BigMart Sales Prediction - Documentation

Overview

This notebook is designed for preprocessing and modeling sales data for a Big Mart sales prediction problem. The workflow includes data cleaning, feature engineering, exploratory data analysis, and training multiple regression models to predict sales.

Data Preprocessing

1. Loading Data

The dataset is loaded from CSV files (train\_data.csv and test\_data.csv).

Both datasets are combined for uniform preprocessing.

2 Handling Missing Values

Item Weight: Filled using the mean weight of items with the same identifier.

Outlet Size: Imputed based on the most frequent value (mode) within the same outlet type.

3 Feature Engineering

Item Fat Content: Standardized different variations (LF, low fat → Low Fat, reg → Regular).

Adjusted Item Fat Content for non-consumable items by setting it to 'NA'.

Item Category: Extracted from Item\_Identifier (Food, Drinks, Non-Consumables).

Outlet Age: Calculated from Outlet\_Establishment\_Year.

Item Visibility Transformation: Applied square root transformation to reduce skewness.

Sales Transformation: Square root transformation applied to stabilize variance.

4 Encoding Categorical Variables

Label encoding for identifiers (Item\_Identifier, Outlet\_Identifier).

One-hot encoding for categorical columns (Item\_Fat\_Content, Outlet\_Size, etc.).

Data Exploration & Visualization

Feature Distributions: Histograms and boxplots were used to inspect distributions.

Outliers: Addressed in numerical features using transformations.

Model Training & Evaluation

Models Used:

Linear Regression, Ridge Regression, Random Forest Regressor, Gradient Boosting Regressor, LightGBM Regressor

Gradient Boosting Regressor achieved the best performance in the direct models.

Then later used Stacking Regressor and achieved even better performance.

Feature Importance

Got feature importance values from the model as well as displayed using SHAP’s summary plot.

Output & Predictions

The test dataset was processed similarly.

Predictions were made using the best model.

Final predictions were saved for submission.

Conclusion

This notebook systematically cleans and transforms data, applies feature engineering, and evaluates multiple models to find the best one for sales prediction. The Stacking Regressor performed the best and was used to generate final sales forecasts.