List of Experiments

October 25, 2022

0.1 Train and Test Data:

Data is split into 70% Train and 30% Test with stratification. Any trained model is tested on unseen data to see its performance before submitting the model. The random seed is kept at 42

0.2 List of Other Experiments:

0.2.1 Light GBM:

- Dataset1 -> 0.815456 (Denoised data) [196 features]
- Dataset2 -> 0.81627 (Dataset1 + PCA(up to 90% Variance) + Removed Correlated Features (>=0.95)) [272 features]
- Dataset3 -> 0.81474 (Dataset2 + Tree based feature reduction using feature importance once) [247 features]
- Dataset4 -> 0.815423 (Dataset3 + Tree-based feature reduction repeated till all features are used in making the trees) [231 features]
- Dataset5 -> 0.81527 (Dataset4 + Meta Features CatB) [234 features]
- Dataset6 -> 0.81609 (Dataset4 + MetaCatB (Level I only) + MetaLGB (Level I only)) [237 features]
- Dataset_noiseoof -> 0.81636 (Dataset1 + MetaCatB only) [199 features]

0.2.2 Cat Boost:

(There's some randomness associated with CatB and hence two attempts were recorded on the test to see its performance)

- Dataset1 -> Attempt 1: 0.818136 | Attempt 2: 0.818365 (Denoised data) [196 features]
- Dataset2 -> Attempt 1: 0.8175479 | Attempt 2: 0.817852 (Dataset1 + PCA(upto 90% Variance) + Removed Correlated Features (>=0.95)) [272 features]
- Dataset3 -> Attempt 1: 0.8175479 | Attempt 2: 0.817384 (Dataset2 + Tree based feature reduction using feature importance once) [247 features]
- Dataset4 -> Attempt 1: 0.81747 | Attempt 2: 0.817621 (Dataset3 + Tree based feature reduction repeated till all features are used in making the trees) [231 features]
- Dataset5 -> Attempt 1: 0.8174389 | Attempt 2: 0.817689 (Dataset4 + Meta Features CatB) [234 features]
- Dataset6 -> Attempt 1: 0.8181689 | Attempt 2: 0.817972 (Dataset4 + MetaCatB (Level I only) + MetaLGB (Level I only)) [237 features]
- Dataset_noiseoof -> Attempt 1: 0.81843 | Attempt 2: 0.81843 | Attempt 3: 0.81849 [199 features]

0.2.3 Other Experiments:

(Listing down other experiments that were done but didn't work out quite well for us)

- (1st submission ever) Random Forest on columns without missing values on just 50,000 records (Submission score: 0.808)
- Single boosting classifier (either XG/LG/Cat) on
  - Columns with less than 70% missing data
  - Entire Dataset
- Making datasets (one with columns containing less than 30 percent of missing data only, other with missing values, try models on both, then vote/stack it to get predictions)
- Ensembling different models over the different datasets used above
- Stacking the different models on different classifiers (SVC, Linear SVC, RF, Logistic Regression) (Submission score saturated at around 0.8225)

  For this, the data was divided into three segments. Train1, Train2, Test. Train1 was used to train the initial models. Train2 is used to train the stacked classifier weights, Test was to test the model performance
- Neural Networks
  - Addressing missing values (Median/Mean imputation)
  - Try implementing TabNet (the attempt was unsuccessful for this multivariate classification problem)
- Brute Force Feature Engineering (Ratio Features) [Third Best Model]
  - An estimate of numerical columns was obtained by analyzing the number of unique values in each column
  - In the end, around 36 features were selected to account for computation intensity based on missing values, numerical features, correlation, etc
  - Pair-wise ratio is considered a new feature and if the new feature improves the model score by threshold (0.0003), the new feature (colx/coly) is taken
  - Reasoning: In Financial Data, ratios could be a key metric
  - Drawbacks: Searching the entire space requires a lot of computational time. Considering three-pair, four-pair ratio features will require even higher computational time
  - Works best with data knowledge in real life. The black box way of doing it is not good
- Ensemble of XGBoost, CatBoost, LightGBM on Dataset_noiseoof (One layer meta-features) was our second Best Model. It achieved a score of 0.8238 on Submission data.

0.3 Best model:

- Dataset
  - Used dataset1+ metaCATB+ metaLGB (Layer 1) + metaCATB2 + metaLGB2 (Layer 2) [208 features]

- Model contains ensemble of 4 learners(XGB,LGB(DART),LGB(GBDT),CATB) where final output is soft-voted prediction of these 4 classifiers
- Submission score we got with this was 0.8241