**ML Model for Item Outlet Sales Prediction**

Through this document I shall take you through my approach for this problem, underlining the thought process for the steps taken.

1. **Data Preprocessing**

Upon doing the initial checks of the data, I realised there were to columns with missing values, **outlet size** and **item weight** with 28% and 17% values missing respectively.

For treating this missing values, I did the mean imputation for the item weight as the closes approximation and replaced Nans in outlet size with ‘no data,, as there didn’t seem to be a good approximation of the outlet size (although I noticed mostly Tier 2 stores were small, but still not a good case for other tiers)

1. **EDA**

I wanted to understand the distribution and characteristics of all columns, for which I performed univariate analysis for all columns (mainly distributions) and few bivariate (item\_mrp vs item outlet sales). After seeing a relation (item\_mrp vs item outlet sales). , I decide to also produce a correlation map

1. **Feature Engineering**

In my experiments I performed various iterations of creating features but only two stood out as meaningful. Outlet\_age and Item\_Visibility\_Scaled.

**Outlet age** made more sense to me as a derivative of the **Outlet\_Establishment\_Year**  as it seemed more like a direct indicator and also its large values could affect the predictions.

For **Item visibility,**  I could see zero values which logically doesn’t make sense as zero visibility is not possible. Hence, I replace the 0 values with 0.01..to add some information.

1. **Label Encoding**

For categorical variables, I used label encoding. I didn’t feel the need for higher degrees of encoding (target,etc)

1. **Model Building**

I tried various iterations, started with building base models LR & RF. Then tried a few more tree based like xgb and catboost. Upon observing the scores on the leaderboard, I decided t try Ensemble models which in turn resulted in the best performance building on my previous experiment of seeing the RF and XGB running the best.

I decided to build it using Random forest, XGB with Stacking regressor.

1. **Hyperparameter tuning and CV**

Used GridsearchCV for **Random Forest** and **XGBoost** to find the best hyperparameters (like max\_depth, learning\_rate, n\_estimators) using cross-validation and cross validation for Stacking

1. **Feature importance and SHAP**

Upon plotting the Feature importance and SHAP values, it was clear that Item\_MRP, Outlet\_type were amongst the top predictiors

1. **Evaluation – Optimised for RMSE and the used the Analytic vidhya score for comparison**