Project 1 - IEEE-CIS Fraud Detection

Upskill-IM Cohort Program

Abstract: IEEE-CIS works across a variety of AI and machine learning areas, including deep neural networks, fuzzy systems, evolutionary computation, and swarm intelligence. In this closed Kaggle competition they partnered with the world's leading payment service company, Vesta Corporation, seeking the best solutions for the fraud prevention industry.

In this competition, the goal is to benchmark machine learning models on a challenging large-scale dataset. 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 competitors also have the opportunity to create new features to improve their results.

Methodology:

1. *Importing Packages and Libraries:* In the beginning, I imported all necessary packages and libraries in the Kaggle notebook environment.
2. *Importing Data and Reducing Memory Usage:* In this step, I read the csv files of the given dataset, stored them in separate dataframes and also reduced the memory usage of these dataframes using a previously defined function.
3. *Solving Column Mismatch:* I defined a custom function to check if there is any mismatch of columns between the train_identity and test_identity dataframe. After that, I solved the issue by manually renaming the column names.
4. *Joining Transaction and Identity Data:* Here, I joined the transaction and identity dataframes on the basis of the common column 'TransactionID'.
5. *Data Engineering:* In this step, I analyzed the different data types of the feature variables and the missing data.
6. *Feature Engineering:* In this step, I grouped the features on the basis of NAN(missing) values, chose the columns with correlation coefficient value greater than 0.75 to create groups, took the largest subgroups with common elements from them and finally chose the subgroups with the highest number of unique values as features.
7. *Data Splitting for Training:* Here, I split the data into features and targets for each of the train and test purposes.
8. *Feature Encoding:* Now I encoded all the categorical feature variables using LabelEncoder and scaled all the numeric feature variables.
9. *Training:* During training, I trained an xgboost model with referenced hyper-parameters on the training data and then tested the model on the test data.
10. *Confusion Matrix after Training:* Here, I visualized the confusion matrix of the performance of the model on the train and the test data.
11. *Creating a Submission File:* At the final step, I created a csv file to submit to the competition evaluation.

Result Analysis: Submissions are evaluated on area under the ROC curve [1] between the predicted probability and the observed target. For each TransactionID in the test set, I had to predict a probability for the isFraud variable. My model achieved a decent public score of 0.922061.

Conclusion: There are possibilities of future improvements of my current work on fraud detection. For example,

1. Handling Target Class Imbalance: There is a huge imbalance in the dataset as it mostly consists of False values of the isFraud column. It can be improved in order to achieve a better score.
2. More EDA: Better exploratory data analysis can be applied in this work.
3. More In-depth Feature Engineering: Various other methods can be applied in order to find out better features.
4. Application of Ensemble Method: State-of-the-art models can be combined using ensemble method to achieve better prediction.

References:

1. https://en.wikipedia.org/wiki/Receiver_operating_characteristic