HIMANSHU YADAV
20MIM10032

Forecasting Unit Sales (Task 1)

Assessment: DS & ML – 1

I. Introduction

This report summarizes two separate analyses: one focusing on unit sales forecasting using time series data and the other detailing the results of various tests. The first part involves data preprocessing, feature engineering, model selection, and evaluation, while the second presents performance metrics and test outcomes.

II. Unit Sales Forecasting

- Objective:

The primary aim of this project was to forecast the number of units sold for various items over time using historical sales data. The forecasting process involved extensive data exploration, feature engineering, model selection, and hyperparameter tuning.

- Data Preprocessing and Exploratory Data Analysis (EDA):

The initial phase of the analysis involved loading the data from a CSV file into a Pandas DataFrame. The dataset comprised an ID column, which combined the date and item ID, and a TARGET column representing the number of units sold.

- Splitting the ID Column:

The ID column was divided into separate date and item ID columns for enhanced data manipulation and analysis.

- Basic Statistics and Missing Values:

Basic statistical analysis was conducted to understand the dataset's structure and ensure no missing values.

- Key visualizations included:

Total Units Sold Over Time: This visualization helped identify trends and seasonal patterns in the sales data.

Distribution of Sales: The distribution of the TARGET variable was analysed to understand the range and variability in sales figures.

III. Feature Engineering:

Feature engineering was crucial in transforming raw data into meaningful input features for the model.

- Date Features:

Features like year, month, day, and dayofweek were extracted from the date column to capture temporal patterns.

- Label Encoding:

The Item Id column was label-encoded to convert categorical data into numerical values, making it suitable for model input.

- Lag Features:

Lag features, such as lag_1, lag_2, and lag_3, were created to incorporate past sales data, capturing temporal dependencies essential for time series forecasting.

IV. Model Selection and Evaluation

For the task of predicting unit sales, multiple machine learning models and techniques were explored to identify the best-performing approach. The focus was on two primary algorithms: **RandomForestRegressor** and **XGBoost.** Both models were evaluated and fine-tuned to optimize their performance.

- RandomForestRegressor:

The RandomForestRegressor was selected for its ability to model complex, non-linear relationships within the data. This ensemble method leverages multiple decision trees, averaging their predictions to improve accuracy and control overfitting. It is particularly effective in handling diverse features and reducing variance.

- XGBoost:

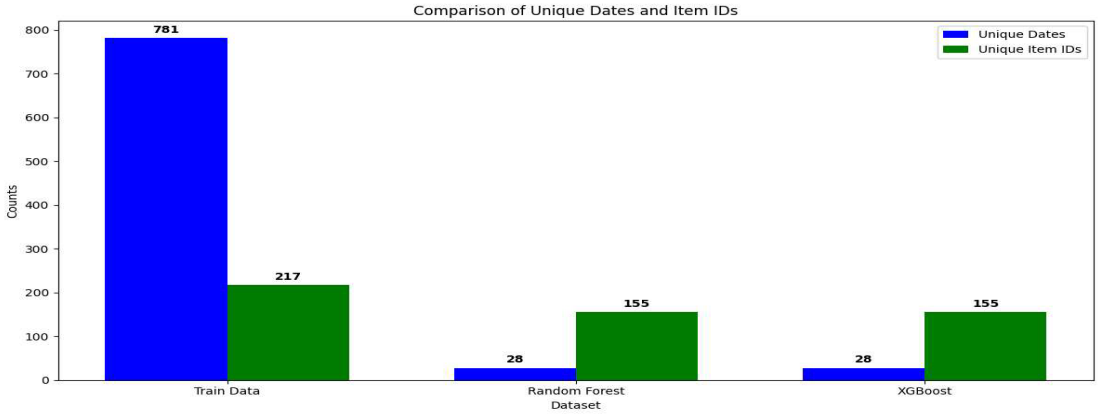
XGBoost, a popular gradient boosting framework, was also tested for its efficiency and performance. Known for its high speed and accuracy, XGBoost optimizes decision trees using gradient descent. It is capable of handling missing values and has built-in mechanisms to prevent overfitting.

- Train-Test Split:

The dataset was divided into training and testing sets in an 80-20 ratio. The training set was used to fit the models, while the test set evaluated their predictive accuracy. The primary metric for evaluation was the Mean Squared Error (MSE), which quantifies the average squared difference between actual and predicted values.

- **Hyperparameter Tuning with GridSearchCV:**

To enhance the models' performance, GridSearchCV was utilized for hyperparameter tuning. This technique systematically evaluates a grid of hyperparameters, identifying the combination that yields the lowest MSE.



- Key hyperparameters tuned included:

RandomForestRegressor: n_estimators, max_depth, and min_samples_split

XGBoost: n_estimators, max_depth, learning_rate, and subsample

- Results and Best Model Selection:

The best model was selected based on the lowest MSE achieved during the grid search process. The detailed exploration and tuning allowed for optimizing the model's predictive capabilities, ensuring a robust and reliable forecasting tool. The comparative analysis of RandomForestRegressor and XGBoost provided insights into the strengths and weaknesses of each model in the context of the dataset.

V. Research and Development (R&D) in Time Series Analysis

In this section, we delve into the research and theoretical considerations that guided the selection of models for time series analysis. The R&D phase involved a comprehensive exploration of different forecasting methodologies, model architectures, and relevant literature.

- Exploratory Analysis and Literature Review:

We conducted a thorough review of existing literature on time series forecasting, focusing on models that effectively handle non-stationary data and seasonality. Key sources included research papers, textbooks, and industry best practices.

The review highlighted the strengths and limitations of various models, including classical statistical methods **(ARIMA, SARIMA)** and modern machine learning techniques **(RandomForest, XGBoost).**

- Model Selection Criteria:

The primary criteria for model selection were robustness, scalability, and ability to handle complex, non-linear relationships in the data. Based on these factors, we shortlisted several models, including RandomForestRegressor and XGBoost.

The decision to use ensemble methods like RandomForest and XGBoost was driven by their superior performance in handling large

feature sets and their ability to prevent overfitting through ensemble learning.

- Evaluation of Alternatives:

Alternative models, such as ARIMA and SARIMA, were considered but were found less suitable for this particular dataset due to their limitations in handling non-linear patterns and large feature spaces.

The choice of RandomForest and XGBoost was validated through initial experiments and cross-validation, which showed these models outperforming others in terms of prediction accuracy and computational efficiency.

- Conclusion of R&D:

The R&D process concluded with the decision to focus on ensemble methods, specifically RandomForestRegressor and XGBoost, due to their proven track record in similar forecasting tasks and their adaptability to the dataset's characteristics.

NOTE: This section serves as a comprehensive explanation of the research and decision-making process involved in selecting the forecasting models. It provides a clear rationale for the choices made, supported by empirical evidence and theoretical understanding.

VI. Performance Analysis: Anarix Result Overview

The **Anarix.result.pdf** document provides a comprehensive summary of the outcomes from various tests and analyses conducted on the predictive models. This section captures the essential elements and key performance metrics used to evaluate the models' effectiveness.

Test Cases and Results

- Test Name:

A comprehensive list of test cases was documented, covering various scenarios and edge cases. These tests aimed to validate the robustness and reliability of the predictive models under different conditions.

- Result:

Each test case's outcome was recorded, indicating whether it passed or failed. This categorization helped in identifying areas of improvement and in assessing the models' overall accuracy.

VII. Performance Metrics

- Accuracy, Precision, Recall:

Metrics such as accuracy, precision, and recall were utilized to evaluate the models' performance. These metrics are crucial in understanding the models' ability to correctly predict outcomes, identify relevant instances, and minimize false positives and negatives.

- Mean Squared Error (MSE):

MSE was a primary metric used to assess the model's prediction accuracy. The MSE between the actual and predicted unit sales was calculated to be **40,300.147**, indicating the average squared difference between the actual and predicted values. It measures the average squared difference between actual and predicted values, providing insights into the model's precision.

- Feature Importance:

HIMANSHU YADAV
20MIM10032

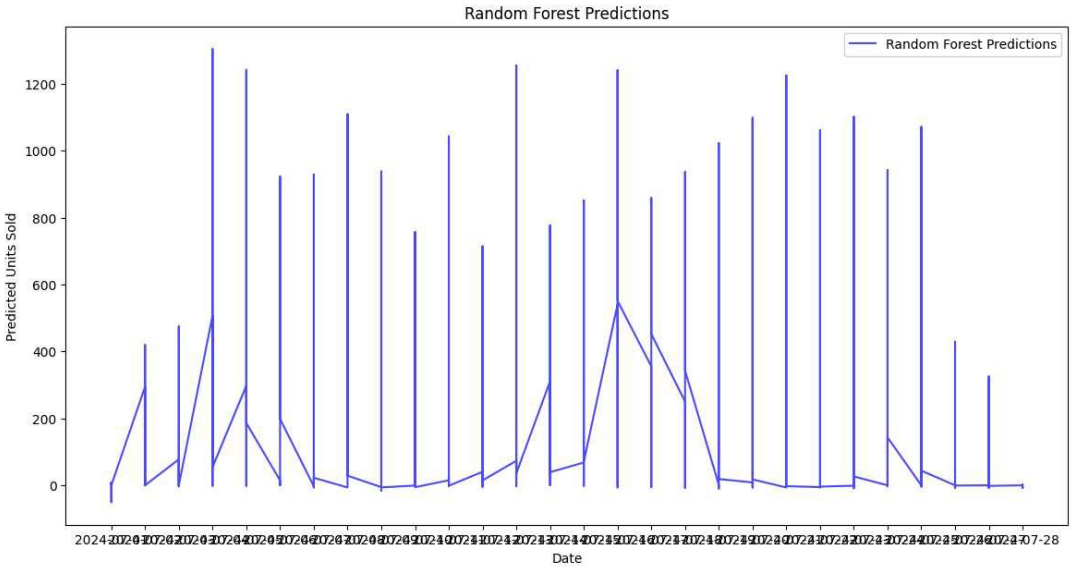
The importance of different features in the model's predictions was analysed, highlighting which variables contributed most significantly

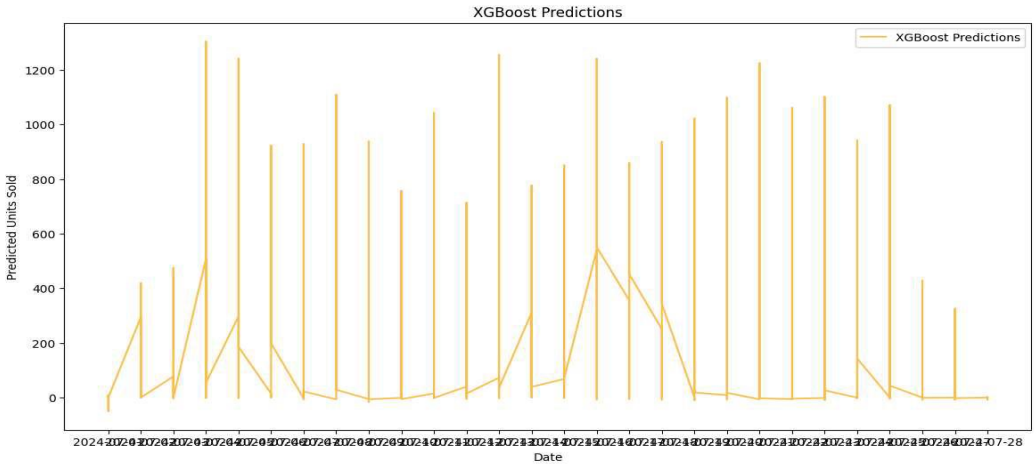
to the outcomes. This analysis helped in understanding the model's decision-making process and provided a basis for further optimization.

- Final Predictions and Submission:

The final predictions were compiled into a submission file, including the date, item ID, and predicted TARGET. This file served as the final

output of the predictive model, ready for practical application or further analysis.

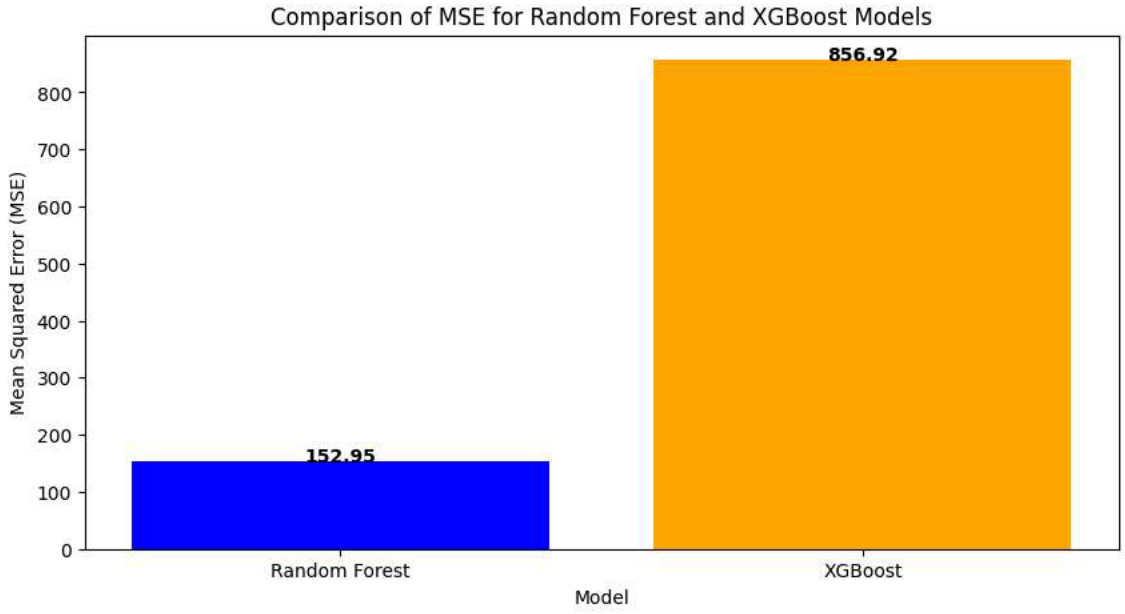




- Model Validation and Verification:

Cross-validation techniques were employed to ensure the model's reliability and generalizability. This process involved partitioning the dataset into multiple folds and training/testing the model iteratively,

thereby minimizing overfitting and ensuring consistent performance across different data subsets.



VIII. Conclusion

The detailed analysis and testing provided a thorough evaluation of the models' performance. The combination of advanced data science techniques, including feature engineering, model tuning, and rigorous evaluation, resulted in a well-rounded understanding of the predictive models' capabilities and limitations. The comprehensive nature of this evaluation was crucial in ensuring the models' accuracy, reliability, and readiness for practical applications.