Machine Learning Engineer Nanodegree

Capstone Proposal

Shakti Misra May 14th, 2019

Proposal

Domain Background

Rise of credit card is on all time high and people use it for various reasons. With this rise in usage the chanches of fraud is also increasing. Credit card and identity are now a days available in black market. The image below shows that these kind of identity threat is much high than any other crime now a days.

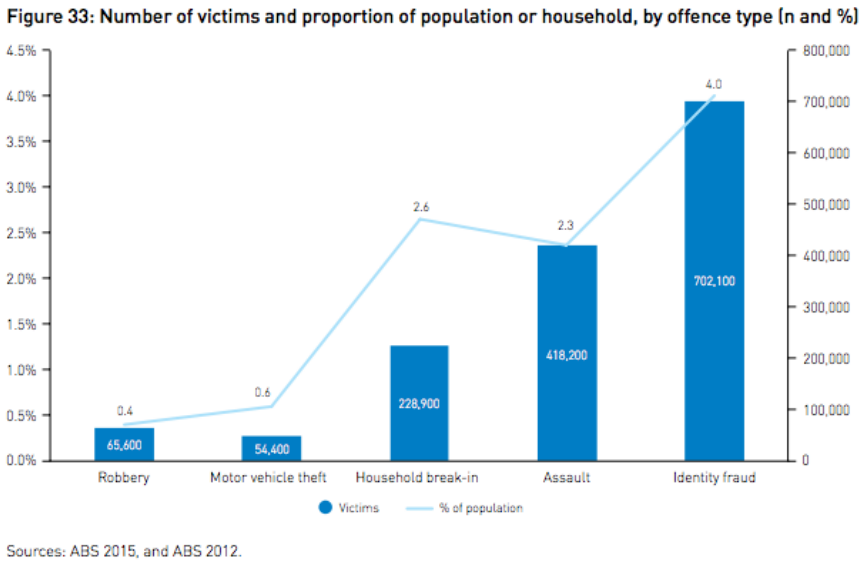

Figure 1: Credit Card Crime

Identity threat (Courtesy wikipedia)

Working in expense management industry it is quite important to help out customers with solutions where these kind of transactions can be identified and flagged. Considering the credit card billing is integrated with our solution and is shown up in our portals it will be a useful feature. Even it will be useful when the credit card is insured. In which case the amount can be claimed from the insurance company if possible.

Problem Statement

Recognizing faudlent credit card transaction from the transaction data so that the customer can be alerted and/or protected from the fraud.

Datasets and Inputs

Getting realtime data and using for these kind of personal information is a challenge. We cannot use company dataset for these. So to create the model and solution I will be using the dataset available on Kaggle.

Brief description The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have `492 fraud / 284807` transactions. The dataset is highly unbalanced, the positive class (frauds) account for `0.172%` of all transactions. It contains only numerical input variables which are the result of a PCA transformation. Features `V1, V2, ... V28` are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'.

Work done Some of the past and future can be found in Researchgate. It contains a list of past research done in this field.

Solution Statement

Problem type This is a classic classification problem, to be more prescise this is a binary classification problem. Given sample data the model has to tell if that is a valid transaction or a fraud transaction.

Technology stack Considring I am familar with a Python stqack for data processing I will be Numpy to understand, matpolit to plot and visualize and Keras or Tensorflow to train and predict the fradunt transaction.

Birds eye view of data

- The dataset is highly skewed with 492 frauds in a total of 284807 whihc is 0.173%.
- The dataset consists of numerical values from the 28 `Principal Component Analysis (PCA)` transformed features, namely V1 ... V28 without any metadate to preserve confidentiality.
- No missing value in the dataset.

High level solution I will be using Adaptive Deep neural network to come up with a solution for detecting the fraud data.

Benchmark Model

Considering this is a binary classification problem I would like to start with a logistic regression. This is simple enough to start as an bench mark. Doing a quick run the logistic regression with random sampling itself gives quite good accuracy ($\equiv 94\%$) on the dataset. So this is a compelling benchmark for the problem and dataset in hand.

Evaluation Metrics

The data is highly imbalanced so I will stick to the suggestion provided by Kaggle to measure the acuracy using `Area Under the Precision-Recall Curve (AUPRC)`1, 2. Confusion matrix will not provide a good measure of the evaluation. The total number of fraud cases is much less (`0.172%`); and variation in the confusion matrix will not be a good measure. So I will not be using this method, rather than stick to the approach provided by Kaggle. We cannot even use accuracy for this case as the data is highly skewed. So, if we predict all values as valid still the accuracy will be quite high. Which is not a desirable trait of evaluation.

Project Design

Technology Stack I will be using python based stack. I will be using the standard kits as Pandas, Numpy, Scikit-Learn, Matplotlib (Seahorn), Keras as they are highly scalable and performant. I will try to see if I can use GPU hours from AWS to perform learning.

Data Preparation

- I will start with preparing the data before I can proceed with the analysis.
- Creating a training data set that will allow our algorithms to pick up the specific characteristics that make a transaction more or less likely to be fraudulent. So I will be using random sampling for the training set.
- I will be analyzing the data for detecting outliers. I will have to consider the trade off between reducing the dataset and keeping outliers in the data. This has to be done carefully so as not to reduce the number of fraudent transactions as they are very less.

Benchmark After these steps I will setup a standard bench mark using Logistic Regression.

Base model training I will be training a base model that is a classical Deep neural network model - 1 input layer - 3 Hidden layers - 1 Output layer - RELU activation

This is the base model that I am planning to use, but as per the prediction I will be doing some adjustments to it. After doing a set of base model training and measurement I will be going with the hyper parameters tuning. Considering my previous expereience I believe this is the step that is going to take a lot of time.

Reference

- Credit card usage in India
- USA credit card ownership and usage pattern
- Growing cyberthreat and identity theft
- Credit card security ID frauds
- ROC Curves
- Area under prescision recall curve