Submitted by **Harshak Bhagavatula**

Forecasting Unit Sales (Task 1)

Assessment: DS & ML - 1

Time Series Analysis and Model Selection for Unit Sales Forecasting

1. Introduction

The goal of this project is to forecast the number of units sold for various items over time using historical sales data. This involves exploring the data, engineering features, selecting a suitable model, and tuning its hyperparameters.

2. Data Preprocessing and Exploratory Data Analysis (EDA)

2.1 Data Loading

The data was loaded from a CSV file into a Pandas DataFrame. The dataset contains a column ID which combines the date and item ID, and a column TARGET which represents the number of units sold.

2.2 Splitting the ID Column

The ID column was split into two separate columns: date and Item Id for better data manipulation.

2.3 Basic Statistics

We checked the basic statistics of the data and verified that there are no missing values.

2.4 Plotting

Two main plots were generated:

- Total units sold over time: This helps to visualize trends and seasonal patterns.
- Distribution of sales: This shows the distribution of the target variable, TARGET.

3. Feature Engineering

3.1 Date Features

We extracted several features from the date column:

- year
- month
- day
- dayofweek

These features help capture any temporal patterns in the data.

3.2 Label Encoding

The Item Id column was label encoded to convert the categorical data into numerical values suitable for the model.

3.3 Lag Features

Lag features (lag_1, lag_2, lag_3) were created to include past sales data as features. This is important in time series analysis to capture the temporal dependencies.

4. Model Selection

4.1 Model Choice

A RandomForestRegressor was chosen for this task due to its robustness and ability to handle non-linear relationships and interactions between features. It also performs well with the engineered features and can handle the complexity of the dataset.

5. Model Evaluation

5.1 Train-Test Split

The data was split into training and testing sets using an 80-20 split.

5.2 Model Training and Prediction

The RandomForestRegressor was trained on the training set and used to predict the test set. The Mean Squared Error (MSE) was used to evaluate the model's performance.

6. Hyperparameter Tuning

6.1 Grid Search

A GridSearchCV was used to tune the hyperparameters of the RandomForestRegressor. The hyperparameters tuned were:

- n_estimators: Number of trees in the forest.
- max_depth: Maximum depth of the tree.
- min_samples_split: Minimum number of samples required to split an internal node.

The best model was selected based on the lowest MSE from the grid search results.

7. Submission Preparation

7.1 Submission File

The final predictions were saved into a submission file which includes the date, item ID, and predicted TARGET.

Conclusion

This approach to forecasting unit sales involved careful preprocessing, feature engineering, and model tuning. The RandomForestRegressor was chosen for its effectiveness in handling complex datasets and was tuned to improve its performance. This structured approach ensures a robust model for forecasting future sales.