Case Study: BigMart Sales Prediction Project

Objective

The goal of this project is to develop a predictive model to estimate product sales across different stores for BigMart. The insights derived will help identify key features impacting sales, enabling data-driven decisions to optimize performance.

Dataset Details

The dataset contains sales data for 1,559 products across 10 stores in different cities, with various product and store attributes, including:

- ✓ **Item_Identifier**: Unique identifier for each product.
- ✓ **Item_Weight**: Weight of the product.
- ✓ **Item_Fat_Content**: Fat content level of the product.
- ✓ **Item_Visibility**: Percentage of total visibility across stores.
- ✓ **Item_Type**: Category of the product.
- ✓ **Item_MRP**: Maximum retail price of the product.
- ✓ **Outlet_Identifier**: Unique identifier for each store.
- ✓ **Outlet_Establishment_Year**: Year the store was established.
- ✓ **Outlet_Size**: Size of the store (e.g., Small, Medium, High).
- ✓ **Outlet_Location_Type**: Location type of the store (e.g., Tier 1, Tier 2, Tier 3).
- ✓ **Outlet_Type**: Type of the outlet (e.g., Grocery Store, Supermarket Type1).
- ✓ **Item_Outlet_Sales**: Target variable, sales of the product in a specific outlet.

Data Exploration and Preprocessing

Exploratory Data Analysis (EDA)

- ➢ **Data Overview**:

- ✓ Checked dataset shape and null values.
- ✓ Investigated numerical features like Item_Weight, Item_Visibility, and Item_MRP.
- ✓ Analyzed categorical features, including frequency distribution for Item_Type, Outlet_Type, etc.

- **Data Cleaning**:
  - ✓ Handled missing values for Item_Weight and Outlet_Size using logic-based imputation.
  - ✓ Standardized categories in Item_Fat_Content.
- **Feature Engineering**:
  - ✓ Created a high-level categorization (Food, Drink, Non_Consumables) from Item_Identifier.
  - ✓ Adjusted Item_Fat_Content to include a Non_Edible category for non-consumables.
  - ✓ Converted Outlet_Establishment_Year into Outlet_Age.
- **Feature Visualization**:
  - ✓ Used histograms and boxplots to understand the distribution and outliers in numerical features.
  - ✓ Count plots to analyze categorical data distributions.

Data Preprocessing

- **Imputation**:
  - ✓ Filled missing Item_Weight values using mappings based on Item_Identifier and Item_Type.
  - ✓ Imputed Outlet_Size using the mode of Outlet_Type.
- **Encoding**:
  - ✓ One-hot encoding for categorical features.
  - ✓ Feature hashing for Item_Identifier.
- **Standardization**:
  - ✓ Standardized numerical features to ensure consistency.

- **Dataset Splits**:
  - ✓ Split data into training (70%) and testing (30%) sets.

Modeling

Models Evaluated

1. **Random Forest Regressor**
2. **Gradient Boosting Regressor**
3. **HistGradientBoosting Regressor**
4. **XGBoost Regressor**
5. **LightGBM Regressor**

Key Evaluation Metrics

- ✓ **R^2 (Coefficient of Determination)**: Indicates the proportion of variance explained by the model.
- ✓ **RMSE (Root Mean Square Error)**: Measures the model's prediction error.

Results

Model	R^2 Mean ± Std Dev	RMSE Mean ± Std Dev
Random Forest Regressor	0.61 ± 0.02	1,156 ± 37
Gradient Boosting Regressor	0.63 ± 0.03	1,132 ± 29
HistGradientBoosting Regressor	0.66 ± 0.02	1,095 ± 25
XGBoost Regressor	0.65 ± 0.03	1,102 ± 32
LightGBM Regressor	**0.67 ± 0.02**	**1,084 ± 22**

Conclusion

The **LightGBM Regressor** achieved the best performance, with the highest R^2 score and lowest RMSE. Key takeaways include:

1. Features like Item_MRP, Item_Type, and Outlet_Type significantly impact sales.
2. Non-consumables have distinct sales patterns due to Item_Fat_Content.