Big Mart Sales Prediction – Case Study Approach Note

The objective of this case study was to predict the sales of products across different Big Mart outlets based on various product and store-level attributes.

1. Exploratory Data Analysis (EDA)

I began by first understanding the data types present in the data, identifying the numeric (4) and categorical (7) features. The next step was to check for any missing values. Missing values were primarily found in Item_Weight and Outlet_Size (both train and test sets). Additionally, Item_Visibility contained a substantial number of zeros, which were treated as missing values, by definition.

Each of these features was imputed logically using relevant relationships within the data rather than simple mean/mode replacements. For example, missing Item_Weight values were imputed using the average weight of the corresponding product type. Categorical features (e.g., in Fat_Content and Item_Type) were dealt with by combining labels to reduce label imbalance in the data.

Univariate and bivariate analyses revealed data skewness, correlations with the target, and several opportunities for meaningful feature transformations.

The most impactful stage for model performance was feature engineering, where several transformations were applied:

- Created MRP groups to capture multimodal pricing behavior.
- Applied log transformation to Item_Visibility to handle skewness.
- Simplified Outlet identifiers into 3 representative groups.
- Converted Outlet Establishment Year into Outlet Age, improving interpretability.
- Categorical columns were label-encoded (where appropriate) to limit dimensionality inflation from one-hot encoding.

Additionally, iterative feature addition/removal/modification was performed by cross-validation results and leaderboard feedback during the model training.

2. Model Training and Optimization

For model training and optimization, I used optuna, a hyperparameter tuning library which proves to be much faster and efficient than traditional search algorithms. A total of 4 models were evaluated – GradientBoosting, LightGBM, XGBoost and CatBoost.

Due to time constraints, a small number of trials were performed on each with basic data after EDA to finalize a model. XGBoost and CatBoost were found to be the best performing model in terms of leaderboard scores. Both were giving close leaderboard scores with basic feature engineering.

With the models finalized, I iteratively added/removed/modified some features in the EDA notebook to increase the predictive power of the model. Finally, the final RMSE score I could achieve was ~1143.4 leading to a rank of **112** out of a total of ~53k registered participants. This was ultimately achieved by using CatBoostRegressor and performing a statistically backed output clipping to reduce error.

3. Key Learnings

The project reinforced the importance of domain-driven feature engineering — combining retail business insights (e.g., store type and location, product visibility, pricing tiers) with technical modelling rigor significantly improved predictive power. Iterative experimentation and systematic validation were key in refining the final solution. I believe it is possible to further improve the rank, using methods such as stacking different models to capitalize on the strengths of different types of regressors.