CREDIT CARD FRAUD DETECTION ON E-COMMERCE WEBSITE DATA

Kazeem B. Tijani
KTIJANI@MY.APSU.EDU

December 12, 2018

Austin Peay State University
Clarksville TN, U.S.A

ABSTRACT:

With compromised credit cards and data breaches dominating the headlines in the past couple of years, institutions and individuals are being robbed of their peace and forced into financial debts and institutions losing the trust of their customers. The use of technology and machine learning to help alleviate the menace to the barest minimum is of great interest to financial institutions as well customers. The algorithm used in this project aims at using binary classification on over 300,000 anonymized credit card transaction details, due to the confidentiality of the data, the information was not released in the details but using principal component analysis (PCA) the most important details that are necessary to build our model has been deduced and made available on Kaggle which is the largest data repository for machine learning and data science. Deep learning models and Random Forest was used to do the binary classification of this credit card fraud detection. The dataset is unbalanced since most of the transactions will be non-fraudulent so I used SMOTE (Synthetic Minority Oversampling Technique) to randomly oversample the dataset to help improve the model, and compared the results in my confusion matrix.

Author Keywords
Deep learning, Machine learning, Random Forest, Keras

INTRODUCTION

Imagine the pains that one goes through when his credit card details are stolen and used for transactions, this will incur unexpected debts and recurrent difficulties for the credit card owner as well as the credit card issuing company since the customer will definitely get fed up of continuing transactions with them due to the rate of dissatisfaction if he eventually gets out of the debts. Hence, It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

It was reported by NASDAQ that most card fraud occurs in the United States. In fact, a 2015 research note from Barclays stated that the U.S. is responsible for 47 percent of the world's card fraud despite only accounting for 24 percent of total worldwide card volume.

The high level of debit and credit card fraud in the United States also impacts other countries. Among U.K.-issued cards in 2015, 35 percent of fraud-related losses occurred in the United States, compared to 10 percent in France and Australia, 9 percent in Canada and 6 percent in Germany.

Cross-border fraud occurs when criminals use a consumer's credit or debit card data in one country to make fraudulent transactions in another country. In 2014, 47 percent of fraudulent cross-border transactions on U.K. credit cards took place in the United States.

U.S. credit card fraud is on the rise, too. About 31.8 million U.S. consumers had their credit cards breached in 2014, more than three times the number affected in 2013.

This paper describes our approach to using both Deep learning and Machine learning models to do the binary classification of detecting whether a transaction is fraudulent or not. 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 frauds out of 284,807 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. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. 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'. 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 example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

OVERVIEW OF MACHINE LEARNING ALGORITHM USED

The problem to be solved is a binary classification type since our goal is to detect if a transaction is fraudulent or not, and for this I used both the deep learning and machine learning models to classify the anonimized dataset.

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. It is mainly sub-divided into three categories, namely; Supervised, Un-supervised and Semi-Supervised. Supervised learning is further subdivided into Regression techniques and classification techniques in which Random Forest belongs.

Random forests an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. The general method of random decision forests was first proposed by Ho in 1995.

Deep learning is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts.

Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation.

METHODOLOGY:

The approach used to solve this problem are as follows:

- Data Pre-processing
- Feature Extraction
- Feature Engineering
- Model Building
- Result

CHALLENGES FACED:

Problem 1: Had great difficulties installing pandas after upgrading Anaconda, despite reading all the documentations and checking possible solutions on stack overflow, I used the pip install pandas, pip3 install pandas all to no avail as it claimed it had successfully installed but didn't read my dataset which is the first stage of my data pre-processing.
Solution: I had to upgrade my Conda environment to match up with the upgraded Anaconda and then installed pandas and eventually I pulled through.

Problem 2: While using Random forest, I had an issue with fitting my data because I was using random_forest.fit(X_train,y_train.values.ravel()) which couldn't fit my data as Python 3.7 had been upgraded and we didn't need to use values.ravel
Solution: I found out the updated version of the fit method and used it in place of the old; random_forest.fit(X_train,y_train.ravel()) instead of random_forest.fit(X_train,y_train.values.ravel())

Also, earlier version of splitting data into training and test set used cross_validation package in the scikit learn library but it had been updated into sklearn.model_selection

FUTURE PERSPECTIVE

Based on my interest in using machine learning to solving real life business challenges, I seek to get access to the details of what factors were considered in the principal component analysis and then build softwares that can really help against credit card fraud detection.

ACKNOWLEDGEMENT

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. More details on current and past projects on related topics are available on http://mlg.ulb.ac.be/BruFence and http://mlg.ulb.ac.be/ARTML

I want to express my gratitude to our Professor, Dr. Mayo who exposed me the rudiments of machine learning and challenged me with his passion when delivering the lectures. I also appreciate my learned colleagues who actually made the class lively and fun-filled learning experience.

REFERENCES

1. Wikipedia contributors, "Deep learning," *Wikipedia, The Free Encyclopedia,* https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=873508864 (accessed December 13, 2018).
2. Wikipedia contributors, "Random forest," *Wikipedia, The Free Encyclopedia,* https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=871977243(accessed December 13, 2018).
3. Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015
4. https://www.kaggle.com/mlg-ulb/creditcardfraud/kernels
5. https://superdatascience.com
6. https://www.nasdaq.com/article/credit-card-fraud-and-id-theft-statistics-cm520388
7. Gemalto's 2014 Breach Level Index
8. Barclays' Security in Payments: A Look into Fraud, Fraud Prevention, & the Future, May 22, 2015
9. Financial Fraud Action UK's Fraud The Facts 2015
10. FICO press release, June 25, 2015
11. Javelin Strategy & Research 2015 Data Breach Fraud Impact Report
12. https://keras.io/
13. https://www.sas.com/en_us/insights/analytics/machine-learning.html