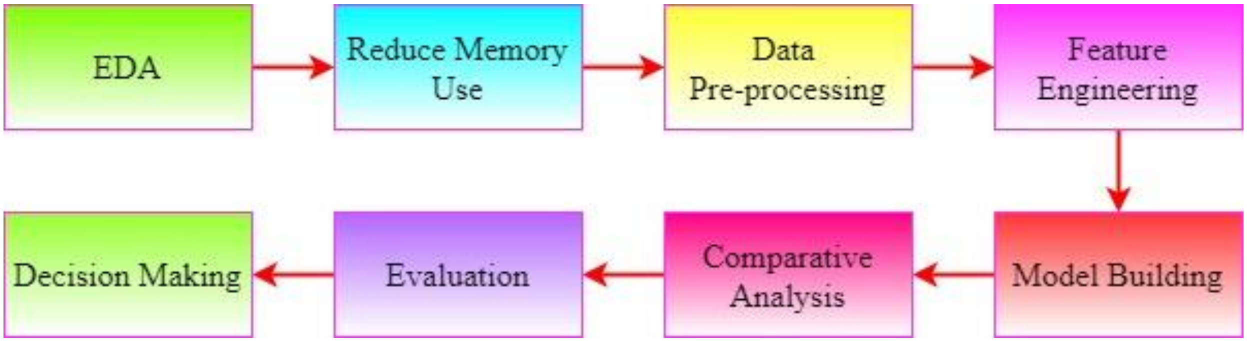
Project 1

IEEE-CIS Fraud Detection

1. Abstract:

This project is inspired from IEEE-CIS Fraud Detection Kaggle competition. This competition introduces a binary classification problem. The target variable is a binary attribute and this task aims to classify users into "fraud" or "not fraud" as well as possible. In this project, I have used the mentioned dataset to observe various tree based models' (xgboost, lgbm, random forest) performance. Consequently, a tree based architecture made out of xgboost emerged triumphant with the public Score of 92.95%. This work proposes a model that focuses the need for most feature engineering to get optimal performance.

2. Methodology:



This figure includes the logical diagram of the entire workflow. EDA, Reduce memory use, Data pre-processing, Feature engineering, Model building, Comparative analysis, Model Evaluation, and Decision Making are the eight main segments of the workflow.

2.1 Exploratory Data Analysis(EDA)

After understanding the problem domain, it is important to know about every pros and cons of the dataset. EDA has done to uncover the underlying structure of the dataset. EDA has exposed important patterns, and relationships that were not readily apparent.

2.2 Reduce Memory Use

I have observed after merging transaction & identity data file the dataset size arises 1 GB+. While running the notebook on Kaggle, this size exceeded memory limit. For this reason, memory size has reduced by changing datatype to manage this large volume of data. Moreover, python's built in 'del' function has also used to delete the data frames those were not needed anymore till the final execution.

2.3 Data Pre-processing

- The train and test dataset is separated into two data file (transaction and identity). In consequence, the dataset has concatenated on TransactionID for training and testing.
- There was some mismatch in train and test features. Hence, the mismatched columns of the test dataset were renamed to match the columns of the train dataset.
- The original dataset contains huge number of null cells. To handle these null cells, firstly the features containing more than 100000 null values have removed.
- The null values of the feature containing expected amount of null values have filled with mode or mean considering the particular feature's requirement.

2.4 Feature Selection

- Total number of null values in a particular column has taken into account primarily to select the features. Features containing more than 1 lac null values have removed.
- Correlation has measured related to Fraud to find out the best feature. Features with less correlation have dropped from the train & test dataset.

2.5 Model Building

Three tree based classifiers have been utilized to build model such as XGBoost, LGBM, Random Forest.

2.6 Comparative Analysis

Hyper parameter tuning has done to compare the effect of different parameters of the mentioned classifiers. Fine-tuned hyper-parameter has used to train the dataset.

2.7 Evaluation

Model built with three tree based classifiers was evaluated with the full test dataset which contains more than 560000 rows.

2.8 Decision Making

XGBoost based model performs better than LGBM and Random Forest based models.

3. Result Analysis

Classifier	Public Score
XGBoost	92.95
LGBM	90.25
Random Forest	87.62

- According to the accuracy (Descending order):

  XGBoost > LGBM > Random Forest
- According to Time Complexity (Descending order):

  Random Forest > XGBoost > LGBM

Conclusion:

The 0.929595 score is decent but definitely there is more room to improve. Better selection of feature, feature engineering, hyper parameter tuning and model ensemble methods will help to push the result.

About Author

Khadija Akter Lima
Recent Graduate
Highly interested to get involved with Data Science related project.
Previous Area of Work: Information Extraction System, Name Classification, Image Classification.
Current Interest: Fintech
E-mail: khadijalima2017@gmail.com