Upskill ISA Project - 1

Md. Abrar Jahin

Department of Industrial Engineering and Management

Khulna University of Engineering and Technology

June 10, 2021

Abstract

After getting shortlisted among tons of candidates for the Upskill ISA Program with Intelligent Machines 1.5 months long Machine learning course, the former Kaggle Competition dataset **IEEE-CIS Fraud Detection** has been selected as project-1. After working on it for a week, I learned to tackle the challenges with gigantic dataset with huge number of features and develop a sophisticated model that performs well to predict a transaction as fraudulent one or not.

1 Methodology

There are 590540 observations and 435 features in train transactions data, 144233 observations and 41 features in train identity data. Our target column is **isFraud** in which binary 0 and 1 represents not fraud and fraud respectively. There are redundant V columns (V1, V2,..V339) among which most of them contain equal number of null values. All columns sharing equal number of null values were grouped into distinct groups. Then those groups were further investigated how much they are correlated. Surprisingly, they all follow a correlation pattern and the big idea was to subgroup the V-columns which are highly correlated. Then the challenge was to find one suitable column from each subsets and it was done on the basis of the most unique value. Choosing a single column containing the highest number of unique values refers to containing more new information for the model training, because it's better to remove duplicates and provide correlated features containing new information.

The D Columns are 'time deltas' from some point in the past. The D Columns were transformed into their point in the past. This will stop the D columns from increasing with time. The formula is D15n = Transaction_Day - D15 and Transaction_Day = TransactionDT/(24*60*60). Afterwards it was multiplied by negative one.

After adding 28 new features, 219 V-Columns were removed from correlation analysis. Then each of the 242 features were checked for "time consistency". 242 models were built to be trained on the first month of the training data and then

This reported was presented at the Upskill ISA Program with Intelligent Machines-2021.

predicted the last month of the training data. Both training AUC and validation AUC were expected to be above AUC = 0.5. It turned out that 19 features failed this test, so they were removed from the training data. Additionally, 7 D-columns that are mostly NaN were removed.
XGBClassifier had been used with GPU to train the model and the metric was roc_auc_score. The learning rate was 0.02 and of our model and Used GroupKFold Cross Validation by splitting the data into 6 folds where 75% was train data and 25% was test data.

2 Result Analysis

Upon submission on Kaggle, this model obtained public score: 0.950844 and private score: 0.922507 which is not quite within the range of competence while considering the scores of the top performers but still decent. The main differentiator in the score is feature selection. Those who will choose or modify the features more wisely can obtain significantly better score. In case of Kaggle, 20% of test data is selected as public and the rest of the 80% is private data. In this ISA program, only public score is to be submitted, so basically it is evaluated over 20% of the test data.

3 Conclusion

The model can be developed more robust utilizing GridSearch or RandomSearch Cross Validation but the limitation is resource. If the machine possesses more GPU, CPU, RAM then it's convenient trying out different models to obtain better result. Moreover, the features can be reduced in number carefully without losing much new information. For instance, the V columns can be further shortlisted to avoid confusing the model to learn new information.

Acknowledgements

To the StackOverflow, for all the required Python environment solutions and predefined handy functions posted by the developers.