Machine Learning Engineer Nanodegree

Capstone Proposal

IEEE-CIS Fraud Detection Kaggle competition

Mukesh Yadav Oct 10th, 2019

Proposal

Domain Background

Credit card fraud is a wide-ranging term for theft and fraud committed using or involving a payment card, such as a credit card or debit card, as a fraudulent source of funds in a transaction. The purpose may be to obtain goods without paying or to obtain unauthorized funds from an account. Credit card fraud is also an adjunct to identity theft. According to the United States Federal Trade Commission, while the rate of identity theft had been holding steady during the mid-2000s, it increased by 21 per cent in 2008. However, credit card fraud, that crime which most people associate with ID theft, decreased as a percentage of all ID theft complaints for the sixth year in a row.

Anomaly Detection can be utilized to solve such problems where we need to find transaction which could be a fraud. This field is being transformed as computers get better and better at analyzing large datasets. See Chandola et al 2009[1] for a general overview of anomaly detection techniques and applications.

Problem Statement

Although incidences of credit card fraud are limited to about 0.1% of all card transactions, they have resulted in huge financial losses as the fraudulent transactions have been large value transactions. In 1999, out of 12 billion transactions made annually, approximately 10 million—or one out of every 1200 transactions—turned out to be fraudulent. Also, 0.04% (4 out of every 10,000) of all monthly active accounts were fraudulent. Even with tremendous volume and value increase in credit card transactions since then, these proportions have stayed the same or have decreased due to sophisticated fraud detection and prevention systems. Today's fraud detection systems are designed to prevent one-twelfth of one percent of all transactions processed which still translates into billions of dollars in losses.

Creating a successful machine learning would “improve the efficacy of fraudulent transaction alerts for millions of people around the world, helping hundreds of thousands of businesses reduce their fraud loss and increase their revenue.

Datasets and Inputs

The data will be taken from Kaggle competition which is organized by IEEE Computational Intelligence Society (IEEE-CIS). The data include product names, card numbers, addresses, emails, a device purchased from, and high dimensional transaction data.

The data is broken into two files identity and transaction, which are joined by TransactionID. Not all transactions have corresponding identity information.

Categorical Features - Transaction:

- ProductCD
- card1 - card6
- addr1, addr2
- P_emaildomain
- R_emaildomain
- M1 - M9

Categorical Features - Identity:

- DeviceType
- DeviceInfo
- id_12 - id_38

The TransactionDT feature is a timedelta from a given reference datetime (not an actual timestamp).

Files:

- train_{transaction, identity}.csv - the training set
- test_{transaction, identity}.csv - the test set (you must predict the isFraud value for these observations)
- sample_submission.csv - a sample submission file in the correct format

Solution Statement

In this project, we need to identify fraudulent transactions from the given training and testing data set. Upon our finding, we have to return a Boolean isFraud variable which will say if respective transaction is fraud or not.

Benchmark Model

Benchmark model would be to come closer to the top 10% from the leaderboard[2] of this competition.

Evaluation Metrics

Area under the ROC curve, where the y-axis is the true positive rate, TPR = TP / (TP+FN) and the x-axis is the false positive rate, FPR = FP / (FP + TN).

Project Design

Usually, all the data we get for prediction requires some sort of preprocessing. In this data we have 2 files and both have transactionID, hence both pairs need to be joined based on the transaction ID.

We need to do Manual feature engineering for the better result which may include normalizing data, transforming variables (log or otherwise), creating interaction variables, or creating entirely new features. Additionally, the transaction data is very high dimensional and will require PCA to reduce dimensions.

I would try to follow the template outlined in the population segmentation project previously completed as part of the MLND. There is also the issue with class imbalance. I will also follow the previous outline of how to balance (weight) the positive class more to account for imbalance. Since the training data is very large, I plan on splitting the set into training and validation sets in order to try different models and hyperparameter tuning.

I would also try to do automate Feature engineering by using featuretools[3]. Also, I would be referring Discussion forum and Notebooks to optimize the model.

Refrences

1. http://cucis.ece.northwestern.edu/projects/DMS/publications/AnomalyDetection.pdf
2. https://www.kaggle.com/c/ieee-fraud-detection/leaderboard
3. https://www.featuretools.com