Big Mart Sales Prediction

1. Executive Summary

This project, a solution to the Big Mart Sales III challenge, focuses on building a predictive model to forecast retail product sales across different stores. The primary objective is to identify key product and outlet attributes that influence sales and to predict Item_Outlet_Sales for a test set. The model's performance is evaluated using the Root Mean Squared Error (RMSE) metric. The project follows a standard data science methodology, beginning with extensive exploratory data analysis (EDA), followed by feature engineering, and concluding with a predictive modeling pipeline.

2. Data Overview

The analysis was performed on a combined dataset comprising 8,523 records for training and 5,681 records for testing. The data contains 11 features including both numerical (e.g., Item_Weight, Item_MRP) and categorical attributes (e.g., Outlet_Type, Item_Fat_Content), along with the target variable, Item_Outlet_Sales, in the training set. A preliminary inspection revealed several key data quality issues that needed to be addressed before modeling.

3. Data Cleaning and Preprocessing

A significant effort was dedicated to cleaning and preparing the data to ensure the model's robustness and accuracy. Key steps included:

- **Handling Missing Value**s: The dataset had missing values in the Item_Weight and Outlet_Size columns.
    - Item_Weight: Missing values were imputed by calculating the mean Item_Weight for each unique Item_Identifier. This method assumes that a product's weight is consistent regardless of the store.
    - Outlet_Size: Missing Outlet_Size values were imputed with the mode of the Outlet_Size for their corresponding Outlet_Type. This was a more effective strategy than a simple global mode, as it leverages the relationship between an outlet's type and its typical size.
- **Addressing Data Inconsistencies**: The Item_Fat_Content column contained inconsistent entries such as "low fat", "LF", and "reg". These were standardized to a consistent "Low Fat" and "Regular" to prevent the model from treating them as separate categories.
- **Handling Zero Values**: Unrealistic zero values were present in the Item_Visibility column. These were imputed with the mean visibility for the corresponding Item_Identifier, as it is highly unlikely for a product to have zero visibility.

4. Feature Engineering

To extract more predictive power from the raw data, several new features were engineered:

- **Outlet_Age**: A new numerical feature was created by calculating the difference between the current year (2013) and the Outlet_Establishment_Year. This feature captures the age of the outlet, which can influence sales.
- **Item_Category**: A new categorical feature was derived from the Item_Identifier by extracting the first two letters. This allowed for the creation of new high-level categories like "Food," "Drinks," and "Non-Consumable."

5. Exploratory Data Analysis (EDA) Insights

The EDA phase, driven by a series of visualizations, provided critical insights that guided the modeling approach.

- **Target Variable Distribution**: A histogram of Item_Outlet_Sales confirmed a highly right-skewed distribution, with a long tail extending to higher sales values. This highlighted the need for a logarithmic transformation to improve the model's performance.
- **Numerical Feature**s: A correlation analysis revealed that Item_MRP had a strong positive correlation (~0.57) with Item_Outlet_Sales. Other numerical features like Item_Weight, Item_Visibility, and Outlet_Age showed very weak linear relationships.
- **Categorical Feature**s: Box plots and a feature importance analysis indicated that Outlet_Type was the most powerful categorical predictor. Supermarket Type3 outlets showed significantly higher sales than any other type, while Grocery Store outlets consistently had the lowest sales. Outlet_Size and Outlet_Location_Type were also found to be important features.

6. Predictive Modeling & Evaluation

A machine learning pipeline was constructed to streamline the workflow, from data preprocessing to model training and prediction.

After the initial exploratory data analysis and feature engineering, a **boilerplate XGBoost model was trained with basic hyperparameters** to establish a performance baseline. This initial model provided a solid starting point for the task.

To further improve the model's accuracy, a **Bayesian optimization process was conducted using the Optuna framework**. This method intelligently searches the hyperparameter space to find the optimal combination of settings for the XGBoost model. By running numerous trials, Optuna identified the best values for key parameters like learning rate, tree depth, and regularization terms.

The final model used for prediction was **the result of this Bayesian optimization**, ensuring that the best-performing set of hyperparameters was used to deliver the most accurate predictions possible.

7. Conclusion & Recommendations

The project successfully delivered a robust predictive model for Big Mart sales. The key takeaways from the analysis are:

- **Product Pricing is Key**: A product's Item_MRP is the most influential factor in determining sales.
- **Store Type Matters**: The Outlet_Type is a critical determinant of a store's sales performance.
- **Data Quality is Crucial**: Addressing missing values and inconsistencies significantly improved the quality of the dataset for modeling.

Recommendations for Future Work:

- **Advanced Tuning**: Further performance gains can be achieved by performing a more extensive hyperparameter search on the best-performing model (XGBoost) using a wider range of parameters.
- **Feature Engineering**: Exploring more complex feature interactions could uncover hidden patterns.
- **Ensemble Modeling**: Combining the predictions from multiple models (e.g., XGBoost and Random Forest) through averaging or stacking could yield a more accurate and stable final prediction.