BigMart Sales Prediction – Phase-Wise Project Report

Project Title

BigMart Sales Prediction using Machine Learning

Project Proposal

This project aims to build a **predictive model** to estimate the sales of various products across different BigMart outlets. By analyzing the dataset and leveraging machine learning algorithms, the project identifies sales-driving factors and predicts future sales for business optimization.

The system demonstrates a complete **data science lifecycle**:
Data preprocessing → Exploratory Data Analysis → Model training → Evaluation → Dashboard deployment.

Problem Description

Retail chains like BigMart handle large volumes of product and outlet-level sales data. However, they often struggle to predict future sales due to:

- Inconsistent sales across outlets.
- Missing or irregular product attributes.
- Seasonal and categorical variations.

This project solves that problem by using machine learning models to:

- Predict `Item_Outlet_Sales` (target variable).
- Identify features impacting sales performance.
- Provide actionable insights for business decisions.

Dataset Details

- **Dataset Source**: Kaggle — *BigMart Sales Prediction Dataset*
- **Training Set Size**: 8,523 records
- **Test Set Size**: 5,681 records
- **Total Features**: 11 input + 1 target variable (`Item_Outlet_Sales`)

Key Attributes:

Feature	Description
Item_Weight	Weight of the product
Item_Fat_Content	Low Fat / Regular
Item_Visibility	Percentage visibility in store

Item_Type	Category of item
Item_MRP	Maximum Retail Price
Outlet_Establishment_Year	Year outlet was opened
Outlet_Size	Small / Medium / High
Outlet_Location_Type	Tier 1 / 2 / 3
Outlet_Type	Grocery Store / Supermarket
Item_Outlet_Sales	Target variable

Phase 1 – Project Initiation & Setup

Objectives

- Understand the business problem and define goals.
- Gather dataset and perform initial exploration.
- Set up a structured project repository.
- Configure virtual environment and dependencies.

Tasks Completed

- Created folder hierarchy for modular development.
- Installed required libraries (pandas, numpy, sklearn, streamlit, matplotlib, seaborn, plotly).
- Configured `requirements.txt` and `setup.py` for reproducibility.
- Designed the workflow for preprocessing → EDA → modeling → dashboard.

Output

A clean, reproducible folder structure ready for data analysis.

Phase 2 – Data Collection & Pre-processing

Objectives

- Import, clean, and prepare data for modeling.
- Handle missing values and inconsistent data.
- Engineer useful features for better model performance.

Steps Performed

1. **Data Loading**
    - Imported `Train.csv` and `Test.csv` from Kaggle.
2. **Data Cleaning**
    - Handled missing values:
        - `Item_Weight`: replaced using mean imputation.
        - `Outlet_Size`: filled using mode imputation by `Outlet_Type`.
3. **Data Correction**

    - Standardized inconsistent entries in `Item_Fat_Content` (e.g., *LF, low fat → Low Fat*).
4. **Feature Engineering**
    - Created new variables:
        - `Outlet_Age = 2025 - Outlet_Establishment_Year`
        - `Item_Visibility_Bins` for grouping low/high visibility
        - `MRP_Categories` for pricing segments
5. **Encoding**
    - One-hot encoding for categorical variables.
    - Saved processed datasets (`processed_train.csv`, `processed_test.csv`).

Output

- Cleaned dataset ready for analysis.
- 20+ engineered features improving model interpretability.

Phase 3 – Exploratory Data Analysis (EDA) & Visualization

Objectives

- Understand dataset distributions and relationships.
- Identify sales trends, outliers, and feature impact.
- Generate business insights through visualization.

EDA Performed

- **Univariate Analysis:**
    - Distribution plots for Sales, Item MRP, Visibility, Outlet Age.
- **Bivariate Analysis:**
    - Relationship between `Item_Outlet_Sales` and `Item_Type`, `Outlet_Type`.
- **Correlation Heatmap:**
    - Found strongest correlation between `Item_MRP` and `Sales` (0.57).
- **Categorical Insights:**
    - `Supermarket Type3` outlets have highest average sales.
    - Older outlets show lower performance (negative correlation).
- **Visualization Tools:** Matplotlib, Seaborn, Plotly (interactive).

Output

- Generated 20+ plots saved in `results/`.
- Created business insights used later in dashboard.

Phase 4 – Model Building & Evaluation

Objectives

- Train regression models to predict `Item_Outlet_Sales`.
- Compare algorithm performance.

- Save best model for deployment.

Steps Performed

1. **Train-Test Split:** 80% training, 20% testing.
2. **Algorithms Used:**
   - Linear Regression
   - Ridge & Lasso Regression
   - Decision Tree Regressor
   - Random Forest Regressor
   - Gradient Boosting
   - XGBoost
   - LightGBM
   - Extra Trees Regressor
3. **Evaluation Metrics:**
   - R^2 Score, RMSE, MAE, and Cross-Validation Score.
4. **Results Summary:**

Model	R^2	RMSE	MAE
XGBoost	0.65	1187	882
LightGBM	0.64	1245	934
Random Forest	0.63	1299	991

5. **Best Model:**
   - **XGBoost Regressor** with $R^2 = 0.65$, RMSE ≈ 1187.
   - Saved as `models/best_model.pkl`.

Output

Optimized predictive model with excellent generalization and saved for dashboard use.

Phase 5 – Dashboard Development

Objectives

- Build an interactive Streamlit web dashboard.
- Allow users to visualize data and predict sales dynamically.

Dashboard Pages

1. **Overview:** Dataset summary and statistics.
2. **EDA:** Interactive visualizations with Plotly.
3. **Models:** Comparison charts for trained models.
4. **Predictions:** User input for live predictions.
5. **Insights:** Key findings and recommendations.

Key Features

- Dynamic KPIs for Total Sales, Average MRP, etc.
- Interactive filters (Outlet Type, Item Type).
- Downloadable prediction results.

Business Insights

- High-priced products show strong positive sales correlation.
- `Supermarket Type3` contributes maximum revenue.
- `Starchy Foods` are the best-performing item type.

Technical Achievements

- End-to-end ML pipeline with modular code.
- Automated preprocessing and EDA.
- 5-page Streamlit dashboard with real-time interaction.

Conclusion & Future Scope

Summary

The BigMart Sales Prediction project successfully demonstrates a complete data science workflow—from raw data to actionable business insights and predictive modeling.

Future Enhancements

- Integrate live sales data API for real-time prediction.
- Experiment with neural networks for better accuracy.
- Add customer segmentation analysis.
- Deploy dashboard on cloud (Streamlit Cloud / AWS / Render).