HSBC Hackathon

- Vikas Solanki, 14113131
B.Tech 4th year

Data Exploration - EDA was done by analysing the statistics (mean,median, percentile values) of each feature. The values of “mid” are plotted against “datetime” to see the overall trend and seasonality. Variations of data with hour,minute and second are also observed. Moving average and variance of the series are determined and transformations such as log and difference are done in order to make the series stationary.

Feature Extraction - Trained the initial models with variables such as hour, minute, second, imbalance, mid and realised the importance of variable “mid”. Lagged values of “mid” are extracted at intervals of 10s. Further, the intervals are reduced sequentially and value of RMS of both train and test dataset are calculated. Finally, lags at interval 1s are found to give the lowest RMS value with out-sample data.

Prediction with ML model - Different models such as linear regression, tree-based algorithms,ARIMA, FB Prophet, simple NN and LSTM are tried on the dataset. Following observations are made -

1. Linear Regression - Found to be the best model. An intuition for expecting good results is the same loss function in Linear Regression and our problem. High variance was present in the predictions since all the lag-features are highly correlated with each other.
2. Tree-based Models - Tried XGBoost, Random Forest and Light GBM. Trees were prone to overfitting and tuning parameters to reduce the same made the model extremely slow to work with.
3. Neural Networks and LSTM - Sequence based NN gave much higher value than the models mentioned earlier even after tuning several parameters. One reason might be the lack of data in both sequence length and training data as the latter keeps getting reduced as more lag fetaures are formed (more nan values to be dropped).

In order to reduce the variance and overfitting of linear regression model, two methods are adopted i.e selecting subset of features and using regularised version such as Ridge and Lasso. Ridge regression is selected and its parameters are tuned to give best results on out-sample data. Next, backward selection of features are done with the help of sklearn’s RF Estimator.

Final results are as follows -

Training parameters:
'futMid7S','futMid8S’,'futMid10S','futMid11S','futMid12S','futMid13S','futMid14S','futMid15', 'futMid20S','futMid22S', 'futMid23S','futMid24S','futMid25S','futMid26S','futMid27S' ,'futMid28S','futMid29S'

RMS of out-sample data : 1.1736 with #predictions = 52223
RMS of in-sample data : 0.9118