IEE CIS Fraud Detection

Project 01 Report by

Md Raisul Islam

Email: shuvo714@gmail.com

Abstract

The data comes from Vesta's real-world e-commerce transactions and contains a wide range of features from device type to product features. The goal is to predict the probability that an online transaction is fraudulent, as denoted by the binary target isFraud.

The Dataset was given to the Kaggle IEEE-CIS Fraud Detection Competition.

The data is broken into two files identity and transaction, which are joined by TransactionID. Not all transactions have corresponding identity information.

The training dataset consists of more than 400 features and 5.9 million samples. This is supervised binary classification problem and goal is to predict if a credit card transaction is Fraud based on input features mentioned below.

1. Methodology

I have used CatBoost Classifier, Light GBM -weighted, Cat_5fold models and appended all together at the end. This gives me an accuracy of 0.942334 Public score and 0.912191 Private score in Kaggle submission.

CatBoost is a high-performance open-source library for gradient boosting on decision trees.

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel, distributed, and GPU learning.
- Capable of handling large-scale data.

2. Result Analysis

I have used Single CatBoost Modeling method, trained the data and then fit the model with CatboostClassifier Parameters.

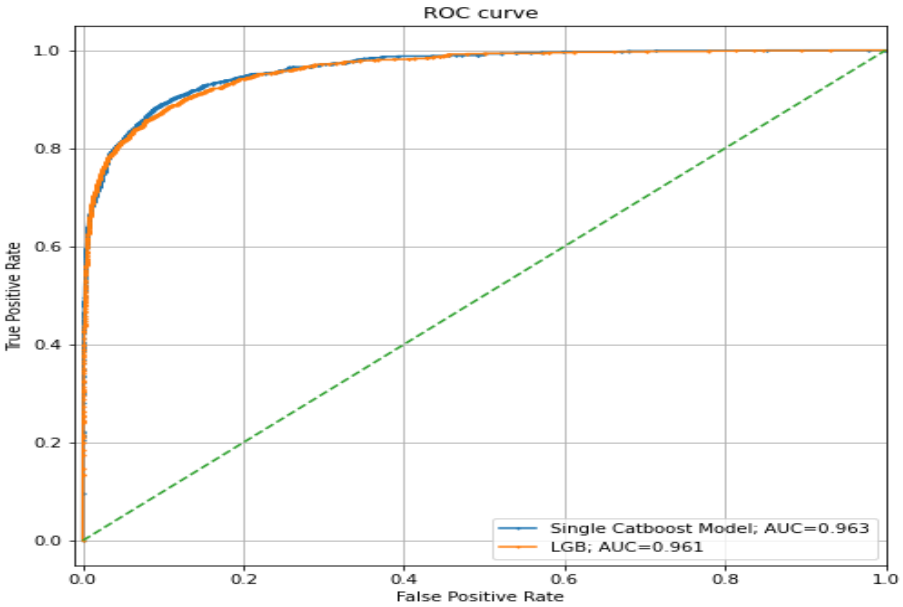
```
model_single = CatBoostClassifier(**params)
model_single.fit(train_data, eval_set=holdout_data, plot=True, verbose=False)
```

Then used Light GBM with selective parameters, trained the model & finally after 736 iterations the best AUC score, I got was 0.96051.

```
bst.best_score
defaultdict(collections.OrderedDict,
            {'valid_0': OrderedDict([('auc', 0.9605100655661709)])})
```

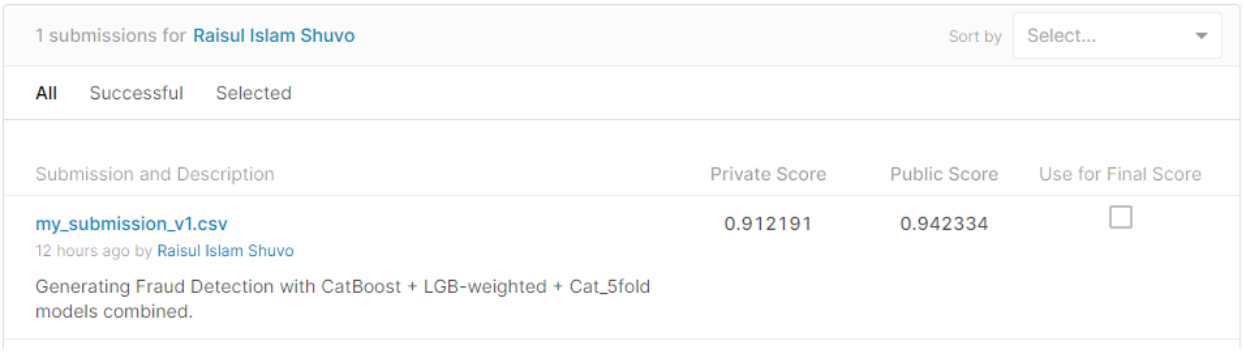
Then I have used the CV approach with 5-fold using the full dataset. After that, I have appended the CatBoostClassifier with the K-fold model and used mean of the model's predictions for the final submission test dataset.

3. Performance Evaluation



The ROC plot compares the Single CatBoost model with the Light GBM AUC scores. It is clear that Single CatBoost model performed better.

The Kaggle submission for the competition gives the following evaluations.



4. Conclusion

The models used for this notebook can be improved with further Hyperparameter tuning. Although the current Score for the final model evaluation was 0.96051 & 0.942334 in Kaggle submission, it can be imrpoved with Tree based Gradients like XGBoost, Neural Network, etc. It seems having a powerful machine support like higher CPU, GPU & Memory can also play a significant role in improving the model's performance.

References

[1] Vesta Corporation. *IEEE-CIS Fraud Detection.* IEEE Computational Intelligence Society (IEEE-CIS), 2019

[2] Lyalikov Artyom. *Kaggle User, IEEE Dataset Notebook (lgbm_baseline + small fe(no blend)).* Kaggle, 2019

[3] CatBoost. *catboost.ai* , Yandex.

[4] Microsoft Corp. *LightGBM's documentation*

[5] Zak Jost. *Github User (zjost/cc_fraud_proj).* Github.