BigMart Sales Prediction Project - Detailed Report

Objective:

The primary goal of the BigMart Sales Prediction Project is to build a machine learning model capable of predicting the sales of individual products at different outlets. This model will assist BigMart in understanding which product and store attributes contribute to higher sales, thus enabling the company to optimize its strategies.

Table of Contents

---

1. Problem Statement

BigMart wants to predict the sales of each product across various outlets based on historical sales data and product attributes. By doing so, they aim to: - Understand the key drivers of product sales. - Optimize product placement and store configurations. - Predict future sales trends based on product/store characteristics.

Business Problem:

- **Target Variable:** `Item_Outlet_Sales`
- **Type of Problem:** Supervised Regression Problem

---

2. Hypothesis Generation

To build a meaningful predictive model, several hypotheses regarding factors that may affect product sales were generated. These hypotheses will guide the analysis and feature engineering process.

Potential Factors Influencing Sales: - **Product MRP (Maximum Retail Price):** Higher-priced products may result in higher sales. - **Product Weight:** Heavier products might have higher sales due to larger packaging or premium nature. - **Product Category:** Some product categories might sell better in certain outlets (e.g., food items vs. non-food items). - **Outlet Size:** Larger stores may have more customers and, consequently, higher sales. - **Outlet Location Type:** Stores in urban locations may have higher sales compared to rural stores due to higher footfall. - **Outlet Age:** Older outlets might have more loyal customers, contributing to higher sales.

---

3. Data Understanding

The dataset contains 1559 products across 10 outlets, with various attributes recorded for both products and outlets.

Key Features:

- **Item_Identifier:** Unique identifier for each product.
- **Item_Weight:** Weight of the product.
- **Item_Fat_Content:** Fat content of the product (low fat, regular, etc.).
- **Item_Visibility:** The percentage of the product's visibility in the store.
- **Item_Type:** Type/category of the product (e.g., dairy, snacks).
- **Item_MRP:** Maximum retail price of the product.
- **Outlet_Identifier:** Unique identifier for the outlet/store.
- **Outlet_Establishment_Year:** Year the outlet was established.
- **Outlet_Size:** Size of the outlet (small, medium, large).
- **Outlet_Location_Type:** Type of location (urban, rural, suburban).
- **Outlet_Type:** Type of store (supermarket, grocery store, etc.).
- **Item_Outlet_Sales:** The target variable representing sales.

---

4. Data Preprocessing and Cleaning

Steps Taken:

1. **Handling Missing Values:**
   - Missing values were observed in `Item_Weight` and `Outlet_Size`. These were handled by:
     - Imputing the mean weight for missing values in `Item_Weight`.
     - Filling missing values in `Outlet_Size` based on `Outlet_Type` correlation.
2. **Handling Outliers:**
   - Outliers were detected in features like `Item_Visibility` (some products had zero visibility, which isn't realistic). This was corrected by replacing zero visibility with the mean visibility of that product type.
3. **Encoding Categorical Variables:**
   - **Label Encoding** was used for binary categorical variables (e.g., `Item_Fat_Content`).
   - **One-Hot Encoding** was applied to multi-class categorical features such as `Item_Type` and `Outlet_Location_Type`.
4. **Normalization:**
   - Certain numerical variables (e.g., `Item_MRP`, `Item_Visibility`) were normalized to ensure all features were on a similar scale before modeling.

---

5. Exploratory Data Analysis (EDA)

EDA was conducted to identify patterns in the data and relationships between features and the target variable (`Item_Outlet_Sales`).

Key Findings:

- **Sales Distribution:** Sales are skewed with a long tail, indicating that most products have moderate sales while a few products have very high sales.
- **Price vs. Sales:** There is a positive correlation between `Item_MRP` and sales, as expected. Higher-priced items generally have higher sales.
- **Outlet Type:** Supermarket outlets tend to have higher sales than grocery stores, likely due to a broader range of products and higher footfall.
- **Outlet Age:** Older outlets tend to have higher sales, which could be due to established customer loyalty.

Visualizations:

1. **Sales vs. MRP Scatter Plot:** Shows the positive relationship between price and

sales.
2. **Sales by Outlet Type (Bar Chart):** Highlights the difference in sales between supermarket types and grocery stores.
3. **Sales by Item Type (Box Plot):** Displays how different product categories perform in terms of sales.

---

6. Feature Engineering

To enhance model performance, new features were created based on domain knowledge:

1. **Outlet_Age:** Calculated as the difference between the current year and the outlet establishment year.
2. **Item_Type_Grouped:** Grouped item types into broader categories like 'Food,' 'Drinks,' and 'Non-Food.'
3. **Price_Per_Unit_Weight:** Created a feature by dividing `Item_MRP` by `Item_Weight` to understand the relationship between weight and pricing.

---

7. Model Building and Evaluation

Multiple machine learning models were trained and evaluated using the preprocessed and engineered dataset.

Models Used:

1. **Linear Regression:** Baseline model to understand the linear relationship between features and sales.
2. **Ridge Regression:** A regularized version of linear regression to handle multicollinearity.
3. **RandomForest Regressor:** A powerful ensemble model to capture non-linear relationships in the data.
4. **XGBoost:** A gradient boosting technique that often yields high accuracy in regression tasks.

Evaluation Metrics:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors in the predictions.
- **R^2 Score:** Indicates the proportion of variance in the target variable that is predictable from the independent variables.

Model Performance:

- **Linear Regression:**
  - MAE: 1130.6
  - R^2: 0.56
- **Random Forest:**
  - MAE: 865.3
  - R^2: 0.78
- **XGBoost:**
  - MAE: 825.4
  - R^2: 0.82

Best Model:

- XGBoost had the best performance with the lowest MAE and the highest R^2

score, making it the final model for deployment.

---

8. Model Deployment

The final model (XGBoost) was deployed using a REST API built with **Flask**. The following steps were taken for deployment:

- The trained model was saved using `joblib`.
- An API was created where users can input product and store details and get real-time sales predictions.
- The API is designed to handle both single product predictions and batch predictions via file uploads (CSV).

---

9. Results and Insights

- **Key Drivers of Sales:** The most important features contributing to sales predictions were `Item_MRP`, `Outlet_Type`, `Outlet_Age`, and `Item_Visibility`.
- **Price Influence:** Higher-priced products consistently resulted in higher sales.
- **Outlet Characteristics:** Supermarkets, particularly larger ones, had much higher sales compared to smaller, grocery-type stores.

---

10. Conclusion and Future Work

The BigMart Sales Prediction model successfully predicts sales based on product and store attributes, with the XGBoost model yielding the most accurate results. This model can help BigMart optimize its store layout, pricing strategy, and product placement.

Future Work:

- **Additional Features:** Incorporate more features such as promotional campaigns, store-specific holidays, and seasonal effects.
- **Time-Series Modeling:** Integrate a time-series approach to capture the temporal trends in sales over months or years.
- **Real-time Prediction Pipeline:** Extend the deployment to include real-time predictions based on streaming data from stores.

Appendix

- **Code for Model Training:** See attached files.
- **Full EDA Report:** See attached plots and visualizations.
