Udacity Machine Learning Nanodegree
Mobile Payments Fraud Detection

Venkat Maddi

July 5th, 2018

I. Definition

Project Overview

Nowadays, people are extensively using mobile devices to handle financial transactions. Banks and Financial industry experts are predicting that customers 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 customers are more comfortable with executing transactions from mobile devices [3]. The mobile based payment transactions will grow exponentially in the next few years.

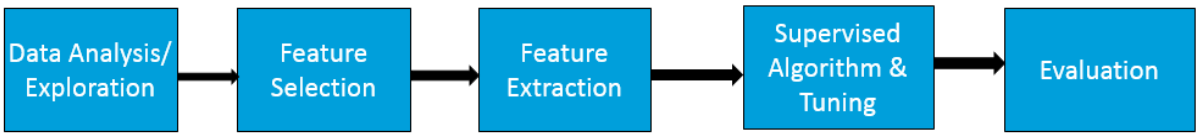
Problem statement

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 hourly and 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.

Normally, the financial institutions do not publish mobile money transactions. Kaggle platform [4] has provided a synthetic dataset generated using the simulator called PaySim as an approach to detect the fraudulent transactions. [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.

I have decided to work on machine learning algorithm to detect the mobile payments' fraud. My project will implement Supervised Learning Classification techniques to detect fraudulent transactions. Also, I will utilize fraud detection dataset available on Kaggle website at https://www.kaggle.com/ntnu-testimon/paysim1.

The high-level design activities and workflow while implementing the Supervised Learning models to predict fraudulent mobile payments are as follows:



Workflow Diagram

1) **Data Analysis/Exploration**: Data Analysis is a very important and key activity of the Machine Learning model creation. The activities within this phase are as follows:
   - Review and understand Data
   - Identify Data thresholds like minimum, maximum and etc
   - Determine Data mean, standard deviation
   - Load required libraries and data files

2) **Feature Selection:** The dataset contains multiple data fields/columns. The data field is considered as a feature.
   - Review all features included in the datafile.
   - Identify Feature Dependency. Some of the features are critical to determine the prediction. If a feature is dependent on another primary feature, then the primary feature must exist in the model.
   - Reduce features to a reasonable number. Eliminate least important features which do not cause major impact to the prediction.
   - Select Best Features from dataset.

3) **Feature Extraction**
   - Review the feature's data distribution and ranges. If the data range (difference between min and maximum values) is wide, then Normalize the data using Logarithm transformation.
   - Normally, Numeric values tend to tune models more effectively. Kaggle Paysim dataset contains features with non-numeric values. I will apply "One-Hot encoding" technique to convert non-numeric values to numeric values.
   - Split data into training and testing
   - Assign a portion of training data to the Validation activity

4) **Supervised Learning Algorithms & Tuning**
   - Create Accuracy and F-score bench marks using Naïve Predictor model.
   - Implement at least five (5) Supervised models.
   - Train the Model with the Training data
   - Tune the Algorithm by modifying Hyper parameters

5) **Evaluation**
   - Execute Model using the Testing data
   - Analyze the results and Model performance
   - Identify a best supervised Model which provides the maximum performance results, high accuracy, and high F-score.

Metrics

Each Supervised Model performance is calculated using twp statistical concepts, Classification Accuracy, and F-Score. The following table provides confusion matrix definitions.

	Predicted as Fraud	**Predicted as Genuine**
Fraud Transaction	True Positive (TP)	False Negative (FN)
Genuine Transaction	False Positive (FP)	True Negative (TN)

Table 1: Confusion Matrix

True Positive (TP): Transaction is Fraud and Model has predicted as Fraud accurately.
False Negative (FN): Transaction is Fraud and Model has predicted as Genuine incorrectly
False Positive (FP): Transaction is Genuine, and Model has predicted as Fraud incorrectly
True Negative (TN): Transaction is Genuine, and Model has predicted as Genuine accurately

The **Accuracy** measures how often the Model makes the correct prediction. It's the ratio of the number of correct predictions to the total number of data points.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

The **Recall(sensitivity)** indicates what proportion of actual fraud transactions is predicted by the Model as fraud. It is a ratio of True Positives to True Positives Plus False Negatives.

$$\textbf{Recall} = \frac{TP}{TP+FN}$$

The **Precision** tells us what proportion of transactions Model predicted as fraud, actually were fraud. It is a ratio of True Positives to True Positives Plus False Positives.

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

The **F-beta** score is a metric that considers both precision and recall:

$$F\beta = (1+\beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall}$$

II. Analysis

Data Exploration

The input dataset financial mobile based transactions provided by Kaggle platform. The dataset contains input attributes (AKA features) and the Fraud attribute (Target). The file contains more than six million records. Each record consists of both input attributes (features) and output variable. The classification goal is to predict whether mobile payment is a fraudulent or not.

Input Variables (features):

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).
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.

Table1: Feature Details

Output Variable (Target)

The 10th attribute, *isFraud*, is an output variable. The output variable valid values are either zero (0) or one (1). If the output variable value is zero, then the data record categorized as a genuine transaction. If the output variable value is one, then the data record is categorized as a Fraudulent transaction.

Data Attribute #	Attribute Name	Description
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

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 account old balance and destination new balance attributes are zero.

Categorical Features: The second column in the datafile is Type and it explains the transaction category. There are six types of transactions available in the dataset. These transaction Types are CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.

The dataset input file contains 6,362,620 rows and 11 columns. It represents 10 features (input variable) and one target column (output). The number of Fraud records count in the dataset is 8,219. It translates 0.1291% of records are fraud payments. The percentage of fraud records are less than one percent.

Few sample records from the dataset are as follows:

type	amount	nameOrig	oldbalanceOrg	nameDest	oldbalanceDest	isFraud	isFlaggedFraud
PAYMENT	9839.64	C1231006815	170136.00	M1979787155	0.0	0	0
PAYMENT	1864.28	C1666544295	21249.00	M2044282225	0.0	0	0
TRANSFER	181.00	C1305486145	181.00	C553264065	0.0	1	0
CASH_OUT	181.00	C840083671	181.00	C38997010	21182.0	1	0
PAYMENT	11668.14	C2048537720	41554.00	M1230701703	0.0	0	0
PAYMENT	7817.71	C90045638	53860.00	M573487274	0.0	0	0
PAYMENT	7107.77	C154988899	183195.00	M408069119	0.0	0	0
PAYMENT	7861.64	C1912850431	176087.23	M633326333	0.0	0	0
PAYMENT	4024.36	C1265012928	2671.00	M1176932104	0.0	0	0
DEBIT	5337.77	C712410124	41720.00	C195600860	41898.0	0	0

Table 2: Sample Records

The following table explains the Statistical info. The amount and balance features contain a higher standard deviation.

step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	isFraud
count	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06
mean	2.433972e+02	1.798619e+05	8.338831e+05	8.551137e+05	1.224996e+06
std	1.423320e+02	6.038582e+05	2.888243e+06	2.924049e+06	3.674129e+06
min	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	1.560000e+02	1.338957e+04	0.000000e+00	0.000000e+00	0.000000e+00
50%	2.390000e+02	7.487194e+04	1.420800e+04	0.000000e+00	2.146614e+05
75%	3.350000e+02	2.087215e+05	1.073152e+05	1.442584e+05	1.111909e+06
max	7.430000e+02	9.244552e+07	5.958504e+07	4.958504e+07	3.561793e+08

Exploratory Visualization

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 set aside at least 30% of Training data for the Data validation. I will also verify the quality of the data. I will verify which features which cause the major impact on detecting the fraud. Some of the features may not cause major impact to the fraud detection. I will eliminate the least important 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 are needed. Here are some of the Supervised Learning Algorithms I will try to apply during the implementation:

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

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

Benchmark model

I am going to start the model with Naïve Predictor 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 a 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 Leaderboard.

Project design

After executing these tasks, I will prepare a conclusion and document observations at the end.

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