|  |  |
| --- | --- |
| **Capstone Project**  **Machine Learning Engineering**  **Nanodegree** | **Christopher Kim**  **October 6, 2019** |

**IEEE-CIS Fraud Detection**

**Definition**

**Project Overview**

At any given moment, millions of credit card transactions occur in the world. With the rise of electronic commerce businesses, the credit card transactions became even more common in the past couple of decades. With such a massive number of compromised credit card numbers, fraudulent credit card transactions are not so rare occurrences any more. In 2018, the Federal Trade Commission processed 1.4 million fraud reports totaling $1.48 billion in losses.

Preventing credit card fraud is an essential part of credit card payment processing. Improving fraud detection system will save a lot of money and improve consumers' experience.

In this project, I explored the data set provided by IEEE-CIS Fraud Detection competition and apply feature engineering to prepare for machine learning algorithms. Using the prepared data, I applied machine learning algorithms to train the fraud detection model and evaluate the model form meaningful metrics.

**Problem Statement**

Using Machine Learning technique to detect and prevent fraudulent transactions is already saving millions of dollars a year. Researchers from the IEEE Computational Intelligence Society (IEEE-CIS) want to improve this figure, while also improving the customer experience. On the Kaggle platform IEEE-CIS has opened a competition IEEE-CIS Fraud Detection competition. Building a machine learning model that scores high accuracy on detecting fraud in transactions will be the problem to solve.

**Scoring**

As I explored the provided datasets, it turned out about 3.5% of the dataset is labeled as fraudulent transactions. If the system predicts every transaction to be normal transaction, the accuracy score would be 96.5%. Therefore, scoring high on the accuracy wouldn’t necessarily be a good model.

Detecting fraud in transactions, scoring high on Recall in Confusion Metrics is a better metrics because we want to prevent fraudulent transactions as much as possible. Also, high Precision score would also provide better consumer experience as it would flag fewer false positive on transactions.

Using the Precision and Recall score, I evaluated the model on F1 Score.

As the competition submissions are evaluated on area under the ROC curve between the predicted probability and the observed target, the Area Under the Curve in the Receiver Operation Characteristics curve.

**Analysis**

**Data Exploration and Visualization**

# Two sets of data are provided for the competition. One is the identity data and the other is the transaction data, and they are matched by TransactionID values. The transaction data set has the target label value of isFraud on each row, and not every transaction has a matching row in the identity data set. There are about ¼ number of identity rows to the transaction data set.

Along with the mismatching transaction rows, most of the features have null values. Some of the features have 99% of rows have null values.

However, the data sets have many feature columns, especially in the transaction data has 394 columns, and the identity data set has 41 columns. As the identity and the transaction data sets have been joined on the TransactionID column, the resulting data set has 434 columns.

# **Identity Data**

Variables in this table are identity information – network connection information (IP, ISP, Proxy, etc) and digital signature (UA/browser/os/version, etc) associated with transactions. They're collected by Vesta’s fraud protection system and digital security partners. (The field names are masked and pairwise dictionary will not be provided for privacy protection and contract agreement)

**Categorical Features:**

The following features are identified as categorical features in the Identity data set, and the others are numerical values.

* DeviceType
* DeviceInfo
* id12 - id38

# **Transaction Data**

The data in the Transaction data set has transaction related columns such as transaction date, amount, product code, payment card information, engineered columns, and etc.

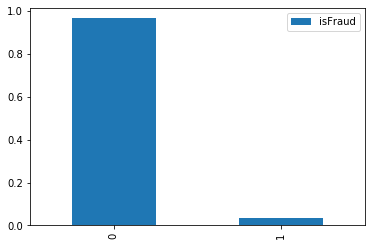
* TransactionDT: timedelta from a given reference datetime (not an actual timestamp)
* TransactionAMT: transaction payment amount in USD
* ProductCD: product code, the product for each transaction
* card1 - card6: payment card information, such as card type, card category, issue bank, country, etc.
* addr: address
* dist: distance
* P\_ and (R\_\_) emaildomain: purchaser and recipient email domain
* C1-C14: counting, such as how many addresses are found to be associated with the payment card, etc. The actual meaning is masked.
* D1-D15: timedelta, such as days between previous transaction, etc.
* M1-M9: match, such as names on card and address, etc.
* Vxxx: Vesta engineered rich features, including ranking, counting, and other entity relations.

**Categorical Features:**

In the transaction data set, the following features are identified as categorical data.

* ProductCD
* card1 - card6
* addr1, addr2
* Pemaildomain Remaildomain
* M1 - M9

**Ratio of normal transaction vs. fraud transactions**



About 3.5% of the transactions are fraud.

**Null values**

Some of the identity features sudh as id\_07, id\_08, id\_21 - id\_27 have null values for 99% of the data.

Id\_07

NaN 0.991271

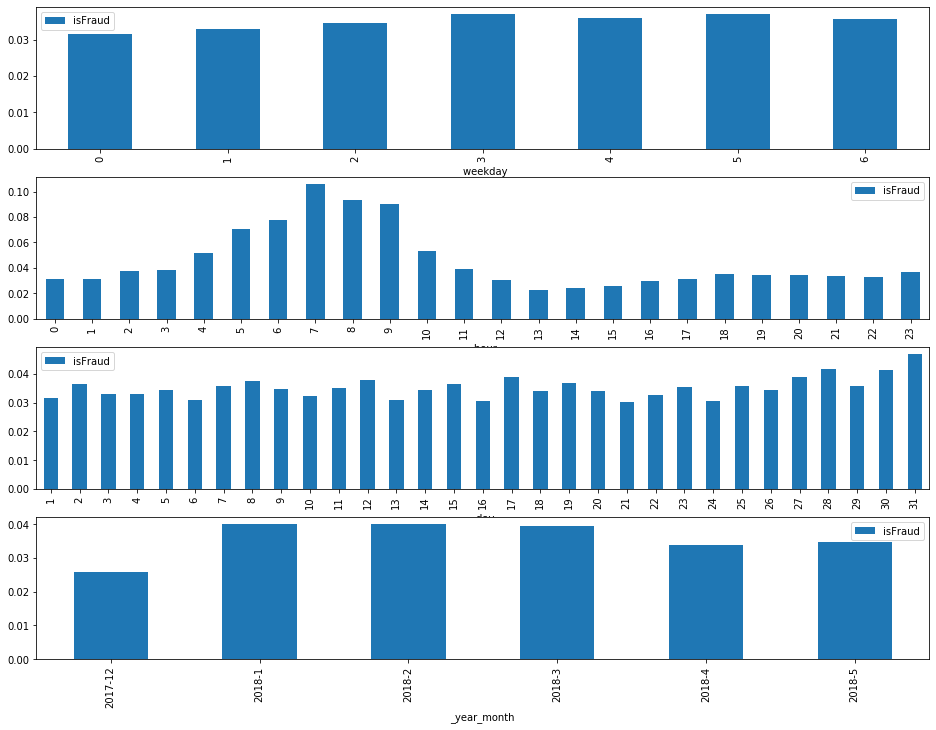
0.0 0.000693

16.0 0.000415

…

**Transaction Date**

The following graphs show the distributions of the fraud transactions by weekdays, time of day, day of month, and month of year.



**Algorithms and Techniques**

**Benchmark**

**Methodology**

**Data Pre-processing**

**Null Values**

The null values for the numeric features are filled with the mean values in preparation for the modeling.

**Transaction Date**

As the graph above in the Transaction Date of the Data Exploration and Visualization section shows. The fraud transactions are not uniformly distributed in each time frame. This makes the transaction time frame be a useful feature for the learning model. I have decided to use the day of week, hour of day and day of month as features for the training.

Since, the data set only includes transactions for 6 months out of a year, the month of year feature cannot be generalized for the rest of the months in a year. Therefore, I didn’t include the feature for the learning model.

The original feature, TransactionDT has been dropped from the training features because the specific time wouldn’t be a feature found common in the data for the model.

**OS Data**

An identity feature, id\_30 is an identification of device’s Operating System with version information. The version information of the OS would vary over the time, and new versions will appear in the future data. Particular version information may not match the version in the learning model that the feature may not be properly evaluated for the prediction.

In order to properly evaluate the feature for the learning model, the values are simplified to ‘Windows’, ‘iOS’, ‘Mac’ and ‘Android’.

**Browser Agent**

The feature, id\_31 has browser agent information with version number and platform information. The reasoning for not to include the version number of the OS is also applicable for the version number and the platform information that are included in the browser information feature.

The feature values have been generalized into browser names, ‘Chrome’, ‘Firefox’, ‘Safari’, ‘Edge’, ‘IE, ‘Samsung’, ‘Opera’ and ‘Others’.

**E-mail Domains**

The features P\_emaildomain and R\_emaildomain have many common email service providers, country specific top-level domains. New features are created by categorically divided by email service providers, and top-level domains

**Implementation**

**Utilities**

**Refinement**

**Results**

**Model Evaluation and Validation**

**Conclusion**

**Reflection**

**Improvement**

**References**

IEEE Fraud Detection competition: <https://www.kaggle.com/c/ieee-fraud-detection>

Data Set: <https://www.kaggle.com/c/ieee-fraud-detection/data>

<https://www.creditdonkey.com/credit-card-fraud-statistics.html>

https://www.kaggle.com/c/ieee-fraud-detection/discussion/101203#latest-628660

<https://www.consumeraffairs.com/finance/identity-theft-statistics.html>