Machine Learning Engineer Nanodegree

Capstone Project: Credit Card Fraud Detection Model

Akbarali Shaikh July 19, 2020

Table of Content

1. Project Overview

Banking industry works on credit and income from interest which can be in a form of mortgage, other types of loan or from Credit Card. As cost of having credit card reduced it got broadly accepted in society and many people started having credit card. As the credit card reached the masses, new ways of fraud also came to existence. Today, CC fraud is a very big problem for a banking industry and it is plaguing financial industries for years, some countries it is more less problem than others

Solving this problem could bring about reduced volume of fraudulent transactions in financial industries hereby saving them a huge volume of money lost and this problem can be avoided by using predictive analytics where machine learning algorithms are used to detect fraud patterns and determine future probabilities and trends.

The dataset to be used for this project was gotten from data.world

2. Problem Statement

Just by looking at the transaction, it is often difficult know which is fraudulent or genuine. We have a binary classification problem i.e. either a transaction is fraudulent or genuine. Objective is to create a model that accurately predicts whether a transaction is fraudulent or not.

To solve this problem, first we will baseline the model and than run a predictive model on different machine learning algorithms on same dataset to see which will give the best performance which is better than the others,

These algorithms include Naïve Bayes Classifier, SVC, GaussianNB, LogisticRegression, KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier, BaggingClassifier, ExtraTreesClassifier, XGBClassifier and AdaBoostClassifier.

Dataset is normalized and all columns names changed before it is shared, it could also be a case where some of the features were reduced from original dataset, no other form of reduction would be made going forward. All features provided would be used in building the model.

3. Metrics

In binary classification model, accuracy with minimal type 1 and 2 error is an important metric.

Type I error also referred as False Positive: model predicts positive when it isn't.

Type II error also referred as False Negative: model predict negative when it isn't.

Accuracy is an important measure of binary classification model it shows how accurate the model predicts true and false.

In our case, we have a highly imbalanced dataset, other evaluation metrics consider such class imbalance:

1. Precision
2. Recall

Recall measures the fraction of fraudulent transactions that are correctly predicted while measures that fraction of cases predicted to be fraudulent that are truly fraudulent.

Precision and Recall are defined as follows:

Recall: When the focus in catching all fraudulent transactions even in case if few genuine transactions are also categorized as fraudulent then this is a go-to metric. When the cost of failures is high, we want to recall the rate to be high.

Precision: How accurate the model performs e.g., correctly categorizing true as true. Ideally. When raising false alerts is high we want to have

high precision

F beta score: It combines precision and recalls into one metric. The higher the score the better model. Value high than 1 means more emphasis on Recall as compared to Precision and if less than 1 it is vice versa.

F1 score (beta=1) It's the mean of precision and recall

F2 score (beta=2) It's a metric that combines precision and recall, putting 2x emphasis on recall - consider using it when recalling positive observations (fraudulent transactions) is more important than being precise.