Udacity Machine Learning Nanodegree
Mobile Payments Fraud Detection

Venkat Maddi

June 1st, 2018

Domain Background

Nowadays, people are extensively using mobile devices to handle financial transactions. Banks and Financial industry experts are predicting that people will utilize mobile devices to initialize payments extensively this year [1]. Financial institutions are constantly working on various methods to improve the customer experience, execute the payments faster and safer [2]. Banks have introduced Reward Points to encourage customers to complete the payments electronically.

Accenture Consulting Study indicates that the Gen Z, new generation adults and young people today, will make up to 40 percent of USA population by 2020. The Gen Z people are more comfortable with executing transactions from mobile devices [3]. The mobile based payment transactions will grow exponentially in the next few years.

Problem statement

It is very important to detect fraudulent transactions while processing mobile payments. It is not possible to detect the fraudulent transactions manually because of huge volume of transactions banks handle daily. Researches and Data scientists are creating new algorithms and introducing new processes to detect the fraud as soon as fraudulent transaction hits the financial institutes.

Kaggle platform has provided synthetic financial datasets to develop innovative algorithms for detecting the fraud [4]. I have decided to work on machine learning algorithm to detect the mobile payments fraud. Kaggle fraud detection dataset details can be found at https://www.kaggle.com/ntnu-testimon/paysim1.

Datasets and inputs

Normally, the financial institutions do not publish mobile money transactions. Kaggle platform has provided a synthetic dataset generated using the simulator called PaySim as an approach to such a problem. [5]. PaySim uses aggregated data from the “private dataset” to generate a synthetic dataset that resembles the normal operation of transactions and injects malicious behavior to later evaluate the performance of fraud detection methods. The private dataset is based on real transactions from a mobile money services implemented in African country.

The dataset was generated by PaySim for Kaggle platform. The input dataset file contains more than one million records. Each input record consists of 11 attributes. The data attribute details as follows (Table 1):

Data Attribute #	Attribute Name	Description
1	Step	It maps a unit of time in the real world. In this case, step 1 represents First hour of transactions
2	type	Transaction Type, CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER
3	amount	Transaction Amount in local currency
4	nameOrig	The customer who initiated the transaction
5	oldbalanceOrg	The initial balance before the transaction
6	newbalanceOrig	The new balance after processing the transaction.
7	nameDest	The customer who is the recipient of the payment
8	oldbalanceDest	The initial balance in the recipient account before the transaction. Note that there is not information for customers that start with M (Merchants).
9	newbalanceDest	The new balance in the recipient account after processing the transaction. Note that there is not information for customers that start with M (Merchants).
10	isFraud	Value values are either 0 or 1. The value 1 indicates that this transaction was created by the fraudulent agent inside the simulator
11	isFlaggedFraud	If a transfer amount is more than 200,000 then single transaction flags as illegal attempt. The business model flags the transaction as “illegal Attempt” for higher denominations.

Missing Attributes: The recipient account’s old balance and new balance attributes do not have values for all records. If the recipient (destination customer) name starts with M(Merchants), then destination accont old balance and destination new balance attributes are empty.

Output Variable: The 10th attribute, *isFraud*, is an output variable. If the output variable value is zero, then the data record categorized as a real transaction. If the output variable value is one, then the data record is categorized as a Fraudulent transaction.

Solution statement

I will prepare the data by splitting feature and target/label columns. I will split the data into training and testing datasets. I will allocate 80% to the training data and 20% of datasets to the testing to verify the accuracy of the model. I will also verify the quality of the data. I will verify which features which cause the major impact on identifying a fraud. Some of the features

may not cause major impact to the fraud detection. I will eliminate the redundant features that cause minor impact to the fraud detection.

The dataset contains multiple non-numeric columns like transaction type. The non-numeric feature columns will be converted to 1/0 binary values. As described in above section, there are several non-numeric columns that need to be converted. The amount and balance columns will be normalized using scaling technique to have a reasonable data range.

I am not sure which algorithms would be fit for this problem or what hyper parameters configurations to use. Here are some of the Supervised Learning Algorithms I will try to apply during the implementation:

- Gaussian Naive Bayes
- Logistic Regression
- K-Nearest Neighbors
- Random Forests
- Decision Trees
- Support Vector Machines

I will implement above Supervised models and identify a best model applicable to the fraud detection.

Benchmark model

I am going to start the model with Naïve Bayes algorithm and make it as a benchmark score. I will slowly implement other Supervised models and tune the hyper parameters to improve the score. Also, Kaggle platform has created Leaderboard to determine the best models. I will try to publish my model results to the Kaggle leaderboard. I will try to compare my ranking with other competitors in the leader board.

Evaluation metrics

Project design

References

[1] Mobile Payment Trends in 2018: https://www.paymentvision.com/blog/2017/12/26/7-trends-that-prove-mobile-payments-are-here-to-stay-in-2018

[2] Mobile Payments Safer and Faster: https://www.mobilepaymentstoday.com/blogs/3-trends-for-2018-safer-data-faster-payments-better-experiences/

[3] Banking Future Payments- Accenture Consulting Study : https://www.accenture.com/us-en/insight-banking-future-payments-ten-trends

[4] Kaggle Financial Fraud Detection dataset: https://www.kaggle.com/ntnu-testimon/paysim1

[5] PaySim Simulator: E. A. Lopez-Rojas , A. Elmir, and S. Axelsson. "PaySim: A financial mobile money simulator for fraud detection". In: The 28th European Modeling and Simulation Symposium-EMSS, Larnaca, Cyprus. 2016