Machine Learning Engineer Nanodegree

Capstone Proposal

Sai Sumanth

Domain Background

- Credit Card transactions has been a larger share of US payment system. In today's increasingly electronic society and with the rapid advances of electronic commerce on the Internet, the use of credit cards for purchases has become convenient and necessary.
- Credit card transactions have become the de facto standard for Internet and Web based e-commerce. The US government estimates that credit cards accounted for approximately US $13 billion in Internet sales during 1998. This figure is expected to grow rapidly each year.
- However, the growing number of credit card transactions provides more opportunity for thieves to steal credit card numbers and subsequently commit fraud. When banks lose money because of credit card fraud, cardholders pay for all of that loss through higher interest rates, higher fees, and reduced benefits. Hence, it is in both the banks' and the cardholders' interest to reduce illegitimate use of credit cards by early fraud detection.
- Credit Card companies are approaching data scientists in order find a better solution to this problem. To solve this problem we need to build a model which flags the fraud transactions and gives an alert to the companies and the cardholders. This model can be designed using both supervised learning and unsupervised learning methods. For supervised learning methods, we need to have labelled data to train our algorithm. Whereas for unsupervised learning methods, we can tag the outlier transactions as fraud.
- A research collaboration of Worldline and the Machine Learning Group of ULB on big data mining and fraud detection have worked on the datasets provided by Kaggle in order to find a better model to tag the fraud credit card transactions.

Academic Papers on Credit Card Fraud Detection using Supervised and Unsupervised Models

1) https://ieeexplore.ieee.org/document/5159014/

2) https://pdfs.semanticscholar.org/1752/a117dec81740c1d5516be15a3395a6d74a3c.pdf

3) http://ijcttjournal.org/Volume4/issue-7/IJCTT-V4I7P143.pdf

Problem Statement

- The goal of this project is to find out whether a cardholder transaction is fraud or not.
- In supervised learning, the model is trained on data which consists of label feature, it gives an information about a transaction, 1 if the transaction is fraud, and 0 if the transaction is good.
- In unsupervised learning, the model is directly trained on the transaction data without the label feature. The transactions which lie farther away (i.e., outliers) from the normal transactions are considered as fraud transactions.
- The model performance is evaluated using recall score rather than accuracy score. Since, we care about tagging the fraud transactions, we need to have less number of false negatives.

Datasets and Inputs

- For this project, I am going to use the datasets which contains transactions made by credit cards in September 2013 by European cardholders which are available on Kaggle.
- The dataset contains 492 frauds out of 284,807 transactions that occurred in 2 days. Due to confidentiality issues, the dataset contains only numerical input variables which are the result of a PCA transformation. It consists 28 Principal Components, along with Time and Amount of the Transaction.
- Since, we have only 492 frauds, the dataset is considered as class imbalanced. In order to overcome this, we have to perform undersampling or upsampling on the dataset to make it balanced.
- Feature Time contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature Amount is the transaction Amount, this feature can be used for dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. The remaining 28 Principal components gives us the transformed personal information about the cardholder.

Solution Statement

- Initially, I will visualize the distribution of the data and the skewed features are transformed using logarithmic transformation or square root transformation.
- The features are then standardized in order to achieve zero mean and equal variance.
- I'm going to build models using both supervised and unsupervised learning methods, and compare their performance results.
- For supervised learning, the dataset with Label feature is fitted to a supervised classifier and the Area under the Precision-Recall Curve and Recall score are calculated to evaluate its performance.
- For unsupervised learning, the dataset without the Label feature is fitted to unsupervised algorithm, and the outliers are tagged as fraud transactions, and they are compared with the actual Label results to evaluate the performance of the model.

Benchmark Model

- I plan to compare my models with Basic Logistic Regression Classifier as my bench mark model, since it is considered as a go to method for binary classification problems.
- I want to fit Logistic Classifier to the labelled dataset and calculate the Area under the Precision-Recall Curve and Recall scores.
- I would like to choose these performance metrics as a threshold to evaluate my supervised and unsupervised models.

Evaluation Metrics

- For supervised model, I'm going to evaluate my model using Area under the Precision-Recall Curve and Recall Score to evaluate its performance.
- For unsupervised model, the outliers are tagged as 1 and the remaining instances are tagged as 0. The predicted results are then compared with the actual results to calculate Area under the Precision-Recall Curve and Recall Score to evaluate its performance.
- The precision-recall curve shows the trade-off between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. For our problem statement, the model should have area under precision-recall curve nearly equal to 1.
- Recall Score is the ratio of True Positives and Sum of (True Positives and False Negatives). For our problem statement, the model should have a recall score nearly equal to 1.

Project Design

- Initially, the data should be clear of any missing values, I would achieve this by either replacing them with mean or median of the features or a feature with high percentage of missing values is removed.
- The distribution of the features is observed in the data. The features with skewness are dealt by applying logarithmic or square root transformation.
- The relation between the features are observed using visualizations techniques.
- The data is standardized in order to achieve zero mean and equal variance.
- Supervised Learning Method, Random Forest classifier is fitted to the dataset and performance metrics are calculated.
- The model is tuned accordingly in order to achieve high scores.
- I would also like to find the important features and use them to fit the model in order to reduce the complexity and increase the performance.
- Unsupervised Learning Methods, K-means algorithm is fitted to the dataset by removing the Label feature.
- The distance between the cluster centroid and the instances are calculated and the instances which are higher than 95% percentile distance from the cluster centre are considered as outliers.
- These outliers are compared with actual Label results and the accuracy of the model is calculated.
- Towards the end, I will compare the performance results of both the supervised and unsupervised learning methods and declare the best model for credit card fraud detection.

References:

1) Kaggle:

https://www.kaggle.com/mlg-ulb/creditcardfraud

2) Research Paper on Credit Card Fraud Detection:

https://pdfs.semanticscholar.org/7456/adc49f8b3d2bfba97064284aa81364ad7dbc.pdf

3) Random Forest:

http://ai.ms.mff.cuni.cz/~sui/nezvalovapopelinsky.pdf

4) K-Means:

http://pmg.it.usyd.edu.au/outliers.pdf