BigMart Sales Prediction – Approach Note

Problem Statement and Objective

The objective of the given problem statement was to predict product-level sales across multiple BigMart outlets using historical sales data. The problem was formulated as a supervised regression task where the goal was to minimise prediction error while ensuring the model generalises well to unseen outlet–item combinations. In addition to predictive accuracy, understanding the influence of product and outlet characteristics on sales was an important consideration.

Data Understanding and Exploration

The dataset consisted of product attributes, outlet characteristics, and historical sales values for the training dataset, while the test dataset contained similar features without target values. Initial exploratory data analysis was performed to identify missing values, inconsistent categorical labels, skewed feature distributions, and potential data quality issues.

Key observations included missing values in item weight and outlet size, inconsistent representations in fat content categories, and zero values in item visibility which were not practically meaningful. Sales distribution analysis indicated nonlinear relationships between price, outlet characteristics, and sales, suggesting that linear assumptions alone may not be sufficient.

Data Preprocessing and Feature Engineering

Missing values in 'Item_Weight' were imputed using item-level averages to preserve product-specific characteristics. Missing values in 'Outlet_Size' were imputed using the mode within outlet types to maintain structural consistency across stores. Inconsistent categorical labels were standardised to avoid duplication during encoding.

Feature engineering steps included deriving outlet age from establishment year and grouping item identifiers into broader product categories to improve predictive signal. Zero visibility values were replaced with item-wise average visibility to reflect realistic shelf exposure.

To ensure consistent feature representation, training and test datasets were combined temporarily for preprocessing and one-hot encoding, preventing column mismatch during model inference.

Model Experimentation and Evaluation

Multiple models were evaluated in an iterative manner. During experimentation, Linear Regression produced unstable predictions with extremely high RMSE, which led to further investigation into multicollinearity and adoption of Ridge Regression. Ridge Regression was subsequently applied to introduce regularisation, resulting in improved stability and reasonable performance.

Tree-based ensemble models, including Random Forest and XGBoost, were then explored due to their ability to capture nonlinear relationships and interaction effects between features. Model performance was evaluated using Root Mean Squared Error (RMSE) on a validation dataset created through train-validation split. Random Forest achieved the lowest RMSE among all models, indicating stronger generalisation capability for the given dataset.

Final Model Selection and Prediction

Based on validation performance and model stability, Random Forest was selected as the final model. The model was retrained on the complete training dataset before generating predictions for the test dataset. The final submission file was prepared by combining predicted sales values with original item and outlet identifiers.

Conclusion

The final solution followed a structured and iterative machine learning workflow involving data cleaning, feature engineering, model experimentation, and performance-driven model selection. The approach prioritised robustness and generalisation over aggressive tuning, resulting in reliable sales predictions and a strong leaderboard performance within the top percentile of participants.