High-Level Document (HLD) for BigMart Sales Prediction Project

---

1. Introduction

Project Name: BigMart Sales Prediction
Objective: The goal of this project is to build a predictive model to estimate the sales of different products at various BigMart stores. This will help BigMart understand the factors that drive product sales and allow them to optimize their store and product management to increase profitability.

---

2. Project Overview

Business Goal:
Develop a machine learning model to predict **Item_Outlet_Sales**, which represents the total sales of a particular product at a given outlet. Using this model, BigMart will gain insights into the product and store features that significantly impact sales performance.

Scope of the Solution: - Predict sales for 1559 products across 10 stores. - Handle missing data and feature engineering. - Build a model that accounts for both product and store-specific attributes. - Provide insights into which features (e.g., product type, outlet location) influence sales. - Save the final model for use in forecasting future sales.

---

3. Architecture Overview

3.1 Input Data: - **Data Sources:** Historical sales data for 2013 - **Data Attributes:** - **Product-Level:** Item Identifier, Item Weight, Item Fat Content, Item Visibility, Item Type, Item MRP, etc. - **Store-Level:** Outlet Identifier, Outlet Size, Outlet Location Type, Outlet Type, Outlet Establishment Year, etc. - **Target Variable:** Item_Outlet_Sales

3.2 Workflow:

1. **Data Collection and Loading:**
   Load the dataset containing product and store details along with historical sales data.

2. **Data Preprocessing:**

   - Missing Value Treatment
   - Outlier Detection and Handling
   - Feature Engineering

3. **Exploratory Data Analysis (EDA):**
   Perform univariate and bivariate analysis to discover trends, correlations, and distributions within the data.

4. **Data Transformation:**

   - Categorical Encoding (Label Encoding and One Hot Encoding)
   - Data Normalization/Standardization

5. **Modeling:**

   - Build and test multiple regression models:
     - Linear Regression
     - Regularized Linear Regression (Ridge, Lasso)
     - Random Forest Regressor
     - XGBoost Regressor
   - Hyperparameter Tuning for optimized performance

6. **Model Evaluation:**

   - Metrics used:
     - Mean Absolute Error (MAE)
     - R-squared Score (R^2)
   - Compare performance across models and select the best-performing one.

7. **Model Deployment:**
   Save the final model and prepare it for future use in predicting sales for new products/outlets.

---

4. Data Flow Diagram

```
+---------------------------------------+
|          Historical Sales Data        |
| +-----------------------------------+ |
| |  Product Data   |  Store Data     | |
| +-----------------------------------+ |
+---------------------------------------+
             |
             v
+---------------------------------------+
|        Data Preprocessing             |
|  - Missing Value Treatment            |
|  - Feature Engineering                |
|  - Categorical Encoding               |
|  - Data Transformation                |
+---------------------------------------+
             |
             v
```

```
+---------------------------------------+
|        Exploratory Data Analysis      |
|  - Univariate Analysis                |
|  - Bivariate Analysis                 |
+---------------------------------------+
             |
             v
+---------------------------------------+
|            Model Building             |
|  - Linear Regression                  |
|  - Random Forest Regressor            |
|  - XGBoost Regressor                  |
+---------------------------------------+
             |
             v
+---------------------------------------+
|       Model Evaluation & Selection    |
|  - Compare MAE, R²                    |
|  - Select Best Model                  |
+---------------------------------------+
             |
             v
+---------------------------------------+
|          Final Model & Insights       |
|  - Save Model for Future Use          |
|  - Provide Key Insights               |
+---------------------------------------+
```

---

5. Data Preprocessing

- **Missing Value Treatment:**
  Impute missing values in features like `Item_Weight` using median imputation, and fill in missing values for categorical variables with the mode.

- **Outlier Handling:**
  Identify outliers in numerical variables like `Item_Visibility` and handle them by capping values.

- **Feature Engineering:**

  - Calculate `Outlet_Age` from `Outlet_Establishment_Year`.
  - Group product types into broader categories (e.g., food, non-food).
  - Create interaction features between products and stores.

- **Categorical Variable Encoding:**
    - Apply **Label Encoding** to ordinal features (e.g., `Outlet_Size`).
    - Apply **One-Hot Encoding** to nominal categorical variables (e.g., `Item_Type`, `Outlet_Location_Type`).

---

6. Modeling Approach

- **Models Considered:**
    - **Linear Regression:** To build a simple baseline model and interpret coefficients.
    - **Ridge and Lasso Regression:** To regularize the model and handle multicollinearity.
    - **Random Forest Regressor:** To capture non-linear relationships between the features.
    - **XGBoost Regressor:** For highly optimized and scalable tree-based modeling.
- **Model Evaluation:**
    - **Training and Testing Split:** 70/30 split between training and test data.
    - **Metrics:**
        - **Mean Absolute Error (MAE):** Measures average error in prediction.
        - **R² Score:** Explains the variance captured by the model.

---

7. Deliverables

- **Predictive Model:** A trained machine learning model to predict `Item_Outlet_Sales` for new data.
- **Insights Report:** Summary of key findings regarding product and store attributes that drive sales.
- **Final Model File:** Saved model in `.pkl` or `.joblib` format for deployment.

---

8. Risks and Mitigation

- **Risk:** Data inconsistency or missing values.
  Mitigation: Imputation techniques and data validation.

- **Risk:** Overfitting models.
  Mitigation: Regularization (Ridge, Lasso), Cross-validation.

- **Risk:** Model complexity may lead to slow training times.
  Mitigation: Use of efficient algorithms like XGBoost and RandomForest.

---

9. Timeline

Task	Duration
Data Collection and Preparation	1 week
Data Preprocessing & Feature Engineering	1.5 weeks
Exploratory Data Analysis	1 week
Modeling	2 weeks
Model Evaluation & Optimization	1 week
Report Preparation & Deployment	1 week

10. Conclusion

The BigMart Sales Prediction project will provide actionable insights into which factors most influence sales performance at each store. By using a predictive model, BigMart can make data-driven decisions to enhance product placement, store management, and marketing strategies.