Machine Learning Engineer Nanodegree

Capstone Proposal

Christopher Kim

September 9th, 2019

IEEE-CIS Fraud Detection

Domain Background

Recent years, we've been hearing major identity data breach stories in the news so much often. ACI Worldwide (an electronic payment systems company) estimates that 46% of Americans have had their card information compromised at some point in the past 5 years. Of those compromised information, a lot of the cases are credit card information that may be used in fraudulent transactions. Fraudulent credit card transactions cost financial damages for consumers and businesses.

At any given moment, millions of credit card transactions occur in the world. With the rise of electronic commerce businesses, the credit card transactions became even more common in the past couple of decades. With such a massive number of compromised credit card numbers, fraudulent credit card transactions are not so rare occurrences any more. In 2018, the Federal Trade Commission processed 1.4 million fraud reports totaling $1.48 billion in losses.

Preventing credit card fraud is an essential part of credit card payment processing. Improving fraud detection system will save a lot of money and improve consumers' experience.

Problem Statement

Using Machine Learning technique to detect and prevent fraudulent transactions is already saving millions of dollars a year. Researchers from the IEEE Computational Intelligence Society (IEEE-CIS) want to improve this figure, while also improving the customer experience. On the Kaggle platform IEEE-CIS has opened a competition IEEE-CIS Fraud Detection competition. Building a machine learning model that scores high accuracy on detecting fraud in transactions will be the problem to solve.

Datasets and Inputs

The dataset is provided by Vesta Corporation, a payment service company as a part of IEEE-CIS Fraud Detection competition. The data is broken into two files identity and transaction. The identity and the transaction sets are joined by TransactionID.

The training set of the identity data has 144K rows with 41 features, and the transaction data has 507K rows with 393 features. Most of the features of the identity data has been masked as “id_” fields to be compliant with the security partner T&C. The transaction data includes Vesta engineered rich features, including ranking, counting, and other entity relations.

Solution Statement

Building a machine learning model that scores high accuracy on detecting fraud in transactions will be the solution. It will be a classification model that determines the probability of a transaction being fraudulent. I will begin with exploring the dataset and examine the relationship between feature values and the labels, and make necessary transformations on features to be used for the learning algorithm. The intuition on the model that might perform well would be DNN or Gradient Boosting model. I will try those models and select the model that performs well in terms of accuracy and efficiency.

Benchmark Model

The Decision Tree would be one of the most basic models for classification binary decision making. The selected model should perform significantly better than Decision Tree model. I will compare the accuracy of the prediction of the solution model with the Decision Tree model.

Evaluation Metrics

The competition submissions are evaluated on area under the ROC curve between the predicted probability and the observed target. I will use the accuracy of the prediction and the area under the ROC curve for the evaluation metrics.

Project Design

The project will be developed using Python language and using Jupyter Notebook. The following packages/frameworks will be used.

- Numpy, MatPlotLib, Seaborn, Pandas
- SciKit-Learn, TensorFlow, Keras

The project workflow will follow the following steps.

- Explore the data set and analyze the features
- Evaluate features and relevance with the desired prediction model
- Cleanse and transform data set to be used for machine learning algorithm
- Evaluate metrics on Decision Tree algorithm
- Evaluate metrics on DNN model
- Evaluate metrics on Gradient Boosting model
- Evaluate metrics on other learning model if necessary
- Select well performing learning model
- Tune hyperparameters to improve accuracy of the model
- Finalize the model

References

- IEEE Fraud Detection competition: https://www.kaggle.com/c/ieee-fraud-detection
- Data Set: https://www.kaggle.com/c/ieee-fraud-detection/data
- https://www.creditdonkey.com/credit-card-fraud-statistics.html
- https://www.consumeraffairs.com/finance/identity-theft-statistics.html