ABB Senior Data Scientist Assignment Summary

Project Summary: Big Mart Sales Prediction

The objective was to build and evaluate a robust machine learning model to predict retail sales for a variety of products across different store outlets.

Methodology

The process can be broken down into the following steps:

1. Imported Relevant Libraries
2. Data Cleaning
   - 2.1. Formatting & Item_Fat_Content
     The formatting of all the columns in train.csv was done. Replaced 'low fat' &'reg'
   - 2.2. Check Column Type, NaN, Blanks and Zeros
     Item_Weight, Item_Visibility & Outlet_Size had some issues.
3. EDA
   - 3.1. Target Variable Analysis
     Item_Outlet_Sales is right skewed.
   - 3.2. Non-Target Variable Analysis
     Numerical & Categorical Column analysis
4. Feature Engineering
   Item_weight had 6.17% values 0 which were replaced by 1 before log transformation applied to columns, One-hot encoding,
5. Model Fitting
   - 5.1. XGBoost -> Hyperparameter tuning
   - 5.2. Recreated Feature Engineered train & test datasets with more features (Min-Max Scaling, Log Transform,
   - 5.3. XGBoost -> Hyperparameter tuning with lot more features
   - 5.4. Cross validation
   - **5.5.** Loop of Creating 'Better' Feature Engineered train & test + Model fitting **(Linear Regression, Ridge, Lasso, Random Forest, Gradient Boosting, and XGBoost) +** Hyperparameter tuning **(3 times in total from 5.1 to 5.5)**

Conclusion

Gradient Boost is the best model. To improve the performance, feature engineering should be done in more detail.

Note: Total time spent on the problem 12 hours. The rank improved from 5500 to 3478