Transaction Fraud Detection using Random Forest and Logistic Regression Methods

Zengxiaoran (Shawn) Kang, Kazy Atausha

December 9, 2021

1. Introduction

1.1 Importance of Study

Since the explosion of wire transactions and online banking, financial fraud has been on the rise. According to a report by TransUnion [10], a consumer credit reporting agency, global digital fraud attempts in the financial services sector have risen 149% in the first four months of 2021 compared to the last four months of 2020. In the U.S., digital fraud attempts have risen 109% in the first four months of 2021. Furthermore, a recent study from Juniper Research [5] stated that digital money transfer fraud is growing, especially in the emerging eCommerce market, which would expect a loss of over $200 billion between 2020 and 2024. Due to advances in technology, fraudsters have increased their avenues of attempting fraud as well as increased sophistication in their attempts. As the financial services industry continues to find new ways to interact with consumers, fraud analytics is an increasing area of importance.

Our study is designed to explore transaction fraud in the context of banking services by promoting a statistical model that can make binary decisions on whether transactions are fraudulent based on key characteristics.

1.2 Aims & Hypothesis

Our first aim is to identify which factors are most important for detecting fraudulent transactions. Our hypothesis is that type of transaction, time, and the balance of the sender and the receiver before and after the transaction will be the most predictive characteristics of fraudulent transactions. We suspect that the types of transactions most susceptible to fraud are transfer and credit transactions. Timing can play a crucial role in identifying fraudulent transactions since there might be certain periods during the day where fraudulent behavior is more prevalent.

Our second aim is to create a statistical model that can predict whether a transaction will be fraudulent based on the transaction's characteristics.

2. Data

To secure customers' confidentiality and privacy, large financial institutes tend not to release public data that simultaneously prevents academic researchers from understanding the money transaction domain in financial services. The lack of public resources is a cause of the intrinsically private nature of financial transactions. Therefore, we will import data from Kaggle's blog [4], a synthetic dataset generated by the PaySim mobile money simulator.

We use PaySim's aggregated data that resembles real transactions with some artificial noise. We are aware that the malicious behavior is manually injected so that the response can be treated as the ground truth. We also want to clarify that the dataset has a sequential feature that maps a unit of time in the real world. For example, one unit of step is equal to one hour with a total of 744 steps in a 30-days simulation. In an overview, the dataset has features describing time, payment type, documented transaction amount, source balance, destination balance, and the fraud flag.

3. Research Strategy

3.1 Exploratory Analysis

In our exploratory analysis, we will observe if there exists an imbalance between classified fraud and not fraud data on a bar graph. We will first figure out what kinds of transactions are most used by fraudsters. Then, we will compare the difference in transaction amounts between source and destination balance to discover if discrepancies exist. Furthermore, the distribution of transaction amount for fraud or not fraud can be studied with a histogram. Finally, we will study the relationships among features with visual supports such as the correlation matrix plot.

3.2 Statistical Analysis Strategy 1

For our first model approach, we will use a random forest model. A random forest model uses a large number of decision trees that work together by bucketing observations by key characteristics. We will also evaluate the model's predictive ability to classify transactions as fraudulent or not.

3.3 Statistical Analysis Strategy 2

We will apply logistic regression, which is a binary classification model that estimates the probability of fraud based on the weights given to transaction characteristics. In addition, we will use two different variable selection methods: Lasso regression and recursive feature elimination. We will also evaluate the model's predictive ability to classify fraud transactions based on different probability cutoffs.

4. Exploratory Analysis

4.1 Imbalance of Fraud in Transaction Type

In our exploratory analysis, we found that transaction fraud only occurs in cash-out and transfer transactions. Out of the 2.8 million total cash-out and transfer transactions that occur, only about 0.3% are fraud. In addition, the dataset contains information on whether the transactions were legitimately fraudulent or not and if the bank noticed the fraud by flagging these transactions. The bank only classified 0.2% of fraudulent transactions as fraud, indicating a large area in which our analysis could improve the results. The cash-out and transfer transactions have roughly equal amounts of fraudulent cases because a fraudulent transfer is always followed by a fraudulent cash-out transaction within an hour. About 99% of fraudulent cases follow that pattern. There can be several reasons as to why the Bank does not recognize these fraudulent transactions, but it is evident that there is a weak point in the transaction process.

4.2 The Discrepancy Between Transaction Amounts and Changes in Senders' Balances

About 85% of all transaction records have a discrepancy between the bank's documented transaction amount and the change in the sender's balance during the transaction. Interestingly, in the graph below we observed that there are a large number of legitimate transactions that have larger discrepancies than that of fraudulent transactions. In the case of fraudulent transactions, there seems to be a ceiling to how large the discrepancy between the bank's records and the sender's balance is tolerable. Perhaps if the discrepancy is too large, it will trigger an alert to the bank, so fraudsters do not go over the limit. These discrepancies will be influential in detecting fraud transactions, and we will take this into consideration in our analysis.

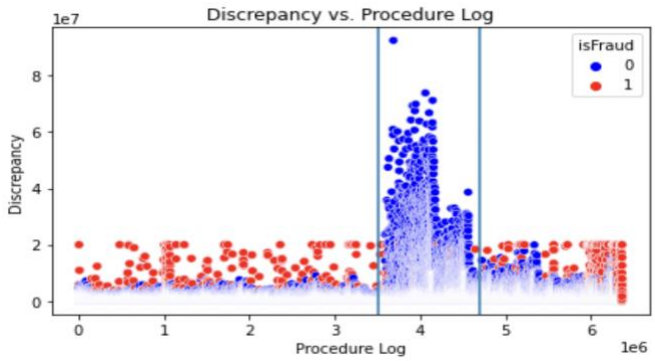

Figure 1. Effects of discrepancy on fraud or genuine transactions. The discrepancy is the sum of the absolute value of error in the sender's balance and the absolute value of error in the receiver's balance.

4.3 Correlation Between Variables

We calculated the correlation coefficients between variables, and unsurprisingly, we found a strong relationship between the transaction amount and the receiver's balance before the transaction, the transaction amount and the receiver's balance after the transaction, and the transaction amount and the discrepancy. The relationships between other variables are weak.

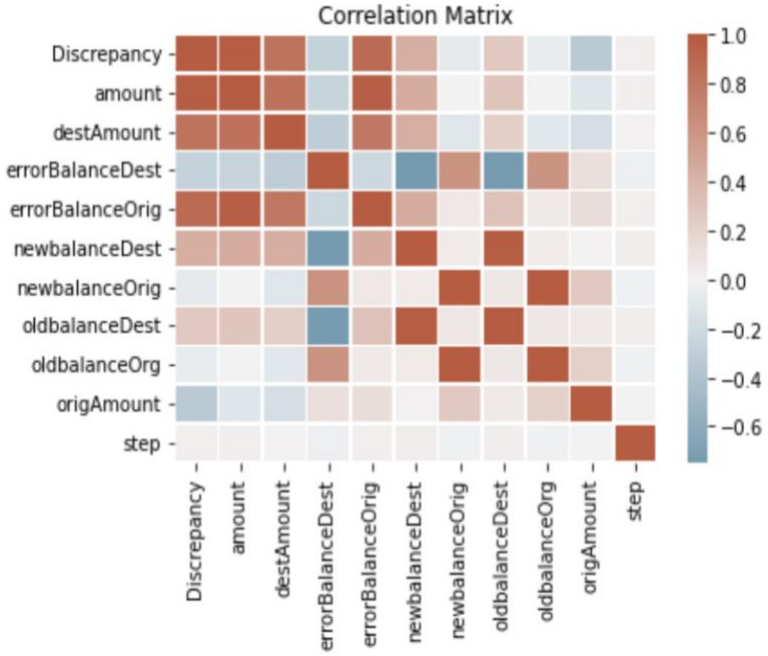

Figure 2. Correlation matrix among all variables.

4.4 Outliers on Variables

Since the distributions of all variables are highly skewed, we applied the Robust Z-Score and Winsorization [9] method to look at the outliers. We primarily checked on the bank's documented transacted amount because other variables have similar outliers. Robust Z-Score uses median and absolute deviation from the median to determine outliers, which detects about 14% outliers in fraud data and 0.9% in non-fraud data. On the other hand, Winsorization finds outliers if the data falls outside the 1st or 99th percentile, and it detects about 6% outliers in fraud data and 2% in non-fraud data. Since both detected outliers are mostly large-scale transfer transactions between businesses, we concluded not to exclude them because they support practical objectives.

5. Random Forest Model Approach

5.1 Methods

We fit the training data into a single decision tree to estimate the optimal tree size and applied the same level to each tree in the random forest classifier model built with 500 decision trees. Then we evaluated the result with the testing data and tuned the threshold with AUC-ROC and confusion matrix.

5.1.1 First Attempt in Single Decision Tree

We attempted using the Rpart library in R [6] to compute a single decision tree. The complexity parameter (CP) is used to control the size of the decision tree and we selected the optimal tree size with 6 level depth according to the minimum error described by CP.

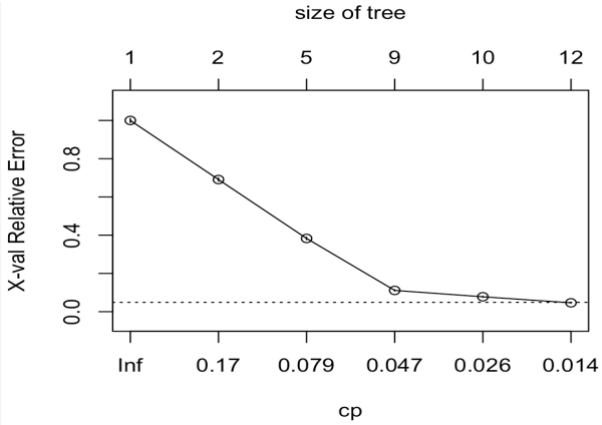

Figure 3. Selection of a single tree depth based on CP.

The single classification tree as illustrated in Figure 4 had 85.4% accuracy but resulted in a nearly 38% false-positive rate. In the tree, we observed that the sender's change in balance and the receiver's balance after transaction have leading impacts on classification decisions.

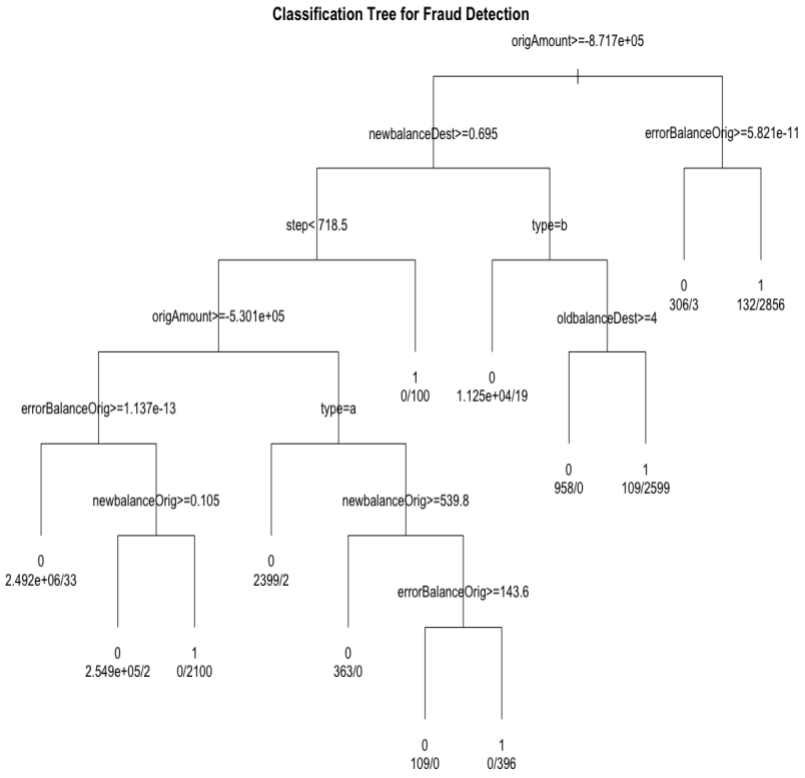

Figure 4. Single Decision Tree

5.1.2 Random Forest

The random forest approach builds each tree with bagging and incorporates randomness to create an uncorrelated random forest. The random forest predictions are typically more accurate and stable than a single tree [2]. Therefore, we advanced to the random forest, which is better suited to the non-parametric data. We created 500 decision trees and each with a maximum 6-level depth based on the level of the previous single decision tree.

5.2 Results

5.2.1 Variable Importance

When we fit the training data using the random forest classifier, we observed variables with high importance that were drivers of the outcomes. As shown in Figure 5, we confirmed that the sender's change in balance, the sender's new balance after the transaction, and the error in the receiver's balance after the transaction are the leading variables in classifying fraud transactions.

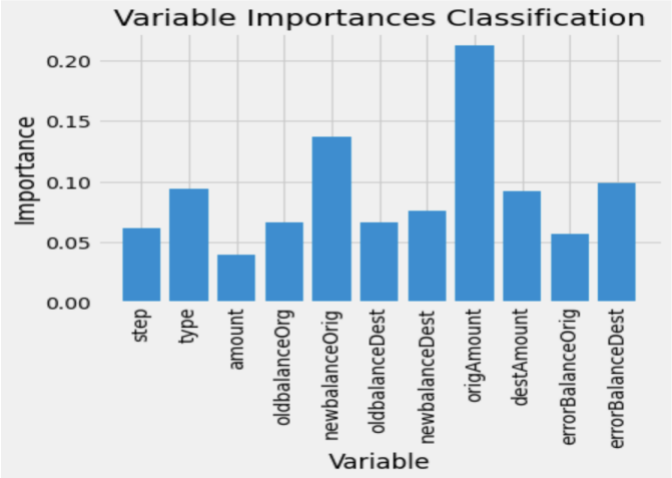

Figure 5. Variable Importances In Random Forest.

5.2.2 Out of Bag (OOB) Score

According to the Scikit-Learn library [7], the OOB score is the average error for each decision tree, which was calculated using data that do not contain each tree's respective bootstrap sample with a 0.5 cutoff. We used the OOB score to evaluate the classification accuracy of our model, which was higher than 99%, and overall, the score stated that the model is capable of distinguishing between genuine and fraudulent transactions.

5..3 Confusion Matrix and Adjusting Threshold with the Area Under the ROC curve (AUC)

We evaluated the random forest classifier model using testing data. Since we had imbalanced numbers of genuine and fraudulent transactions in the data set (i.e., only 2% of million transactions were fraud), the 99% OOB score was not effective due to the dominant number of true negative classifications.

The AUC-ROC curve measures the classification performance at various threshold settings. We dived deeper into the confusion matrix and tuned the threshold to optimize our detection model based on the AUC-ROC curve. We tuned the threshold to 0.205, which indicated marking the transaction as fraudulent if the probability of a transaction being fraud exceeds 0.205. The 0.205-threshold had better performance than the default 0.5-threshold as illustrated in Figure 6.

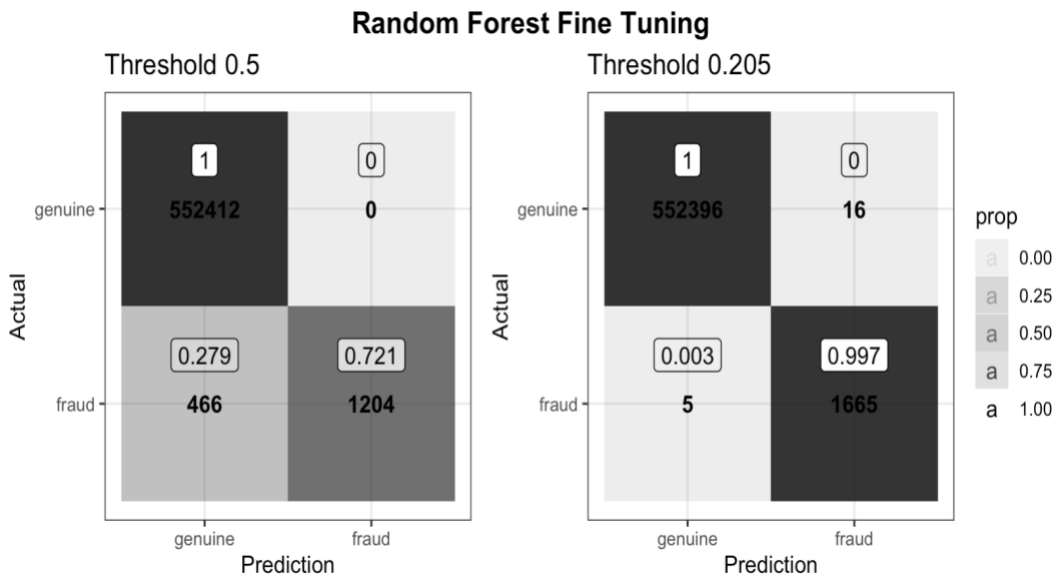

Figure 6. Confusion matrices of random forest classification results with standard and tuned thresholds. The number in the box is the percentage of classified cases in response to the total number of actual cases in the testing data. The number below is the number of cases being classified correspondingly.

6. Logistic Regression Model Approach

6.1 Methods

We transformed some variables and fit the training data using a logistic regression [8] model with variables selected from two variable selection methods: recursive feature elimination (RFE) and least absolute shrinkage and selection operator (Lasso) regression. Then we evaluated the result with the testing data and tuned the threshold with the AUC-ROC curve and confusion matrix.

6.2 Variable Transformation

Before we fit the training data with a logistic regression model, we transformed some variables to have a better fit for our model. Previously, the distributions of some of the most important variables were highly skewed, and transforming them reduced the impacts of outliers and skewness.

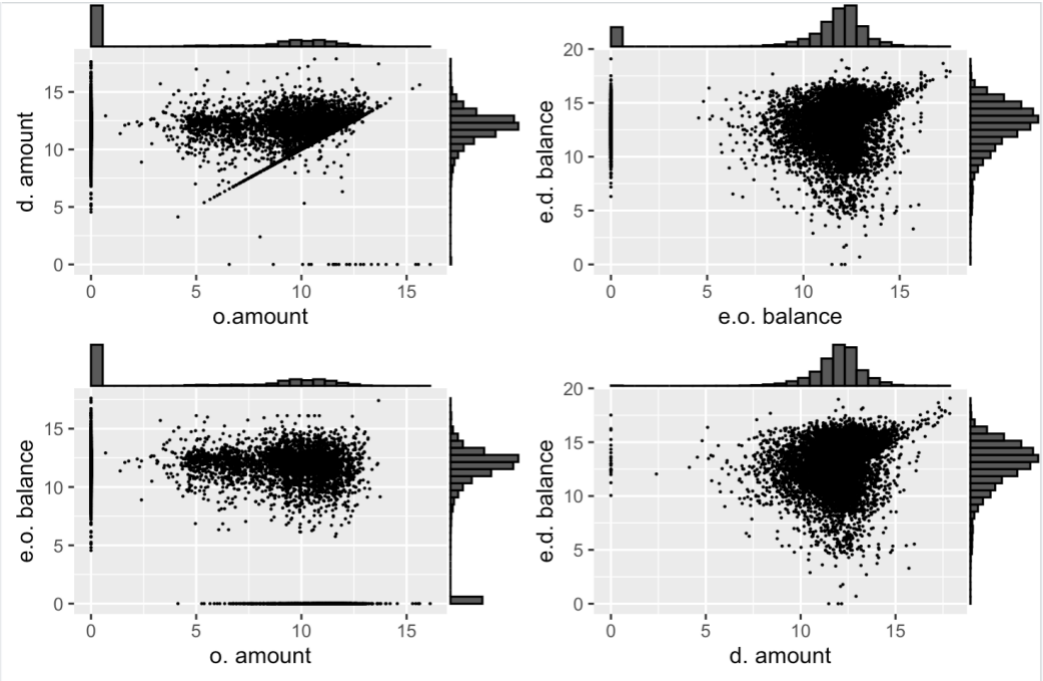

Figure 7. Joint distributions of some important variables after transformation. We take the absolute value so we can account for negative values, add 1 to account for zeroes, and finally we take the log transformation. In the image, 'e' stands for error, 'd' stands for the destination (i.e., transaction receivers), and 'o' stands for the origin (i.e., transaction senders).

6.3 Variable Selection

To create an efficient model with accurate results, we looked into two approaches for selecting the most relevant variables and discarding noise variables: RFE and Lasso regression.

RFE [1] is a feature selection algorithm wrapper-style feature selection algorithm that also uses filter-based feature selection internally. It takes input on the number of features one chooses for the model and ranks the features that are the most important out of the entire pool of variables. In addition, the importance calculation is model-based (e.g., the random forest importance criterion) [1]. We paired this with a wrapped algorithm that uses RFE to select the best variable for a single variable model to a model that includes all the features. We discovered that a model that uses 8 features had a 99.86% accuracy classification score. The RFE algorithm selected the time, type of transaction, amount being transacted, the sender's balance before and after a transaction, the receiver's balance after a transaction, and the error in the balance of the sender and receiver.

Lasso regression is very similar to linear regression. However, the cost function includes a penalty term where the goal is to impose a constraint on the model parameters that causes regression coefficients for some variables to shrink towards zero [3]. Through this penalty term, Lasso regression drives the coefficients of irrelevant variables to 0, and thus automatically selects important features. We found that Lasso selected the amount being transacted, time, type of transactions, and errors in the balance of the receiver and sender to be the most important features.

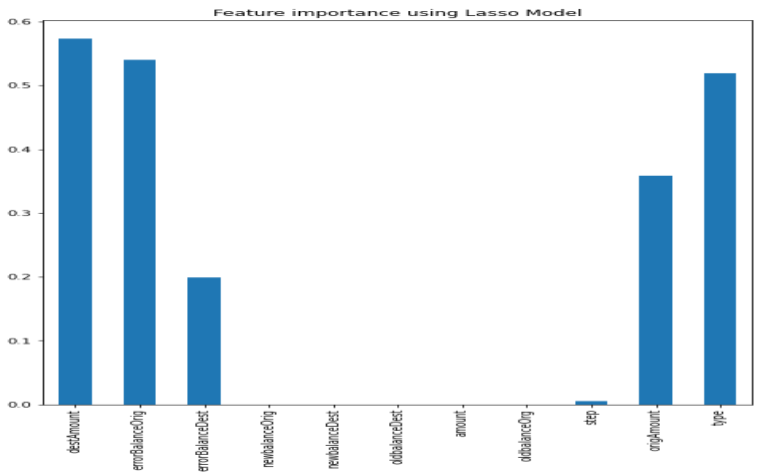

Figure 8. Lasso regression coefficients. The size of each feature coefficient is indicated with a blue bar.

The main difference between the variables selected by Lasso compared to RFE is that there are two fewer variables selected: the balance of the sender before and after the transaction. Given that the two variables were highly correlated and that the error in balance of the receiver was also highly correlated with the two variables, we concluded that the most important variable out of the three was the error in the balance of the receiver.

6.4 Results

6.4.1 Logistic Regression on Threshold 0.5

For our logistic regression models, we evaluated the results using testing data and compared the difference between models whose variables were selected via RFE and Lasso. The model that used the 8 variables selected from the RFE had a 99.85% accuracy classification score. However, the model had significant numbers of false positives and false negatives. As shown in Figure 9, the model misclassified about 46% of actual fraud transactions as not fraudulent. The model that used the variables selected through Lasso had better results with about a 99.9% accuracy classification score. Moreover, the model misclassified about 31% of actual fraud cases as not fraudulent as compared to 46% from the model that used variables selected through RFE.

6.4.2 Adjusting Threshold with the Area Under the ROC curve (AUC)

We further explored ways we could improve the models with variables selected through RFE and Lasso. We used the AUC-ROC curve to select the best threshold for classifying transactions as fraud or genuine. Under the model with variables selected through RFE, we found that the best threshold for our model was 0.225, which resulted in about 36% of actual fraud transactions being classified as not fraudulent.

The model that used variables selected through Lasso outperformed the one with variables selected through RFE. By adjusting the threshold to 0.105, the Lasso model only misclassified 23% of actual fraud transactions as compared to 31% when using a 0.5 threshold.

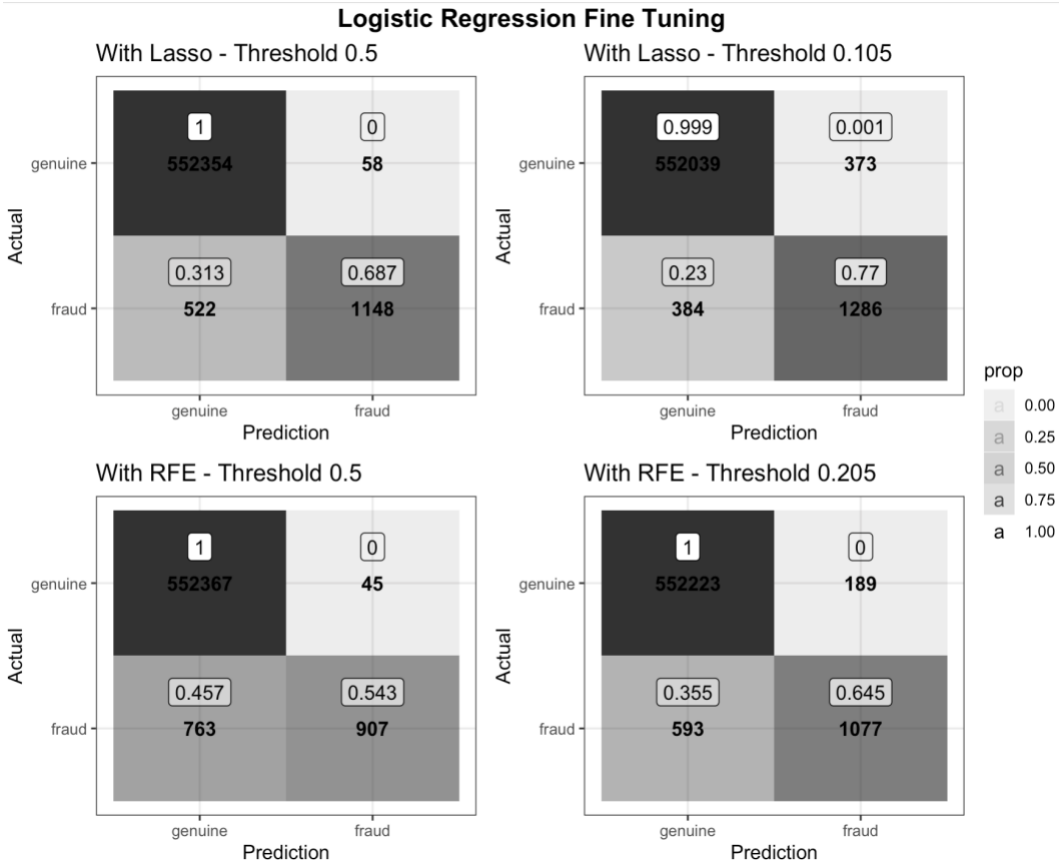

Figure 9. A set of confusion matrices for the logistic regression classification result with features selected from Lasso and RFE under different thresholds. The number in the box is the percentage of classified cases in response to the total number of actual cases in the testing data. The number below is the number of cases being classified correspondingly.

7. Discussion

One of the key outcomes of this paper was to determine which variables are most important to classify whether a transaction is fraudulent or not. Between RFE and Lasso for logistics regression, the variables selected via Lasso were better at predicting fraudulent transactions than the logistics regression model that used variables selected by RFE. The key variables selected by Lasso were the amount being transacted, time, type of transactions, and errors in the balance of the receiver and sender to be the most important features. In comparison to the random forest model, virtually the same features that were selected through Lasso for the logistic regression model were the most important variables for the random forest model.

Despite having better performance than the logistic regression model, the random forest model does not explicitly define the relationship between the response and the variables. Banks have many regulatory requirements that may make it more difficult to implement such a model as we do not fully understand the underlying relationship between the variables and the response. In comparison to the random forest model, the logistic regression model clearly defines the relationship between variables. However, in a practical sense, the main drawback of the logistic regression model was that the model misclassified many fraudulent transactions. The consequence of the logistic regression model is that the

bank is losing money and therefore losing income. Both models have barriers that make it difficult to apply practically due to either regulatory constraints or simply the model not being able to classify to a more satisfactory level.

In industrial applications, banks would have thousands of different variables to help identify whether transactions are fraudulent, such as credit score, debt to income ratio, total debt, location, geography, age, and more. Since our dataset did not contain these variables, those may be the limiting factors for improving our models. In addition, we attempted two different approaches out of many, yet a more complete exploration would include a plethora of classification models.

References

[1] Brownlee, J. (2020, August 27). *Recursive feature elimination (RFE) for feature selection in Python*. Machine Learning Mastery. Retrieved December 8, 2021, from https://machinelearningmastery.com/rfe-feature-selection-in-python/.

[2] Deng, H. (2021, April 26). *Why random forests outperform decision trees*. Medium. Retrieved December 8, 2021, from https://towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5.

[3] *Least absolute shrinkage and selection operator (lasso)*. Search the website. (n.d.). Retrieved December 8, 2021, from https://www.publichealth.columbia.edu/research/population-health-methods/least-absolute-shrinkage-and-selection-operator-lasso.

[4] Lopez-Rojas, E. (2017, April 3). *Synthetic financial datasets for fraud detection*. Kaggle. Retrieved December 8, 2021, from https://www.kaggle.com/ealaxi/paysim1.

[5] *Online payment fraud losses to exceed $200 billion by 2024*. Online Payment Fraud Losses to Exceed $200 Billion By 2024. (n.d.). Retrieved December 8, 2021, from https://www.juniperresearch.com/press/online-payment-fraud-losses-to-exceed-200-billion.

[6] *Rpart: Recursive partitioning and regression trees*. RDocumentation. (n.d.). Retrieved December 8, 2021, from https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart.

[7] *Sklearn.ensemble.randomforestclassifier*. scikit. (n.d.). Retrieved December 8, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[8] *Sklearn.linear_model.logisticregression*. scikit. (n.d.). Retrieved December 8, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[9] Stephanie. (2020, December 11). Winsorize: Definition, examples in easy steps. Statistics How To. Retrieved December 8, 2021, from https://www.statisticshowto.com/winsorize/.

[10] TransUnion. (2021, June 2). Suspected Financial Services Digital Fraud attempts rise nearly 150% worldwide as prevalence of digital transactions increase. Retrieved December 8, 2021, from https://newsroom.transunion.com/suspected-financial-services-digital-fraud-attempts-rise-nearly-150-worldwide-as-prevalence-of-digital-transactions-increase/.