Sales Forecasting Project Report

1. Project Objective & Business Problem

Accurate sales forecasting is vital for retail businesses to optimize inventory, reduce waste, maximize revenue, and enhance customer satisfaction. This project aims to predict monthly sales revenue for individual products across multiple store outlets using historical sales data enriched with product and store-specific attributes. The forecasting system is designed to guide operational planning and strategic decision-making at Big Mart outlets.

2. Dataset Description & Data Preparation

- **Dataset Source:**
  Big Mart Sales Prediction dataset from Kaggle, featuring approximately 8,523 sales records.
- **Key Features:**
  - Product Attributes: `Item_Identifier`, `Item_Weight`, `Item_Fat_Content`, `Item_Type`, `Item_MRP`
  - Store Attributes: `Outlet_Identifier`, `Outlet_Establishment_Year`, `Outlet_Size`, `Outlet_Location_Type`, `Outlet_Type`
  - Target Variable: `Item_Outlet_Sales` (monthly sales revenue)
- **Preparation Steps:**
  - Imputed missing values for `Item_Weight` using item-level mean.
  - Filled missing `Outlet_Size` values with the mode per `Outlet_Type`.
  - Standardized categorical entries (e.g., `Item_Fat_Content`) for consistency.
  - Capped outliers in sales (`Item_Outlet_Sales`) based on the 99th percentile.
  - Encoded categorical variables using label encoding and one-hot encoding where appropriate.

3. Exploratory Data Analysis (EDA) & Insights

- **Sales Distribution:**
  Sales data show a right-skewed distribution with most sales clustered at moderate levels and some extreme high-sales outliers.

- **Revenue Drivers:**
    - Product category `Food` contributes the largest share of total sales, followed by `Clothing` and `Household`.
    - Outlet location tier and store type significantly influence sales volume, with Tier 1 stores and Supermarket types generating higher sales.
- **Correlations and Patterns:**
    - Strong positive correlation found between `Item_MRP` (price) and sales.
    - Seasonal and promotional effects inferred from sales spikes align with known festival and marketing periods.
    - Outlet establishment year inversely correlates with sales, suggesting newer outlets may have fresher marketing or stock strategies.
- **Outliers and Anomalies:**
  Validated sales outliers correspond with promotional events and seasonal demand surges.

4. Feature Engineering & Model Selection

- **Feature Engineering:**
    - Converted categorical variables (`Item_Type`, `Outlet_Type`, `Outlet_Location_Type`, etc.) via one-hot encoding and label encoding.
    - Generated derived features:
        - Sales lags and rolling averages for outlets and products (where time-sequenced data exists).
        - Price buckets based on item MRP.
        - Age of outlet (`Current Year - Outlet_Establishment_Year`).
    - Reduced dimensionality by dropping redundant or low-variance features based on correlation analysis.
- **Model Selection:**
    - **Baseline:** Mean Sales Model, to provide a simple benchmark.
    - **Advanced Models:**

        - Random Forest Regression: For robust handling of non-linear relationships and categorical data.
        - Gradient Boosting Models (XGBoost, LightGBM): For high accuracy and feature importance analysis.
    - Time-series models like SARIMA / LSTM not applied due to lack of continuous time-stamped data in this dataset.

5. Model Performance Evaluation

Model	RMSE	MAPE	R^2 Score
Mean Sales Baseline	~2750	~45%	~0.00
Random Forest	~2130	~33%	~0.41
XGBoost	~1850	~28%	~0.53
LightGBM	~1770	~27%	~0.55

- Advanced models significantly outperform the baseline.
- Gradient boosting models show superior accuracy and efficiency.
- Feature importance analysis indicated `Item_MRP`, `Outlet_Type`, and `Outlet_Location_Type` among top predictors.

6. Business Insights & Recommendations

- **Sales Trends:**
  Seasonal peaks correlate with promotional periods and holidays, requiring adaptive inventory and marketing strategies.
- **Pricing Impact:**
  Item price is a powerful sales driver; price optimization can further enhance revenue.
- **Outlet Strategy:**
  Focus efforts on Tier 1 locations and supermarket formats for maximum sales impact.
- **Inventory Planning:**
  Increase stock ahead of high-demand periods identified through forecast patterns, especially for Food items and high-volume outlets.

- **Promotions:**
  Plan targeted promotions during historically lower sales periods to stabilize revenue.

7. Challenges & Future Improvements

- **Data Limitations:**
  Absence of transaction-level timestamps limits time-series modeling and trend extraction.

- **Feature Gaps:**
  Inclusion of external factors like marketing campaigns, competitor activities, and economic indicators would enhance predictive power.

- **Model Advancements:**
  Hybrid models combining time-series and feature-based forecasting could improve accuracy.

- **Automation and Retraining:**
  Implement automated hyperparameter tuning and periodic model retraining with fresh data to maintain relevancy.

Prepared by: Mayank Lalwani
Date: 31/07/2025