COURSE NAME - Risk & Fraud Analytics

Fraud Detection Project Log

Name: Mukhtadir Syed (Full time elective- A)	Best Gini Score: 0.536672
Email: mukhtadir@student.ie.edu	Std Deviation: 0.09232798438021768
Username: Mukhtadir	Total Submissions- 51

Things that were tried and worked:

- The data cleaning measures were slightly impactful and did not move the score by too much
- Filling Missing values on the test set- imputed categoricals with mode and continuous with median
- Dropping Columns that were not adding too much value to the model because mostly they were of same value (marginal improvement in score)
- SMOTE was fairly effective in pushing the score well beyond 0.5
- A lot of research was conducted on SMOTE technique and the different types
- SMOTE-NC worked best which is for both categorical and continuous data
- RandomForest base score was 0.47, I was able to push it to 0.5366
- RandomForest randomized search CV was the most effective technique in this case
- The final score differs due to the stochastic nature of the predictions and the score varies from 0.51 - 0.5366

Things that were tried and did not work:

Different algorithms that were tried:

- ❖ XGBoost:

This algorithm is always a bit tricky to train because of its sensitivity to hyper parameter tuning. Grid search is always a disaster and never gives good results from past experience. Randomized search CV was not able to beat the score of Random Forest using the same technique. Another hyper parameter technique that was tried i.e. Hyperopt (a bayesian based hyperparameter technique).

- ❖ TPOT Classifier

A fair bit of research was done on this genetic algorithm and its Python library. However, more time was needed to configure this one properly. Also, the basic configuration itself took a lot of time to train and the scores given by 5 generations were not that great. Lack of knowledge and experience definitely played a part here. Genetic algorithms are definitely very interesting and in future more research will be conducted on this.

- ❖ Different SMOTE techniques

These are the SMOTE techniques that were tried after a lot of research:

- SMOTE
- SMOTE-NC
- Borderline SMOTE
- SVMSMOTE

Of all this SMOTE-NC which is for both categorical and continuous data, worked best. Borderline SMOTE and SVM smote are quite interesting but the results were not as good as SMOTE-NC. Also, scaling the dataset before applying SMOTE did not make much difference.

- ❖ Scaling the data- Even though SMOTE uses K nearest neighbor which is based on distance, scaling before SMOTE did not make any difference. Also Random forest being a tree based algorithm, so scaling does not matter too much for it

Things that were not tried but would like to try in future (Wishlist):

- ❖ Neural Networks- I have just started deep learning as an elective and would love to try it on this dataset once I have gained more experience and knowledge
- ❖ More Genetic algorithm techniques
- ❖ Using Robust Scaler instead of StandardScaler (this was missed)
- ❖ Catboost and LightGBM- Due to lack of time, these were not tried
- ❖ More feature engineering techniques- would have helped to know the columns here
- ❖ Better ways to generate synthetic data (if there are any)
- ❖ Usually boosting beats bagging in most cases, may be more time to tune hyper parameters on XGBoost would have given a better score. However, boosting also requires more data and this dataset was fairly small with less than 1000 examples