**Quan Nguyen Hien, Vinh Nghiem To**

**DATA 325: Applied Data Sciences**

**Instructor Colby Long**

**Final Project**

**Predicting Fraud Detection in Online Transactions using Data Mining techniques**

**Abstract**

*Recently, the rise in digital crime is also more popular, especially fraudulent online transactions have been a threat of lowering the credibility of the banks and costs the economy millions of dollars per year. This study aims to examine the significant features as well as attempting to build a predictive model for fraud detection. The effect of some variables such as email domain and distance between billing addresses are discussed. The accuracy of some models is reported to be above 95%.*

1. **Introduction**

With the fast development of 4.0 technology, more and more transactions are preferred to be completed online. However, the rise in digital crime is also more popular, especially fraudulent online transactions, which is defined as a case when a person’s card is used without the owner’s authority, lowering the banks’ credibility. In general, credit card fraud is classified into two types, card-present fraud and card-not-present fraud. The former situation occurs when a person’s card’s information such as card number, expiration date, and CVC are taken physically. The second case occurs when the card’s details are known and exploited without the physical presence of the card in the transaction. This usually happens with online transactions. Although the introduction of chip-payment cards decreased point-of-sale fraud by 28%, card-not-present fraud increased significantly by 106% from 2015 to 2018 (Harrow, 2018). Therefore, it is necessary to figure out a solution to minimize and prevent the fraudulent practice, saving customers and financial companies time and millions of dollars each year. Data Science and Machine Learning techniques can take part in the data mining process to figure out the solution to such detection problems.

This paper focuses on using different Data Mining techniques to make predictions on a transaction to determine if it were fraudulent or not based on those factors from transaction-based and identity-based information. We use the statistical models to find the answer for the following two questions:

* **Are there some typical tracks likely to occur in fraud transactions (the number of decimal places in the transaction amount, the email domain, the distance, etc.)?**
* **Are there any factors’ threshold that can serve as essential metrics in detecting fraud?**

In a research paper published in 2016, Abdallah and his colleagues (Abdallah et al., 2016) provide the model for state-of-the-art fraud detection systems to collaborate with the fraud prevention systems. This gives a brief review of data mining techniques used in each topic of fraud detection: anomaly-based fraud detection and misuse-based fraud detection. In 2019, Cheng and colleagues proposed a method of detecting fraud on online transactions based on user behaviors (Cheng et al., 2019), which generalized each user’s transaction to find the optimal risk threshold to avoid the fraud issues. This covers the amount, time, location of the transaction, and more detailed information on each transaction, such as frequency, trends, states, etc., which can be a good start for our feature engineering process before finding the optimal results for the model.

In the research, we want to upgrade the techniques of using three popular Machine Learning methods: Support Vector Machines (SVM), Random Forest, and the probabilistic binary classification with Logistic Regression (Bhattacharyya et al., 2011) by applying those techniques to the problem of detecting online transaction fraud, figuring out each advantage and disadvantage of each algorithm by using k-fold cross-validation in the training process. Moreover, we also want to test and upgrade Random Forest’s methods on this problem since it was stated that it outperforms the Logistic Regression and Decision Tree (Patil and Soni, 2018). However, it leaves a problem of overfitting the trees-based memory as data increases, which might not be reliable for real-time data loading.

1. **Data Collection and Processing**
2. **Overview**

Our dataset was collected from IEEE-CIS Fraud Detection Contest, which was hosted on the Kaggle competition. There are 394 factors related to online tracked transactions and 41 factors associated with the identity of the transactions. Since the dataset was significant, we decided to use Python notebooks to pre-process the dataset, involving partitioning the data by its relationship with the response variable and figuring out a way of filtering N/A variables by its definition and collection process.

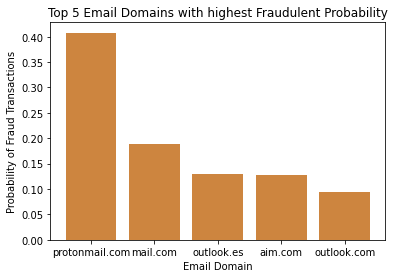
1. **Data Pre-Processing**

In terms of filtering data by its assumption, we decided to choose the factor that represents the distance between the purchaser and the receiver, email domain, transaction amount, and PCA-masked card, customer, and recipient information.

Prior to the process of analyzing and fitting data to the models by its assumption, we found that there were only 24.42% observations in the Identity data that matched the key (TransactionID) with the transactions in the Transactions main data, which contains the factors that represent the status of the Fraudulent transactions or not. Since this was a small number regarding all observations, we decided not to use the factors in the Identity data regarding a transaction to build a statistical model due to the imbalance of data, which might create a large bias for the model. Variables *C1-C14* are the counts of how many addresses are found to be associated with the payment card, phone number, email, device, billing, etc. The actual meaning is masked due to confidentiality.

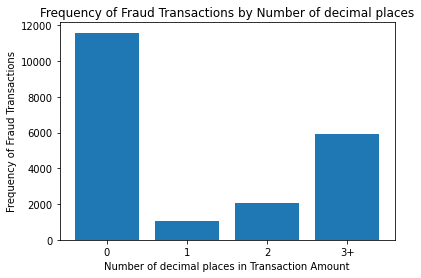
The N/A values would be removed when making aggregated calculations or statistical summary on those factors since they are specific to each transaction, making them individually independent of each other. Thus, we cannot replace those values with the nearest non-N/A value or regarding them as 0, which makes the analysis seem to be vague.

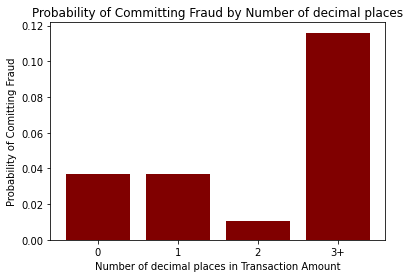
1. **Exploratory Data Analysis:**
2. **Email domain**



Firstly, we investigate whether some specific email domains are correlated with committing fraud since less common email domains may have weaker online security and verification methods. The table above shows the top 5 of the purchaser email domains that have the highest probability of committing fraud, which is approximately 10%, be at least. *Protonmail.com* purchasers perpetrate credit card fraud the most frequently, with 40.7% of them doing so. As researching further, *protonmail.com* is a free encrypted email domain, making it much more difficult to track users’ information and thus make it easier to perform fraud. Other domains include *mail.com* and *aim.com*, both of which are relatively obscure with little ìnformation found on the internet. On the other hand, a well-known domain such as Microsoft’s Outlook also appears in the top 5 is quite surprising. Still, one of them is located in Spain, which is far away from the receiver in the U.S., which in turn is more challenging to follow than other U.S.-based domains. The rest of the investigated email domains show only below 6% probability of being a fraud transaction. Thus, the purchaser email domain is a significant predictor in this research model.

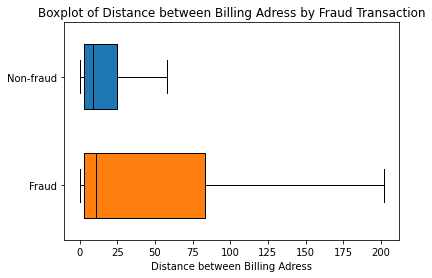
1. **The number of decimal places in transaction amount:**





A fascinating insight of this dataset is that there might be a relationship between the decimal places of the transaction amount and whether that transaction is fraudulent. In order to simplify and avoid bias, transactions with more than three decimal places will also be counted as three decimal places since they scarcely happen and might affect the overall model if it is too specific (the longest decimal is 30). According to the graphs above, those with three decimal places in transactions are three times more likely to commit fraud than those with only none and one decimal place in transactions. It is infrequent to have standard transactions and online payments to have over two decimal places in their value. Therefore, the intuition to have three or more decimal places in their transactions is suspicious, except for foreign exchange circumstances. In conclusion, the number of decimal places is also vital to include in our fraud detection model.

1. **Distance between sellers and buyers**



Another variable that is worth investigating is the distance between the purchaser and the receiver. Note that the outliers are removed from the plot because the variable ranged widely, so that it shrinks the boxplot into thick lines. According to the boxplots, the fraudulent transactions are located much further than normal transactions, which also emphasizes the result that transactions from foreign email domains have a higher probability of committing fraud. As a result, the distance from the purchaser to the recipient is essential to the predictive model.

1. **Modeling and Results**
2. **Assumptions**

During the process of analysis, we came up with the following assumptions:

* The distance from the site’s location where items were sold (location of the seller) to the transactions’ location was completed is positively correlated with the probability of a fraudulent transaction.
* Some unpopular email domains used by the purchasers might be suspicious of a fraudulent transaction.
* Number decimal in transaction amount can provide suspicious information for determining a fraudulent transaction.
* There should not be any difference among different cards used for transactions in determining the probability of a fraudulent transaction.

1. **Data Mining Techniques**

We decide to use the following Machine Learning and Deep Learning techniques to find the best model predicting the chance for a transaction to be fraudulent based on its factors. All of those Data Mining techniques, except for Neural networks would be evaluated using k-fold cross-validation. Afterward, the final model would be determined based on the one with the highest macro F1-score, which is the harmonic mean of average precision and average recall metrics since the data is imbalanced with the majority of non-fraud observations.

* **Logistic Regression**

In terms of Logistic regression, the set of chosen factors would be used to predict whether the transaction was fraudulent or not using the probability that those transactions would be fraudulent.

* **Decision Tree**

The Decision Tree splits the classification questions into sub-questions, in which the ultimate response would occur when no further gain of information could be made from a tree. We used Gini impurity to assess the training process of the model. On the other hand, the top 10 most important features to the response variables would be chosen after running the decision trees.

* **Support Vector Machines**

Support Vector Machines classification would find a hyperdimensional space that represents a boundary of partitioning two parts of the data, fraudulent transaction or not.

* **Neural Network**

Moreover, to find a better optimal solution for the problem, we decide to use a Simple Neural Network with the first hidden layer contains the number of nodes equal the dimension of the input factor matrix in which each node process an exploration on a specific feature), then using the Sigmoid activation function to get the probability of being fraudulent of a transaction for classification process in last hidden layer. In terms of setting up the neural networks, we use ADAM optimizer, which is the combination of Momentum and RMSprop optimizer, for our optimization problem, updating the weights with a batch size of 64 for 200 epochs.

**c. Results**

Examining the results found by the EDA process, we use different models to explore those relationships among factors and the response variable that indicates whether a transaction was fraudulent or not. All of those models use the same threshold of 0.4 (in terms of the probability of being classified as a fraudulent transaction).

|  | *Average Accuracy* | *Average Precision* | *Macro F1-score* |
| --- | --- | --- | --- |
| **Logistic Regression** | 96.4% | 0.736 | 0.52 |
| **SVM** | 95.6% | 0.714 | 0.58 |
| **Decision Tree** | 96.2% | 0.723 | 0.55 |
| **Neural Networks** | 97.1% | 0.690 | 0.70 |

From the above results table, we can see that although all of the chosen methods gave us high accuracy in predicting the fraudulent transaction. However, due to the skewness of the dataset (mostly non-fraud transitions), the highest macro F1-score of 0.7 regarding the usage of Neural Networks would be determined as our final predictive model. On the other hand, the results of each factor’s coefficient impact would be examined from the Decision Tree methods.

**5) Discussions and Conclusions**

1. **Discussions**

Card payment fraud is an act of criminal dishonesty, without a doubt. This article has listed the data mining-based fraud detection methods and their accuracy.

Using the importance score of the decision tree model, we have selected the top 10 most impactful features to detect fraud. These include distance, the purchaser’s email domain, transaction amount, and the payment card, purchaser, and recipient information. The model results also reconfirm that the top 5 suspicious email domains include those discussed in the EDA section. The transaction’s number of decimal places is a noticeable insight found from the EDA, and the threshold of 3 shows a significant climb in fraudulent probability. The results match the assumption that the larger the transaction value, the more likely it is a fraud. Therefore, transferring more money requires more verification documents in practice. However, we also found that distance has a negative relationship with fraudulent cases, which is the opposite of what we expected, yet it is the least impactful among the top 10.

The data mining models are run on a credit card dataset, and the accuracy of analytical models is evaluated with the help of a confusion matrix. Among the three models, logistic regression shows the best model in terms of precision, while Neural Network is the most effective regarding accuracy score and macro F1-score.

1. **Limitations and Suggestions**

Despite getting a good accuracy when using Neural Network as a final model for predicting the chance of a fraud transaction based on selective predictors, some bias on the assumptions might occur during the factor selection since some of the critical information in the dataset were masked for the privacy protection of the customers. Therefore, further research can be done to improve the model by doing an investigation on the masked factors to figure out a better solution. As a result, banks will build fraud detection models within their fraud risk infrastructure, using it as an end-to-end machine learning system in real-world problems.

**Works Cited:**

1. Abdallah, Maarof, Zainal. “Fraud detection system: A survey.” *Journal of Network and Computer Applications*, no. 68, 2016, pp. 90-113.
2. Bhattacharyya, Jha, et al. “Data mining for credit card fraud: A comparative study.” *Decision Support System,* no. 50, 2011, pp. 602-13.
3. Chen, Zhang, et al. “A Method for Online Transaction Fraud Detection Based on Individual Behavior.” *ACM TURC-AIS 2019 Conference on Artificial Intelligence and Security,* 2019.
4. Patil, Soni. “Predictive Modelling for Credit Card Fraud Detection using Data Analytics.” *Procedia Computer Science*, no. 132, 2018, pp. 385-95.
5. Harrow. “Is Your Credit Card Less Secure Than Ever Before?” *Forbes*, April 2018. <https://www.forbes.com/sites/robertharrow/2018/04/20/is-your-credit-card-less-secure-than-ever-before/?sh=1b5c3b105c8d>