Server was employed for managing and querying large datasets efficiently.
- **Pros:**

  i. **Query Performance:** Efficient handling of large datasets through optimized query performance.
  ii. **Scalability:** Scalability to accommodate growing data needs.

- **Cons:**

  i. **Licensing Costs:** Depending on the version, SQL Server may involve licensing costs.

VIII. **Kaggle:**

- **Definition:** Kaggle provided access to a diverse range of datasets and community-driven insights for model training and benchmarking.
- **Pros:**

  i. **Diverse Datasets:** Access to diverse datasets enriched the model training process.
  ii. **Community Insights:** Community discussions and competitions offered valuable insights and benchmarking opportunities.

- **Cons:**

  i. **Limited Control:** The datasets on Kaggle may not always align with specific project requirements.

IX. **YouTube:**

- **Definition:** YouTube served as a resource for tutorials and expert talks, enhancing understanding and implementation of complex forecasting concepts.
- **Pros:**

i. **Visual Learning:** Video tutorials facilitated visual learning for complex concepts.
ii. **Expert Insights:** Access to expert talks provided deeper insights into advanced forecasting techniques.

- **Cons:**

i. **Quality Variability:** The quality of tutorials may vary, impacting the consistency of learning.

X. **ChatGPT:**

- **Definition:** ChatGPT assisted in providing theoretical knowledge, answering queries, and offering guidance on data analysis techniques.
- **Pros:**

i. **Information Assistance:** ChatGPT served as a valuable resource for theoretical knowledge and guidance.
ii. **Query Resolution:** Provided real-time answers to specific queries during the project.

- **Cons:**

i. **Limited to Text-Based Interaction:** Interaction limitations to text-based queries may be a constraint for some users.

These tools, in concert, facilitated every stage of our sales forecasting process, from the initial stages of data gathering and processing to the intricate phases of analysis, visualization, and interpretation. Their collective strengths and considerations formed a robust foundation for a comprehensive and successful sales forecasting endeavor.

Summary:

Store sales forecasting is a crucial process in retail, predicting a store's future sales by analyzing historical data and market trends. This optimizes operations like inventory management. The chapter introduces the significance of forecasting, a reference study on key models (linear regression, ARIMA, triple exponential smoothing), and explores main concepts. Time Series Analysis and Machine Learning are highlighted, with models offering unique strengths and limitations. Then comes the part that talks about data, it’s important to understand and leverage data for Store Sales Forecasting. The tools used in the process, such as Python, Google, and Kaggle, are outlined. Together, they form a robust foundation for sales forecasting, aiding in data analysis, model development, and decision-making in the retail sector.

Chapter Three

Analytical Study & Entry-Level Models

Embarking on sales forecasting requires adept navigation through a meticulously crafted roadmap, where each step enhances precision in predictive models. Initial model curation strategically selects from diverse forecasting models to align with specific business goals. Core to this process is the meticulous dataset selection, considering pertinence, thoroughness, and alignment with business objectives. Visualizing datasets through graphics unveils critical patterns, guiding subsequent modeling decisions. The methodical model-building phase, from preprocessing to parameter tuning, ensures a model adept at capturing business intricacies. Culminating in presenting model outcomes through graphics, this iterative process underscores the dynamic relationship between each phase, achieving a refined and accurate sales forecasting model.

In this chapter, we will be looking at:

1. **Gantt Chart.**
2. **Model selection.**
3. **Dataset Selection**
4. **Dataset visualization (Old Dataset).**
5. **Dataset visualization (New Dataset).**
6. **Visualization by Power BI.**
7. **Models building.**
8. **Models results.**
9. **Summary.**

Gannt Chart:

This Gantt chart provides a visual timeline of the project's key milestones and activities planned out over the coming months. It outlines the schedule for seminars, phases of research, model development, and the stages of report preparation. Each task is carefully plotted to reflect its start date, duration, and end date, offering a clear overview of the project's trajectory. This chart serves as a dynamic tool to track progress and ensure that all components of the project are completed in a timely and systematic fashion.

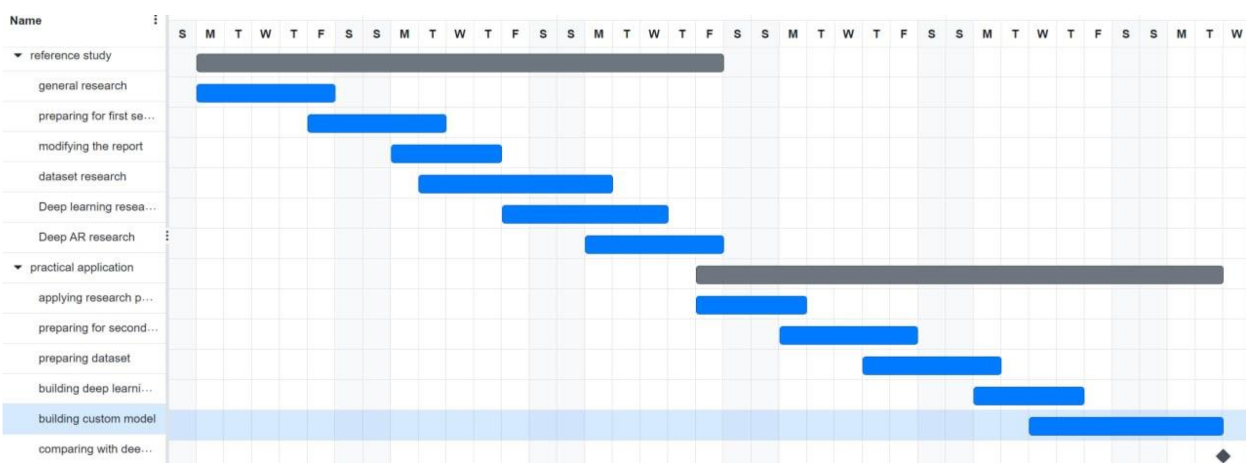


Figure 3. 1. Gantt Chart.

Model selection:

The steps for model selection in sales forecasting typically include:

1. **Define Business Objectives:** Understand what you want to achieve with the forecasting model (e.g., short-term vs. long-term forecasting).
2. **Analyze Data Availability:** Review the data you have, considering its volume, variety, and quality.
3. **Understand Data Characteristics:** Identify patterns in the data like seasonality, trends, and cyclicality.
4. **Choose Relevant Models:** Based on the data and objectives, select potential models (like time series, and machine learning algorithms).
5. **Test and Evaluate Models:** Implement models on a subset of data to assess their accuracy and efficiency.
6. **Consider Model Complexity vs. Performance:** Balance the complexity of the model with the level of accuracy needed.
7. **Iterate and Optimize:** Continuously refine the model based on testing results and business feedback.

The selection of Linear Regression, ARIMA, and Holt-Winters Triple Exponential Smoothing models in sales forecasting follows a structured approach. This involves defining business objectives, analyzing data availability, understanding data characteristics, and selecting models that align with these factors. Linear Regression is chosen for its simplicity in interpreting variable relationships, ARIMA for handling complex, non-seasonal data, and Holt-Winters for its proficiency in seasonal trends. These models are then rigorously tested and evaluated for accuracy, balancing complexity against performance, and continually optimized based on results and feedback. This methodology ensures the chosen models are well-suited for diverse forecasting needs.

Dataset selection:

Choosing the right dataset for sales forecasting based on previous steps is crucial for effective sales forecasting. These steps, including defining objectives, assessing data relevance, and checking data quality, ensure the chosen dataset aligns with specific forecasting goals. This methodical approach led us to Kaggle, a platform known for its diverse and quality datasets. On Kaggle, we identified a dataset that met our defined criteria, including historical sales data with relevant external factors, thereby providing a robust foundation for our forecasting models. This careful selection process is essential for accurate and reliable forecasting outcomes.

In the initial stages of the project, I meticulously examined the dataset to extract valuable insights. This examination was crucial to understanding the nature of the data and informing subsequent decisions. I then split the data into training and testing sets, a standard practice in model development to evaluate performance and generalizability. During the model development phase, I experimented with various modifications to the dataset. This process involved tweaking and adjusting data inputs and model parameters to test and enhance the model's predictive accuracy.

Dataset Visualization (Old Dataset):

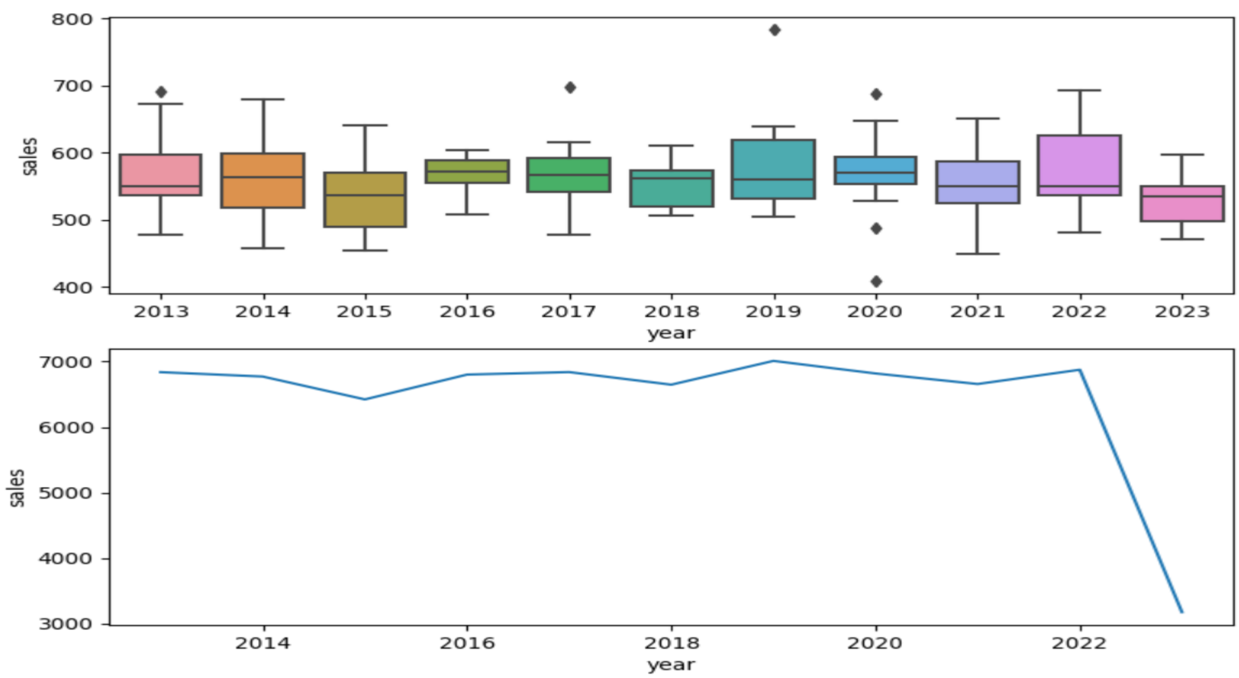


Figure 3. 2. Yearly Sales for Store Items (Old Dataset).

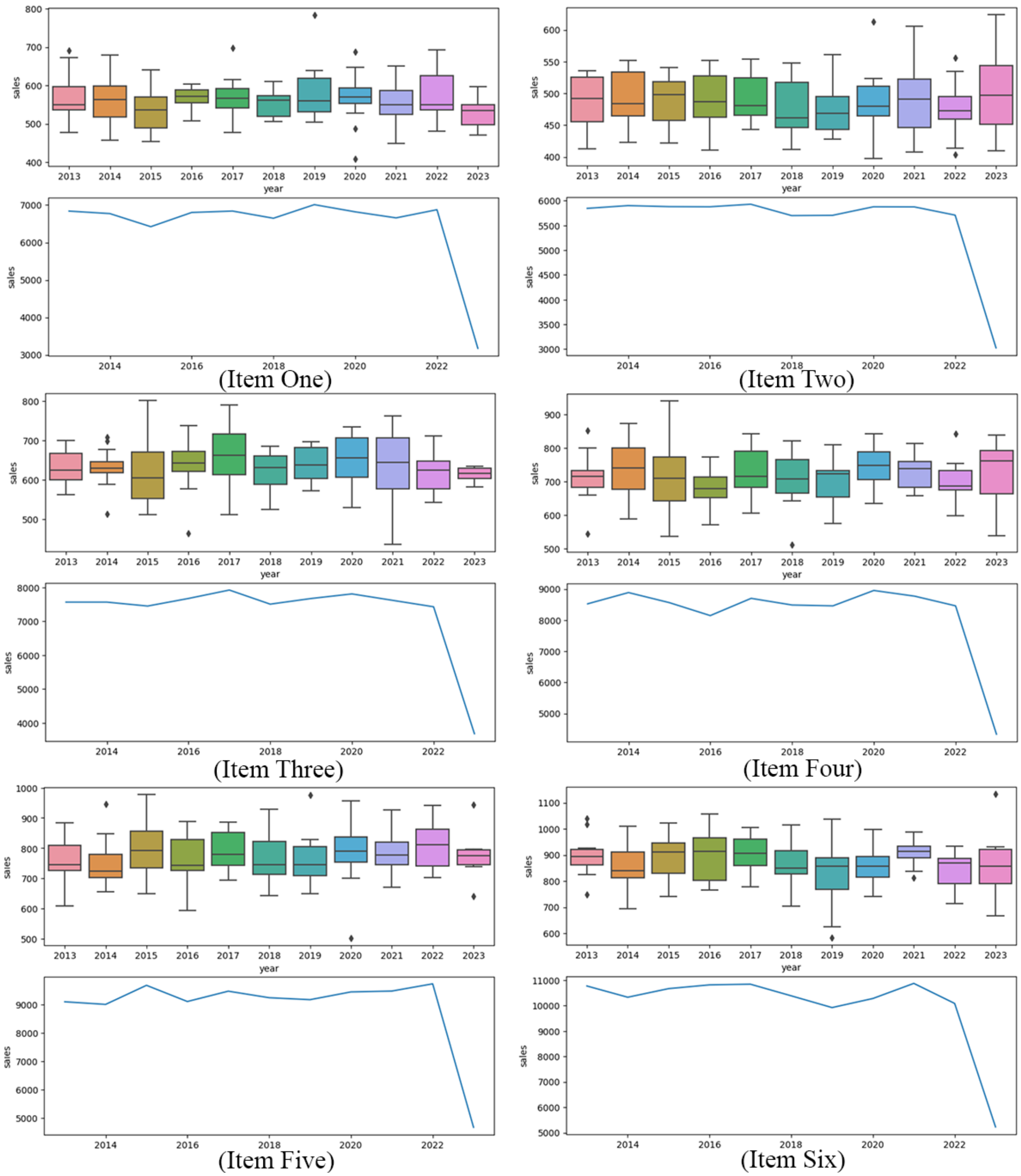


Figure 3. 3. yearly Sales for Each Item (Old Dataset).

Dataset Visualization (New Dataset):

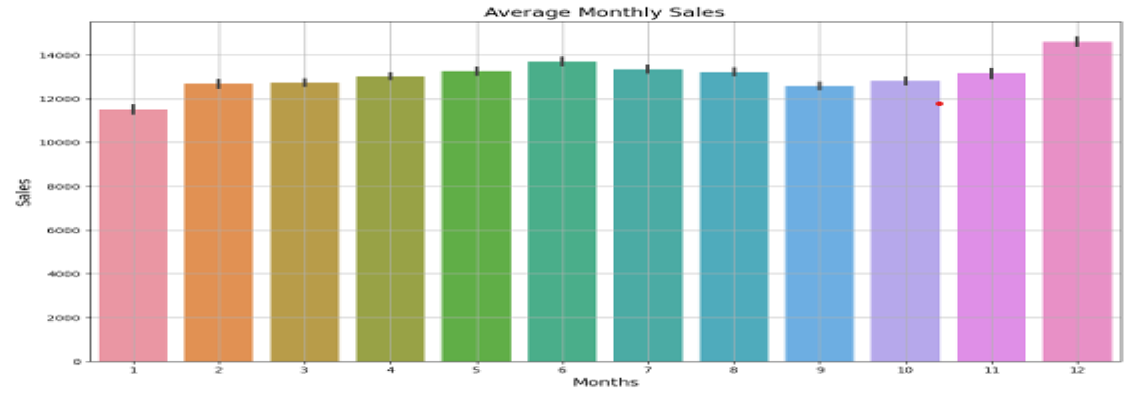


Figure 3. 4. Average monthly Sales (New Dataset).

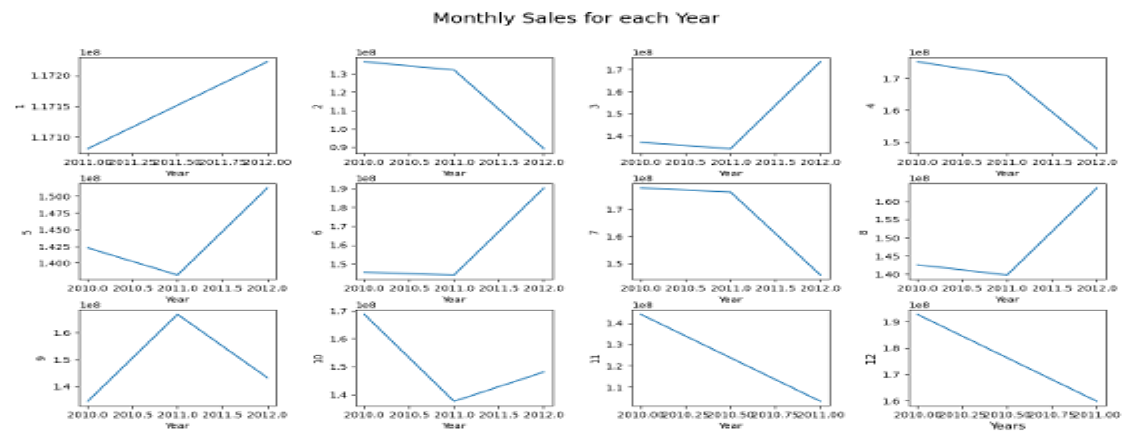


Figure 3. 5.Average Monthly Sales (New Dataset).

Figure 3. 6. Average sales Per Store (New Dataset)

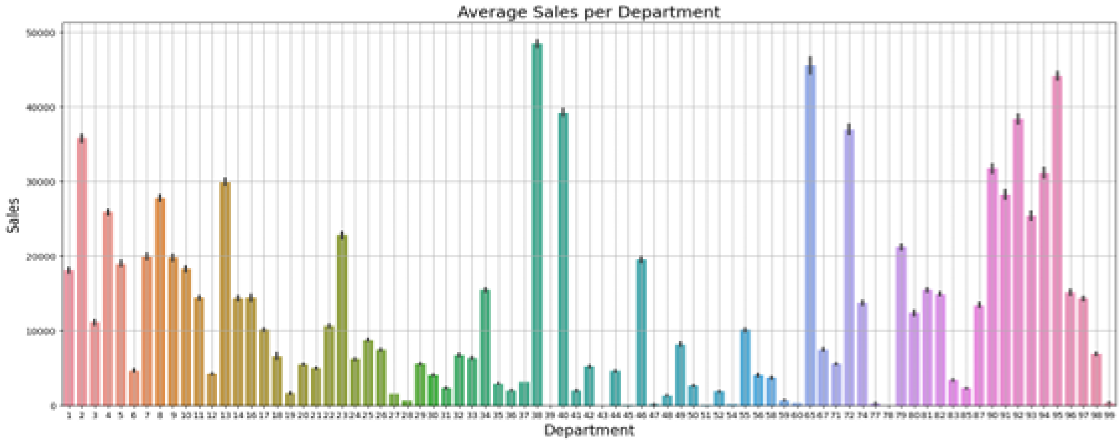


Figure 3. 7. Average Sales Per Department (New Dataset).

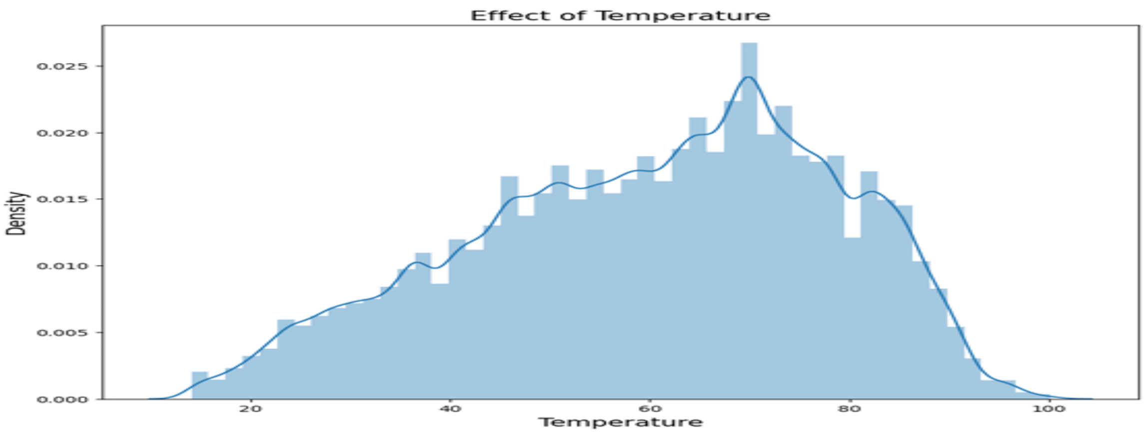


Figure 3. 8. Effect of Temperature (New Dataset).

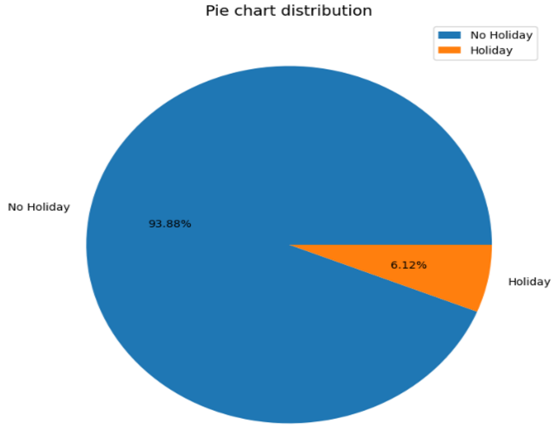


Figure 3. 9. Pie Chart of Temperature (New Dataset).

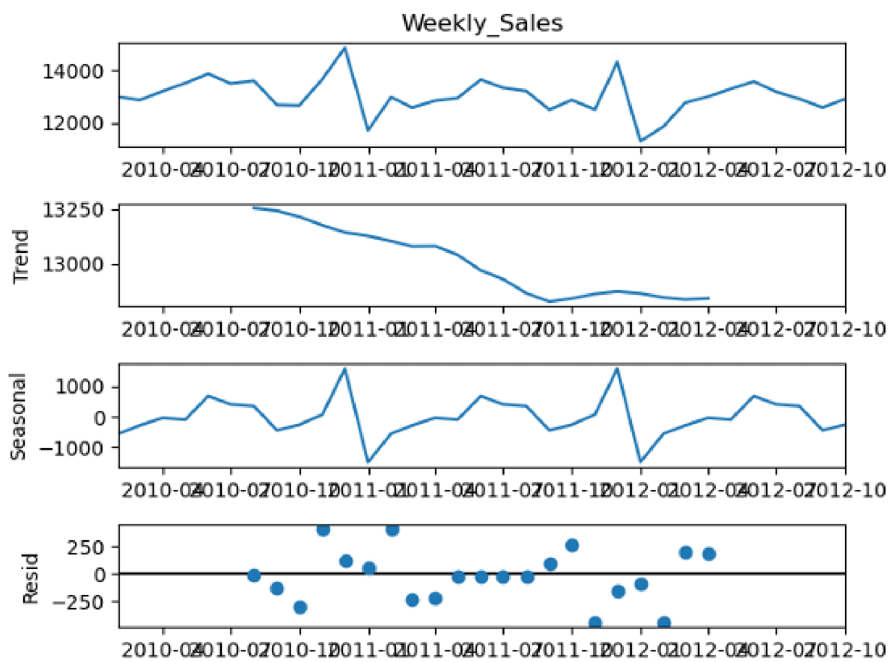


Figure 3. 10. Seasonal Components (New Dataset).

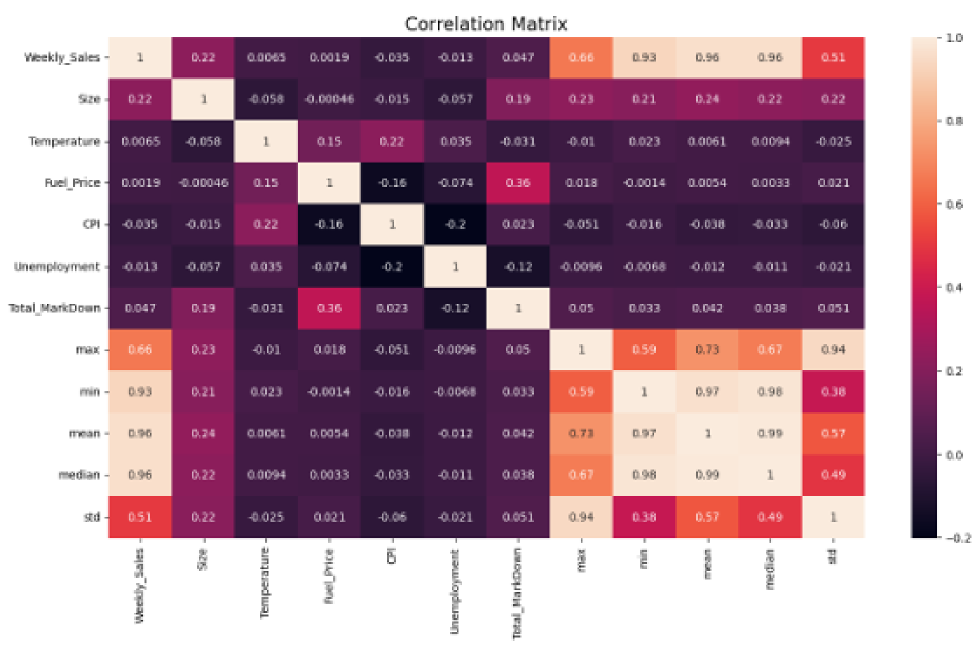


Figure 3. 11. Correlation Matrix (New Dataset).

Visualization by Power BI:

In the realm of data analysis and business intelligence, the ability to forecast sales is crucial for making informed decisions that drive company growth. The process of analyzing store sales encompasses several technical steps, from the initial data upload to the final visualization. This report documents a comprehensive walkthrough of uploading a sales dataset into Power BI through SQL Server, structuring the data within a defined schema, and leveraging Power BI's robust visualization tools to forecast store sales. The following visual aids illustrate each pivotal stage in the data's journey from raw numbers to actionable insights.

With the foundational dataset in place, the subsequent step involves constructing a data warehouse, a centralized repository designed to facilitate reporting and analysis. The importance of a data warehouse lies in its ability to consolidate data from multiple sources, providing a unified platform for robust data operations. This centralization is crucial for comprehensive analytics, as it enables the performance of complex queries and advanced operations on large datasets without impacting the performance of transactional systems.

In this data warehouse, data operations such as cleaning, transformation, and enrichment take precedence. These operations ensure that the data is accurate, consistent, and in a format that is conducive to insightful analysis. Once the data is prepped and stored in the data warehouse, we move to the visualization step using Power BI.

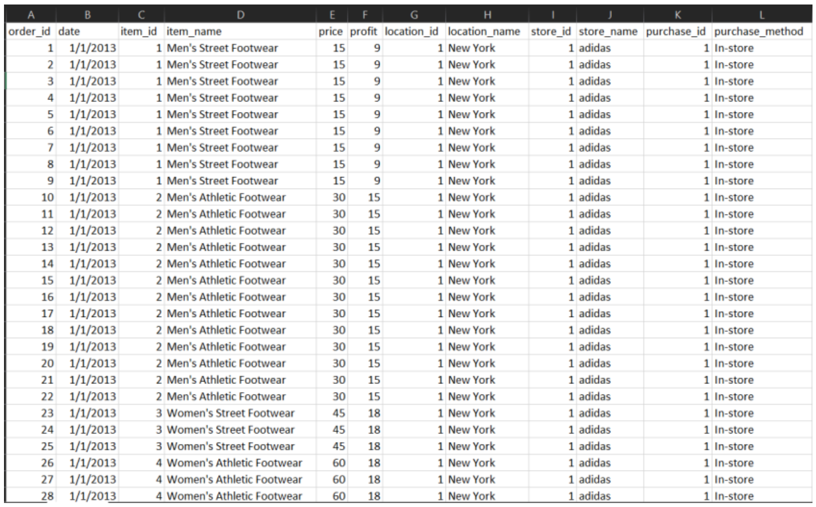
The process of data visualization in Power BI typically follows these steps:

Connect Power BI to the Data Warehouse: Establish a direct connection to the data warehouse, allowing Power BI to access the prepared datasets.

I. **Model the Data:** Use Power BI's data modeling features to define relationships, calculate columns, measures, and create hierarchies.
II. **Design the Visuals:** Craft visual representations such as charts, graphs, and tables to best illustrate the data's story.
III. **Refine the Dashboard:** Apply filters, slicers, and drill-down capabilities to enhance the interactivity of the dashboard.
IV. **Share Insights:** Share the dashboard with stakeholders, providing them with actionable insights and the ability to make data-driven decisions.

By leveraging the power of a data warehouse and the advanced visualization tools of Power BI, organizations can not only predict future trends but also make informed strategic decisions that align with their business objectives.

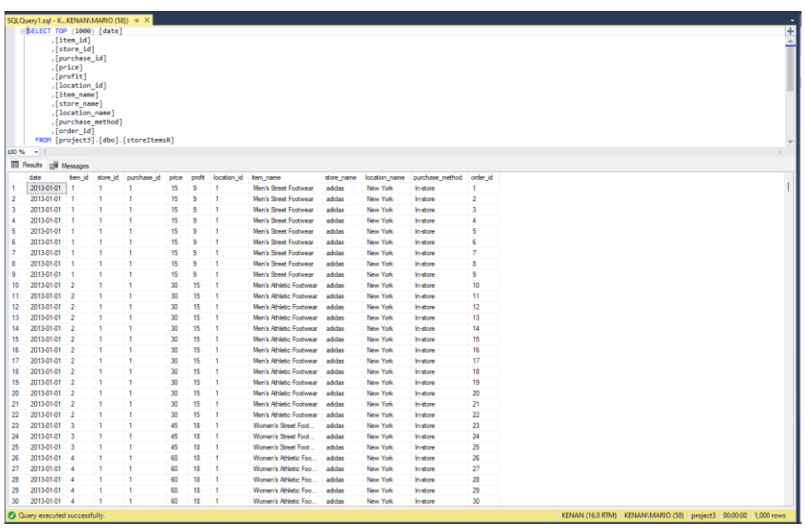
I. **Data Upload Overview:** This image displays a spreadsheet with detailed sales data, including order IDs, dates, item specifics, and profit margins, ready for upload to SQL Server for further analysis.



A	B	C	D	E	F	G	H	I	J	K	L
order_id	date	item_id	item_name	price	profit	location_id	location_name	store_id	store_name	purchase_id	purchase_method
1	1/1/2013	1	Men's Street Footwear	15	9	1	New York	1	adidas	1	In-store
2	1/1/2013	1	Men's Street Footwear	15	9	1	New York	1	adidas	1	In-store
3	1/1/2013	1	Men's Street Footwear	15	9	1	New York	1	adidas	1	In-store
4	1/1/2013	1	Men's Street Footwear	15	9	1	New York	1	adidas	1	In-store
5	1/1/2013	1	Men's Street Footwear	15	9	1	New York	1	adidas	1	In-store
6	1/1/2013	1	Men's Street Footwear	15	9	1	New York	1	adidas	1	In-store
7	1/1/2013	1	Men's Street Footwear	15	9	1	New York	1	adidas	1	In-store
8	1/1/2013	1	Men's Street Footwear	15	9	1	New York	1	adidas	1	In-store
9	1/1/2013	1	Men's Street Footwear	15	9	1	New York	1	adidas	1	In-store
10	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
11	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
12	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
13	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
14	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
15	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
16	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
17	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
18	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
19	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
20	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
21	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
22	1/1/2013	2	Men's Athletic Footwear	30	15	1	New York	1	adidas	1	In-store
23	1/1/2013	3	Women's Street Footwear	45	18	1	New York	1	adidas	1	In-store
24	1/1/2013	3	Women's Street Footwear	45	18	1	New York	1	adidas	1	In-store
25	1/1/2013	3	Women's Street Footwear	45	18	1	New York	1	adidas	1	In-store
26	1/1/2013	4	Women's Athletic Footwear	60	18	1	New York	1	adidas	1	In-store
27	1/1/2013	4	Women's Athletic Footwear	60	18	1	New York	1	adidas	1	In-store
28	1/1/2013	4	Women's Athletic Footwear	60	18	1	New York	1	adidas	1	In-store

Figure 3. 12. Data in CSV File.

II. **SQL Server Query Execution:** Here we see a SQL Server Management Studio (SSMS) window with a query result set, indicating the successful execution of a SELECT statement that retrieves the top records from the dataset.



SQLQuery1.sql - K...KENAN\MARIO (58))

```
SELECT TOP (1000) [date]
      ,[item_id]
      ,[store_id]
      ,[purchase_id]
      ,[price]
      ,[profit]
      ,[location_id]
      ,[item_name]
      ,[store_name]
      ,[location_name]
      ,[purchase_method]
      ,[order_id]
  FROM [project3].[dbo].[storeItemsR]
```

100 %

Results | Messages

	date	item_id	store_id	purchase_id	price	profit	location_id	item_name	store_name	location_name	purchase_method	order_id
1	2013-01-01	1	1	1	15	9	1	Men's Street Footwear	adidas	New York	In-store	1
2	2013-01-01	1	1	1	15	9	1	Men's Street Footwear	adidas	New York	In-store	2
3	2013-01-01	1	1	1	15	9	1	Men's Street Footwear	adidas	New York	In-store	3
4	2013-01-01	1	1	1	15	9	1	Men's Street Footwear	adidas	New York	In-store	4
5	2013-01-01	1	1	1	15	9	1	Men's Street Footwear	adidas	New York	In-store	5
6	2013-01-01	1	1	1	15	9	1	Men's Street Footwear	adidas	New York	In-store	6
7	2013-01-01	1	1	1	15	9	1	Men's Street Footwear	adidas	New York	In-store	7
8	2013-01-01	1	1	1	15	9	1	Men's Street Footwear	adidas	New York	In-store	8
9	2013-01-01	1	1	1	15	9	1	Men's Street Footwear	adidas	New York	In-store	9
10	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	10
11	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	11
12	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	12
13	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	13
14	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	14
15	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	15
16	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	16
17	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	17
18	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	18
19	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	19
20	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	20
21	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	21
22	2013-01-01	2	1	1	30	15	1	Men's Athletic Footwear	adidas	New York	In-store	22
23	2013-01-01	3	1	1	45	18	1	Women's Street Foot...	adidas	New York	In-store	23
24	2013-01-01	3	1	1	45	18	1	Women's Street Foot...	adidas	New York	In-store	24
25	2013-01-01	3	1	1	45	18	1	Women's Street Foot...	adidas	New York	In-store	25
26	2013-01-01	4	1	1	60	18	1	Women's Athletic Foo...	adidas	New York	In-store	26
27	2013-01-01	4	1	1	60	18	1	Women's Athletic Foo...	adidas	New York	In-store	27
28	2013-01-01	4	1	1	60	18	1	Women's Athletic Foo...	adidas	New York	In-store	28
29	2013-01-01	4	1	1	60	18	1	Women's Athletic Foo...	adidas	New York	In-store	29
30	2013-01-01	4	1	1	60	18	1	Women's Athletic Foo...	adidas	New York	In-store	30

Query executed successfully. | KENAN (16.0 RTM) | KENAN\MARIO (58) | project3 | 00:00:00 | 1,000 rows

Figure 3. 13. Running Query to Import Data.

III. **Data Import Wizard:** The third picture shows a step in the SQL Server Import and Export Wizard, where the user specifies the input file, pointing to a CSV file on the desktop, which will be used to create a new table in the database.

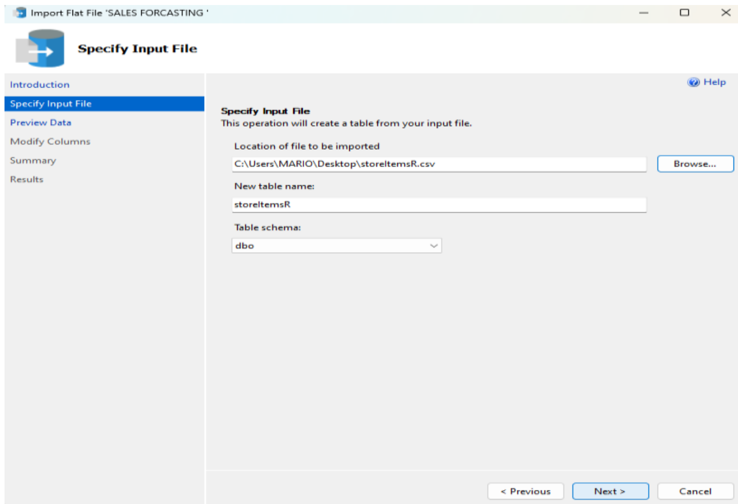


Figure 3. 14. Location of The Dataset for Import.

IV. Data Preview in Import Wizard: This image provides a preview of the data as seen in the SQL Server Import and Export Wizard, highlighting how the data will appear once imported into SQL Server.

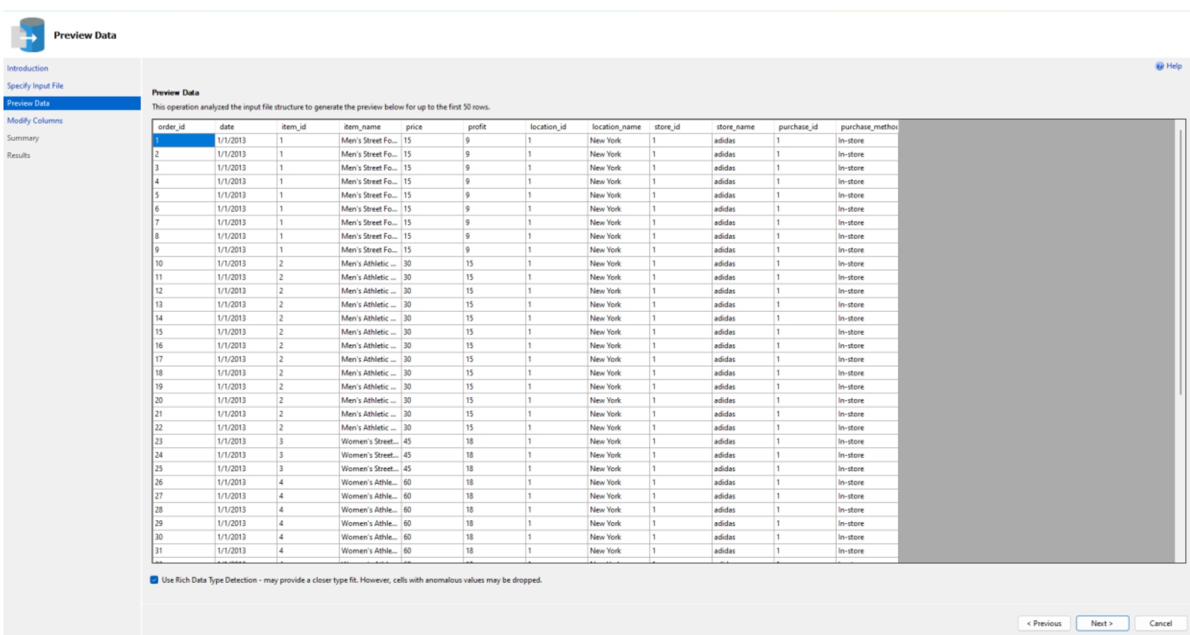
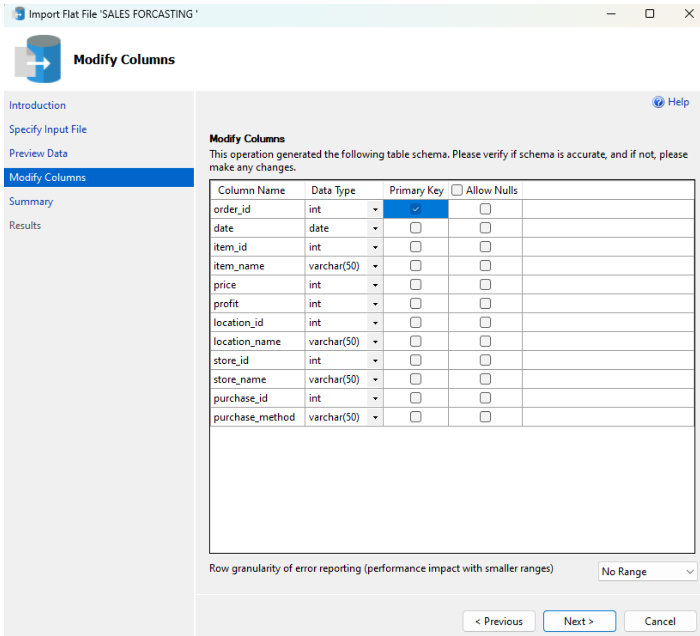


Figure 3. 15. Data in SQL Server.

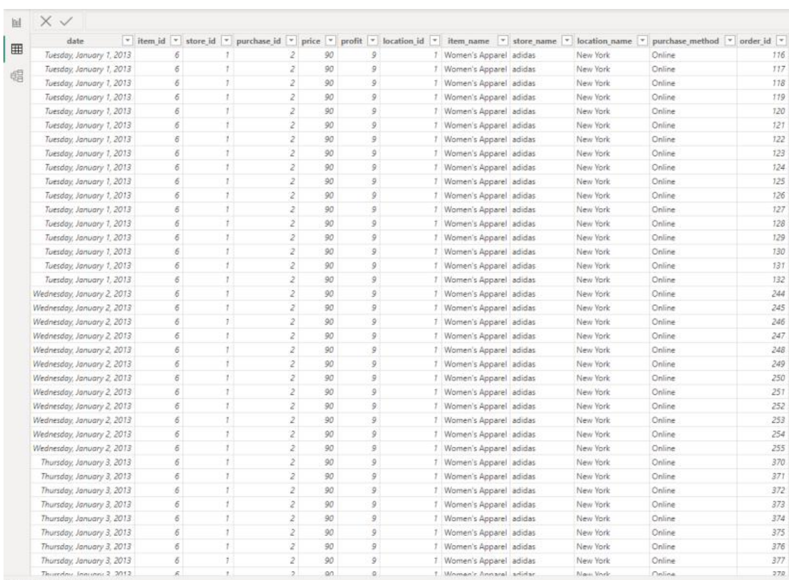
V. Table Schema Confirmation: In this screenshot, we're looking at the Modify Columns step of the SQL Server Import and Export Wizard, where the user can confirm or adjust the data types and properties of the columns in the newly created table.



Column Name	Data Type	Primary Key	Allow Nulls
order_id	int	☑	☐
date	date	☐	☐
item_id	int	☐	☐
item_name	varchar(50)	☐	☐
price	int	☐	☐
profit	int	☐	☐
location_id	int	☐	☐
location_name	varchar(50)	☐	☐
store_id	int	☐	☐
store_name	varchar(50)	☐	☐
purchase_id	int	☐	☐
purchase_method	varchar(50)	☐	☐

Figure 3. 16. Schema Definition and Modification.

VI. Data Integration in SQL Server: This image presents a table with a comprehensive view of the uploaded dataset in SQL Server, showcasing the data ready for analysis.



date	item_id	store_id	purchase_id	price	profit	location_id	item_name	store_name	location_name	purchase_method	order_id
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	116
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	117
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	118
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	119
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	120
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	121
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	122
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	123
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	124
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	125
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	126
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	127
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	128
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	129
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	130
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	131
Tuesday, January 1, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	132
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	244
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	245
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	246
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	247
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	248
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	249
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	250
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	251
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	252
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	253
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	254
Wednesday, January 2, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	255
Thursday, January 3, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	370
Thursday, January 3, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	371
Thursday, January 3, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	372
Thursday, January 3, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	373
Thursday, January 3, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	374
Thursday, January 3, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	375
Thursday, January 3, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	376
Thursday, January 3, 2013	6	1	2	90	9	1	Women's Apparel	adidas	New York	Online	377

Table: storeItemsR (510,608 rows)

Figure 3. 17. Data Successfully Imported.

VII. Database Relationships Diagram: Here we have a visual representation of the database schema in SQL Server, showing the relationships between different tables such as 'store', 'purchases', 'items', and others.

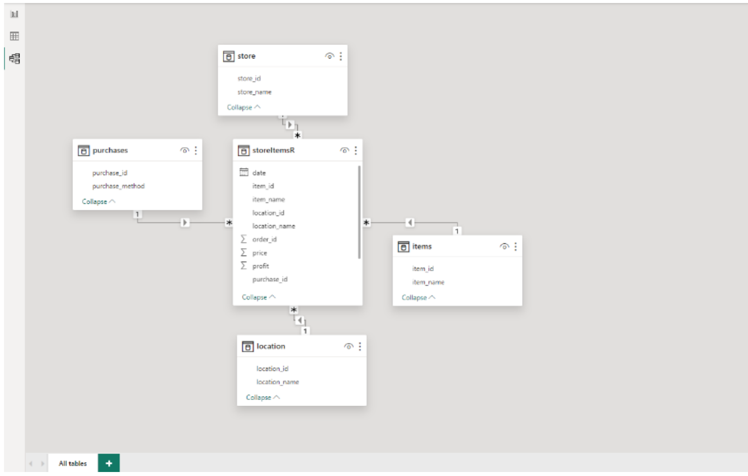


Figure 3. 18. Schema Diagram (Star Flake).

VIII. **Power BI Visualization Dashboard:** The final picture illustrates a dashboard in Power BI, where the data has been transformed into a series of visualizations, including pie charts and line graphs, for sales forecasting purposes **(Old dataset).**

Figure 3. 19. Data Visualization by Power Pi.

IX. **Power BI Visualization Dashboard:** The final picture illustrates a dashboard in Power BI, where the data has been transformed into a series of visualizations, including pie charts and line graphs, for sales forecasting purposes **(New dataset).**

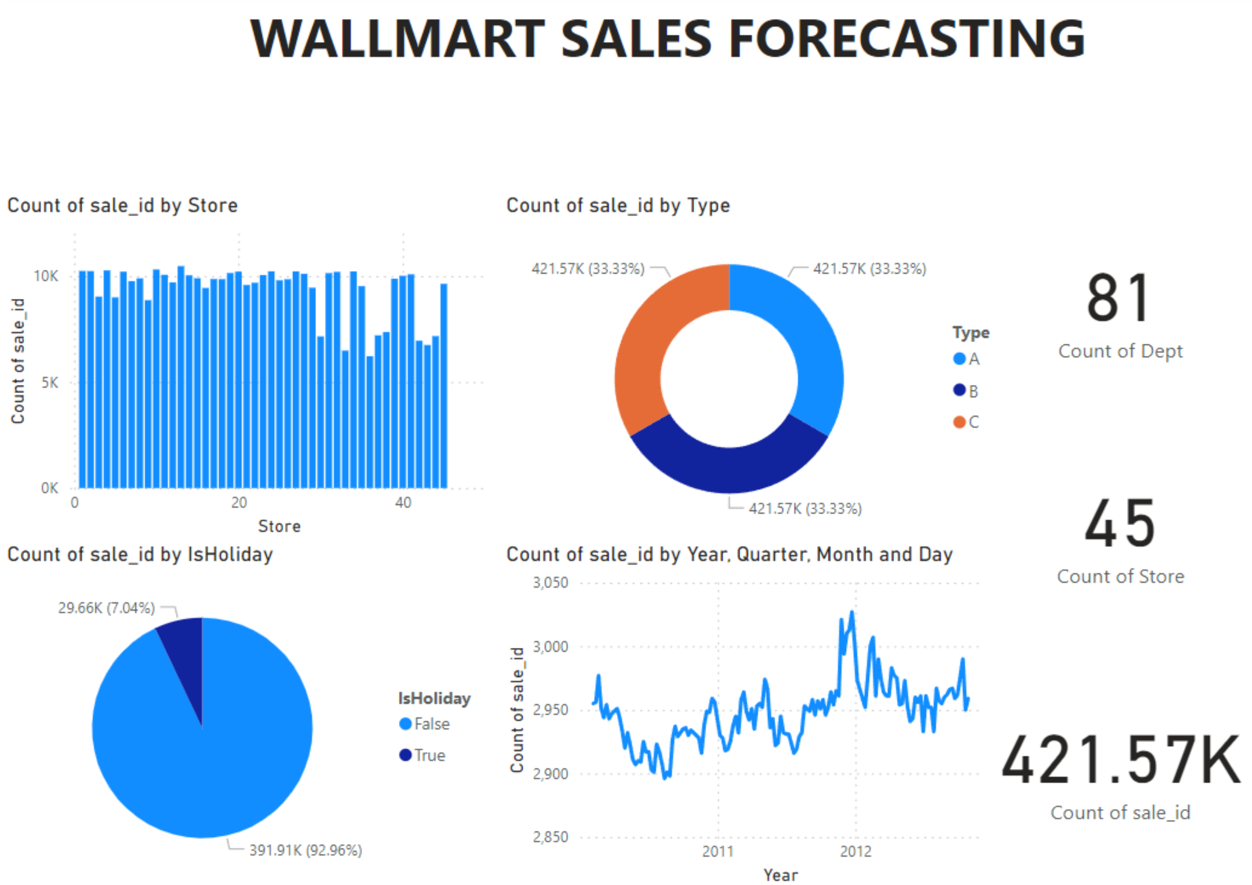


Figure 3. 20. Power BI Dashboard (New Dataset).

Model building:

Building a sales forecasting model typically involves these steps:

- **Data Collection:** Gather relevant historical sales data and potentially influential factors (like market trends, and customer demographics).
- **Data Preprocessing:** Clean the data to handle missing values, outliers, and format inconsistencies.
- **Feature Selection:** Choose relevant variables (features) that influence sales outcomes.
- **Model Selection:** Decide on a forecasting model (like ARIMA, or machine learning algorithms) based on the data's nature and business objectives.
- **Model Training:** Use historical data to train the model, allowing it to learn patterns.
- **Model Validation and Testing:** Test the model on a separate data set to evaluate its accuracy and reliability.

- **Iterative Improvement:** Refine the model based on test results and feedback, adjusting parameters as necessary.
- **Deployment:** Implement the model in a real-world business environment for actual forecasting.
- **Monitoring and Maintenance:** Continuously monitor the model's performance and update it as necessary to adapt to new data and changing market conditions.

To develop the Linear Regression model, ARIMA model, and Holt Winter’s triple exponential smoothing model for sales forecasting, I followed the outlined model-building steps. First, I collected and preprocessed relevant historical sales data, addressing missing values and outliers. Next, I identified key features impacting sales, such as rolling means and lag variables, through feature selection techniques. The model was then selected and trained using this processed data. Subsequent steps involved testing and validating the model's predictions against a separate dataset, followed by iterative improvements based on error analysis. Finally, the model was deployed and continuously monitored for performance, ensuring its adaptability and accuracy over time.

I. The steps in the provided **Linear Regression** model for sales forecasting are:

- **Data Collection:** Gathering relevant sales data along with influential factors.
- **Data Preprocessing:** Cleaning the data, and handling issues like missing values and outliers.
- **Feature Selection:** Identifying significant variables (e.g., rolling means, lag features) that influence sales.
- **Model Selection and Training:** Choosing the Linear Regression model and training it with the selected features.
- **Model Validation and Testing:** Testing the model’s predictions against a separate dataset to evaluate accuracy.
- **Iterative Improvement:** Refining the model based on error analysis and feedback.
- **Deployment:** Implementing the model for actual sales forecasting.
- **Monitoring and Maintenance:** Continuously tracking the model’s performance and making necessary updates.

Each step is designed to ensure the model is accurate, reliable, and adaptable to changing sales trends.

II. This model involves several steps for sales forecasting using **ARIMA** and **SARIMAX** models:

- **Data Preparation:** The script begins by importing necessary libraries and loading the dataset. It filters data for a specific store and formats the 'date' column.
- **Creating Date Features:** Additional features like 'year', 'month', 'day', and 'weekday' are created for more detailed analysis.

- **Data Splitting:** The dataset is divided into training and testing sets based on dates.
- **Stationarity Testing:** The Augmented Dickey-Fuller test checks if the sales data is stationary, crucial for time series modeling.
- **ARIMA Modeling:** An ARIMA model is fitted to the training data. This model is selected based on its ability to handle non-seasonal data.
- **SARIMAX Modeling:** A SARIMAX model, which accounts for seasonality, is also fitted.
- **Forecasting:** The SARIMAX model is used to forecast sales for the test period.
- **Error Calculation:** Errors between the forecasted and actual sales are calculated using metrics like MAE, RMSE, and MAPE.
- **Visualization:** The sales data, along with forecasts and errors, are plotted for visual analysis.
- **Result Aggregation:** Finally, the results are summarized, providing a comprehensive view of the model's performance.

These steps collectively create a robust forecasting model, considering both seasonal and non-seasonal patterns in sales data.

III. This model applies the **Holt-Winters Exponential Smoothing** method for sales forecasting:

- **Data Preparation:** Imports necessary libraries, loads the dataset, filters for a specific store, and formats the 'date' column.
- **Data Splitting:** Divides the dataset into training and testing sets.
- **Seasonal Decomposition:** Uses seasonal decomposition to understand underlying trends and seasonality in the sales data.
- **Holt-Winters Model Implementation:** Two versions of the Holt-Winters model are applied — one without damping and one with a damping trend component.
- **Forecasting:** Both models are used to forecast sales for the test period.
- **Error Calculation:** Calculates forecast errors using methods like MAE, RMSE, and MAPE.
- **Visualization:** Plots the sales, forecasts, and errors for visual comparison.
- **Result Aggregation:** Summarizes the results in tables, comparing the performance of the two models.

IV. This model applies the XGBoost model for sales forecasting:

- **Data Preparation**: Import necessary libraries, load the dataset, and handle missing values and outliers.
- **Feature Engineering**: Create additional features like rolling means, lag variables, and interaction terms that might influence sales.

- **Data Splitting**: Split the dataset into training and testing sets to ensure unbiased evaluation.
- **Model Training**: Train the XGBoost model using the training data, tuning parameters like learning rate, max depth, and number of estimators to optimize performance.
- **Model Validation**: Validate the model on the test set, using metrics such as MAE, RMSE, and R-squared to evaluate accuracy.
- **Feature Importance**: Analyze feature importance scores provided by XGBoost to understand the impact of each feature on sales predictions.
- **Iterative Improvement**: Refine the model based on validation results, adjusting parameters and potentially adding new features.
- **Deployment**: Deploy the trained model in a real-world setting for actual sales forecasting.
- **Monitoring and Maintenance**: Continuously monitor the model's performance and update it as needed to adapt to new data and changing market conditions.

V. This model applies a Custom Deep Learning Neural Network for sales forecasting:

- **Data Preparation**: Import necessary libraries and preprocess the data, including normalization and handling missing values.
- **Feature Engineering**: Create relevant features, including lagged variables, rolling means, and external factors influencing sales.
- **Model Design**: Design a neural network architecture tailored to the problem, selecting the number of layers, neurons per layer, and activation functions.
- **Data Splitting**: Split the data into training, validation, and testing sets to evaluate the model effectively.
- **Model Training**: Train the neural network using the training data, employing techniques like early stopping and learning rate schedules to optimize training.
- **Model Validation**: Validate the model on the validation set, adjusting hyperparameters based on performance metrics such as MAE, RMSE, and MAPE.
- **Iterative Improvement**: Continuously refine the model architecture and hyperparameters based on validation results.
- **Deployment**: Implement the trained neural network for real-time sales forecasting.
- **Monitoring and Maintenance**: Regularly monitor the model's performance and update it as necessary to ensure it adapts to new data and changing conditions.

VI. This model applies the Deep AR method for sales forecasting:

- **Data Preparation**: Import necessary libraries and preprocess the data, ensuring it is in the correct format for time series analysis.
- **Feature Engineering**: Create features that capture temporal dependencies, such as lagged variables, rolling means, and categorical encodings for time components.
- **Data Splitting**: Split the dataset into training, validation, and testing sets, ensuring the test set is representative of future periods.
- **Model Training**: Train the Deep AR model using the training data, leveraging recurrent neural networks to learn complex temporal patterns.
- **Model Validation**: Validate the model on the validation set, fine-tuning hyperparameters like the learning rate, number of layers, and number of epochs.
- **Forecasting**: Use the trained Deep AR model to forecast sales for the test period, generating probabilistic forecasts.
- **Error Calculation**: Calculate forecast errors using metrics such as MAE, RMSE, and CRPS (Continuous Ranked Probability Score).
- **Visualization**: Visualize the actual and forecasted sales along with prediction intervals to understand the model's performance.
- **Result Aggregation**: Summarize the forecasting results, comparing the model's performance with other approaches.
- **Deployment**: Deploy the Deep AR model for real-time sales forecasting, ensuring it can handle new data as it becomes available.
- **Monitoring and Maintenance**: Continuously monitor the model's performance and update it to adapt to new data and changes in sales patterns.

Each step contributes to a nuanced understanding of the model's performance and the impact of incorporating a damping trend in forecasting.

Model results:

I. **Linear regression model results:**

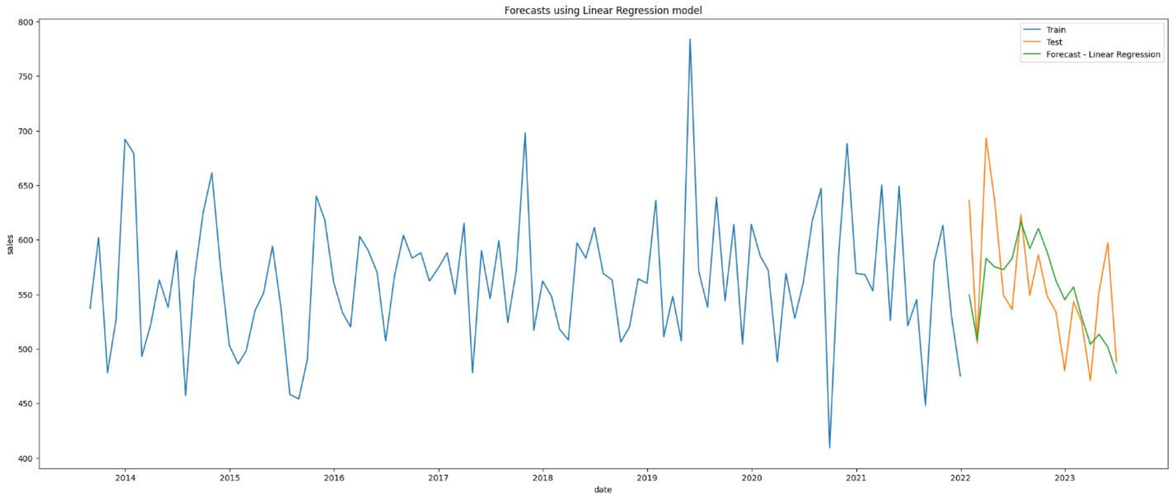


Figure 3. 21. LR-ELM Forecast for Item One.

	total_sales	total_pred_sales
model		
LinearRegression	10051	9966.587652

Figure 3. 22. LR-ELM Sales & Metrics for Item One.

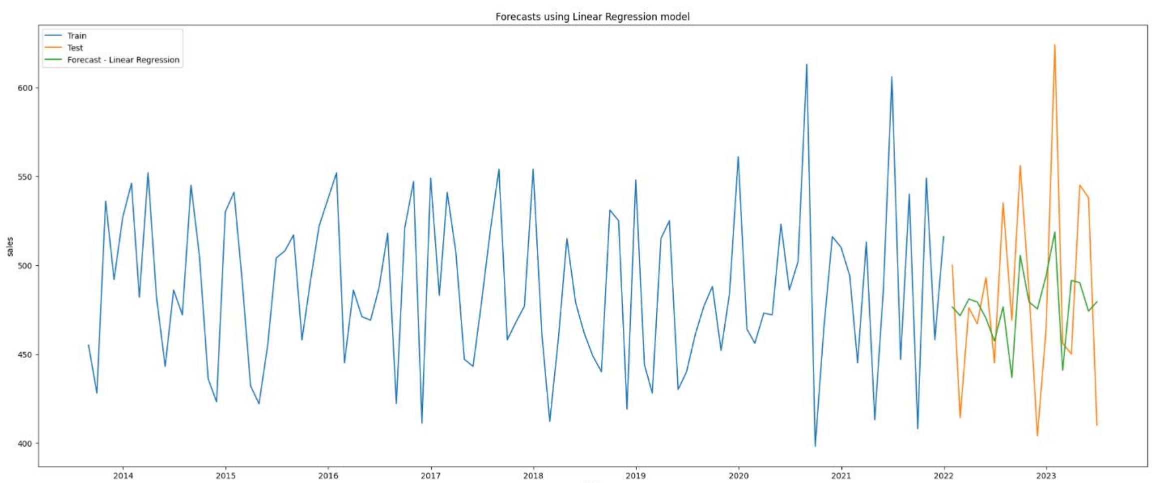


Figure 3. 23. LR-ELM Forecast for Item Two.

	total_sales	total_pred_sales
model		
LinearRegression	8732	8597.405159

Figure 3. 24. LR-ELM Sales & Metrics for Item Two.

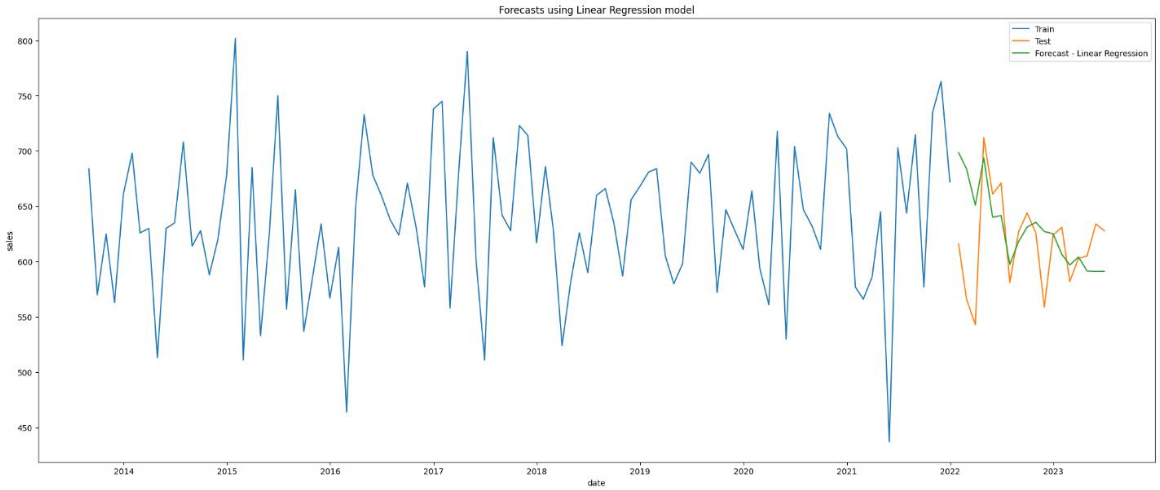


Figure 3. 25. LR-ELM Forecast for Item Three.

model	total_sales	total_pred_sales
LinearRegression	11113	11325.175196

Figure 3. 26. LR-ELM Sales & Metrics for Item Three.

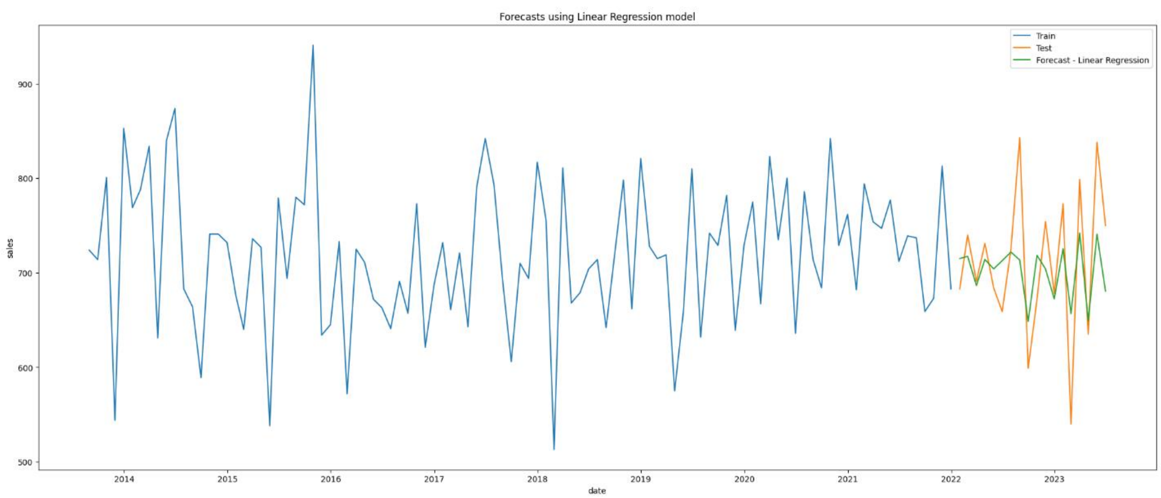


Figure 3. 27. LR-ELM Forecast for Item Four.

model	total_sales	total_pred_sales
LinearRegression	12794	12624.00848

Figure 3. 28. LR-ELM Sales & Metrics for Item Four.

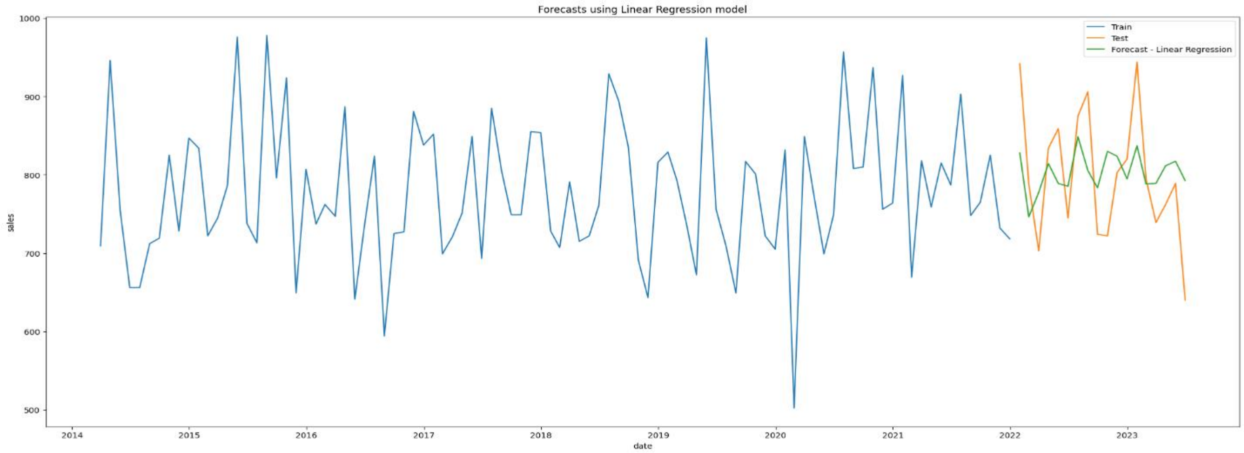


Figure 3. 29. LR-ELM Forecast for Item Five.

model	total_sales	total_pred_sales
LinearRegression	14390	14462.261777

Figure 3. 30. LR-ELM Sales & Metrics for Item Five.

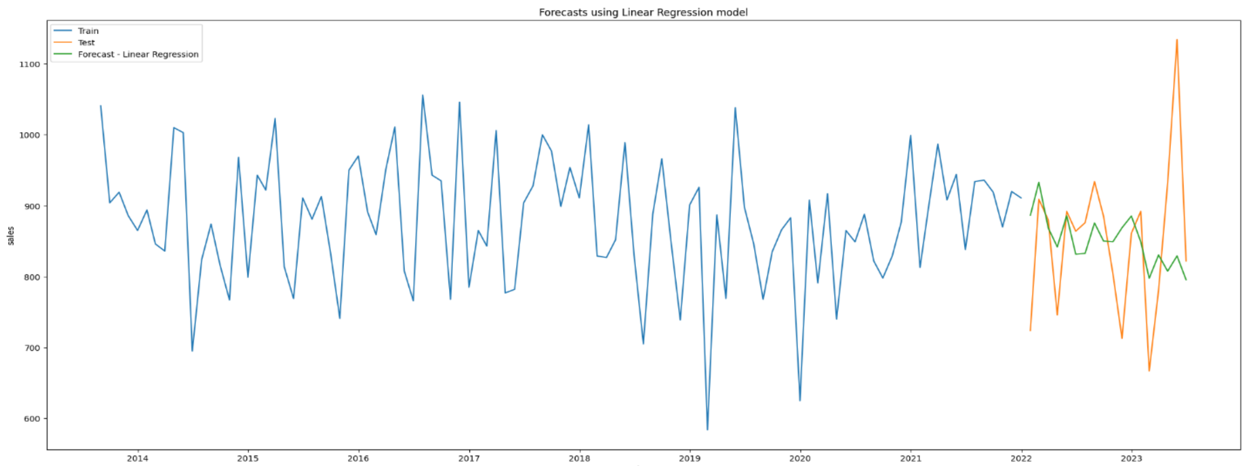


Figure 3. 31. LR-ELM Forecast for Item Six.

model	total_sales	total_pred_sales
LinearRegression	15316	15318.96496

Figure 3. 32. LR-ELM Sales & Metrics for Item Six.

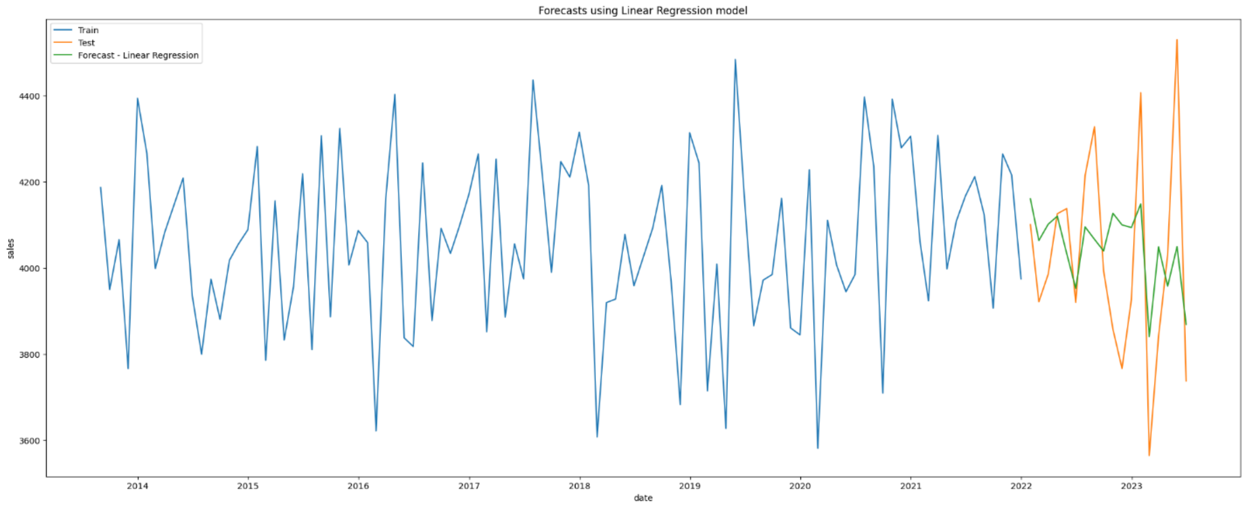


Figure 3. 33. LR-ELM Forecast for Store Items.

model	total_sales	total_pred_sales
LinearRegression	72396	72873.023032

Figure 3. 34. LR-ELM Sales & Metrics for Store Items.

II. ARIMA model results:

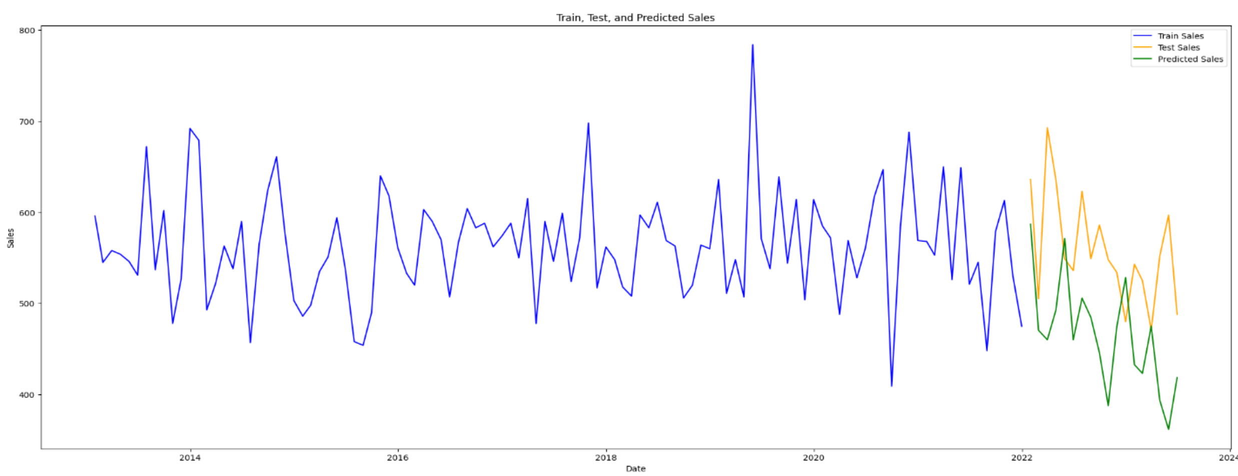


Figure 3. 35. A-ELM Forecast for Item One.

total_sales	total_pred_sales
10051	8372.187641

Figure 3. 36. A-ELM Sales & Metrics for Item One.

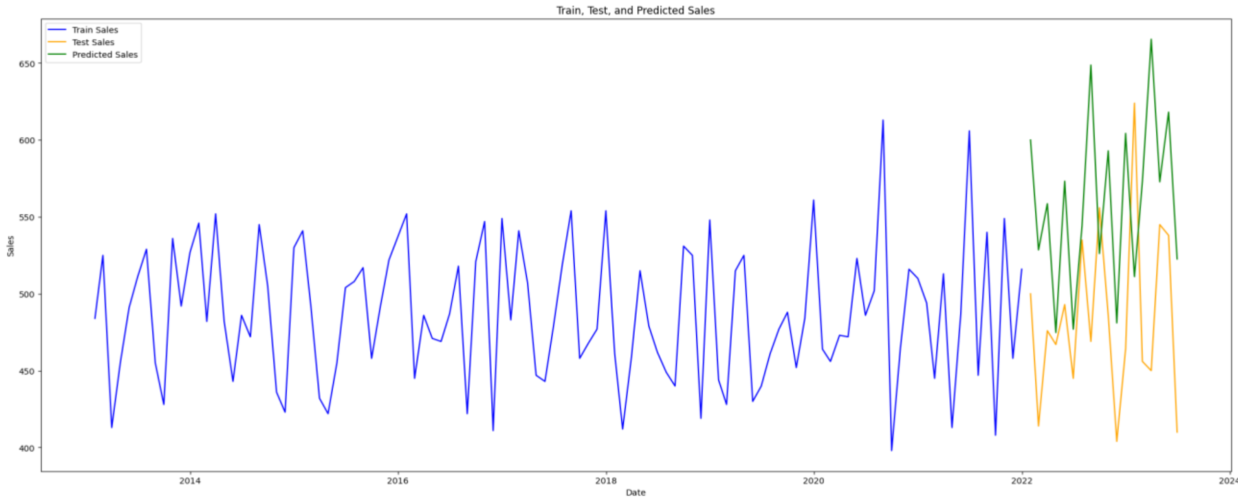


Figure 3. 37. A-ELM Forecast for Item Two.

total_sales	total_pred_sales
8732	10073.034183

Figure 3. 38. A-ELM Sales & Metrics for Item Two.

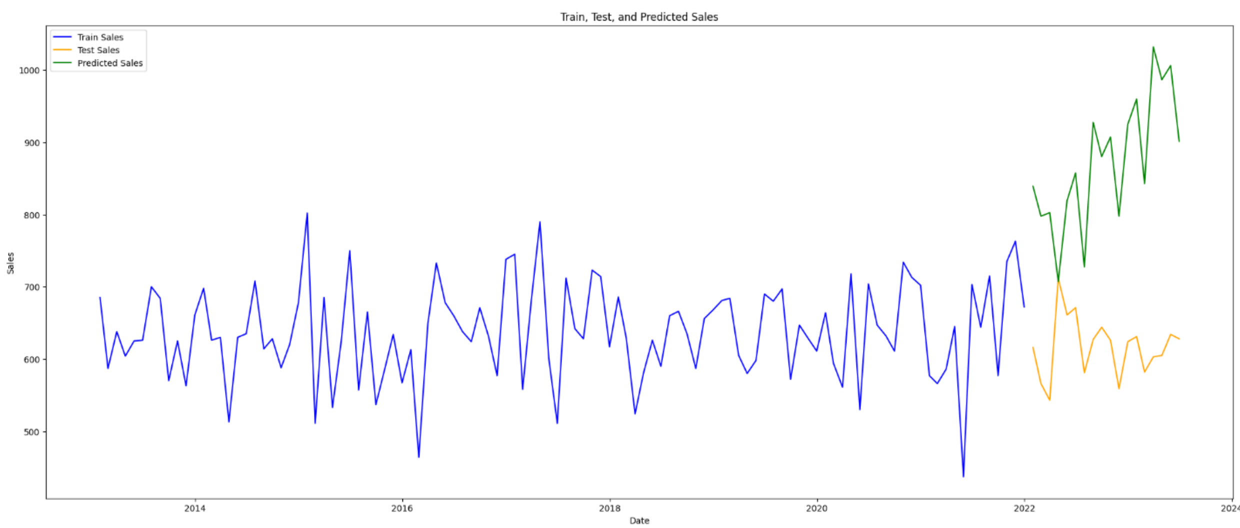


Figure 3. 39. A-ELM Forecast for Item Three.

total_sales	total_pred_sales
11113	15716.177876

Figure 3. 40. A-ELM Sales & Metrics for Item Three.

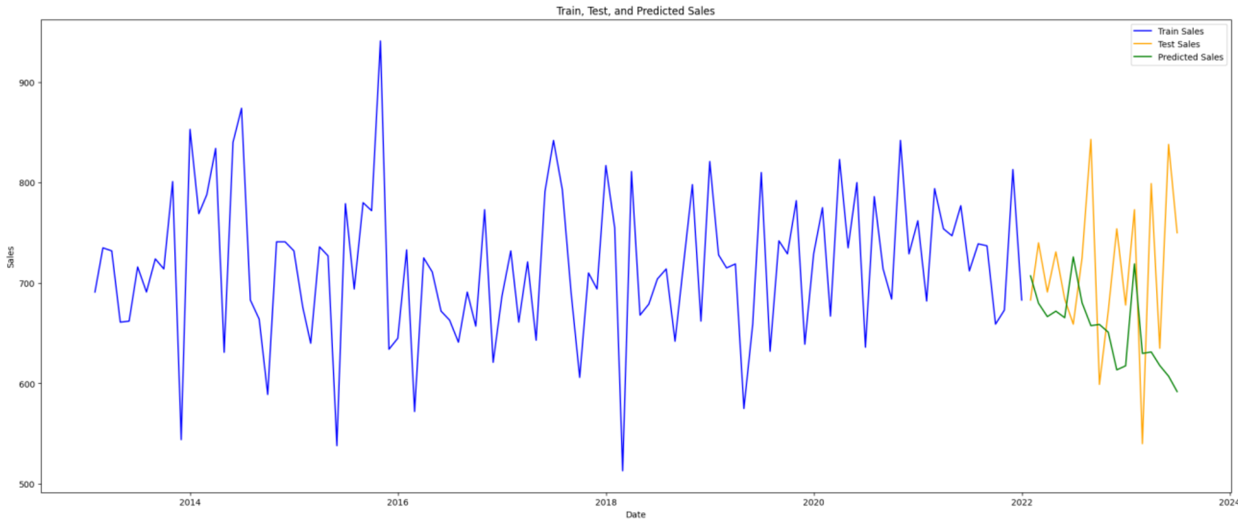


Figure 3. 41. A-ELM Forecast for Item Four.

total_sales	total_pred_sales
12794	11792.533437

Figure 3. 42. A-ELM Sales & Metrics for Item Four.

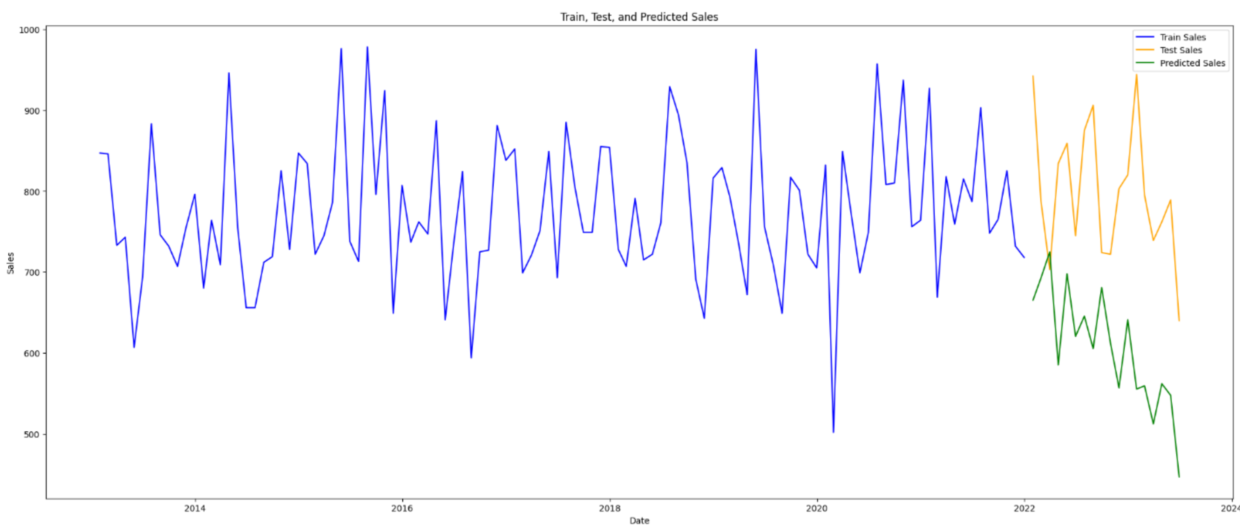


Figure 3. 43. A-ELM Forecast for Item Five.

total_sales	total_pred_sales
14390	10912.578243

Figure 3. 44. A-ELM Sales & Metrics for Item Five.

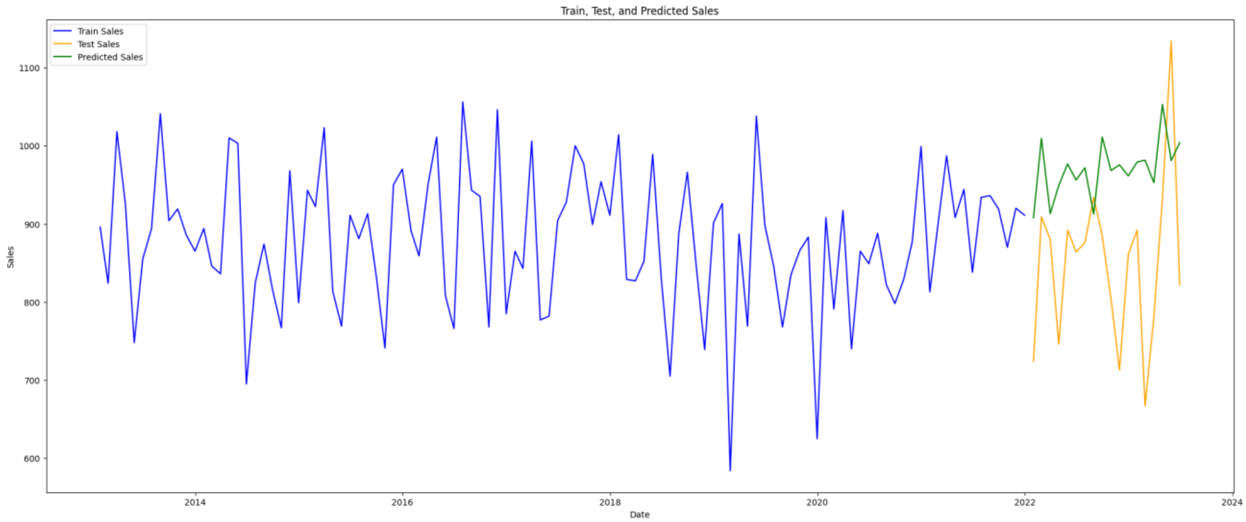


Figure 3. 45. A-ELM Forecast for Item Six.

total_sales	total_pred_sales
15316	17463.492111

Figure 3. 46. A-ELM Sales & Metrics for Item Six.

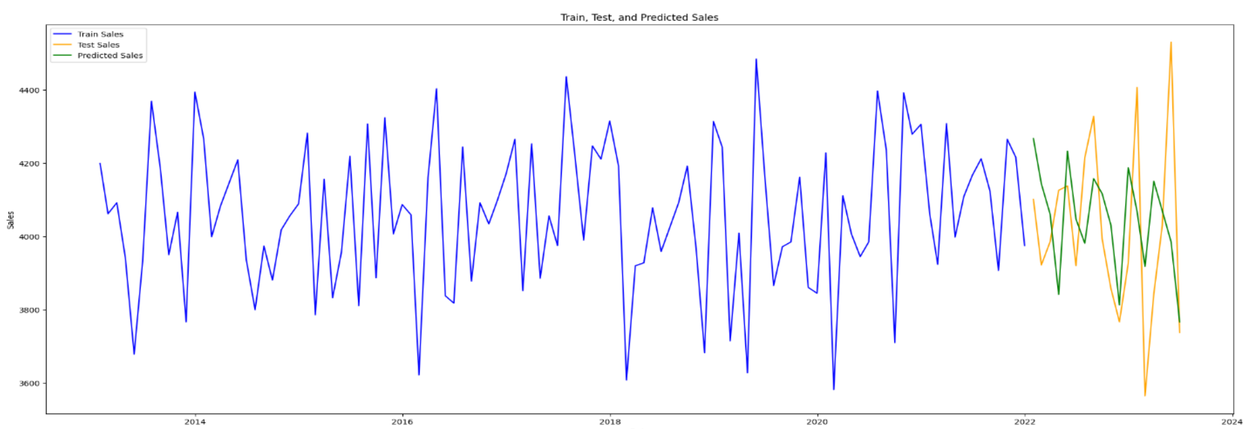


Figure 3. 47. A-ELM Forecast for Store Items.

total_sales	total_pred_sales
72396	72839.622008

Figure 3. 48. A-ELM Sales & Metrics for Store Items.

III. Holt-Winters Exponential Smoothing model results:

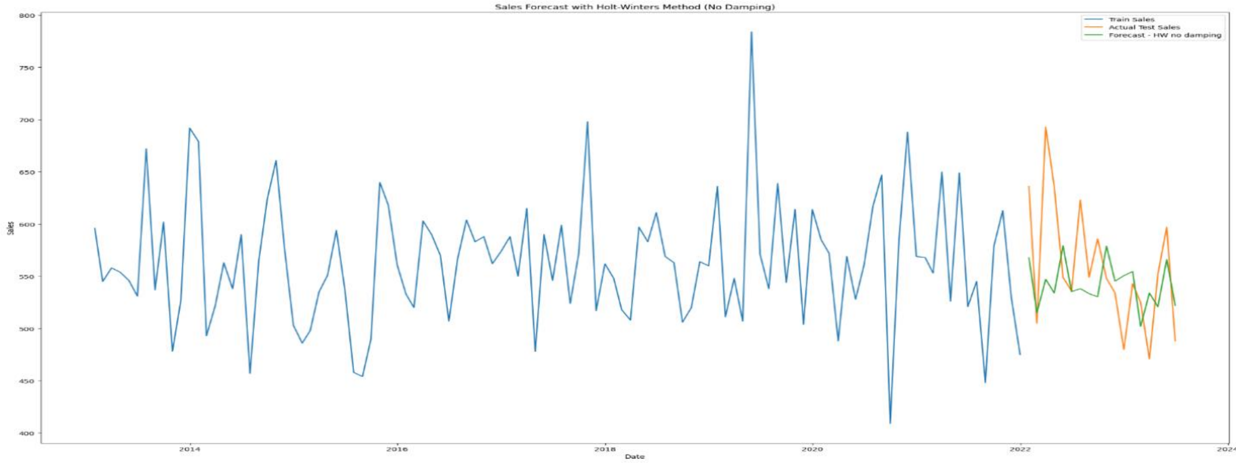


Figure 3. 49. ES-ELM Forecast for Item One (without Dumping).

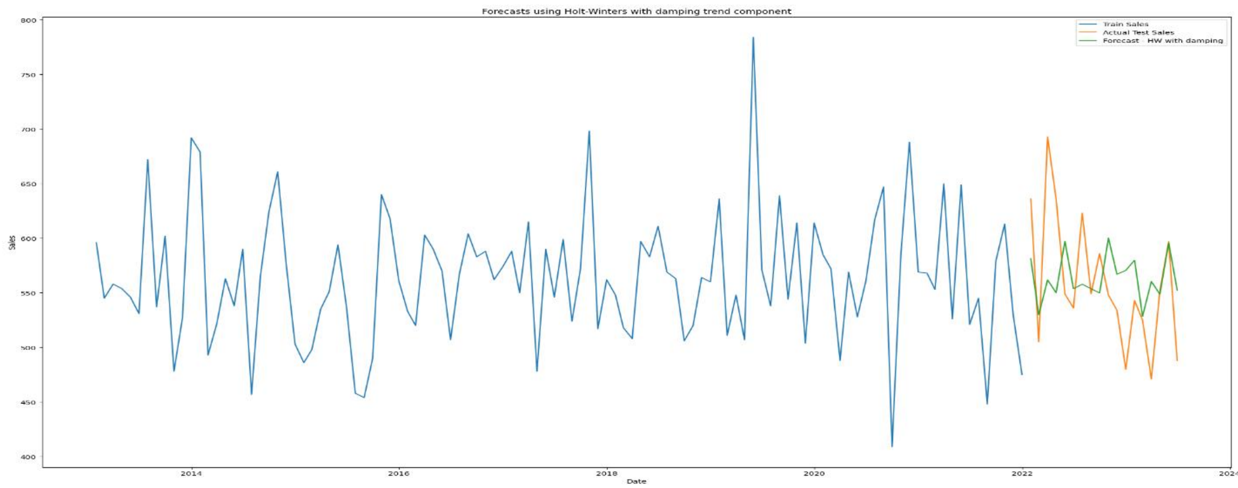


Figure 3. 50. ES-ELM Forecast for Item One (with Dumping).

Results for Holt-Winters Model Without Damping:

	total_sales	total_pred_sales
0	10051	9755.225293

Results for Holt-Winters Model With Damping:

	total_sales	total_pred_sales
0	10051	10138.905228

Figure 3. 51. ES-ELM Sales & Metrics for Item One.

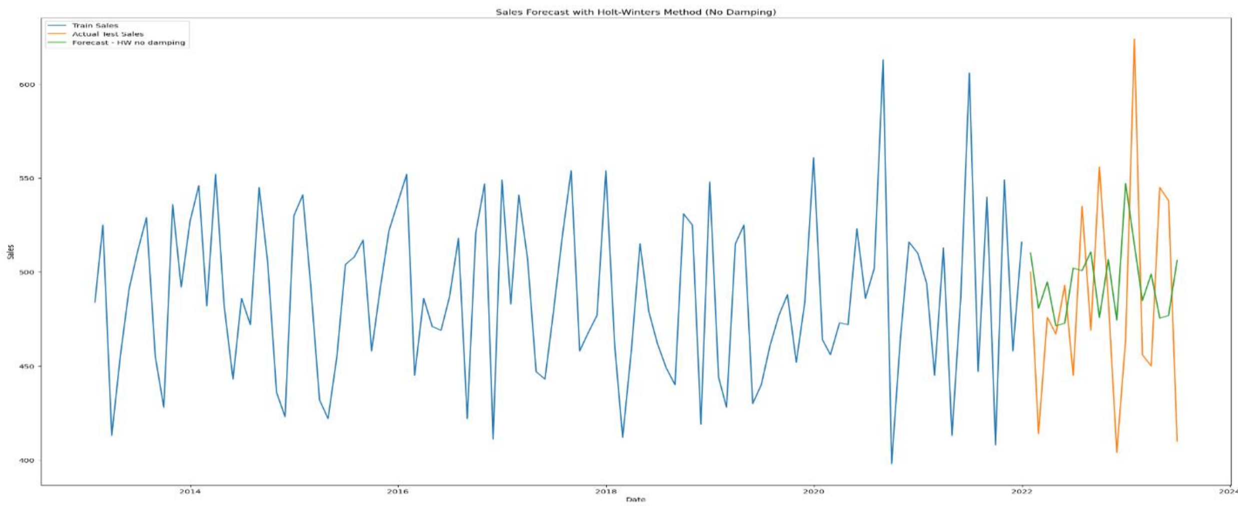


Figure 3. 52. ES-ELM Forecast for Item Two (without Dumping).

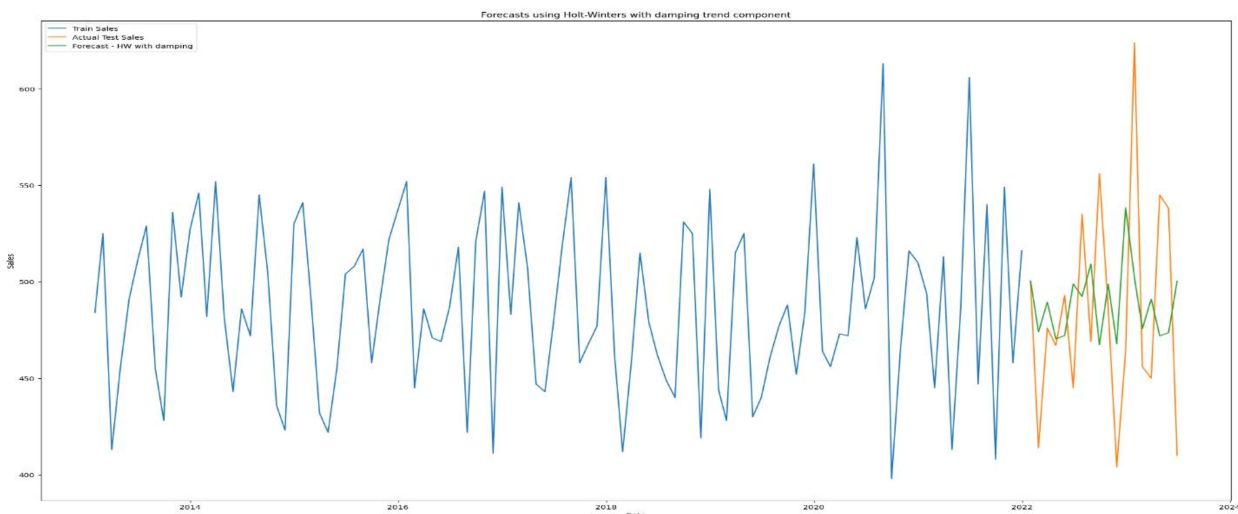


Figure 3. 53. ES-ELM Forecast for Item Two (with Dumping).

Results for Holt-Winters Model Without Damping:

	total_sales	total_pred_sales
0	8732	8904.364157

Results for Holt-Winters Model With Damping:

	total_sales	total_pred_sales
0	8732	8794.239124

Figure 3. 54. ES-ELM Sales & Metrics for Item Two.

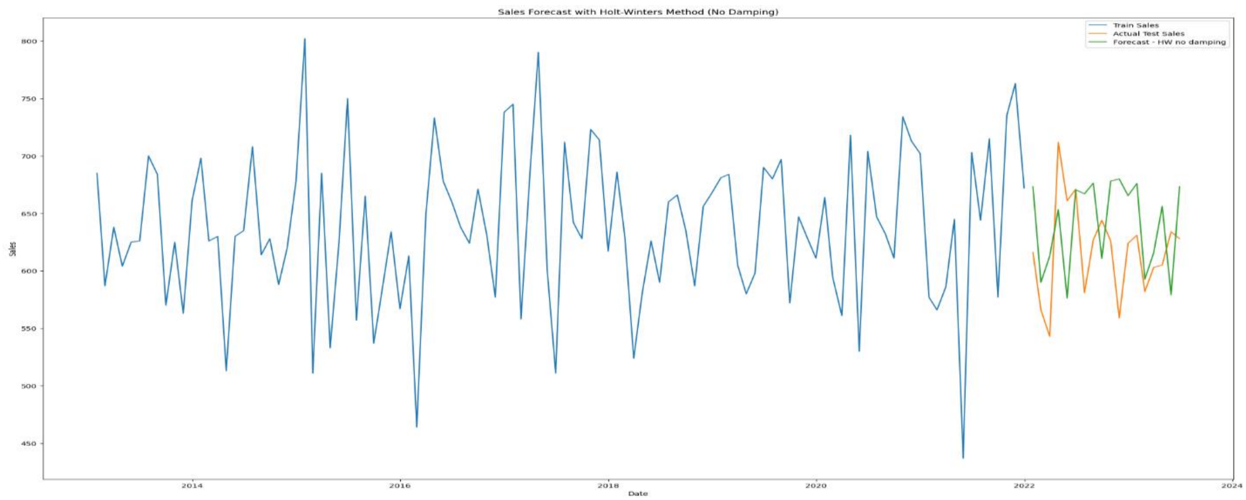


Figure 3. 55. ES-ELM Forecast for Item Three (without Dumping).

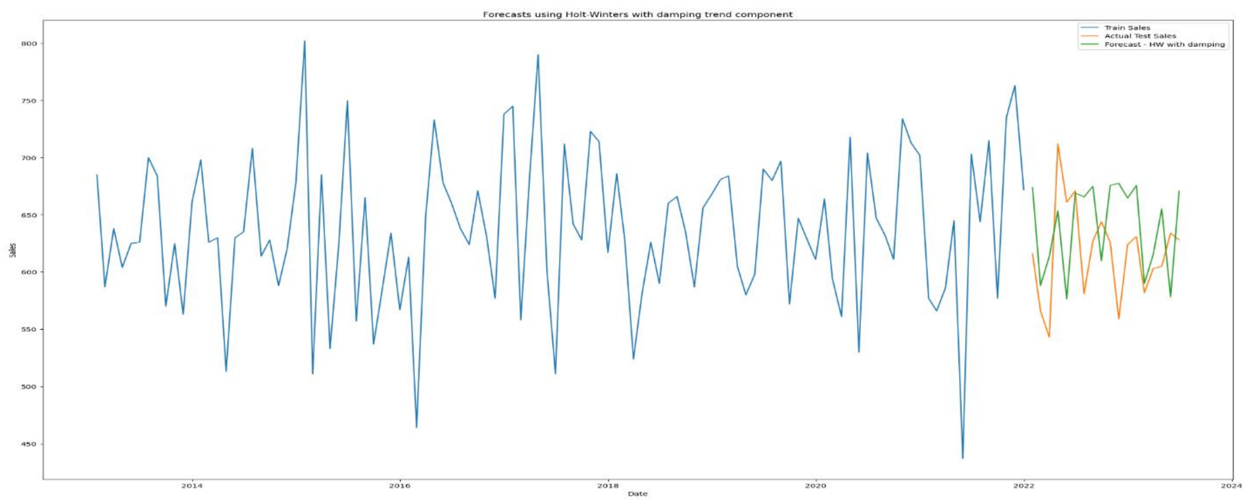


Figure 3. 56. ES-ELM Forecast for Item Three (with Dumping).

```
Results for Holt-Winters Model Without Damping:
```

	total_sales	total_pred_sales
0	11113	11547.258474

```
Results for Holt-Winters Model With Damping:
```

	total_sales	total_pred_sales
0	11113	11528.358823

Figure 3. 57. ES-ELM Sales & Metrics for Item Three.

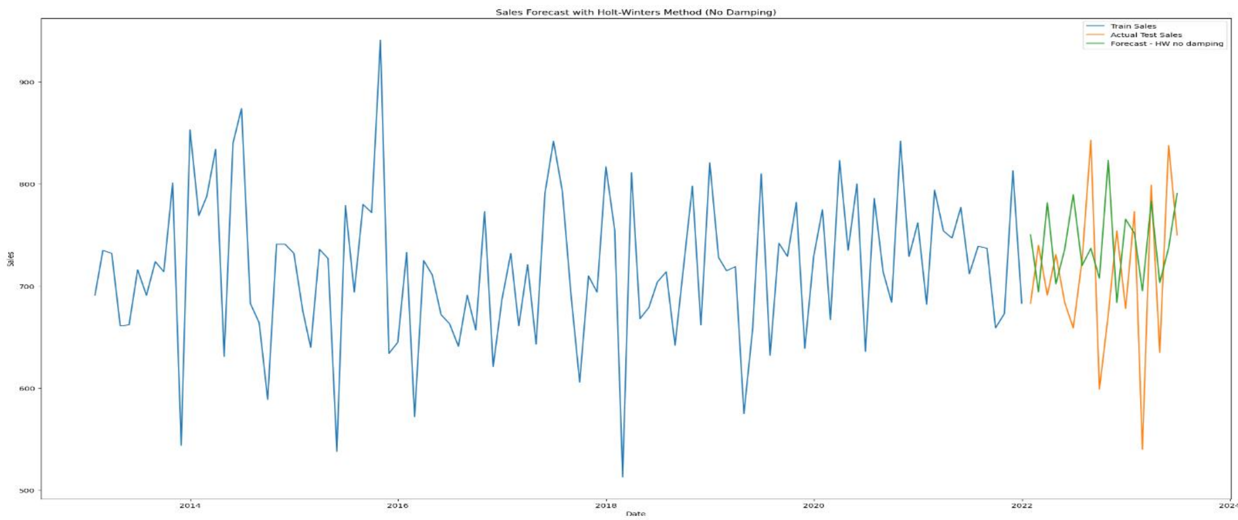


Figure 3. 58. ES-ELM Forecast for Item Four (without Dumping).

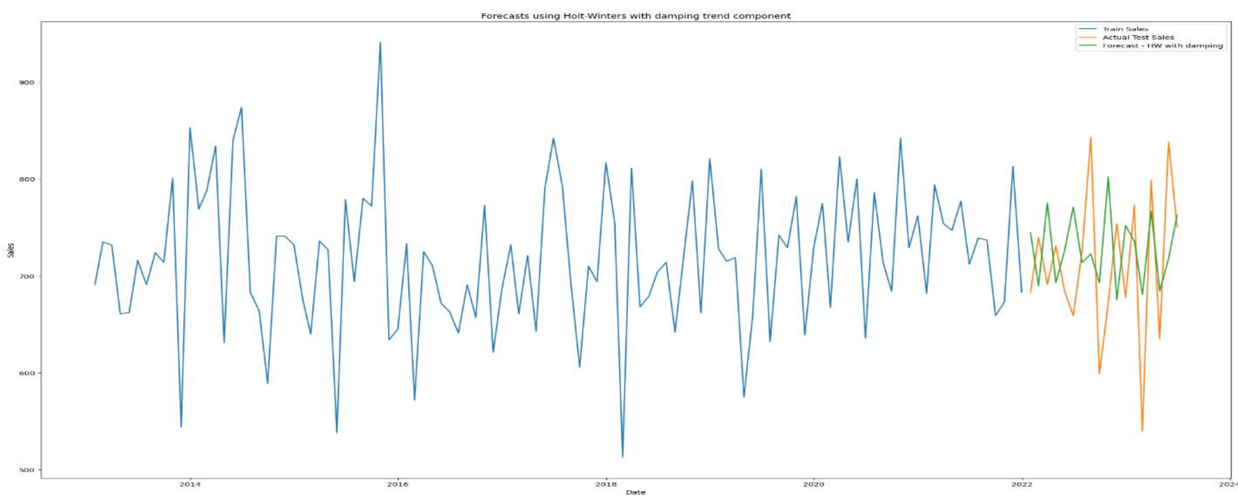


Figure 3. 59. ES-ELM Sales & Metrics for Item Four (with Dumping).

```
Results for Holt-Winters Model Without Damping:
```

	total_sales	total_pred_sales
0	12794	13352.613518

```
Results for Holt-Winters Model With Damping:
```

	total_sales	total_pred_sales
0	12794	13107.394161

Figure 3. 60. ES-ELM Sales & Metrics for Item Four.

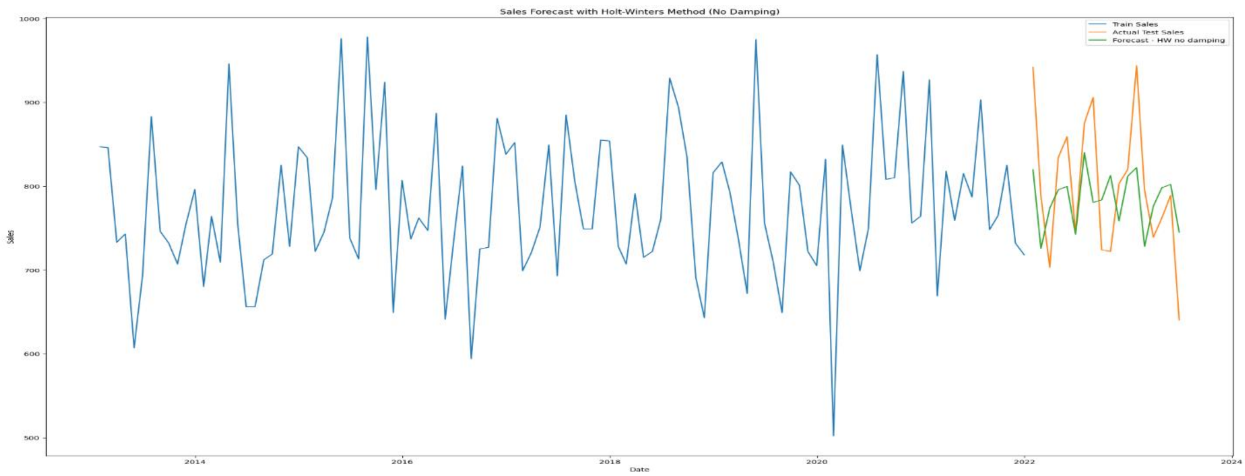


Figure 3. 61. ES-ELM Forecast for Item Five (without Dumping).

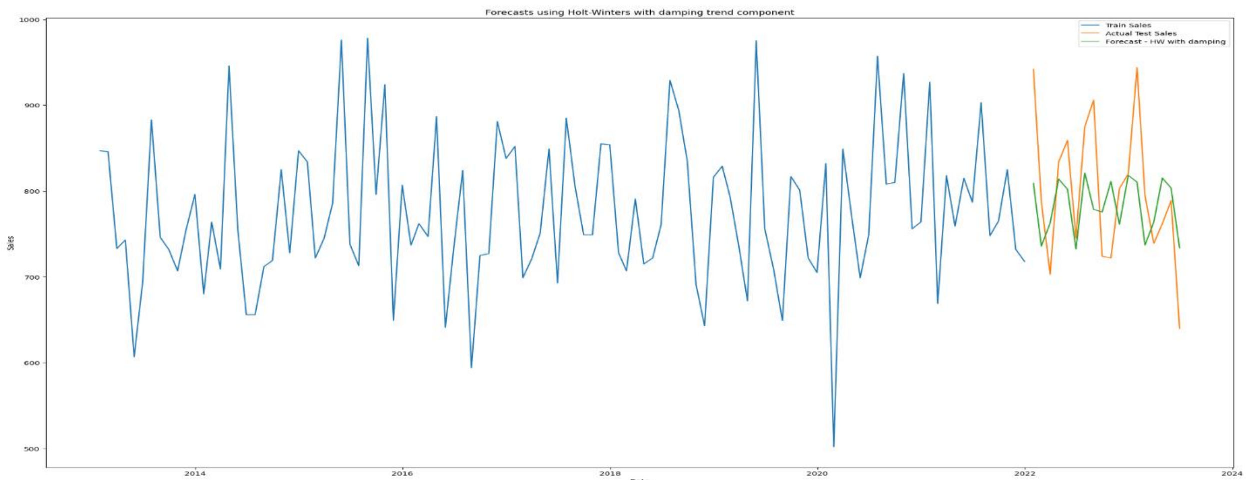


Figure 3. 62. ES-ELM Forecast for Item Five (with Dumping).

```
Results for Holt-Winters Model Without Damping:
```

	total_sales	total_pred_sales
0	14390	14115.423264

```
Results for Holt-Winters Model With Damping:
```

	total_sales	total_pred_sales
0	14390	14086.004132

Figure 3. 63. ES-ELM Sales & Metrics for Item Five.

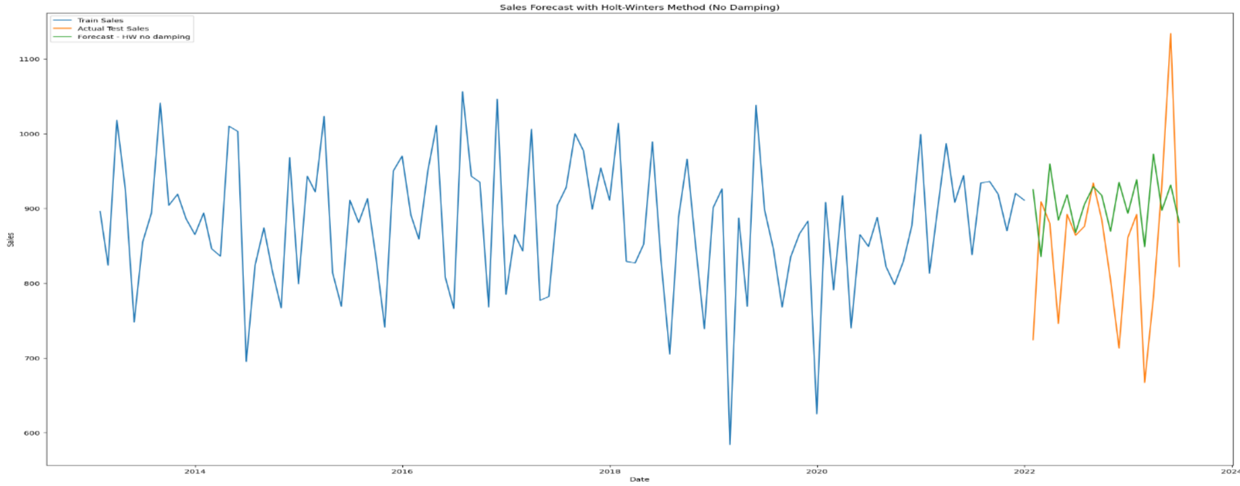


Figure 3. 64. ES-ELM Forecast for Item Six (without Dumping).

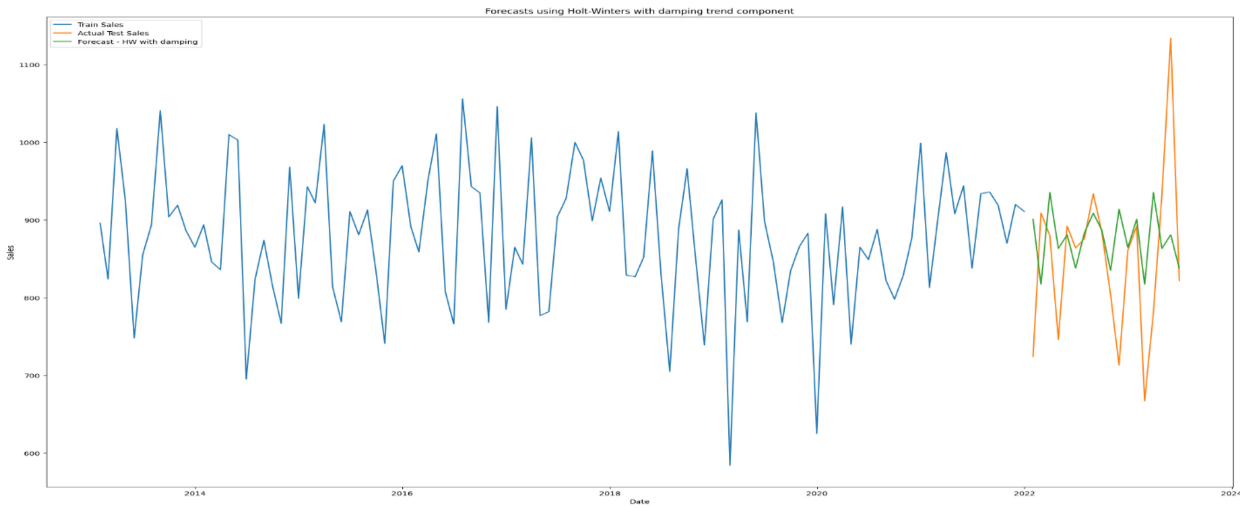


Figure 3. 65. ES-ELM Forecast for Item Six (with Dumping).

Results for Holt-Winters Model Without Damping:

	total_sales	total_pred_sales
0	15316	16309.312141

Results for Holt-Winters Model With Damping:

	total_sales	total_pred_sales
0	15316	15765.820887

Figure 3. 66. ES-ELM Sales & Metrics for Item Six.

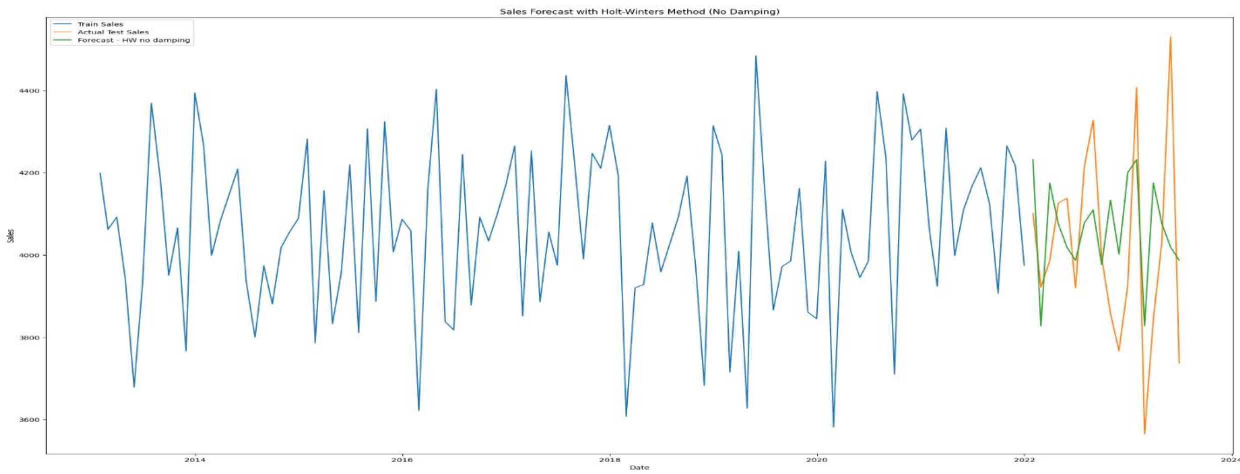


Figure 3. 67. ES-ELM Forecast for Store Items (without Dumping).

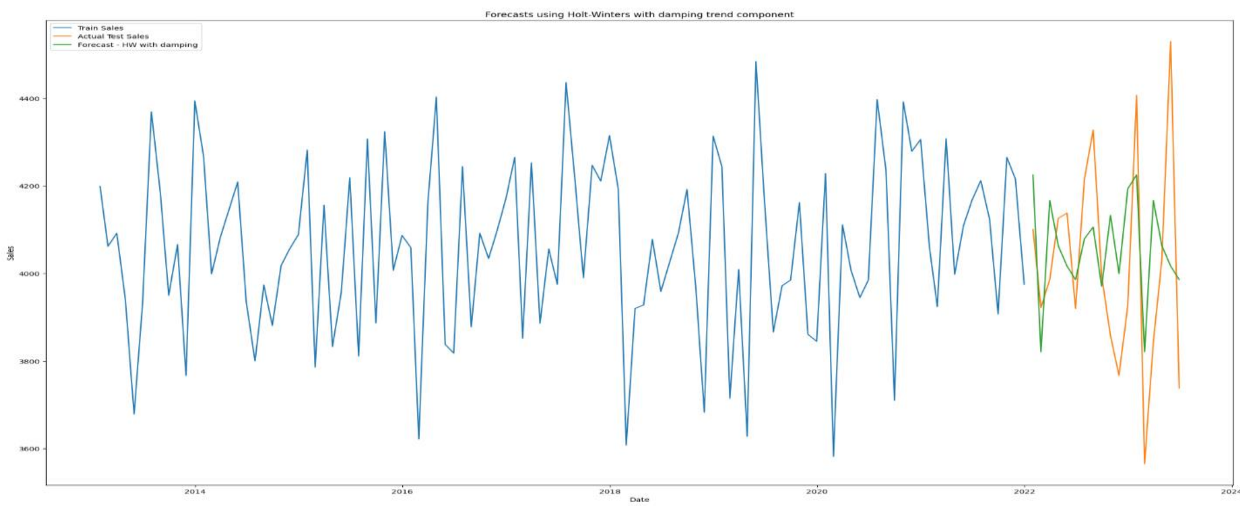


Figure 3. 68. ES-ELM Forecast for Store Items (with Dumping).

Results for Holt-Winters Model Without Damping:

	total_sales	total_pred_sales
0	72396	73130.769822

Results for Holt-Winters Model With Damping:

	total_sales	total_pred_sales
0	72396	73038.93656

Figure 3. 69. ES-ELM Sales & Metrics for Store Items.

Summary:

Embarking on the journey of sales forecasting involves a meticulously crafted roadmap, starting with the critical step of model selection. This guide, comprising multiple strategic steps, navigates through the diverse array of forecasting models to pinpoint the one best suited for the task at hand. Dataset selection is the linchpin, where careful consideration of relevance, completeness, and alignment with business goals ensures the foundation for robust predictions. Transitioning into the realm of dataset visualization, plots become the storytellers, unraveling patterns and insights within the data, and guiding subsequent modeling decisions. The model-building phase unfolds as a systematic guide, encompassing steps from preprocessing to parameter tuning, crafting a model capable of capturing the intricacies of the business environment. The climax of this journey lies in presenting the model results, where plots vividly showcase the forecasting prowess of the chosen models, offering a tangible representation of their performance against defined objectives. This iterative process underscores the symbiotic relationship between each phase, culminating in a refined and accurate sales forecasting model.

Chapter Four

Practical Application

the Practical Section shows the journey of refining sales forecasting, two pivotal elements took center stage: Model Modification and Dataset Manipulation. In the realm of Model Modification, the focus was on a strategic evolution, involving updates, changes, and alterations to simplify, enhance efficiency, and optimize the model's overall performance. Simultaneously, Dataset Manipulation played a crucial role, where a meticulous process of splitting the dataset into individual files and manipulating sales dates at varying intervals aimed to observe the model's dynamic response. This deliberate tinkering sought to uncover insights into how alterations in the dataset structure impact the model's adaptability and, consequently, the resulting forecast outcomes. The following narrative delves into these transformative processes, providing a comprehensive understanding of the deliberate modifications made to both model and dataset.

In this chapter, we will be looking at:

1. **Model Modification.**
2. **Deep AR and Deep Learning Models.**
3. **Models Results.**
4. **Summary.**

Model modification:

I. Comparison table for the entry-level model and primary model made by Linear regression:

Linear regression model		
Comparison	**Entry-level model**	**Primary model**
Model Architecture and Data Preprocessing	▪ Utilizes a linear regression framework on any store item. ▪ specifically filtered for store 1. ▪ Focuses on feature engineering with date-related attributes, lag features, and rolling statistics.	▪ Employs a comprehensive approach to any store item. ▪ filtering for store 1. ▪ Incorporates advanced analyses such as ACF/PACF plots, seasonal decomposition, and correlation matrices, alongside feature engineering.
Feature Selection	▪ Applies SelectKBest for feature selection. ▪ emphasizing simplicity and interpretability in model design.	▪ Leverages SelectKBest for feature selection. ▪ incorporating a broader set of features and analyses. ▪ suitable for capturing nuanced time-series dynamics.
Model Training and Evaluation	▪ Trains the model using linear regression.	▪ Utilizes linear regression for training. ▪ additional evaluation metrics and visualizations, showcasing a more

	▪ accompanied by a rigorous evaluation using metrics like MAE, RMSE, MAPE, and Accuracy. ▪ Visualizations provide insights into model performance.	comprehensive understanding of time-series patterns.
Model Insights and Applicability	▪ Focused on simplicity and streamlined interpretability. ▪ suitable for scenarios where a straightforward approach suffices.	▪ Offers a more intricate solution. ▪ integrate advanced time-series analyses for a nuanced comprehension of underlying patterns. ▪ particularly valuable when detailed insights into time-series dynamics are crucial.
summary	▪ while the **entry-level model** prioritizes simplicity and interpretability. ▪ **The primary model** presents a more detailed and comprehensive approach, integrating advanced time-series analyses for a nuanced understanding of sales forecasting dynamics. ▪ The choice between the two models depends on the specific requirements and characteristics of the sales forecasting task at hand.	

Table 4. 1. Linear Regression Entry-Level Vs. Primary Models.

II. Comparison table for the entry-level model and primary model made by ARIMA:

ARIMA Model		
Comparison	**Entry-level model**	**Primary model**
Model Architecture and Data Preprocessing	▪ Data Loading: Reads data from CSV file and formats the date. ▪ Indexing: Sets the date as the index for time-series analysis. ▪ Feature Engineering: Creates additional date-related features (year, month, day, weekday). ▪ Train-Test Split: Segregates data into training and testing sets. ▪ Modeling: Utilizes SARIMAX with order= (1, 1, 0) and seasonal_order= (1, 1, 0, 7). ▪ Forecasting: Generates predictions for the test dataset. ▪ Evaluation: Calculates metrics such as MAE, RMSE, R2, and Adjusted R2. ▪ Visualization: Displays a time-series plot with forecast and error magnitude. ▪ Summary Statistics: Presents total actual and predicted sales.	▪ Data Loading and Filtering: Reads data from CSV file and filters for store 1. ▪ Indexing and Feature Creation: Sets date as the index and adds date-related features. ▪ Stationarity Test: Conducts Augmented Dickey-Fuller Test for stationarity. ▪ Model Selection and Fitting: Fits ARIMA (6, 1, 1) and SARIMAX models. ▪ Forecasting: Generates predictions for the test dataset. ▪ Evaluation: Computes MAE, RMSE, and MAPE metrics. ▪ Visualization: Displays plots of train, test, predicted sales, and prediction errors. ▪ Summary Statistics: Aggregates total actual and predicted sales, along with error metrics.
Feature Selection	▪ No explicit feature selection; ▪ SARIMAX handles feature importance implicitly.	▪ No specific feature selection methodology was mentioned.

Model Training and Evaluation	▪ Training Method: SARIMAX with specified orders. ▪ Evaluation Metrics: MAE, RMSE, R2, Adjusted R2, and Accuracy.	▪ Training Methods: ARIMA (6, 1, 1) and SARIMAX with specified orders. ▪ Evaluation Metrics: MAE, RMSE, MAPE.
Model Insights and Applicability	▪ Focused on SARIMAX time-series modeling with a detailed statistical summary. ▪ Suited for scenarios requiring in-depth time-series analysis and accurate sales forecasting.	▪ Integrates ARIMA and SARIMAX models, emphasizing stationarity testing. ▪ Suitable for applications where both classical time-series and advanced modeling approaches are valuable.
summary	▪ while both models exhibit a commitment to rigorous time-series analysis. ▪ entry-level model emphasizes SARIMAX modeling with a detailed statistical summary, making it apt for scenarios requiring a nuanced understanding of time-series dynamics. ▪ In contrast, the primary model combines ARIMA and SARIMAX models, highlighting the importance of stationarity testing and providing valuable insights for applications that benefit from a hybrid modeling approach. ▪ The choice between the two models depends on the specific requirements and characteristics of the sales forecasting task at hand.	

Table 4. 2. ARIMA Entry-Level Vs. Primary models.

III. Comparison table for the entry-level model and primary model made by the Exponential smoothing model:

Holt Winter's triple Exponential Smoothing model		
comparison	**Entry-level model**	**Primary model**
Data Preprocessing	▪ Reads data from CSV file. ▪ Converts the 'date' column to a datetime format. ▪ Sets the 'date' column as the index. ▪ Filters data for 'store' equal to 1. ▪ Splits the data into training and testing sets.	▪ Reads data from CSV file. ▪ Filters data for 'store' equal to 1. ▪ Converts the 'date' column to a datetime format. ▪ Sets the 'date' column as the index. ▪ Splits the data into training and testing sets.
Seasonal decomposition	____	▪ Uses seasonal decomposition to visualize the trend, seasonal, and residual components of the training data.
Model training	▪ Uses the Holt-Winters method with additive trend and additive seasonal components. ▪ Fits two models: one without damping and one with damping.	▪ Uses the Holt-Winters method with additive trend and additive seasonal components. ▪ Fits two models: one without damping and one with damping.
Visualization	▪ Plots actual test sales along with forecasts for both models.	▪ Plots train sales, actual test sales, and forecasts for both models. ▪ Plots the forecast errors for both models.

	▪ Plots the error magnitude for both models.	
Evaluation metrics	▪ Defines custom metrics functions for Mean Squared Logarithmic Error (MSLE), Mean Percentage Error (MPE), and Adjusted R-squared. ▪ Calculates and displays metrics such as MAE, RMSE, R2, and Accuracy for both models. ▪ Prints the total actual sales and total predicted sales for both models.	▪ Defines custom metrics functions for Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). ▪ Calculates and displays metrics for both models. ▪ Prints the total actual sales, total predicted sales, and overall error for both models.
summary	▪ Both models follow a similar approach using the Holt-Winters method but differ in terms of data preprocessing, visualization, and custom metrics. ▪ The entry-level model focuses on plotting error magnitude and uses R2 as an accuracy metric, ▪ The primary model has a detailed seasonal decomposition plot and uses MAPE as an additional evaluation metric. ▪ The choice between the two models may depend on specific preferences and requirements for forecasting and model evaluation.	

Table 4. 3. Exponential Smoothing Entry-Level Vs. Primary.

Deep AR and Deep Learning Models:

comparison	Deep Learning Models	Deep AR
Model Architecture	Uses various ML/DL models including XGBoost, Random Forest, and custom DL	Focuses on Deep AR using GluonTS
Data Preprocessing	Extensive steps: handling missing values, merging datasets, outlier detection, normalization	Basic steps: loading data, setting train/test split, highlighting weekends
Feature Selection	Recursive feature elimination with Random Forest Regressor	Uses covariate time series such as price history
Model Training and Evaluation	Trains multiple models (XGBoost, custom DL) Metrics: MAE, MSE, RMSE, R2	Trains a single deep AR model Metrics: MSE, MAPE, Coverage at different quantiles
Model Insights	Detailed feature importance, visualization of correlations	Visualization of forecasts, prediction intervals
Applicability	Suitable for a variety of ML models and datasets	Focused on time-series forecasting for retail sales

summary	Comprehensive guide covering multiple models and techniques for predictive modeling	Specific to deep AR, demonstrating usage with GluonTS library

Table 4. 4. Deep Learning Models Vs. Deep AR models

Model results:

I. Linear regression results:

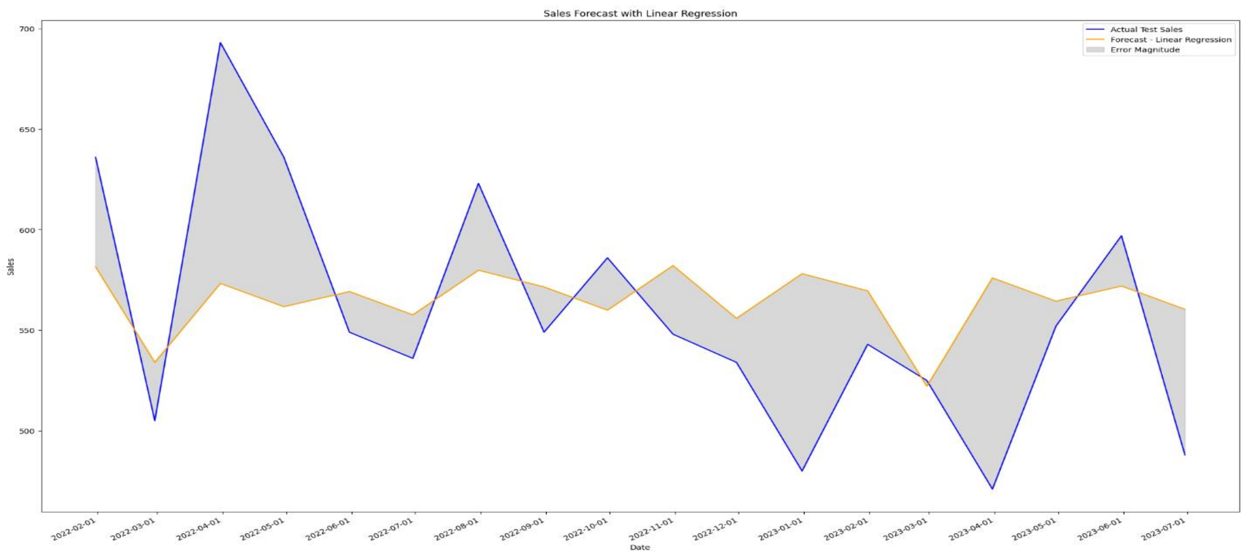


Figure 4. 1. LR-PM Forecast for Item One.

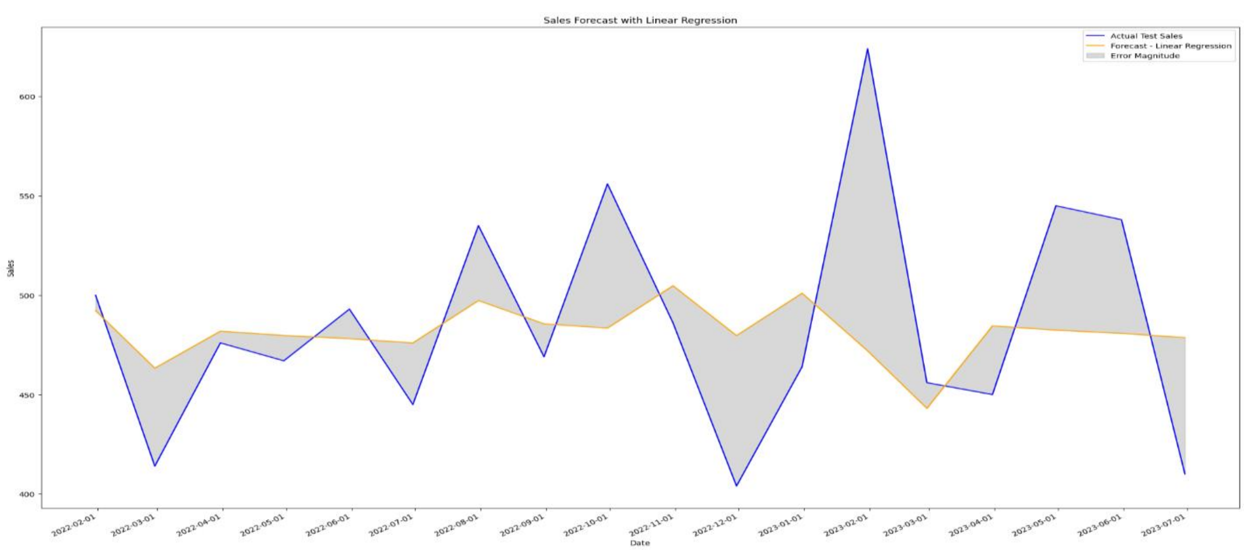


Figure 4. 2. LR-PM Forecast for Item Two.

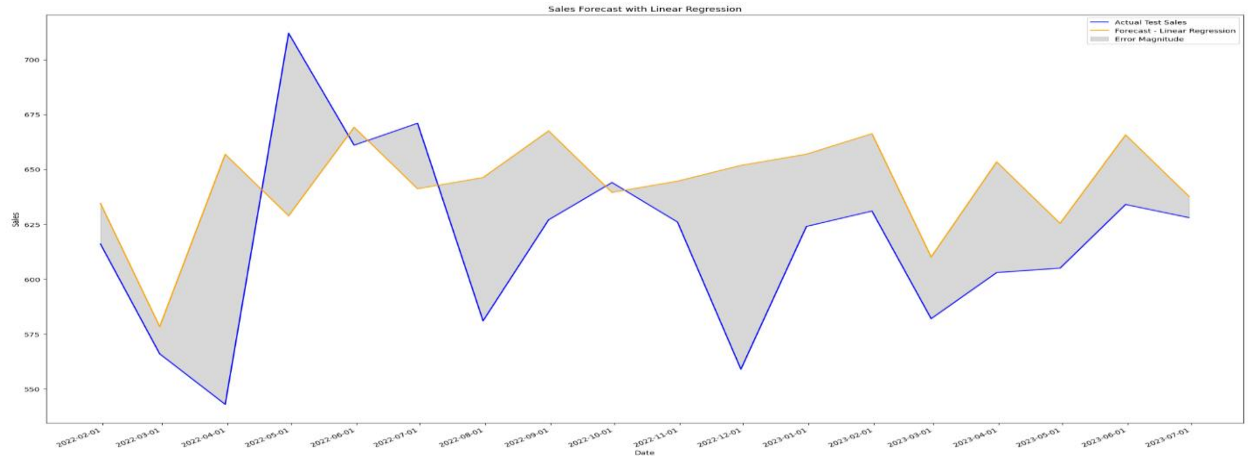


Figure 4. 3. LR-PM Forecast for Item Three.

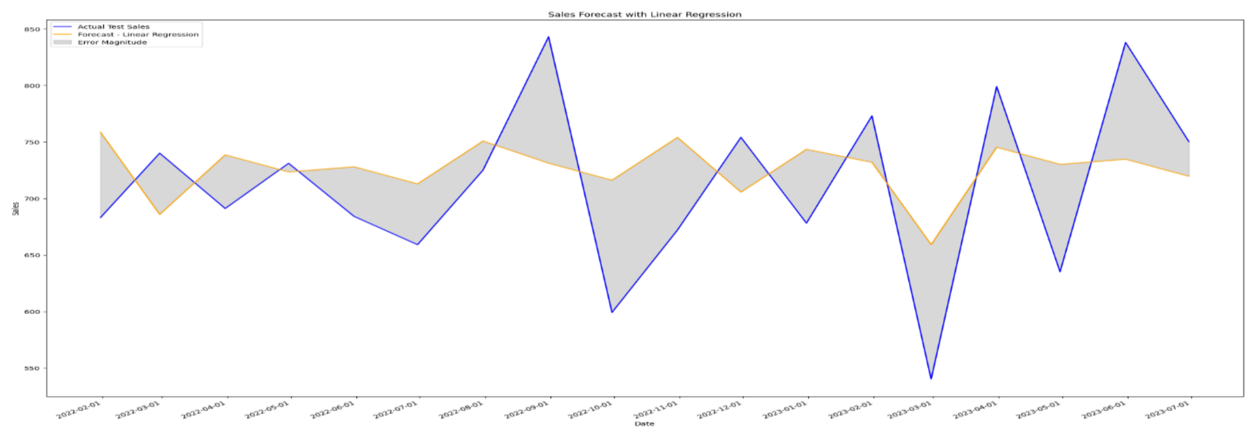


Figure 4. 4. LR-PM Forecast for Item Four.

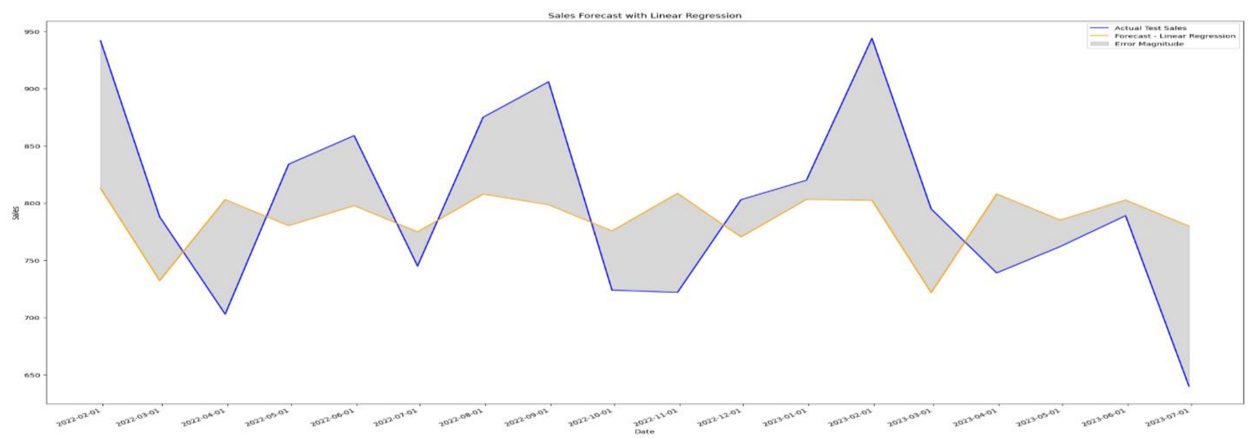


Figure 4. 5. LR-PM Forecast for Item Five.

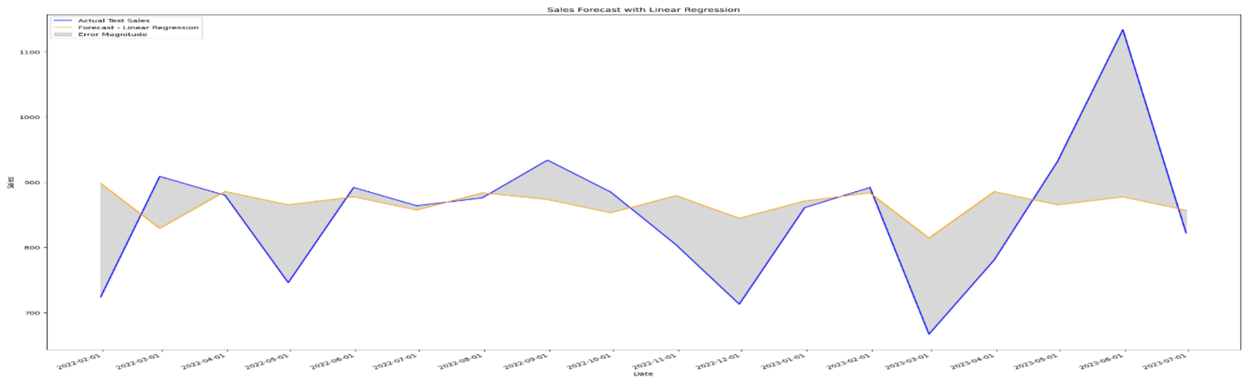


Figure 4. 6. LR-PM Forecast for Item Six.

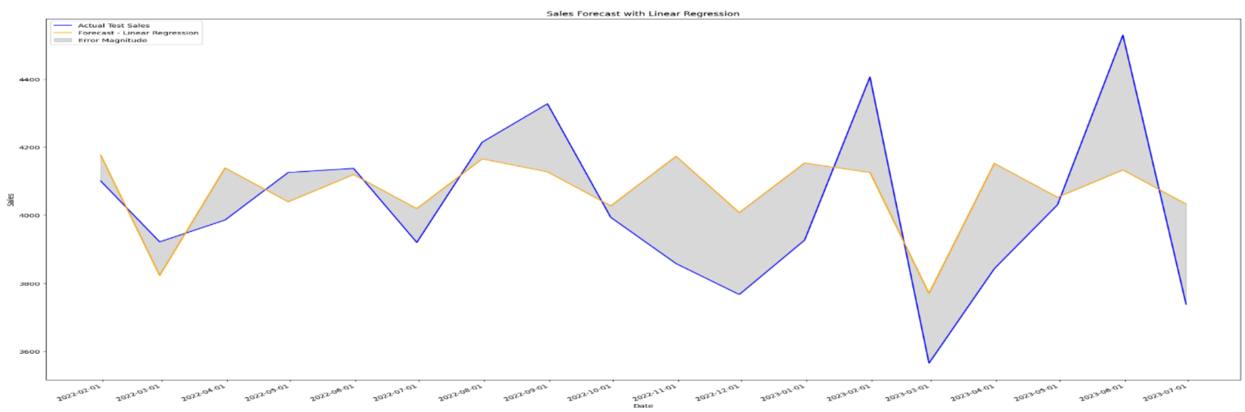


Figure 4. 7. LR-PM Forecast for Store Items.

	Metric	Value
0	MAE	44.981949
1	RMSE	56.092928
2	MAPE	8.126070

Figure 4. 8. LR-PM Metrics for Item One.

	Metric	Value
0	MAE	42.660614
1	RMSE	55.106307
2	MAPE	8.631555

Figure 4. 9. LR-PM Metrics for Item Two.

	Metric	Value
0	MAE	38.678579
1	RMSE	49.100483
2	MAPE	6.415502

Figure 4. 10. LR-PM Metrics for Item Three.

	Metric	Value
0	MAE	65.336911
1	RMSE	72.814165
2	MAPE	9.500125

Figure 4. 11. LR-PM Metrics for Item Four.

	Metric	Value
0	MAE	69.615618
1	RMSE	80.074302
2	MAPE	8.756294

Figure 4. 12. LR-PM Metrics for Item Five.

	Metric	Value
0	MAE	74.309535
1	RMSE	100.992613
2	MAPE	9.028653

Figure 4. 13. LR-PM Metrics for Item Six.

	Metric	Value
0	MAE	172.898211
1	RMSE	207.185579
2	MAPE	4.320232

Figure 4. 14. LR-PM Metrics for Store Items.

```
Total Actual Sales: 10051.00
Total Predicted Sales: 10169.33
```

Figure 4. 15. LR-PM Sales for Item One.

```
Total Actual Sales: 8732.00
Total Predicted Sales: 8664.62
```

Figure 4. 16. LR-PM Sales for Item Two.

```
Total Actual Sales: 11113.00
Total Predicted Sales: 11574.26
```

Figure 4. 17. LR-PM Sales for Item Three.

```
Total Actual Sales: 12794.00
Total Predicted Sales: 13067.91
```

Figure 4. 18. LR-PM Sales for Item Four.

```
Total Actual Sales: 14390.00
Total Predicted Sales: 14165.60
```

Figure 4. 19. LR-PM Sales for Item Five.

```
Total Actual Sales: 15316.00
Total Predicted Sales: 15603.57
```

Figure 4. 20. LR-PM Sales for Item Six.

```
Total Actual Sales: 72396.00
Total Predicted Sales: 73245.29
```

Figure 4. 21. LR-PM Sales for Store Items.

II. ARIMA results:

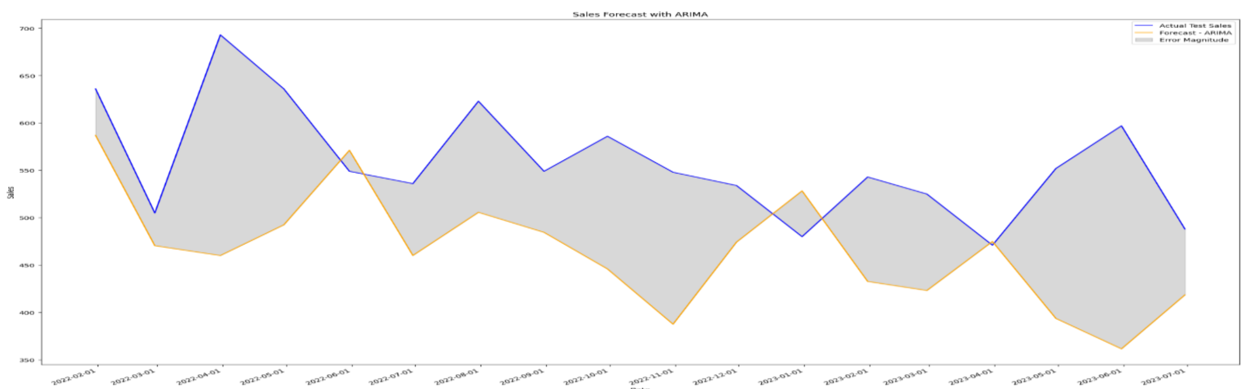


Figure 4. 22. A-PM Forecast for Item One.

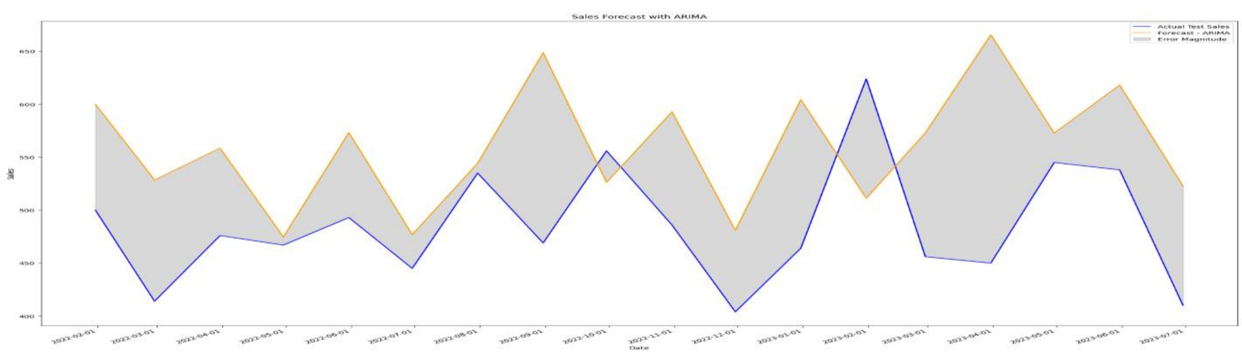


Figure 4. 23. A-PM Forecast for Item Two.

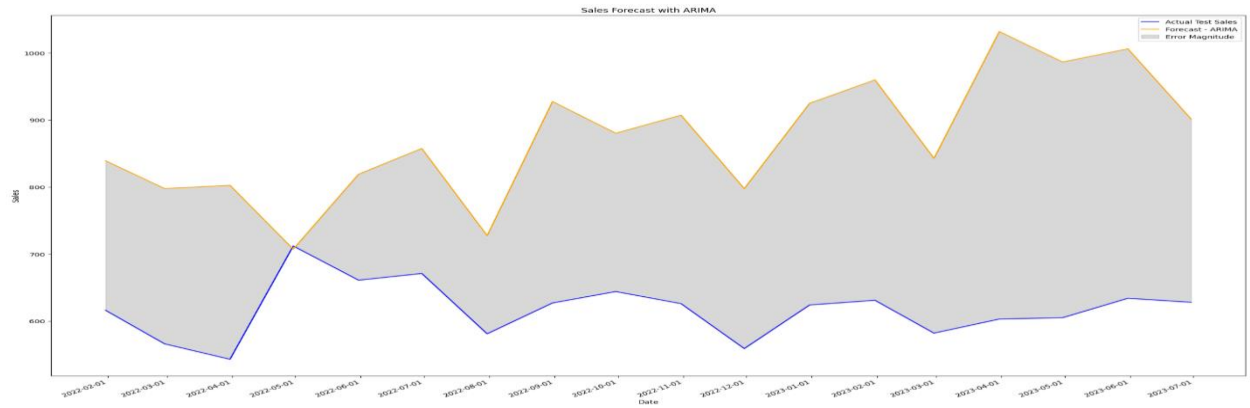


Figure 4. 24. A-PM Forecast for Item Three.

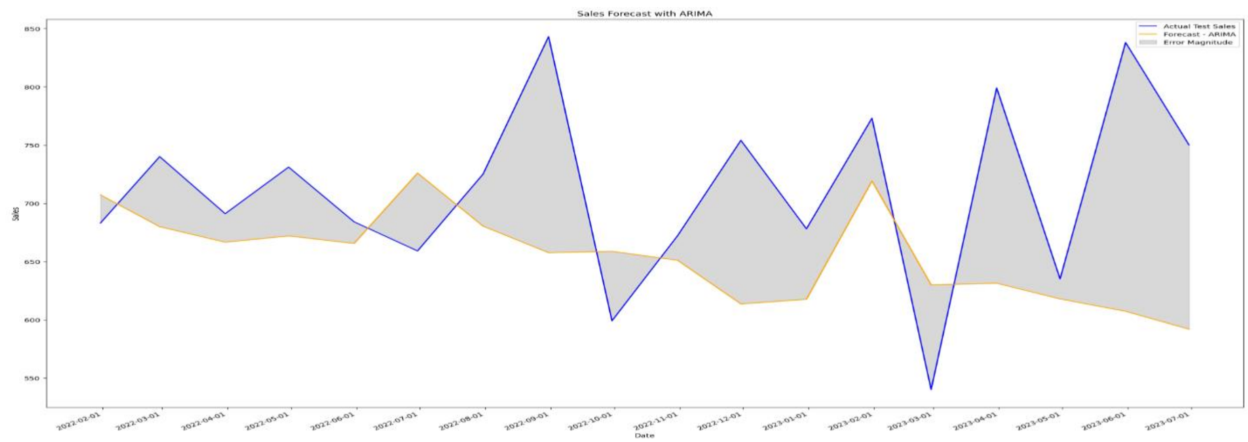


Figure 4. 25. A-PM Forecast for Item Four.

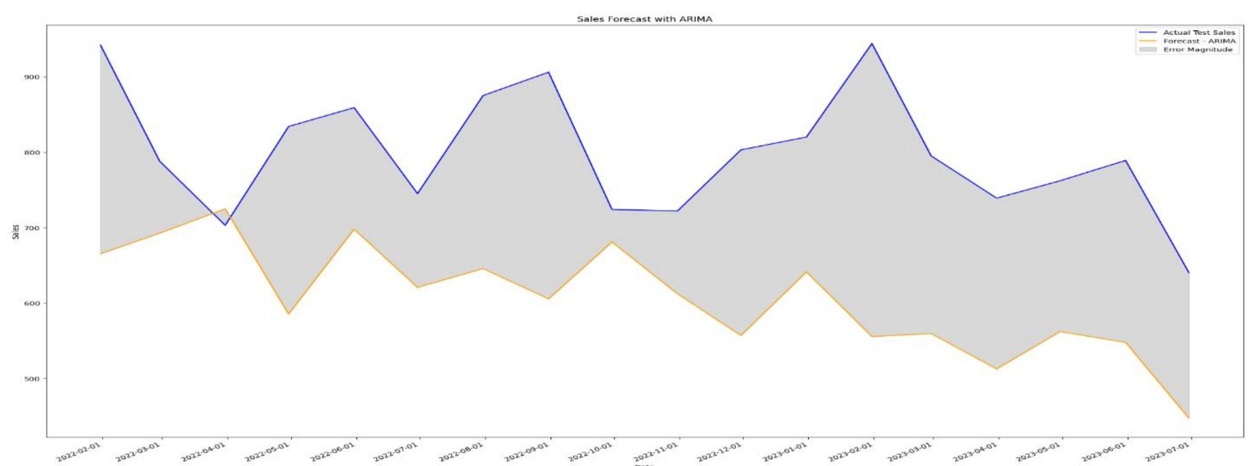


Figure 4. 26. A-PM Forecast for Item Five.

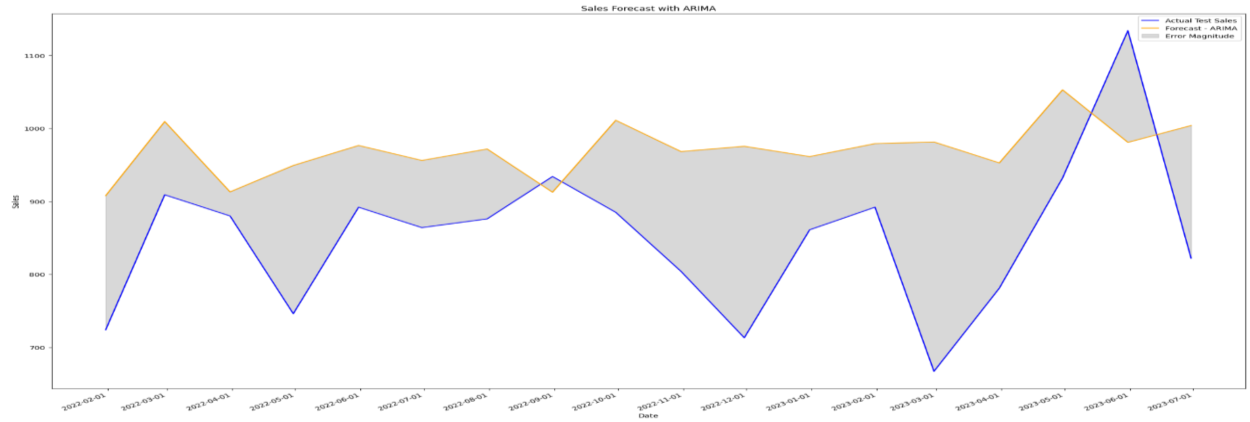


Figure 4. 27. A-PM Forecast for Item Six.

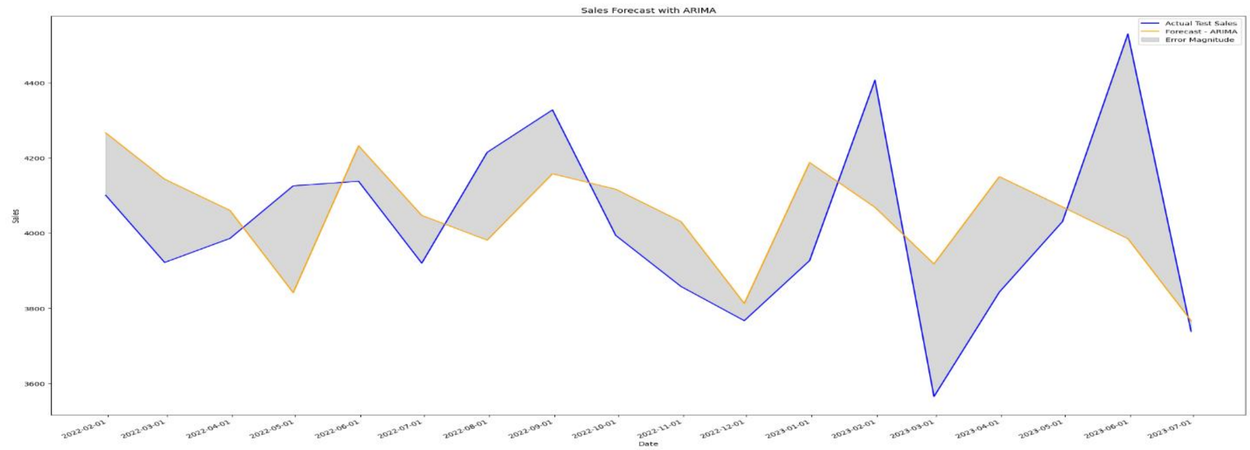


Figure 4. 28. A-PM Forecast for Store Items.

	Metric	Value
0	MAE	101.538058
1	RMSE	120.586068

Figure 4. 29. A-PM Metrics for Item One.

	Metric	Value
0	MAE	90.358858
1	RMSE	105.572899

Figure 4. 30. A-PM Metrics for Item Two.

	Metric	Value
0	MAE	256.259453
1	RMSE	273.183768

Figure 4. 31. A-PM Metrics for Item Three.

	Metric	Value
0	MAE	82.377125
1	RMSE	103.972722

Figure 4. 32. A-PM Metrics for Item Four.

	Metric	Value
0	MAE	195.621399
1	RMSE	215.012054

Figure 4. 33. A-PM Metrics for Item Five.

	Metric	Value
0	MAE	138.701079
1	RMSE	156.423073

Figure 4. 34. A-PM Metrics for Item Six.

	Metric	Value
0	MAE	199.276120
1	RMSE	238.081304

Figure 4. 35. A-PM Metrics for Store Items.

```
Total Actual Sales: 10051.00
Total Predicted Sales: 8372.19
```

Figure 4. 36. A-PM Sales for Item One.

```
Total Actual Sales: 8732.00
Total Predicted Sales: 10073.03
```

Figure 4. 37. A-PM Sales for Item Two.

```
Total Actual Sales: 11113.00
Total Predicted Sales: 15716.18
```

Figure 4. 38. A-PM Sales for Item Three.

```
Total Actual Sales: 12794.00
Total Predicted Sales: 11792.53
```

Figure 4. 39. A-PM Sales for Item Four.

```
Total Actual Sales: 14390.00
Total Predicted Sales: 10912.58
```

Figure 4. 40. A-PM Sales for Item Five.

```
Total Actual Sales: 15316.00
Total Predicted Sales: 17463.49
```

Figure 4. 41. A-PM Sales for Item Six.

```
Total Actual Sales: 72396.00
Total Predicted Sales: 72839.62
```

Figure 4. 42. A-PM Sales for Store Items.

III. Exponential smoothing results:

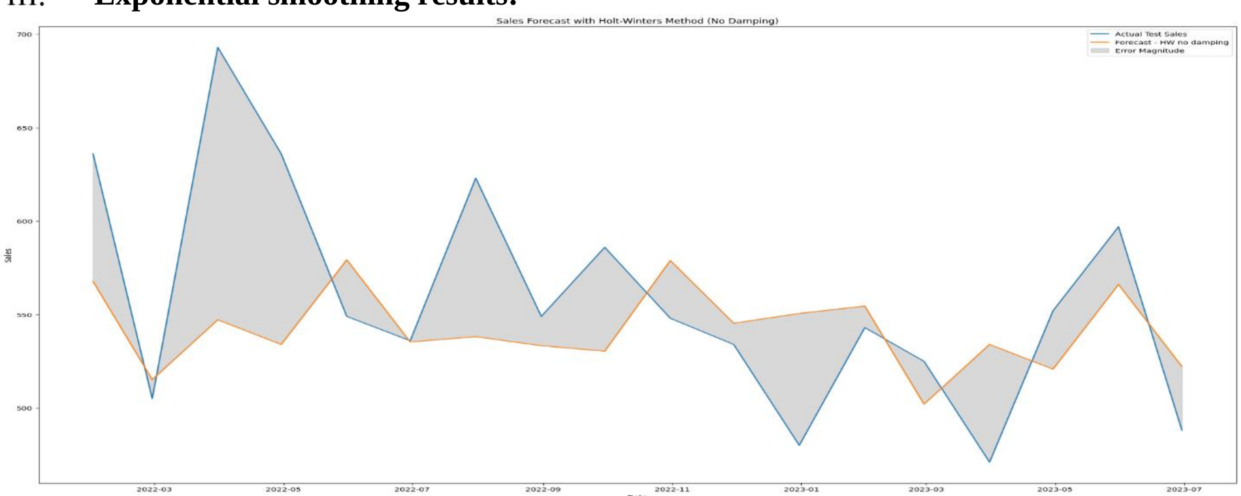


Figure 4. 43. ES-PM Forecast for Item One (without Dumping).

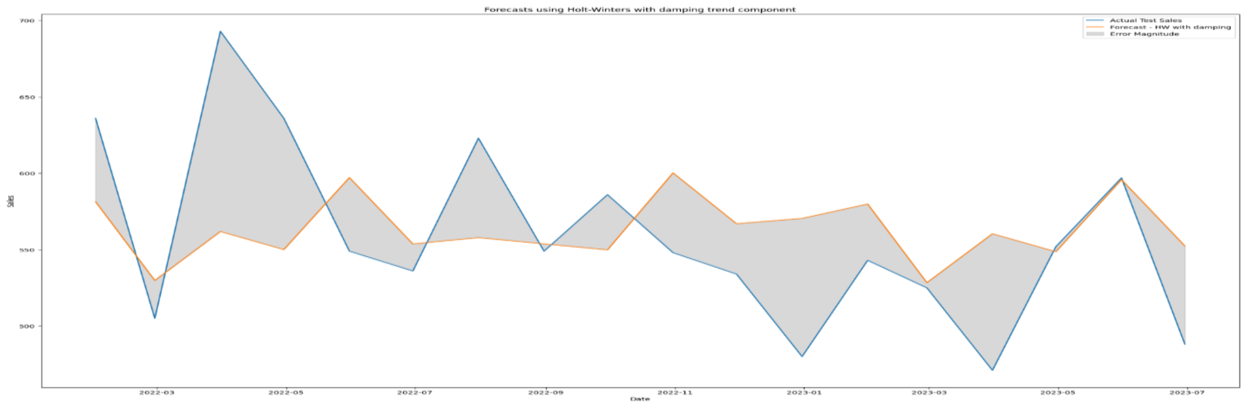


Figure 4. 44. ES-PM Forecast for Item One (with Dumping).

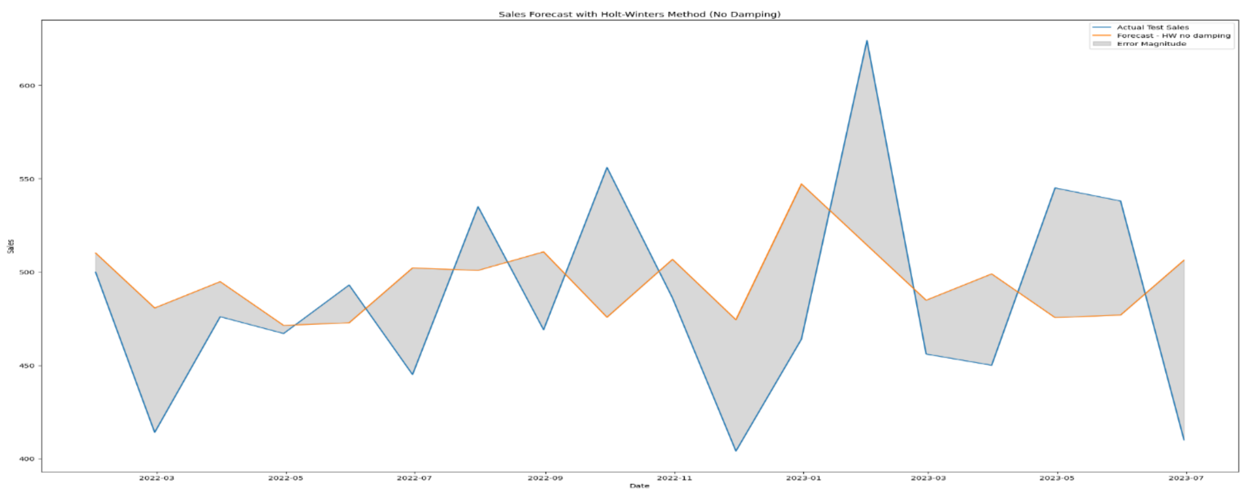


Figure 4. 45. ES-PM Forecast for Item Two (without Dumping).

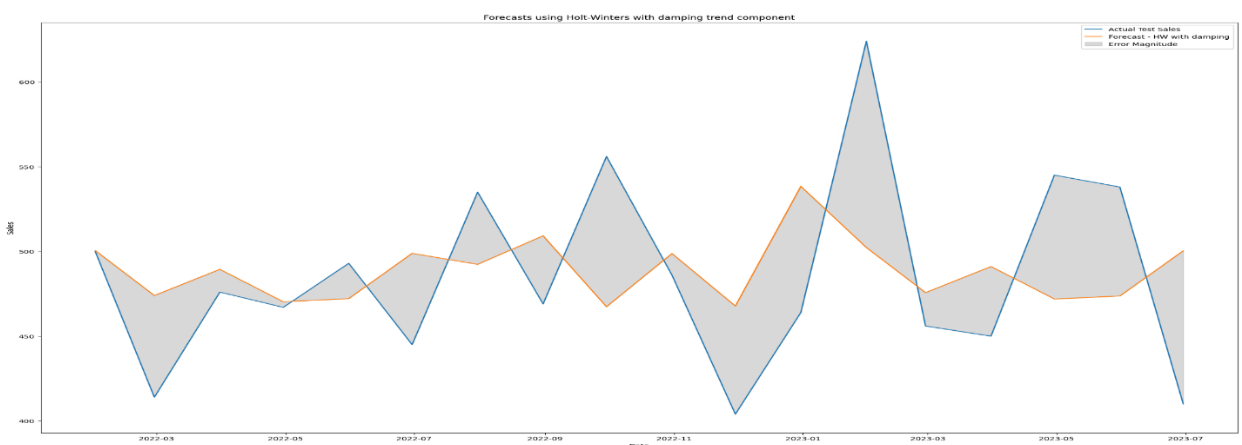


Figure 4. 46. ES-PM Forecast for Item Two (with Dumping).

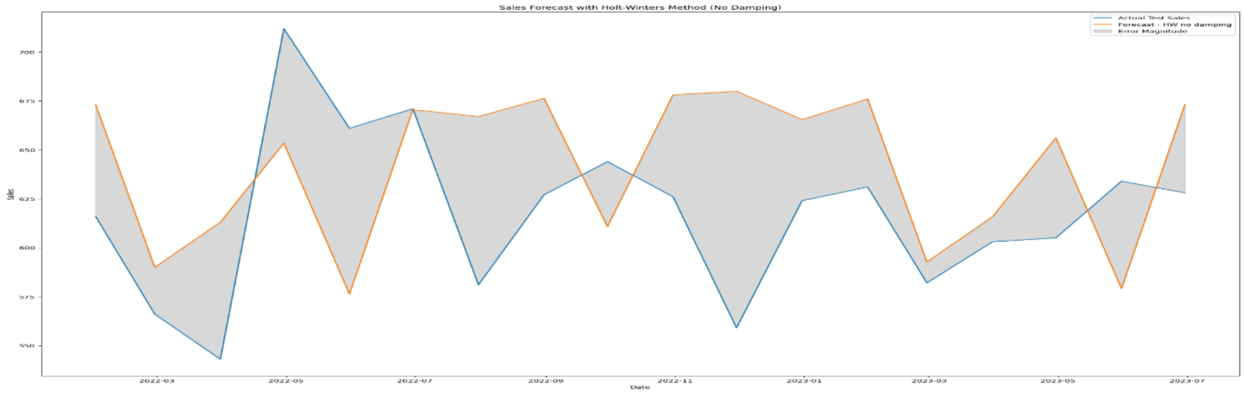


Figure 4. 47. ES-PM Forecast for Item Three (without Dumping).

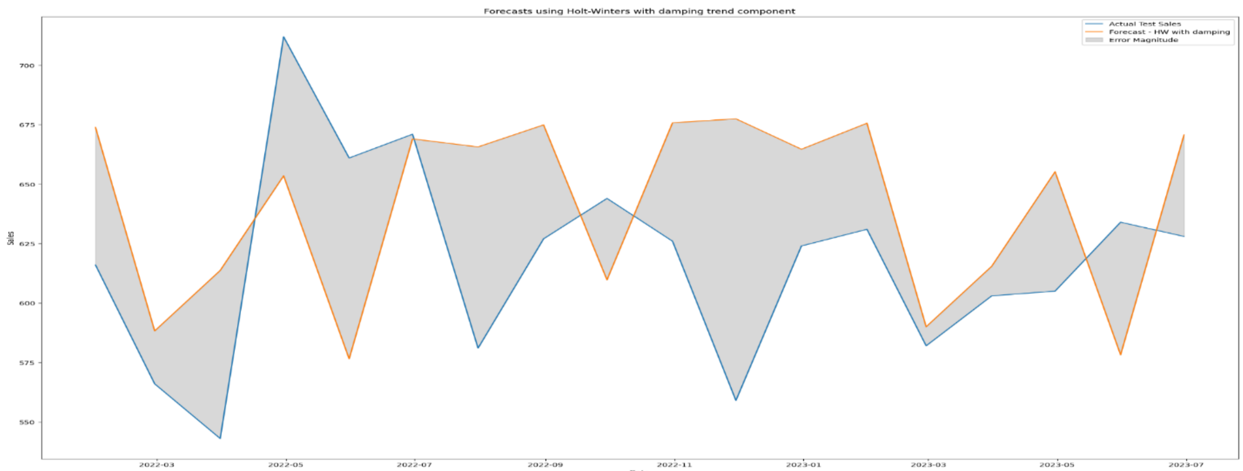


Figure 4. 48. ES-PM Forecast for Item Three (with Dumping).

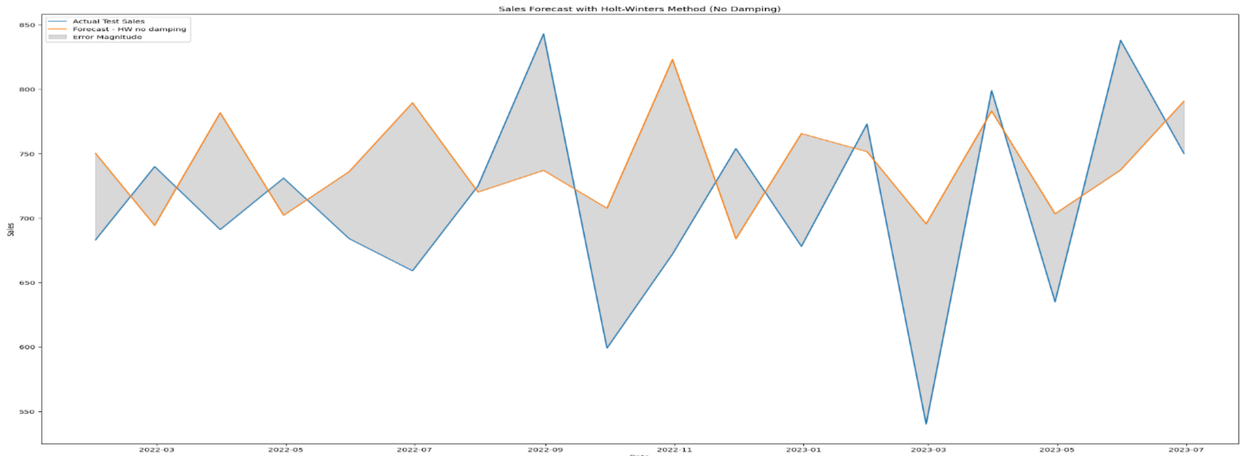


Figure 4. 49. ES-PM Forecast for Item Four (without Dumping).

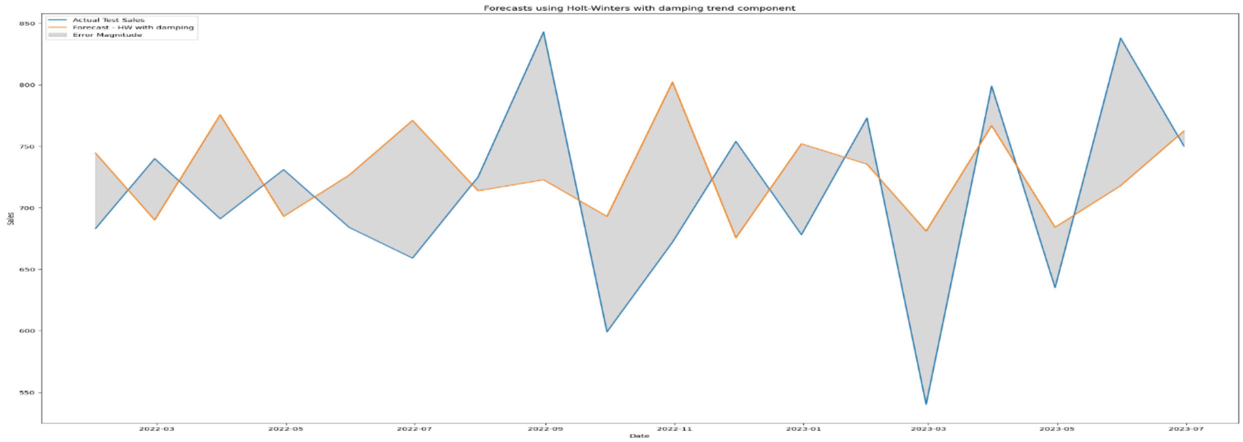


Figure 4. 50. ES-PM Forecast for Item Four (with Dumping).

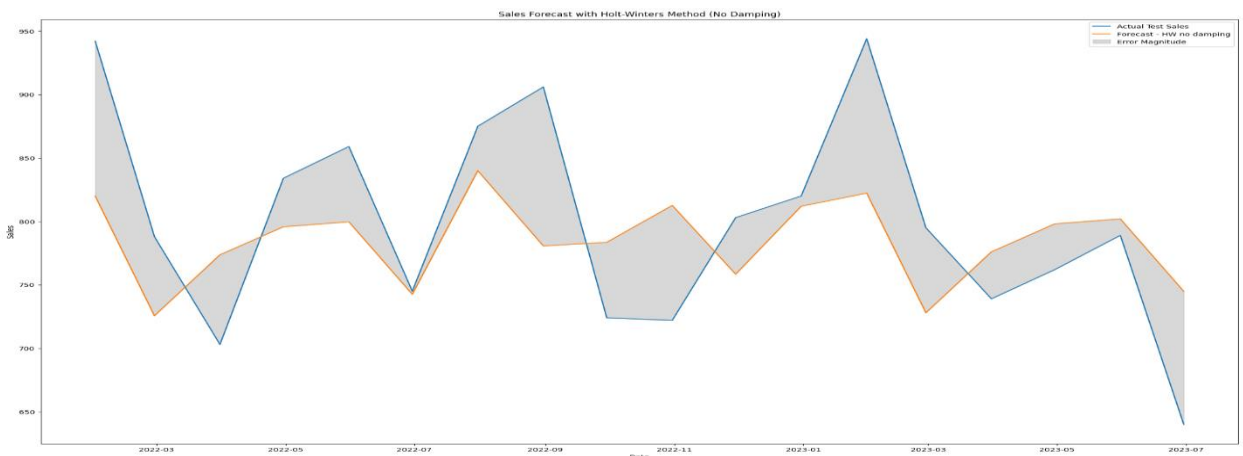


Figure 4. 51. ES-PM Forecast for Item Five (without Dumping).

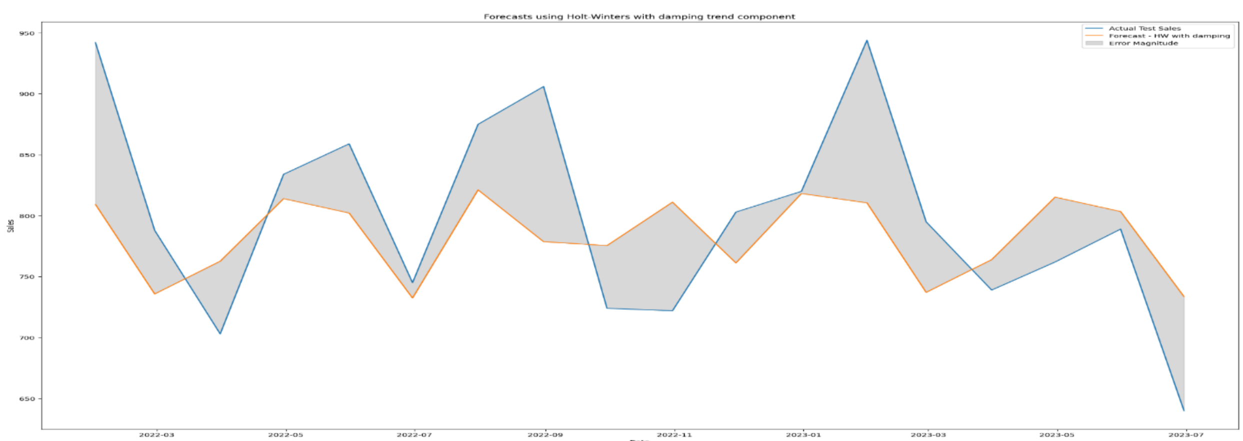


Figure 4. 52. ES-PM Forecast for Item Five (with Dumping).

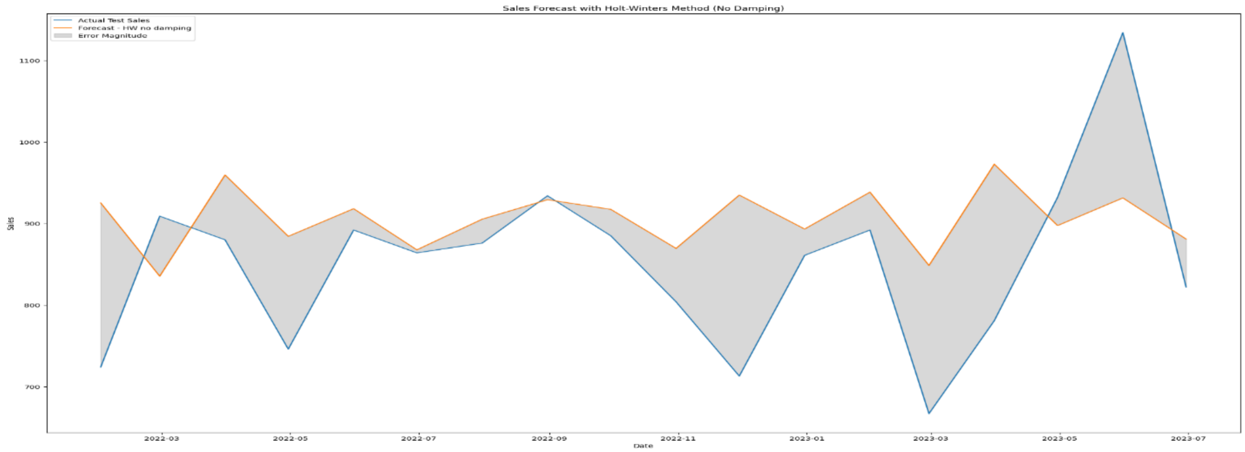


Figure 4. 53. ES-PM Forecast for Item Six (without Dumping).

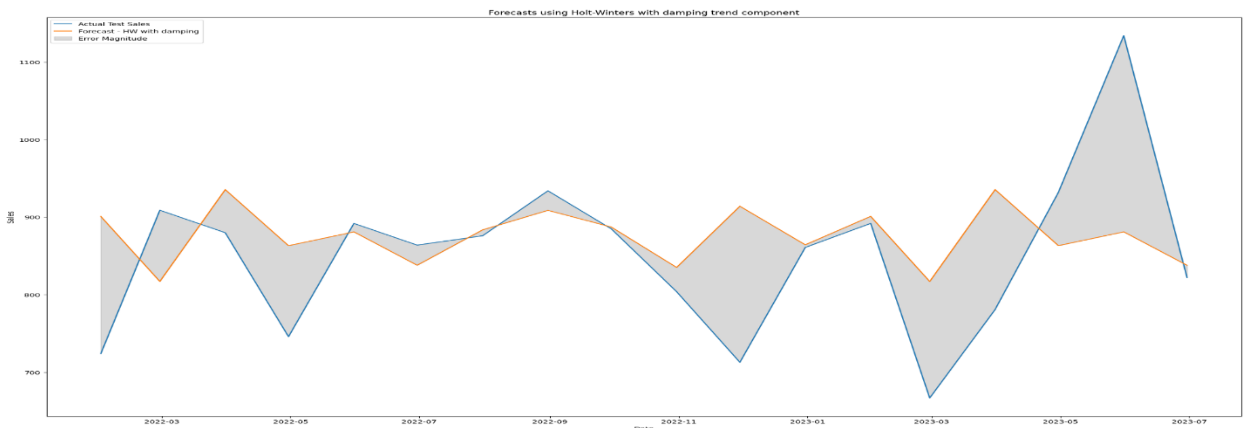


Figure 4. 54. ES-PM Forecast for Item Six (with Dumping).

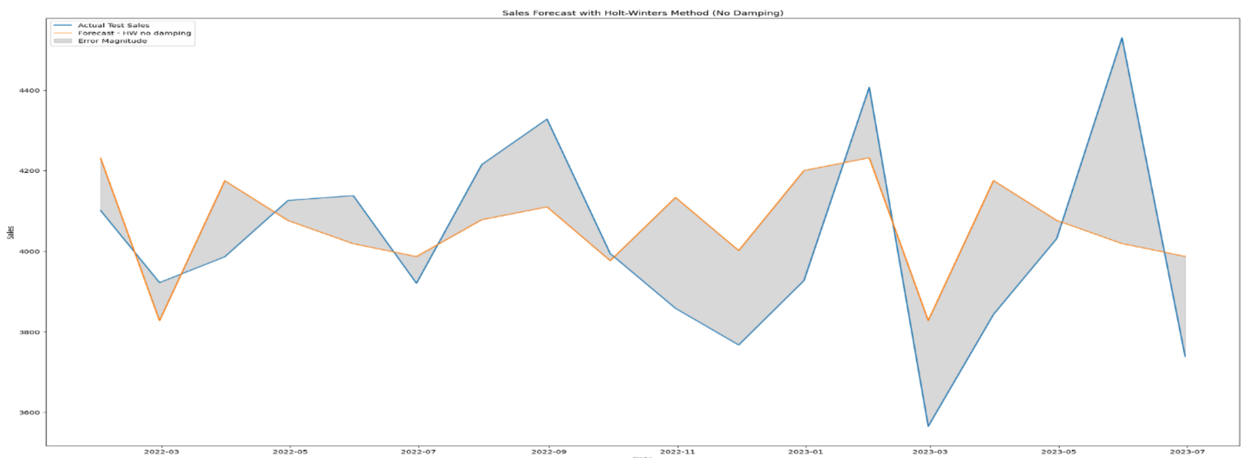


Figure 4. 55. ES-PM Forecast for Store Items (without Dumping).

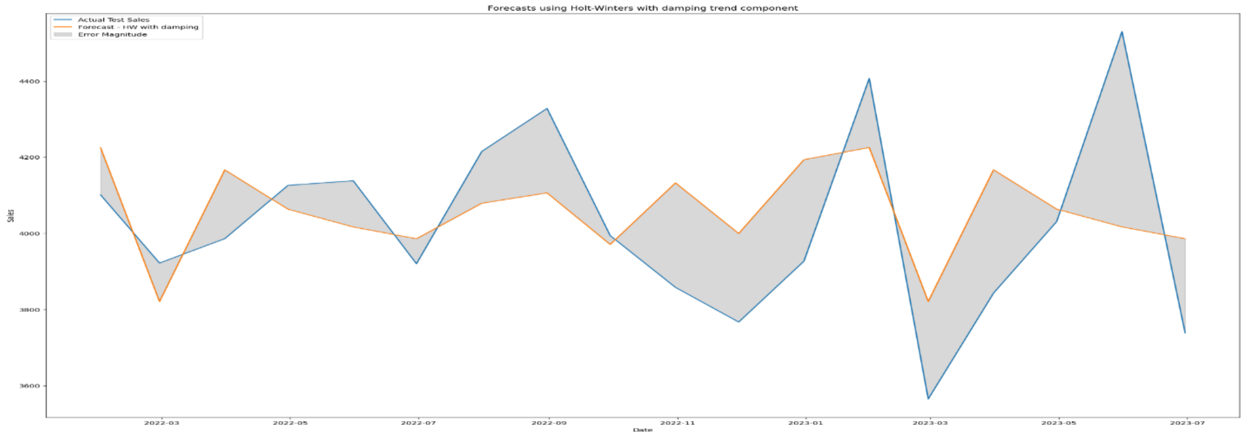


Figure 4. 56. ES-PM Forecast for Store Items (with Dumping).

	Metric	Value (No Damping)	Value (With Damping)
0	MAE	45.555732	46.831468
1	RMSE	58.535455	58.659223

```
Total Actual Sales: 10051.00
Total Predicted Sales (No Damping): 9755.23
Total Predicted Sales (With Damping): 10138.91
```

Figure 4. 57. ES-PM Sales & Metrics for Item One.

	Metric	Value (No Damping)	Value (With Damping)
0	MAE	51.238901	49.156190
1	RMSE	59.319073	59.035779

```
Total Actual Sales: 8732.00
Total Predicted Sales (No Damping): 8904.36
Total Predicted Sales (With Damping): 8794.24
```

Figure 4. 58. ES-PM Sales & Metrics for Item Two.

	Metric	Value (No Damping)	Value (With Damping)
0	MAE	49.899657	49.191046
1	RMSE	57.435192	56.712150

```
Total Actual Sales: 11113.00
Total Predicted Sales (No Damping): 11547.26
Total Predicted Sales (With Damping): 11528.36
```

Figure 4. 59. ES-PM Sales & Metrics for Item Three.

	Metric	Value (No Damping)	Value (With Damping)
0	MAE	74.804405	71.640678
1	RMSE	86.703808	81.855364

```
Total Actual Sales: 12794.00
Total Predicted Sales (No Damping): 13352.61
Total Predicted Sales (With Damping): 13107.39
```

Figure 4. 60. ES-PM Sales & Metrics for Item Four.

	Metric	Value (No Damping)	Value (With Damping)
0	MAE	61.014513	59.848144
1	RMSE	71.763426	71.794554

```
Total Actual Sales: 14390.00
Total Predicted Sales (No Damping): 14115.42
Total Predicted Sales (With Damping): 14086.00
```

Figure 4. 61. ES-PM Sales & Metrics for Item Five.

	Metric	Value (No Damping)	Value (With Damping)
0	MAE	90.251188	77.838703
1	RMSE	116.939997	109.297916

```
Total Actual Sales: 15316.00
Total Predicted Sales (No Damping): 16309.31
Total Predicted Sales (With Damping): 15765.82
```

Figure 4. 62. ES-PM Sales & Metrics for Item Six.

	Metric	Value (No Damping)	Value (With Damping)
0	MAE	187.913858	186.980323
1	RMSE	222.455342	221.017739

```
Total Actual Sales: 72396.00
Total Predicted Sales (No Damping): 73130.77
Total Predicted Sales (With Damping): 73038.94
```

Figure 4. 63. ES-PM Sales & Metrics for Store Items.

IV. Deep AR Results:

- **Plot Details:** Focuses on the sales units in the last 100 days. Weekends are highlighted to indicate patterns related to weekly cycles. Sales units are plotted with grid lines for clarity.
- **Visual Insights:** The graph provides a clear view of sales trends over the full period and the final 100 days. The distinction between training and test periods helps in understanding the data split used for model evaluation. Highlighting weekends aids in visualizing any cyclic patterns in sales data.
- **Graph Summary:** This visualization is crucial for understanding the sales behavior of a specific item in a particular store over time. It effectively sets the stage for further analysis and model training by illustrating how sales fluctuate and how the data is partitioned for training and testing purposes.

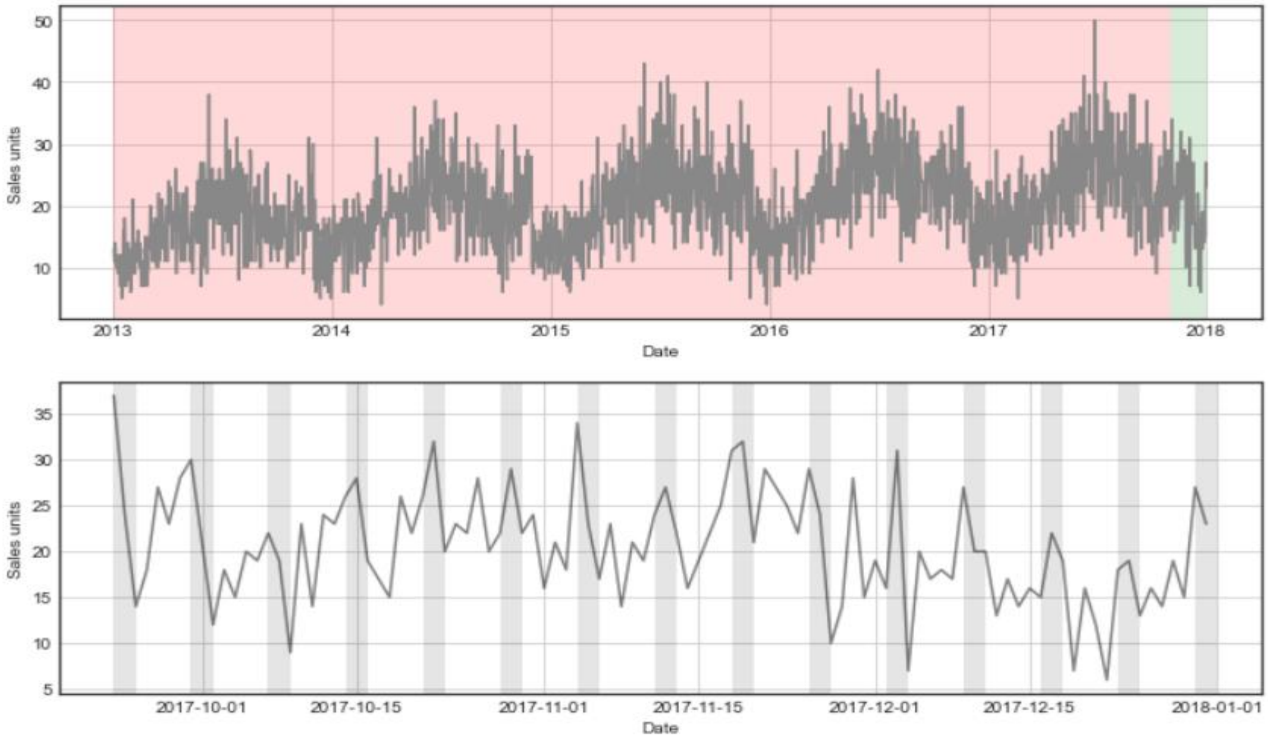


Figure 4. 64. Sales Units in The Last 100 Days (New Dataset)

- **Visual Insights:** These visualizations illustrate the observed sales along with the forecasted sales and prediction intervals for different items in a specific store. The observed sales data and the forecast median are closely aligned, indicating the model's accuracy. The prediction intervals provide a range within which future sales are expected to fall, demonstrating the model's confidence in its predictions.
- **Graph Summary:** These plots effectively communicate the performance of the Deep AR model in forecasting sales for different items. By showing both the observed data and forecast intervals, the visualizations highlight the model's accuracy and reliability in predicting future sales.

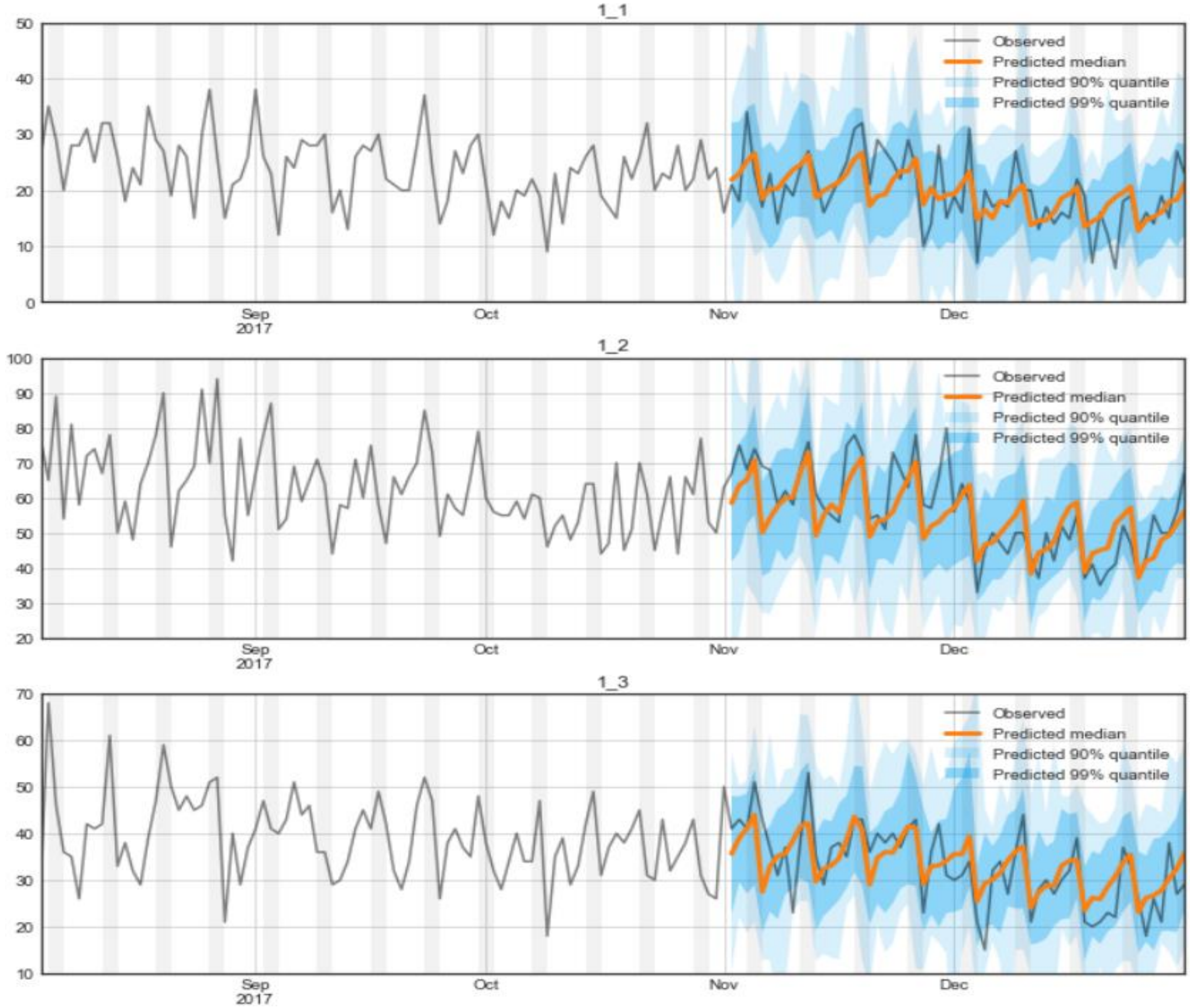


Figure 4. 65. Deep AR Prediction Plot.

V. Deep Learning Models Results:

- **XGboost Model:** for **Visual Insights** The predictions closely follow the real values, indicating the model's accuracy. Minor deviations between predicted and actual values are visible, which is expected in real-world scenarios. And for the **Graph Summary** This plot effectively illustrates the XGBoost model's performance in forecasting sales, demonstrating its high accuracy and alignment with actual sales data.

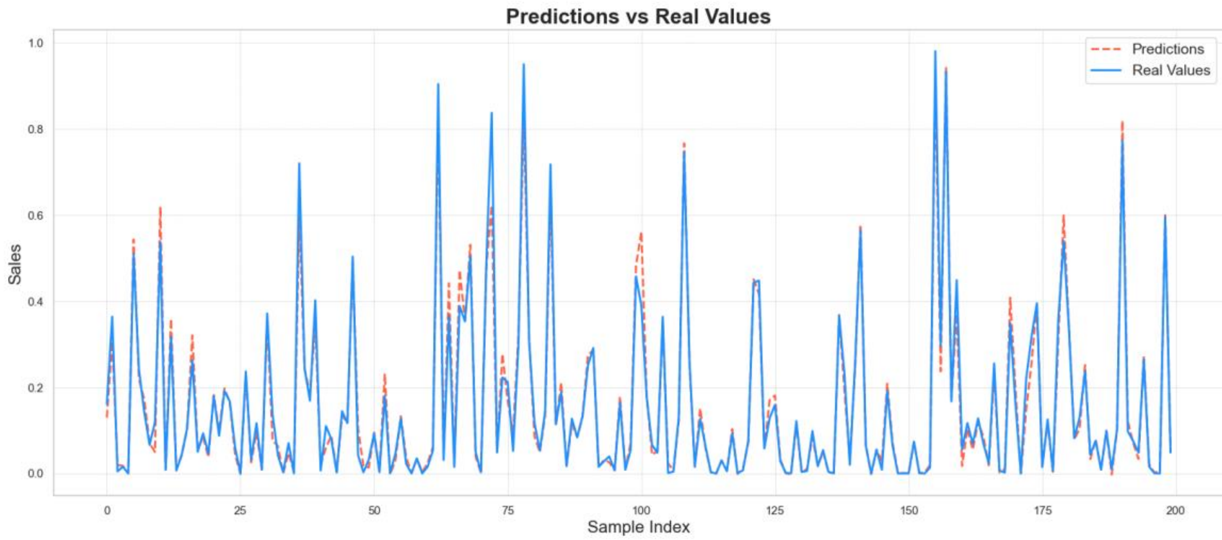


Figure 4. 66. Deep Learning Prediction plot.

- **Custom Deep Learning Neural Network Model:** for the **Visual Insights** The plot demonstrates that the custom deep learning model's predictions are fairly accurate, though there are more noticeable deviations compared to the XGBoost model. The overall trend is captured well, indicating the model's ability to learn from the data. And for the **Graph Summary** This visualization highlights the custom deep learning model's performance, showing its strengths in capturing sales trends but also indicating areas for improvement compared to the XGBoost model.

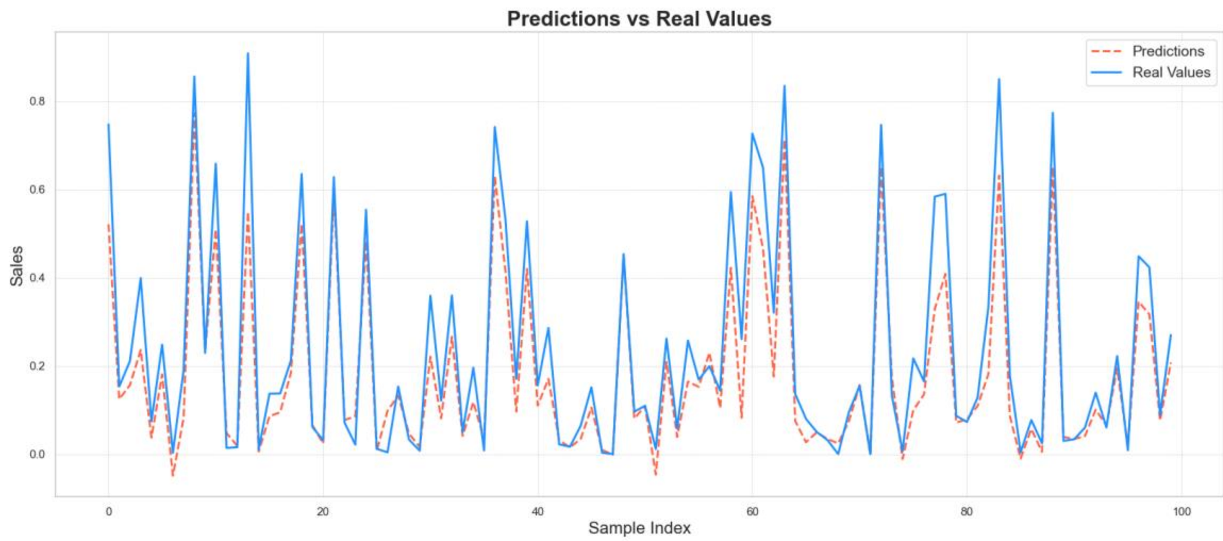


Figure 4. 67. Customs Deep Learning Prediction Plot.

Summary:

the journey of refining sales forecasting models unfolded through strategic Model Modification and Dataset Manipulation. The Model Modification phase witnessed deliberate updates and alterations to simplify, enhance efficiency, and improve overall effectiveness. On the other front, Dataset Manipulation involves the meticulous division of the dataset into individual files and nuanced changes in sales date intervals. This intentional manipulation allowed for a profound exploration of how these alterations influenced the models' responses and, subsequently, the forecast outcomes. The results of each model encapsulated in plots, serve as insightful visual narratives, explaining the intricate dynamics of their performance. This dual-pronged approach, comprising model refinement and dataset manipulation, forms the crux of a comprehensive strategy to elevate the precision and adaptability of sales forecasting endeavors.

Chapter Five

Results

This section marks the culmination of an intricate journey through sales forecasting, where the efficacy and nuances of three prominent models are laid bare. Within these findings, a comprehensive comparison table meticulously dissects the performance of each model, offering a visual roadmap for understanding their strengths and limitations. Beyond mere numbers, the exploration delves into the metrics employed in each model, providing a deeper understanding of their predictive capabilities. Furthermore, this section ventures into the realm of innovation, hinting at the technological advancements poised to shape the next project. As we unravel the outcomes and insights, the Results section serves as a critical juncture, unveiling the transformative power of data-driven decision-making in the realm of sales forecasting.

In this chapter, we will be looking at:

1. **Comparison table for the first three primary models.**
2. **Comparing Deep AR and Deep Learning Models.**
3. **What is next?**
4. **References.**

Comparison table for the first three primary models:

This table compare the forecasted sales from the three primary models Linear Regression, ARIMA, and Exponential Smoothing.

item	Linear regression	Arima	Exponential smoothing
One	▪ **Total Actual Sales:** 10051 ▪ **Total Predicted Sales:** 10169	▪ **Total Actual Sales:** 10051 ▪ **Total Predicted Sales:** 8372	▪ **Total Actual Sales:** 10051 ▪ **Total Predicted Sales (No Damping):** 9755 ▪ **Total Predicted Sales (With Damping):** 10138
Two	▪ **Total Actual Sales:** 8732 ▪ **Total Predicted Sales:** 8664	▪ **Total Actual Sales:** 8732 ▪ **Total Predicted Sales:** 10073	▪ **Total Actual Sales:** 8732 ▪ **Total Predicted Sales (No Damping):** 8904 ▪ **Total Predicted Sales (With Damping):** 8794
Three	▪ **Total Actual Sales:** 11113 ▪ **Total Predicted Sales:** 11574	▪ **Total Actual Sales:** 11113 ▪ **Total Predicted Sales:** 15716	▪ **Total Actual Sales:** 11113 ▪ **Total Predicted Sales (No Damping):** 11547 ▪ **Total Predicted Sales (With Damping):** 11528
Four	▪ **Total Actual Sales:** 12794 ▪ **Total Predicted Sales:** 13067	▪ **Total Actual Sales:** 12794 ▪ **Total Predicted Sales:** 11792	▪ **Total Actual Sales:** 12794 ▪ **Total Predicted Sales (No Damping):** 13352 ▪ **Total Predicted Sales (With Damping):** 13107

Five	▪ **Total Actual Sales:** 14390 ▪ **Total Predicted Sales:** 14165	▪ **Total Actual Sales:** 14390 ▪ **Total Predicted Sales:** 10912	▪ **Total Actual Sales:** 14390 ▪ **Total Predicted Sales (No Damping):** 14115 ▪ **Total Predicted Sales (With Damping):** 14086
six	▪ **Total Actual Sales:** 15316 ▪ **Total Predicted Sales:** 15603	▪ **Total Actual Sales:** 15316 ▪ **Total Predicted Sales:** 17463	▪ **Total Actual Sales:** 15316 ▪ **Total Predicted Sales (No Damping):** 16309 ▪ **Total Predicted Sales (With Damping):** 15765
Store items	▪ **Total Actual Sales:** 72396 ▪ **Total Predicted Sales:** 73245	▪ **Total Actual Sales:** 72396 ▪ **Total Predicted Sales:** 72839	▪ **Total Actual Sales:** 72396 ▪ **Total Predicted Sales (No Damping):** 73130 ▪ **Total Predicted Sales (With Damping):** 73038

Table 5. 1. Comparing Primary Models Results.

From the provided table showcasing the results of three different forecasting models (Linear Regression, ARIMA, Exponential Smoothing) across various store items, several key observations can be made:

- **I.** Across all items, Linear Regression and Exponential Smoothing tend to have predictions closer to actual sales, while ARIMA shows larger variations.
- **II.** The total actual sales and total predicted sales for each model are summarized at the store level.

These observations highlight the importance of selecting the appropriate forecasting model, as different models exhibit varying levels of accuracy in predicting sales for different items.

Comparing Deep AR and Deep Learning Models:

- **I. Deep Learning Models:** The XGBoost model achieved a high R2 score of 0.9726, indicating strong predictive performance. The MSE and RMSE were low (0.001237 and 0.0352, respectively), showing accurate predictions. This document focused on multiple models, with detailed evaluation metrics and various data visualization techniques to support the findings.
- **II. Deep AR**: The Deep AR model focused on time-series forecasting and provided extensive coverage metrics, with a MAPE of 0.155 and good coverage across different quantiles (0.881, 0.937, and 0.982). The training was done over 5 epochs, and detailed visualizations of the forecasts were provided to illustrate the model's performance.

Overall, Document 1 showcased a variety of models and techniques with detailed metrics for model performance, while Document 2 provided an in-depth look at the Deep AR model specifically for demand forecasting, emphasizing time-series prediction and coverage metrics.

What is next?

1. **Development of Interactive Dashboards:**
   To make our forecasting results more accessible and interactive, we plan to create user-friendly dashboards using tools like Dash or Streamlit. These dashboards will enable users to visualize data trends, adjust forecasting parameters, and better understand the impact of different variables on sales predictions.

2. **Automation of Data Handling Steps:**
   We aim to streamline the initial stages of our forecasting process by automating data preparation, preprocessing, and exploratory analysis. This step will reduce manual intervention, minimize errors, and ensure the data quality is consistent, allowing for more reliable forecasts.

3. **Development of an API for Model Deployment:**
   To integrate our forecasting model into broader business systems, we will develop a RESTful API. This will allow external applications to interact with our model directly, facilitating automated and real-time forecasting capabilities.

Implementation Overview

- **Interactive Dashboards:** Design a simple and intuitive interface that displays key forecasting outputs and allows parameter adjustments.
- **Automating Data Processes:** Create scripts to handle common data cleaning and analysis tasks automatically.
- **API for Model Deployment:** Set up a basic API structure that supports data submissions and fetching of forecast results.

Project Enhancement Goals

These enhancements are designed to improve the practical utility of the project without aiming for large-scale changes. By implementing these features, the project will gain modest but significant improvements in user interaction, data handling efficiency, and integration capabilities, making it a more robust tool for forecasting needs.

References:

1. **Data Preparation for Data Mining by Dorian Pyle -** Using and Preparing Datasets for Forecasting.
2. **Data Science for Business by Foster Provost and Tom Fawcett -** Data and Dataset.
   Evaluating Machine Learning Models by Alice Zheng - Model Evaluation Metrics.
3. **Forecasting in Retail and FMCG by Paul Goodwin -** Store Sales Forecasting.
4. **Forecasting: Principles and Practice by Rob J Hyndman and George Athanasopoulos -** Holt-Winters Method.
5. **Introduction to Time Series Analysis and Forecasting by Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci -** Time Series Analysis.
6. **Machine Learning for Time Series Forecasting with Python by Francesca Lazzeri -** Machine Learning.
7. **Pattern Recognition and Machine Learning by Christopher M. Bishop -** Model Modification.
8. **Python Data Science Handbook by Jake VanderPlas -** Python, Anaconda, and Jupyter Notebook.
9. **Python Machine Learning by Sebastian Raschka and Vahid Mirjalili -** Machine Learning.
10. **Retail Sales Forecasting at Walmart from Walmart Corporate -** The Aim of Store Sales Forecasting.
11. **Sales Forecasting: A Practical Guide by Clive Miller -** Sales Forecasting Models.
12. **The Data Warehouse Toolkit by Ralph Kimball and Margy Ross -** Data Warehouse.
13. **The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani, and Jerome Friedman -** Model Selection.
14. **The Visual Display of Quantitative Information by Edward R. Tufte -** Dataset Visualization.
15. **Time Series Analysis: Forecasting and Control by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung -** ARIMA Metrics.
16. **Case studies from the MIT Sloan Management Review -** Insightful case studies on forecasting models.
17. **Demand Forecasting in Retail: Best Practices & Pitfalls by Harvard Business Review -** Pros and Cons of Store Sales Forecasting.
18. **Kaggle's own tutorials and competition case studies -** Kaggle.
19. **Deep AR: Probabilistic Forecasting with Autoregressive Recurrent Networks -** An in-depth look at Deep AR algorithm and its applications.
20. **Amazon SageMaker Deep AR now supports missing values, categorical and time series features, and generalized frequencies -** Enhancements and new features in Deep AR.
21. **Comparison Analysis of Facebook's Prophet, Amazon's Deep AR+ and CNN-QR Algorithms for Successful Real-World Sales Forecasting -** Comparative study of forecasting models.
22. **Robust Sales Forecasting Using Deep Learning with Static and Dynamic Covariates -** Techniques for improving forecast accuracy with Deep AR.
23. **Advanced Techniques in Time Series Forecasting with Deep AR -** Cutting-edge methodologies in Deep AR for time series forecasting.