Big Mart Sales Prediction

Objective

The goal of this project was to develop a machine learning model to predict sales for different products across various stores using historical sales data.

Data Understanding & Preprocessing

The dataset contained information about products, store attributes, and sales figures. Key steps in preprocessing included:

- **Handling Missing Values:** Missing values in numerical columns like Item_Weight were imputed using mean values, and categorical variables were handled using mode imputation.
- **Feature Engineering:** New features such as Outlet_Age were created, and categorical variables were encoded using One-Hot Encoding and Label Encoding.
- **Outlier Treatment:** Anomalies in Item_Visibility were identified and handled by replacing zero values with the median.
- **Feature Scaling:** Numerical features were standardized to ensure uniformity in model training.

Model Selection & Experimentation

Several regression models were tested to determine the best performing approach:

- **Random Forest Regressor**
- **XGBRegressor**

After evaluating model performance using metrics like RMSE (Root Mean Squared Error), it was observed that the **Random Forest Regressor performed best**, yielding the lowest error.

Key Factors Affecting Item Outlet Sales

Based on the visualizations and insights from the analysis, the following factors significantly influence Item Outlet Sales:

1. **Item MRP (Maximum Retail Price)**

   A strong correlation was observed between Item_MRP and Item_Outlet_Sales, indicating that higher-priced items tend to generate higher sales.

2. **Outlet Type**

   Supermarkets recorded higher sales compared to grocery stores, suggesting that store format and product variety impact revenue.

3. **Outlet Location Type**

   Outlets in Tier-1 cities had higher sales than those in Tier-2 and Tier-3 cities, likely due to higher purchasing power and customer footfall.

4. **Outlet Age**

Newer outlets showed lower sales initially, while well-established outlets with more years in operation exhibited higher sales.

5. **Item Visibility**

   Products with very low visibility had lower sales, highlighting the importance of product placement and marketing within stores.

6. **Item Fat Content & Category**

   Certain product categories, especially perishable and high-demand items, contributed more to sales, while fat content variations had a minimal impact.

Findings & Conclusion

- **Random Forest Regressor outperformed XGBRegressor**, likely due to its robustness in handling diverse feature interactions and missing values.
- Further hyperparameter tuning on XGBoost could be explored to improve its performance.
- The final model can be used to make sales predictions for new data, aiding business decision-making for inventory management and revenue forecasting.

This systematic approach ensured a well-prepared dataset and effective model selection, leading to reliable predictions

Date:

19th February, 2025

Submitted By:

Srinivasa Rao Pendela

+91 7093348767