Project Report

Title: Big Mart Sales Prediction

GitHub Link: babutabhavya/big-mart-sales-prediction (github.com)

Dataset Link: BigMart Sales Data (kaggle.com)

Team Members:

1. Arnold Joseph (GitHub Username: Ulysses013, Matriculation No.: 3120369)
2. Bhavya Babuta (GitHub Username: babutabhavya, Matriculation No.: 3120456)
3. Ipshita Singh (GitHub Username: IpshitaSingh, Matriculation No.: 3121217)
4. Shreyaasri Prakash (GitHub Username: shreyaa98, Matriculation No.: 3121429)

Introduction:

The goal of this project is to predict the sales of a supermarket based on its historical data. Sales prediction is vital for effective inventory management, resource allocation, and strategic business planning. This problem is particularly relevant for optimizing operations and maximizing profitability in the retail industry.

Data and Methods:

1. Exploratory Data Analysis (EDA):

Exploratory data analysis (EDA) has been performed to derive insights from the dataset. The datatypes have Been described for all features and performed an initial lotting of the distribution of data

2. Data Visualization:

Data visualization, with Matplotlib and Seaborn, has been done to create various charts to visualize numerical features, uncover patterns, and identify relationships between features.

3. Data Preprocessing/ Feature Engineering:

The project initiated with loading and combining training and test datasets for pre-processing. Feature engineering enhances the dataset by imputing missing values (for features such as ‘Item Weight’ and ‘Outlet Size’) to provide a more complete representation. Dropped unnecessary column such as Item_ Visibility and Identifiers as they were not beneficial for the training. To ensure data consistency, categorical variables have been encoded using one hot encoding and label encoding.

4. Machine Learning:

4.1 Random Forest:

Random Forest Regression has been chosen for its ability to capture intricate relationships within the data. The first Random Forest model was trained on the preprocessed dataset to learn and generalize from historical sales patterns.

Feature Importance and Selection: To improve the generalization, a second Random Forest model is then trained and evaluated on both the training and test sets. For this, feature importances were analyzed to gauge the influence of each feature on sales prediction. The threshold for feature importance was set empirically, balancing model performance and the significance of selected features. Key features such as 'Item_MRP' and 'Outlet_Type' emerged as crucial contributors, shedding light on the most influential factors.

4.2 XGBoost Regressor:

XGboost has been included as the second model but we encountered high error and low r2 score and potential overfitting. Due to which a second XGBoost was trained with only the most important features with respect to first training to XGB, and worsened the model performance with high error and lower r2 score. Lastly, the weak performance of XGB, we trained a third model using ridge regression

4.3 Ridge Regression

We selected range of alpha values for this regression and determines the co-efficient for each alpha. Also, checked three independent values to obtain their co-efficient for the features. Lastly, trained the ridgecv regressor to get the highest score with respect to to the validation data. And encountered that an alpha of 10.xx was given the best fit. Thus, avoiding overfitting and underfitting

5. Model Evaluation:

5.1Random Forest:

Two evaluation metrics were used to assess the model performance, R-squared (R2) Score and Mean Squared Log Error (MSLE). The initial model, trained on original features, demonstrated strong performance with a high R-squared (R2) Score of 0.937 on the training set but struggled with overfitting, as reflected by a lower R2 Score (0.555) on the test set. Post-optimization, the model achieved a slightly lower R2 Score (0.706) on the training set but showed improved generalization, with a closer R2 Score (0.599) on the test set. These adjustments aimed to balance performance and mitigate overfitting, enhancing the model's ability to generalize to new data.

5.2 XGBoostRegressor

There were indications of possible overfitting in the original XGBoost model, as evidenced by a significant discrepancy between training and validation measures. In order to overcome this, a new XGBoost model was created using feature selection based on relevance scores. But on the validation set, the re-trained model performed worse, suggesting that feature selection, which was supposed to reduce overfitting, did not have the anticipated effect. The validation's Mean Squared Error increased from 1,112,401.01 to 1,680,570.87, the R2 Score decreased from 0.59 to 0.38, and the Mean Absolute Error climbed from 730.24 to 906.96. These findings imply that it is still difficult to strike a balance between feature selection and model complexity, and that further work may be required to produce an accurate and widely applicable XGBoost model.

5.3 Ridge Regressor

The Ridge Regression model performs consistently on both the training and validation sets, exhibiting a balanced fit. Satisfactory model generalization is shown by the validation metrics, which include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2 Score. The validation set's R2 Score of 0.579 indicates that the model does a good job of explaining the variation in the data. A small discrepancy between measures used for training and validation suggests strong generalization without overfitting. All things considered, the Ridge Regression model seems appropriate for the task at hand, offering a steady and dependable prediction performance.

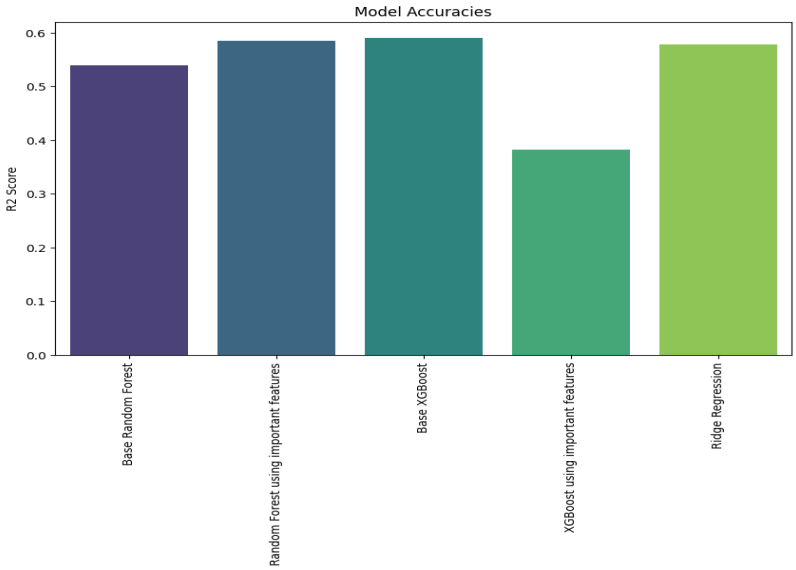
Results:

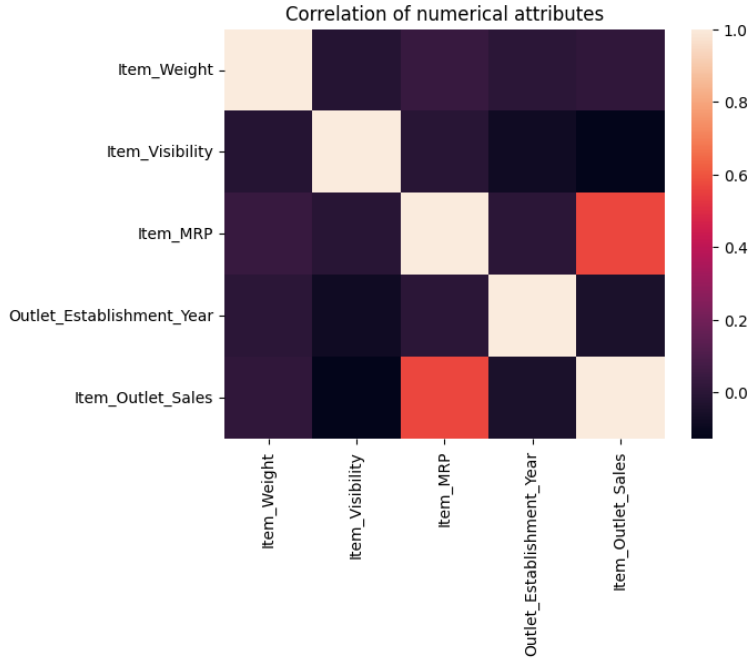
1. **Random Forest:** After optimization based on feature importance, the Random Forest model showed a balanced performance between the training and test sets. Although the R2 score slightly decreased on the training set, the model achieved better generalization on the test set, reducing the overfitting observed in the initial model. This highlights the effectiveness of the feature selection approach in enhancing overall model performance.

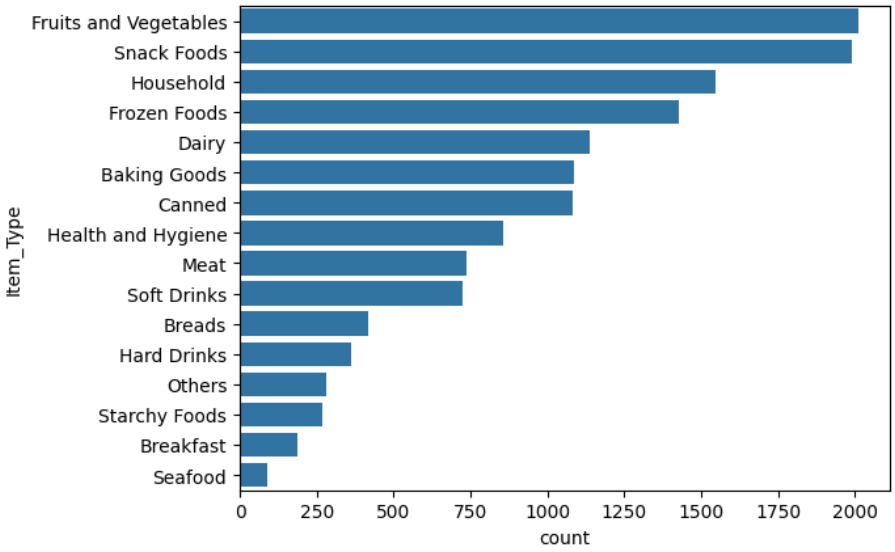
2. **XGBoost Regressor:** Feature selection was necessary when overfitting problems were raised by the first XGBoostmodel. But the performance of the retrained model, which concentrated on significant characteristics, decreased. The overfitting was attempted to be addressed, however the model's prediction accuracy declined on the validation set. It's still difficult to balance feature selection with model complexity

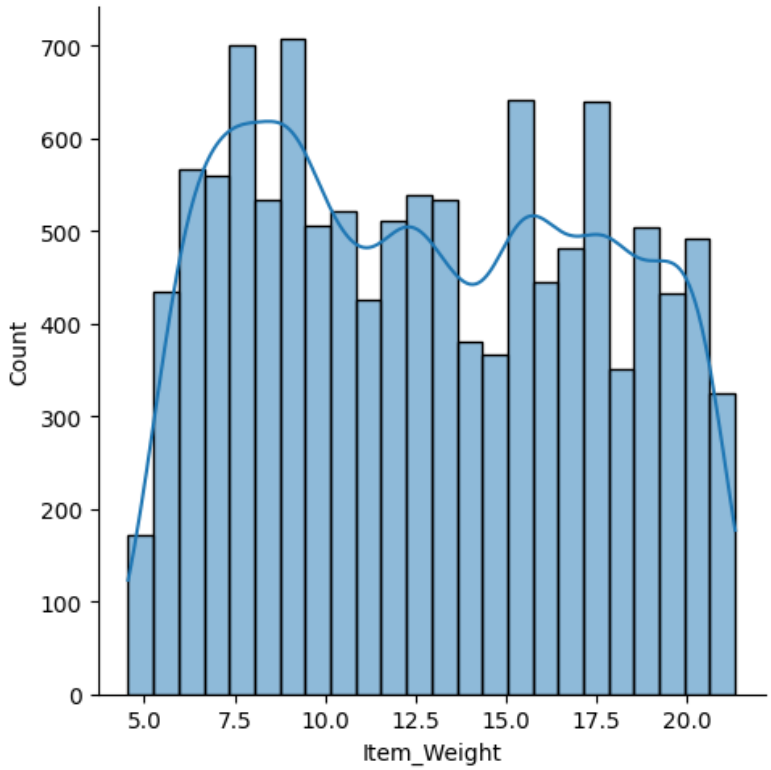
3. **Ridge Regressor:** After optimization, the Ridge Regression model demonstrates a well-balanced fit, maintaining consistency between the training and validation sets. The feature importance-based approach successfully enhances overall model performance, ensuring effective generalization without significant overfitting.

Graphs and Visualization


Model Accuracies
R2 Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Base Random Forest
Random Forest using important features
Base XGBoost
XGBoost using important features
Ridge Regression


Correlation of numerical attributes
Item_Weight
Item_Visibility
Item_MRP
Outlet_Establishment_Year
Item_Outlet_Sales
1.0
0.8
0.6
0.4
0.2
0.0





Conclusion:

The Random Forest model and Ridge Regressor are the highest performing models and XGBoost is not recommended.