Approach for BigMart Sales Analysis

1. Problem Understanding

BigMart operates multiple outlets, and the goal is to predict sales for various products based on historical sales data.

The dataset consists of both store-level and product-level attributes, which impact sales. The target variable is `Item_Outlet_Sales`.

2. Data Preprocessing

- **Handling Missing Values:**

- `Item_Weight`: Imputed using median values based on `Item_Identifier`.

- `Outlet_Size`: Filled based on mode of similar `Outlet_Type`.

- **Feature Engineering:**

- Created a new feature: `Outlet_Age` = `2025 - Outlet_Establishment_Year`.

- Applied **target encoding** for `Outlet_Identifier`.

- Retained categorical features as-is since LightGBM handles categorical variables internally.

3. Exploratory Data Analysis (EDA)

- Identified distribution of sales across different stores and product types.

- Analyzed trends based on `Item_Type`, `Outlet_Location_Type`, and `Outlet_Type`.

- Checked correlation between numerical variables and visualized data distributions.

4. Model Selection & Training

- **Baseline Models:**

- Trained **Linear Regression** and **Random Forest Regressor** for initial benchmarking.

- **Final Model:**

- Used **LightGBM**, which handles categorical variables without additional encoding.

- Applied **hyperparameter tuning** using Grid Search.

- Performed **K-Fold Cross-Validation** to assess model performance.

5. Evaluation Metrics

- **Root Mean Squared Error (RMSE):** Measures prediction accuracy.

- **R^2 Score:** Indicates how well the model explains variability in sales.

- **Feature Importance Analysis:** Used LightGBM's feature importance plot to identify key predictors.

6. Submission Strategy

- **Final Predictions:**

- Generated test predictions using the trained LightGBM model.

- Saved final submission as `submission/submission_final.csv`.

- **Experiment Tracking:**

- Previous trials and different feature engineering approaches stored in `notebooks/experiments/`.

- Intermediate submission files stored in `submission/experiments/`.

7. Key Findings

- **Most influential features:** `Item_MRP`, `Outlet_Type`, and `Outlet_Age`.

- **LightGBM performed best**, achieving the lowest RMSE and highest R^2 compared to other models.

- **Store type and product visibility significantly impact sales predictions.**