**AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE**

**23**

**AIE232M PYTHON FOR AI**

**END SEMESTER REVIEW REPORT**

Marks

Faculty In-charge



PREDICTING FUTURE SALES DATA OF A RETAIL STORE BASED ON HISTORICAL DATA

A THESIS

Submitted by

Baddela Thej Roy-CB.EN.U4AEE23011

Mukuntha V- CB.EN.U4AEE23031

Saurodiip Choudhary- CB.EN.U4AEE23041

Shanu VS- CB.EN.U4AEE23044

Santhosh Kumar- CB.EN.U4AEE23051

Contents

Acknowledgement 3

List of Figures 3

Abstract 6

**1 Introduction**

* 1. Literature Survey 7
  2. Problem statement 7

1.3 Objectives 8

**2 Background 9**

**3 Proposed Work 11**

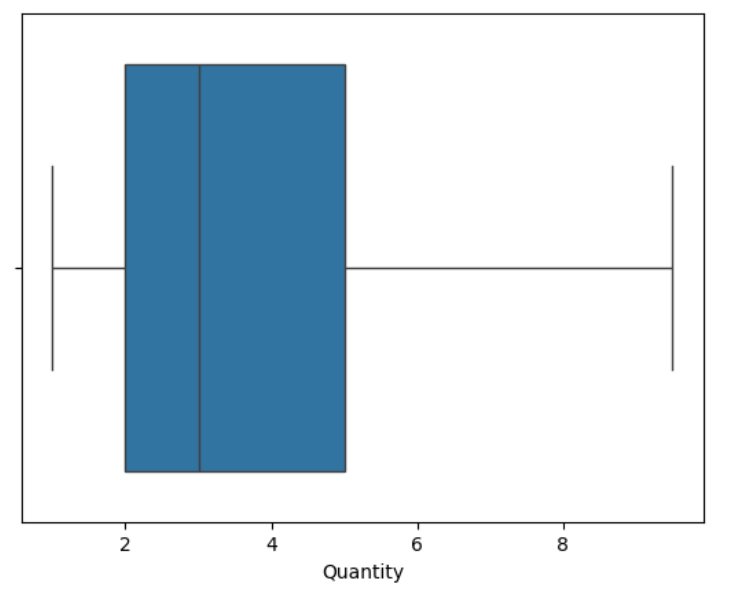
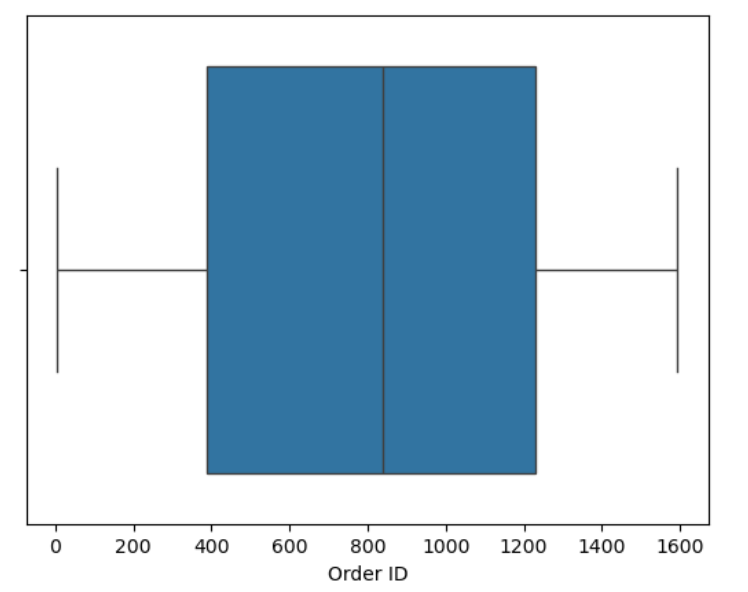
**4 Conclusion 11**

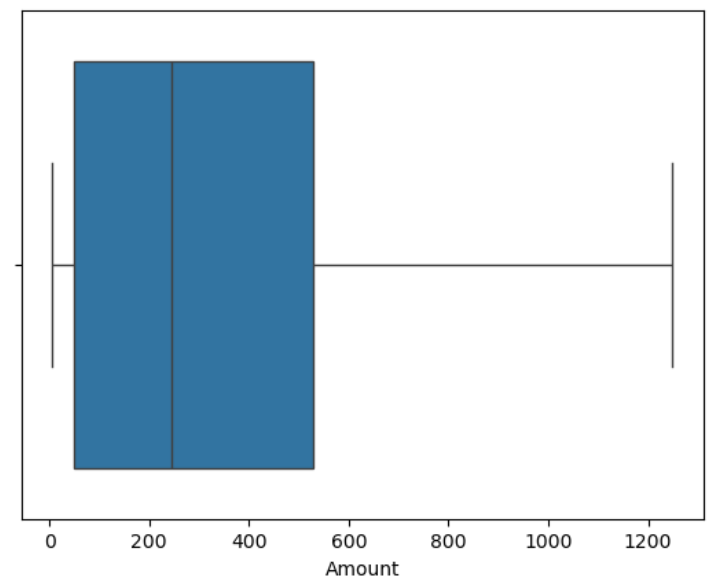
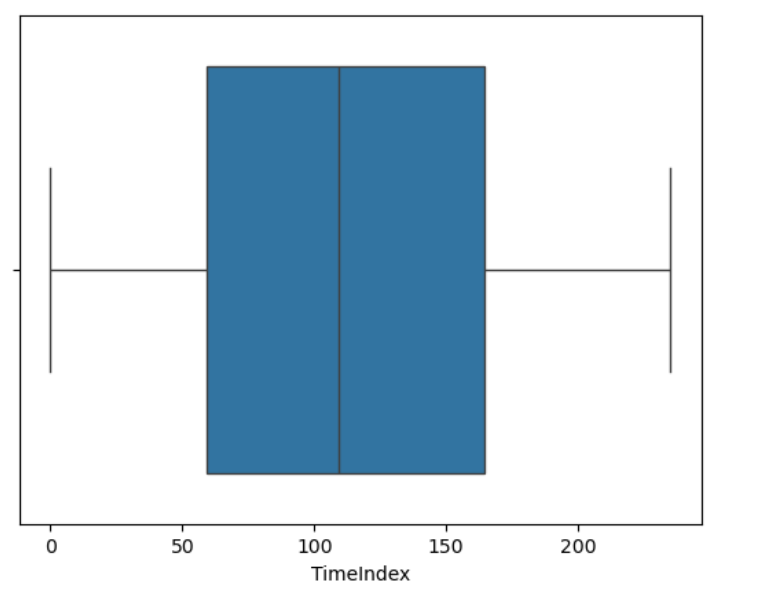
**References**

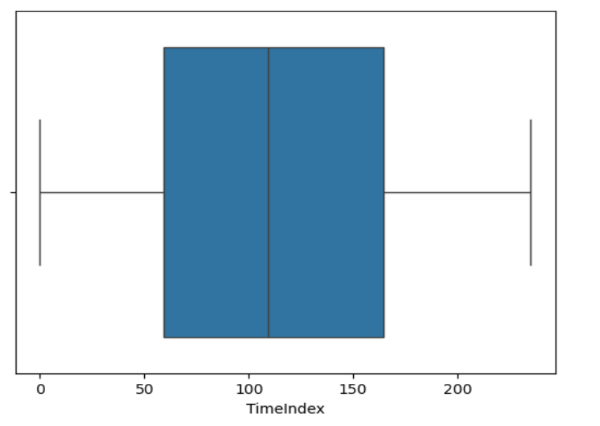
Acknowledgement

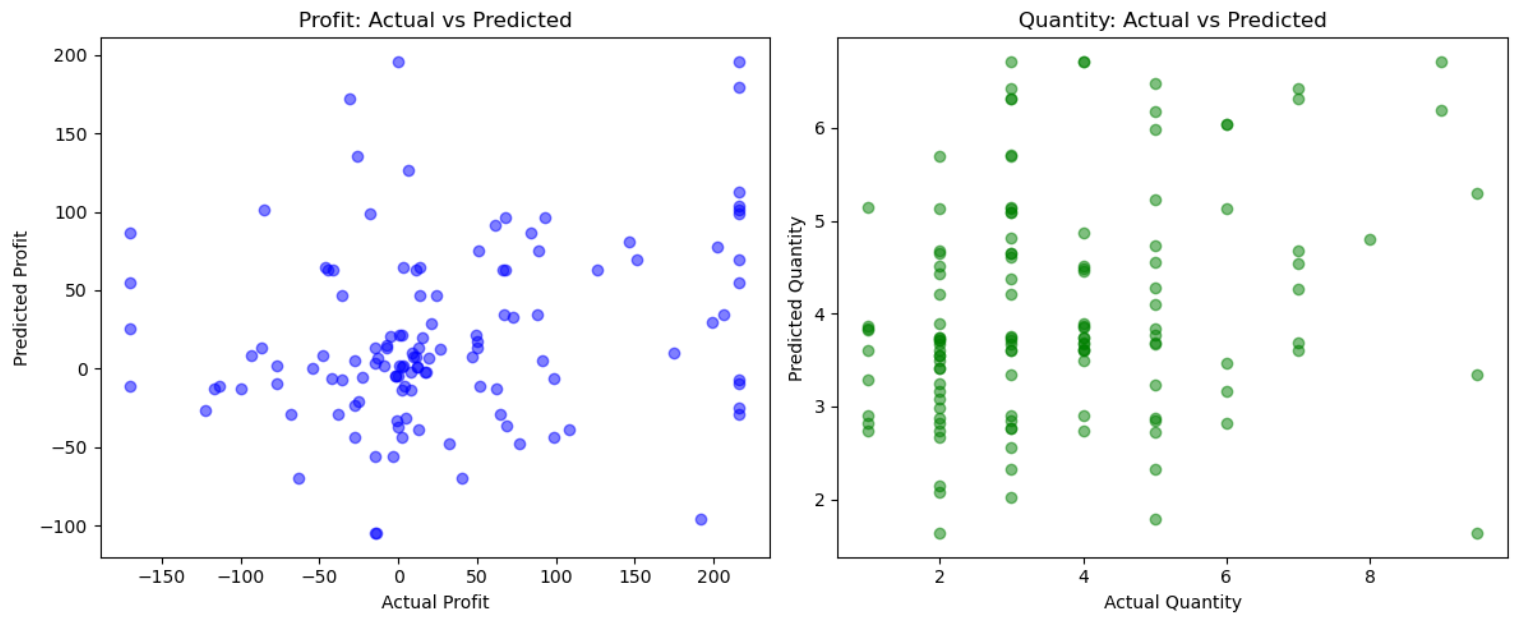
We would like to express our sincere gratitude to our guide, Yadukrishnan Sir and Chandni Mam for their valuable guidance, support, and encouragement throughout the duration of this project. Their insights and continuous feedback played a crucial role in shaping this work.

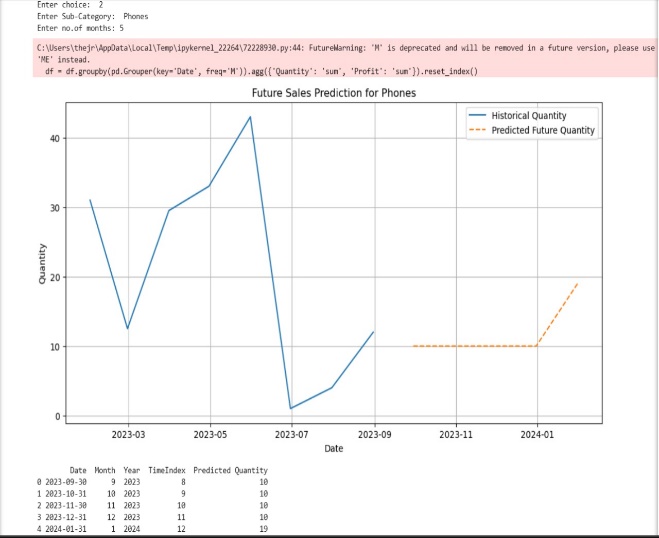
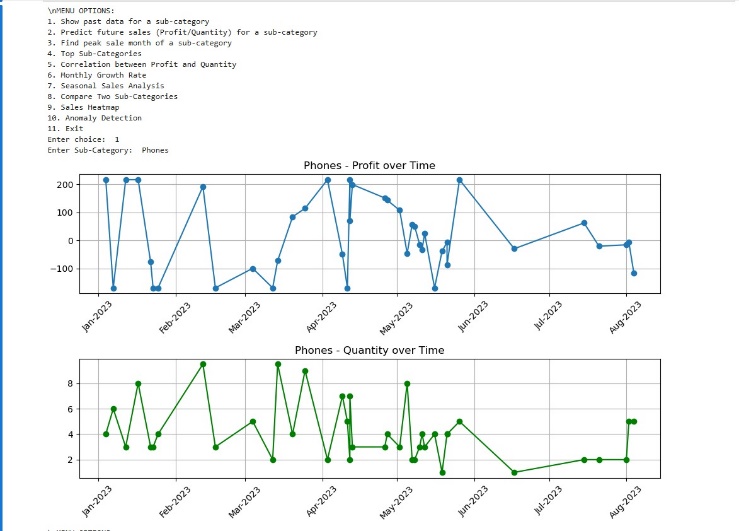
List of Figures









Abstract

This project aims to predict future sales of a retail store using machine learning techniques. The workflow begins with exploratory data analysis (EDA) to understand patterns and detect anomalies within the dataset. Outliers were identified and handled to ensure data integrity, followed by feature engineering steps including one-hot encoding of categorical variables. A Random Forest Regressor (RFR) model was employed for training and prediction. After fitting the model, various visual and analytical techniques were applied to evaluate and interpret the results. The approach demonstrates how machine learning can be effectively leveraged for retail forecasting using a refined and well-preprocessed dataset.

Chapter 1

Introduction

* 1. Literature Survey

Retail forecasting is one of the most common and practical applications of both time-series analysis and machine learning. Traditional models like ARIMA are good for simple trends, but they often struggle with real-world retail data that includes seasonal effects and multiple categories. That’s where machine learning models like **Random Forests** come in — they can handle complex patterns and give better predictions. From what I’ve studied and observed, ensemble models like these are very effective in learning from historical data and improving sales forecasts, especially when the data is noisy or inconsistent.

1.2 Problem statement

To predict future sales quantities and profits for a retail store, based on past trends and patterns in sub-category level transactional data. This includes identifying peak seasons, trends, and enabling business intelligence through predictive modeling.

* 1. Objectives
* Perform data cleaning and preprocessing on retail sales data.
* Explore and visualize sales patterns and outliers.
* Train machine learning models (Random Forest Regressor) on historical data.
* Predict future sales quantities and profits per sub-category.
* Provide user-interactive features to explore and compare performance metrics.

Chapter 2

Background

Retail sales forecasting is a critical aspect of modern business strategy, helping companies optimize inventory, reduce costs, and meet customer demand effectively. With the growing complexity of consumer behavior and seasonal trends, traditional forecasting methods often struggle to deliver accurate predictions. This difficulty arises from several inherent challenges within retail forecasting.

Machine learning offers a powerful alternative by uncovering patterns in historical data that may not be obvious through conventional analysis. Techniques like Random Forest Regression have proven effective in capturing nonlinear relationships and handling categorical data — making them ideal for retail environments. The ability of machine learning to discern complex patterns is essential for effective forecasting.

By integrating machine learning into forecasting, retailers gain a competitive edge through smarter, data-driven decisions, which is why this domain continues to attract significant attention in both industry and research. This highlights the importance of advancing forecasting methodologies.

Retail forecasting poses several challenges, including high variance in customer demand, complex categorical features, and strong seasonal patterns that are difficult to model. Additionally, real-world data often contains noise and outliers, which can mislead traditional forecasting methods. Addressing these issues requires robust models and careful feature engineering to ensure reliable predictions.

This project employs a combination of core machine learning techniques to model and predict retail sales. The foundation is built upon supervised learning, where the model is trained using labeled historical data to predict future outcomes — in this case, profit and quantity sold. At the core is the Random Forest Regressor, an algorithm that constructs multiple decision trees during training and outputs the average prediction of all trees. It is known for its robustness, ability to handle high-dimensional data, and resistance to overfitting — making it well-suited for real-world retail datasets. Additionally, one-hot encoding was applied to transform categorical variables like Sub-Category into a machine-readable format without introducing bias, allowing the model to treat each category independently. Together, these techniques create a powerful and interpretable pipeline capable of capturing complex retail dynamics and producing accurate sales forecasts.

Chapter 3

Proposed Work

This chapter details the methodology and implementation of the retail sales analysis and forecasting system. The system provides functionalities for data exploration, sales prediction, and in-depth analysis of sales patterns, all accessible through an interactive menu-driven interface.

The system is designed around a modular architecture, encompassing data processing, predictive modeling, and analytical functions. The core components include:

* **Data Preprocessing and Feature Engineering:** Functions to clean, transform, and prepare the data for analysis and modeling.
* **Predictive Modeling:** Implementation of Random Forest Regression for sales forecasting.
* **Analytical Functions:** A suite of functions to perform various analyses, including trend analysis, seasonality analysis, correlation analysis, and anomaly detection.
* **User Interface:** A menu-driven interface enabling users to interact with the system and access its functionalities.
* **Functional Components**
* The system's functionality is organized into the following key components:
* **3.2.1 Data Exploration and Visualization**
* **Historical Data Display:** The system allows users to visualize historical sales data (Profit and Quantity) for a given sub-category over time. This is achieved through the main() function (option 1), which filters data by sub-category and generates line plots of Profit and Quantity against Date. This visualization helps in understanding past performance and identifying trends.
* **Top Sub-Category Analysis:** The top\_subcategories() function identifies and visualizes the top N sub-categories based on a chosen metric (Profit or Quantity). This function groups data by 'Sub-Category', calculates the mean of the specified metric, and presents the top N sub-categories in both tabular and bar chart formats. This enables quick identification of best-performing categories.
* **Sales Heatmap:** The sales\_heatmap() function generates a heatmap visualization of sales (Profit or Quantity) by sub-category and month. This heatmap provides a comprehensive overview of sales patterns across different categories and time periods, facilitating the identification of peak sales seasons and category-specific trends.
* **3.2.2 Predictive Modeling**
* **Future Sales Prediction:** The system implements a Random Forest Regression model to predict future sales (Quantity). The predict\_subcategory\_future\_sales() function performs the following steps:
* Filters data for a specific sub-category and aggregates it monthly.
* Performs feature engineering, creating 'Month', 'Year', and 'TimeIndex' features.
* Trains a Random Forest Regressor model on the historical data.
* Generates future dates and corresponding feature values for the prediction period.
* Predicts future sales using the trained model.
* Visualizes historical and predicted sales in a plot and exports the predictions to a CSV file using the export\_prediction function.
* **3.2.3 Analytical Functions**
* **Correlation Analysis:** The correlation\_profit\_quantity() function calculates the Pearson correlation coefficient between Profit and Quantity for a given sub-category. It also generates a scatter plot to visualize the relationship. This analysis helps understand how sales volume is related to profitability within specific categories.
* **Monthly Growth Rate Analysis:** The monthly\_growth() function calculates and visualizes the monthly growth rate of a specified metric (Profit or Quantity) for a given sub-category. This analysis helps in identifying growth trends and potential areas of concern.
* **Seasonal Analysis:** The seasonal\_analysis() function analyzes the average sales (Profit or Quantity) per month for a given sub-category. It generates a bar chart to visualize seasonal patterns, helping businesses understand and prepare for recurring seasonal fluctuations.
* **Sub-Category Comparison:** The compare\_subcategories() function compares the sales (Profit or Quantity) of two sub-categories over time. It generates a line plot to visualize the trends and differences in performance between the categories.
* **Anomaly Detection:** The detect\_anomalies() function identifies anomalies in sales data (Profit or Quantity) for a given sub-category. It calculates the mean and standard deviation of the metric and flags data points that fall outside a specified range (mean ± 2\*standard deviation). This helps in detecting unusual sales patterns that may require further investigation.
* **3.2.4 User Interface**
* **Menu-Driven Interface:** The main() function provides a menu-driven interface that allows users to access all the system's functionalities. Users can select options to display historical data, predict future sales, analyze trends, and perform other analyses. The interface guides the user through each function, prompting for necessary inputs and displaying the results.
*  The system is implemented using Python and leverages libraries such as Pandas for data manipulation, Scikit-learn for machine learning (Random Forest Regression), Matplotlib and Seaborn for data visualization, and dateutil.relativedelta for date calculations.
*  The udf variable likely represents a Pandas DataFrame containing the retail sales data, with columns such as 'Date', 'Sub-Category', 'Quantity', and 'Profit'.
*  Date handling and manipulation are performed using Pandas' datetime functionality and the matplotlib.dates module for formatting date axes in plots.

Chapter 4

Conclusion

This work demonstrated a practical application of machine learning in retail analytics. Through feature engineering and ensemble regression models, it was possible to forecast future sales with reasonable accuracy. The system allows users to interact with the data, generate predictions, and gain actionable insights, thereby supporting strategic decision-making in retail environments.

Future improvements could include:

* Using deep learning models for more complex time patterns.
* Integrating real-time dashboards or stream-based input.
* Applying optimization to inventory or pricing strategies.

References

<https://youtu.be/0B5eIE_1vpU?si=Is_oRfBB88IsbehQ>

<https://chatgpt.com/share/68034146-f45c-800f-a657-3363541d7994>

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>

<https://youtube.com/playlist?list=PLjDarHRb40ZNr2IjP2ZpR5KG1_bDiB8bF&si=bQPPGgTKl_kw4aiv>

<https://youtu.be/al67gXF9cS0?si=tam9wWROZ4RyPwX_>

<https://youtu.be/Dvgjl1ZAfw4?si=GkVnrnmotmJbFuv0>