Credit Fraud and Modeling Predicition of Fraudulent Transactions

Senior Capstone 01

South Dakota State University

Department of Mathematics and Statistics

Advised by Dr. Fred Boehm

Andrew Marchant and Gage Scholting

Spring 2025

Abstract

With the increasing shift from physical to digital transactions, financial institutions face a growing challenge in detecting and preventing fraudulent activity. We leverage a publicly available dataset containing over 1.85 million credit card transactions to identify patterns of fraud and develop a predictive machine learning model. We begin with exploratory data analysis to examine key characteristics of both fraudulent and non-fraudulent transactions. We then implement and compare the performance of multiple machine learning algorithms to assess their effectiveness in detecting fraud. Based on these evaluations, we determine a final model for deployment and discuss potential areas for future improvement to enhance detection accuracy and model robustness.

Credit Fraud and Modeling Prediction of Fraudulent Transactions

1 Background

With the growing rate of digital accounts and payments, fraudulent transactions pose a significant challenge for both financial institutions and cardholders. Credit fraud is typically defined as a form of identity theft where an individual who is an unauthorized cardholder gains access to either a physical credit/debit card or someone else's bank account and attempts to withdraw or spend any amount of money. Many data scientists and financial institutions have attempted to build machine learning models that predict 100 percent of fraudulent transactions, while balancing recall, accuracy, precision, and a high F-1 score. While multiple machine learning models such as decision trees, random forest, and anomaly detection may be used to tackle this kind of imbalanced dataset, settling on the best model and improving it is most important for this scope of work. While modeling this data is important, providing insightful and meaningful information in visualizations is important to help explain what is happening when a transaction is processed. The goal of our analysis and model is to accurately diagnose a fraudulent transaction on the cardholders bank statement to be able to alert the customer of the fraudulent purchase.

2 Introduction to Data

The data used was found in a publicly available credit card transaction data set from Kaggle [4]. Originally, it was compiled by an undisclosed company that measured transaction times, amounts, and associated personal and merchant information on credit card transactions from January 2019, to June 2020. The data set contains over 1.85 million transaction instances with 973 unique cardholders, along with the following key variables:

- **Transaction Details:** Transaction timestamp, merchant name, category, transaction amount, and unique transaction number.
- **Cardholder Information:** First and last name, sex, address, city, state, ZIP code, job, and date of birth.
- **Geographic Data:** Latitude and longitude of both the transaction location and the merchant location.

- **Fraud Indicator:** A binary indicator that specifies whether a transaction is fraudulent.

The Kaggle site gives the following prompts to investigate given the data: fraud detection, customer segmentation, transaction classification, geospatial analysis, predictive modeling, behavior analysis, and anomaly detection. Having an interest and background in the field of geography, along with recognizing the growing need to disrupt the illicit economy, we are keying in on characteristics of fraud and techniques to identify anomalies representing fraud combined with geographic visualizations.

3 Visualizations and Conclusions

To understand fraudulent transactions, we begin by creating visualizations and performing exploratory data analysis. We categorize our findings into four main groups: cardholder characteristics, fraud vs. non-fraud transactions, geographic analysis, and time series exploration.

3.1 Cardholder Characteristics

We start by examining the characteristics of cardholders, trying to spot spending patterns for different demographics. Given the date of birth, we calculate a new column: age. Initially, we want to confirm our dataset is unbiased by looking into age and gender distributions across customers shown in the figures 1 and 2.

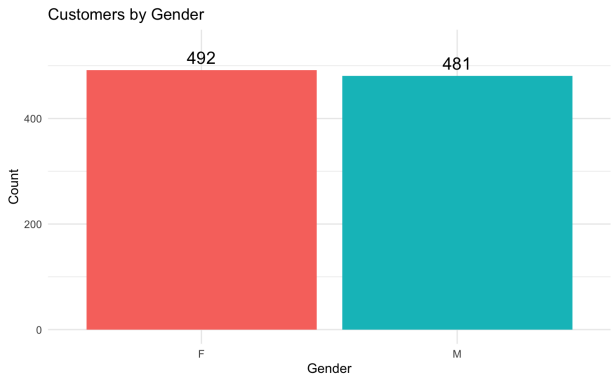


Figure 1: Count of male and female customers [7]

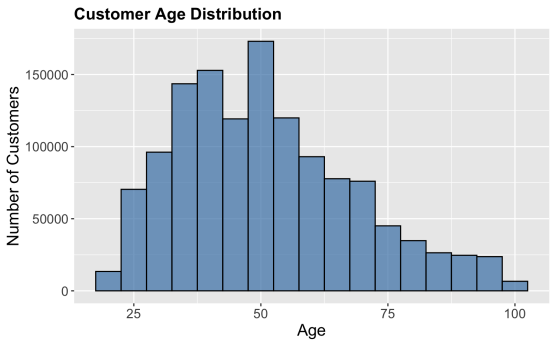


Figure 2: Distribution of age across customers [7]

From Figure 1, we see that our data set is balanced among female and male customers. Figure 2 shows the age distribution is relatively normal, though slightly skewed towards younger customers. Normality and balance among these variables prove beneficial in future modeling. Figures 3 and 4 illustrate the likelihood of fraud among gender and age demographics.

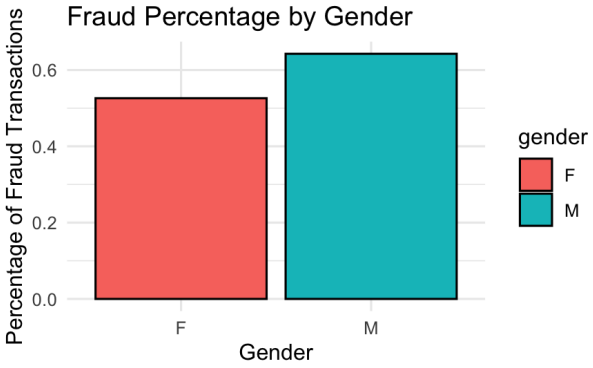


Figure 3: Percentage of fraud transactions among gender demographic [7]

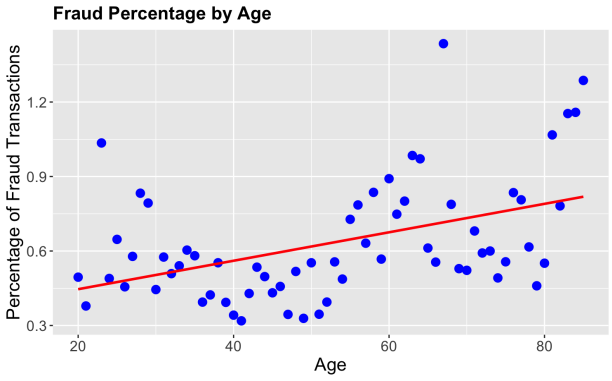


Figure 4: Percentage of fraud transactions among age demographic [7]

Figure 3 shows the percentage of fraud transactions for both male and female customers, with males exhibiting a higher fraud occurrence, suggesting they are more likely to be victimized. Similarly, Figure 4 indicates fraud is more prevalent for ages over 60. From this plot, we estimate fraud percentage is almost double for ages 30-50 to 60+. Examining cardholder demographics helps identify likelihood of fraud. We also compare fraud and non-fraud transactions to identify patterns.

3.2 Comparison of Fraud vs. Non-Fraud Transactions

Directly comparing patterns in fraud and non-fraud transactions helps us identify when fraud is more likely to occur. We begin by looking at the distribution of purchase amounts seen in Figure 5.

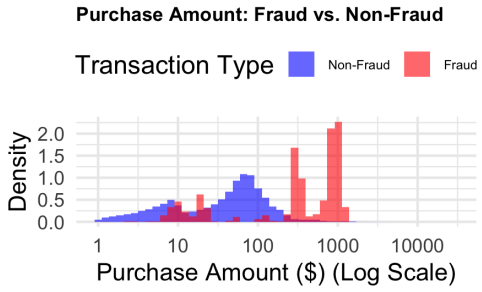


Figure 5: Distribution of purchase amount density for fraud and non-fraud [7]

Figure 5 provides insight into purchase amount trends for fraud and non-fraud transactions. In order to account for the large range of purchase amount values, a log scale was used to better visualize the distribution. Non-fraud purchases are distributed more normally, while fraud transactions seem to be categorized into small and large purchases. Next, we investigate the distribution of the hour of day of purchases for fraud and non-fraud transactions.

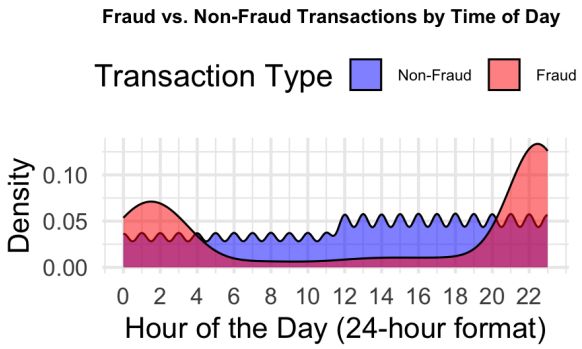
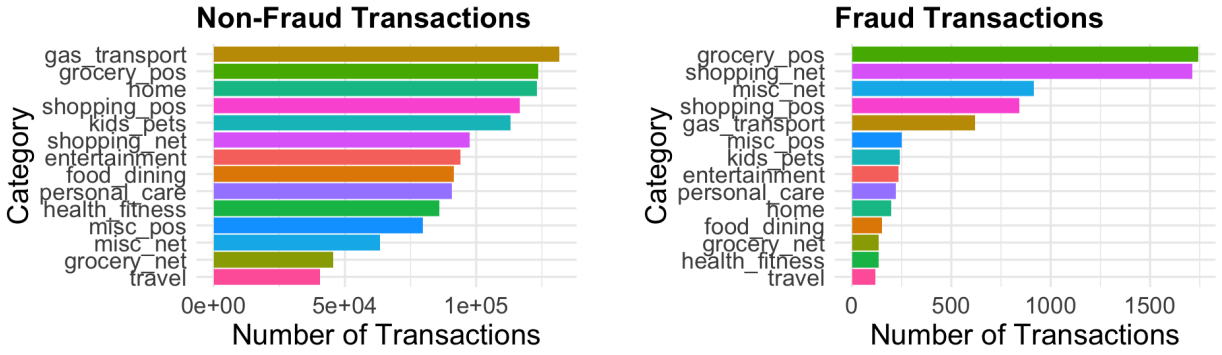


Figure 6: Distribution of hour of purchase for fraud and non-fraud [7]

Figure 6 shows that fraudulent transactions occur at a much higher rate during later hours of the day, while non-fraud transactions are roughly uniform for the hour of purchase. Lastly, we compare how purchase categories differ among fraud and non-fraud transactions.



(a) Non-Fraud Transactions

(b) Fraud Transactions

Figure 7: Comparison of purchase categories for fraud and non-fraud transactions [7]

From Figure 7, non-fraud transactions cover a wide range of purchase categories, with none standing out from the others. On the other hand, fraud purchases show clusters in a few main categories. In particular, fraudulent purchases are concentrated in categories such as grocery, online shopping, point-of-sale shopping, miscellaneous, and gas.

Comparing fraud and non-fraud purchases, we see a common trend among all the figures. Non-fraud transactions seem to have characteristics that are continuously spread between purchase amount, time of day, and category of purchase. Conversely, fraud transactions are more discrete, occurring with specific purchase amounts, times of purchase, and even categories of purchases.

3.3 Geographic Analysis

We examine geographical trends among transactions to gain a sense of the spatial distribution of our dataset. Figure 8 displays the count of customers in each US state.

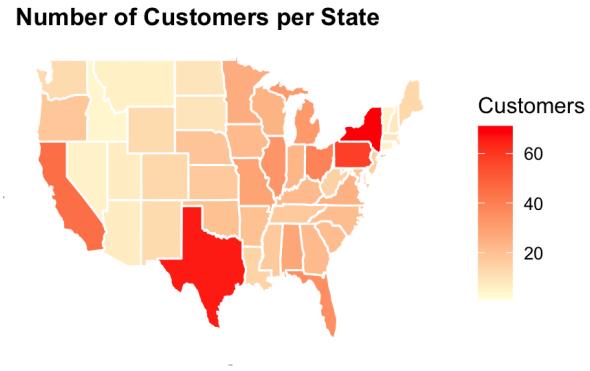


Figure 8: Distribution of customers per U.S state [7]

Figure 8 shows that states such as Texas, California, and New York have a high number of customers, likely due to abnormal population size. Although this figure indicates where the majority of transactions are located, it is also important to understand where fraud is more likely to occur. Figure 9 gives the rate of fraud in each state.

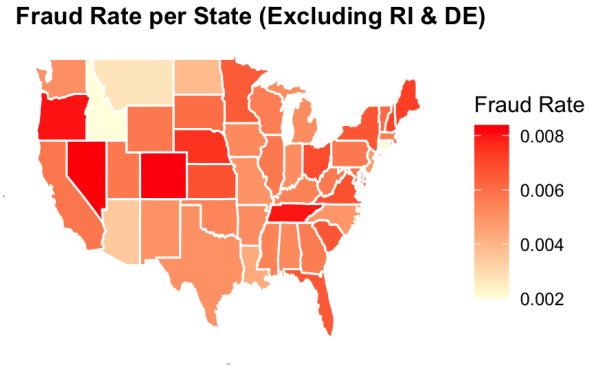


Figure 9: Fraud rate per each U.S state [7]

Figure 9 excludes Rhode Island and Delaware, as these states are outliers, likely due to the insufficient number of transactions. However, this figure gives valuable insight, highlighting states such as Nevada and Oregon, where fraud is more prevalent, compared to states like Idaho, where fraud rates are relatively low.

We also attempt to plot individual customers' transactions to identify geographic differences between fraud and non-fraud activities to help enhance fraud detection. However, we encounter an issue with the longitude and latitude data, as all points form a square around the customer's original address. This prevents us from identifying specific patterns in fraudulent activity. We plan to investigate and address this issue in future work.

3.4 Time Series Analysis

In addition to the above, we perform a time series analysis to examine the dynamics of fraud. Time series visualization is particularly useful for detecting seasonality and trends in fraudulent transactions. Figure 10 shows the time series plot of fraud counts. From the seasonality in the figure, we predict fraud is more likely to occur around the holiday seasons, such as Christmas. This analysis not only helps us understand when fraud is most likely to happen but also informs the development of predictive models for fraud detection.

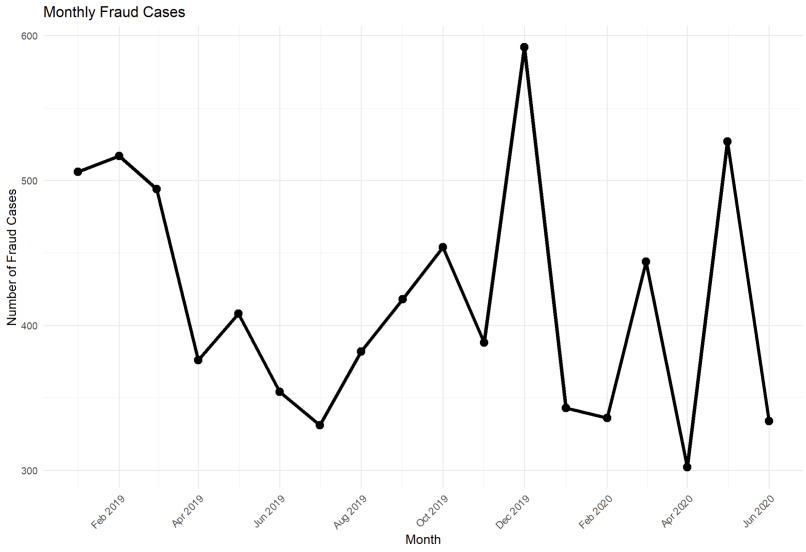


Figure 10: Time series analysis of fraud transactions [7]

In summary, our exploratory analysis and visualizations reveal key patterns that distinguish fraudulent transactions from legitimate ones. The examination of cardholder demographics indicates that fraud tends to be more prevalent among male cardholders and individuals over the age of 60. Comparing fraud and non-fraud transactions shows that fraudulent activities occur more predictably in discrete clusters in purchase amounts, later hours of the day, and within limited

purchase categories. Finally, our geographic analysis, despite some limitations in location precision, identifies regions where fraudulent activity is more common. These insights lay a strong foundation for fraud detection models and more targeted prevention strategies in future work.

4 Techniques

4.1 Introduction to Modeling

Modeling is most often used when trying to predict a certain value or place a value within a certain group. As technology and eventually the creation of machine learning evolved, many different techniques have been created and perfected to fit a multitude of needs in the realm of machine learning. There are many different ways to evaluate and determine the fit of a model, but for the scope of our project a confusion matrix will determine how each of our following 3 models stack up against each other. These 4 models will look at the dataset, build a model based on data used to train each respective model, then finally test the models performance against a test portion of the data to see if the model can correctly predict fraudulent and non-fraudulent transactions. Here are the 4 possible outcomes in the confusion matrix and what each means.

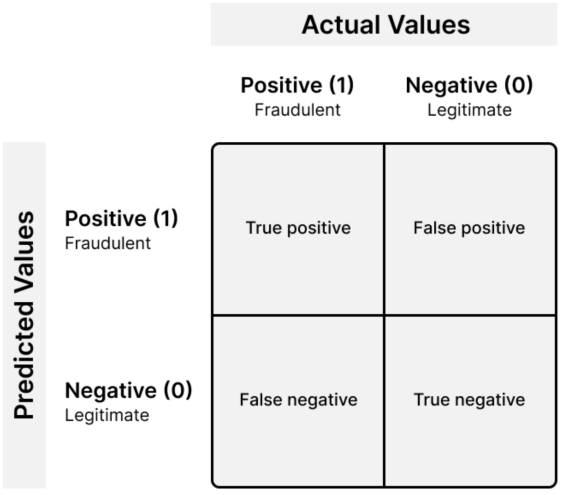


Figure 11: Confusion Matrix for Fraudulent Transactions

As Figure 11 states the 4 different cases in the confusion matrix, here is an english version as to what each box means:

- **True Positives (TP):** Correctly predicted fraudulent transactions.

- **False Positives (FP):** Non-fraudulent transactions incorrectly predicted as fraudulent.
- **False Negatives (FN):** Fraudulent transactions incorrectly predicted as non-fraudulent.
- **True Negatives (TN):** Correctly predicted non-fraudulent transactions.

Looking at the confusion matrix and the true meaning of each box, the goal is to build a model that correctly predicts all True Positive (TP) transactions as well as all True Negative (TN) transactions. A perfect model would only have values in the true positive and true negative boxes, with zero values in the false negative and false positive boxes. With this in mind, three different models will be built trying to minimize the values within the false positive and false negative boxes. It is important to note that a false negative transaction is much worse than a false positive transaction. A customer of a financial institution would much rather be notified of a transaction they made than not be notified that someone else is spending money on their card. The 3 models being built will be Decision Trees, Random Forest, and Anomaly Detection. Each of these models is unique in their own way and have their respective upsides and downsides, but we will ultimately decide on one model to tune and perfect. The 4 key performance indicators are Accuracy, Precision, Recall, and a final F-1 Score, which helps give quantitative comparisons in respect to each model. Here is what each of these categories mean, and how it is calculated:

- **Accuracy:** Measures the proportion of correctly predicted observations out of the total observations.

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

- **Precision:** Measures the proportion of correctly predicted positive observations out of all predicted positives.

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

- **Recall (Sensitivity):** Measures the proportion of correctly predicted positive observations out of all actual positives.

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

- **F1-Score:** Harmonic mean of precision and recall; balances the two in one metric.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

After letting the R Software [7] calculate these KPI's, we are able to compare and contrast the 3 respective models, which will help us decide on which base model will be tuned and refined for our final credit fraud detection model.

4.2 Decision Tree (CART)

The first machine learning technique that we will focus on is a particular decision tree model called CART (Classification and Regression Trees). A CART model builds a tree-like structure of nodes and branches, where each node represents a singular decision point, while the branches represent a possible binary outcome of that singular decision point. These nodes are decided by using classification and regression techniques to examine what is contributing to making the transaction fraudulent, and a series of yes or no decisions will ultimately make the model that determines if a transaction is fraudulent or not. While deciding which columns within the data should contribute to the model, it was noticed that decision trees do an excellent job of handling all types of data, so the model included every column within the dataset. The following are the results of the CART model.

Predicted	**Actual 0**	**Actual 1**
0	386,829	1,959
1	0	215

Table 1: Confusion Matrix of Predictions vs Actuals

Analyzing the Figure **??** CART model's confusion matrix, it is initially extremely impressive. As stated earlier in the paper, one of the biggest boxes to minimize is the false negative box. As the reader can see in the confusion matrix, there is a 0 value in that category, which is what the original goal was. Unfortunately, there are 1,959 cases of a false positive, which is a few

too many for the institution to act on, so this number should be worked on to be much lower than it currently is. Looking at the breakdown of the number within the R output, there is a very high accuracy of 99.5 percent. This high accuracy is to be taken lightly because of the high imbalance of the dataset. Looking at the precision of the R output is a much more telling 9.9 percent, which means that only 9.9 percent of the predicted fraudulent transactions are actually fraudulent, which is something that can be fixed. Finally, the F-1 score comes out to a low 18 percent, suggesting that there needs to be a better balance between precision and recall. Moving forward, precision is something that can be worked on to increase the percentage and produce a much more accurate model.

4.3 Random Forest

The next machine learning model that was explored was a random forest model. Random forest models are extremely common in machine learning due to their ability to avoid overfitting and produce high variability across different trees, which is extremely important when trying to produce models and clusters alike. Similar to a decision tree model, the random forest model produces many trees during the model training stage, but the main difference is that the random forest model randomly subsets the dataset and produces high variability that allows for a much better trained model across all branches. It is common practice for a random forest model to have 100 trees than make up the model, so we decided to also put 100 trees in this random forest model for consistency across research. Running the random forest model, here are the results for a confusion matrix

Prediction	**Reference 0**	**Reference 1**
0	372	0
1	1	19

Table 2: Confusion matrix for the random forest model.

Looking at this confusion matrix is very pleasing, and shows us almost exactly what we want to see. There are 99.8 percent of transactions in the correct spots, with only 1 transaction in the

false negative category. At first glance, this model seems to check all of the boxes and potentially leads us to want to use this kind of model to perfect, but it is important how many instances the model can handle. When trying to use more data points to balance the model and discover the threshold that it has, the model will crash and is unable to process large amounts of data, which is a large problem for this scope of work. The dataset that is being used to train and test these models contains 1,300,000 individual data points, and only being able to test 392 at a time poses a large threat to the integrity of our model, which may cause it to appear too good to be true. As shown in figure 5 it is very evident that the amount of dollars spent in the given transaction is a very telling variable in terms of letting the model know what is fraudulent and what is not. By running a simple visual model in R, the model is able to tell what variables are most significant to a fraudulent purchase.

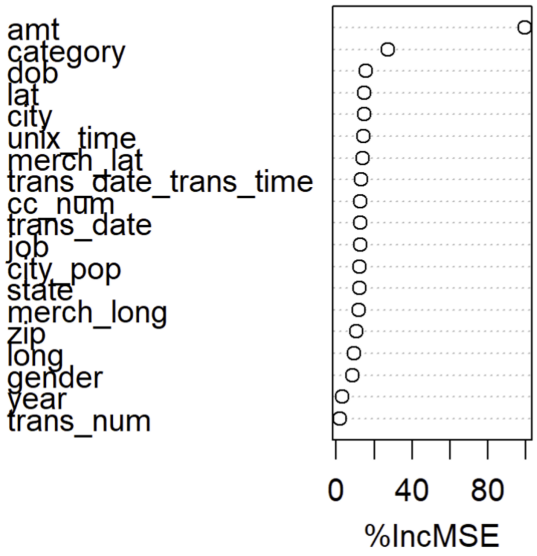


Figure 12: Importance Plot for Random Forest

Looking at this importance plot, the variable "amt" which is amount of money spent in the transaction, is the biggest indicator and most "important" to the model when determining the difference between a fraudulent and non-fraudulent purchase. The conclusion that was drawn from figure 5 and figure 12 was that often times credit thefts spend low amounts of money and stay under the radar until they are confident enough to spend a large amount of money, which is when the model eventually catches them. This results in alerts to the financial institution as well as the card holder that their credit card has been stolen, which is when the fraudulent purchase

is marked. While this model offers some very important and interesting insights, the inability to handle large datasets, and more importantly imbalanced datasets, is extremely costly when handling millions of transactions and is not feasible for this scope of work.

4.4 Anomaly Detection

The final machine learning model that we will examine is the anomaly detection model. Anomaly detection is a special model that attempts to dive into a dataset and find statistical outliers or anomalies and bring them to the users attention. This model is typically used with large datasets that are heavily imbalanced and contain very few outliers, making it a primary target for this scope of work. How this anomaly detection model works is extremely similar to k-means clustering, where the model will group certain points together that share a common trait, and group other points together that may share a different common trait. Eventually there are many groups that fit together in terms of categorical behavior, and the few points that are left outside of those clusters are "outliers" or commonly referred to as "anomalies". In the case of credit fraud an anomaly would be a fraudulent transaction, since fraudulent transactions are the severe minority within this dataset. Running the anomaly detection model with a threshold of 0.55, which means that points with an anomaly score of over 0.6 will be flagged as fraud, and points with an anomaly score of under 0.5 will be flagged as non-fraudulent. Running this model with an optimized threshold, the confusion matrix is produced.

Prediction	**Reference 0**	**Reference 1**
0	2,568,444	1,493
1	989	8

Table 3: Confusion Matrix for Prediction vs Reference

As seen in the confusion matrix, this threshold unfortunately produces an extremely small accuracy, with correctly predicted fraudulent transactions coming in at an extremely small percent of the matrix. However, there are some positives. It is good to see that the model mainly predicts non-fraudulent transactions, as well as the false positive and false negative are fairly balanced.

The major problem with running anomaly detection on giant datasets like this is that barely any data points are "outliers" since there are so many within the dataset, so it is incredibly hard for this kind of model to be able to find a possible fraudulent transaction. It is also extremely hard to find a threshold that satisfies every category at the same time; most often, when the threshold is changed by 0.1, it will completely change the matrix and throw off previous conclusions, which is extremely difficult to work with. Overall, this dataset is too big and unclassified for an anomaly detection model due to the nature of the fraudulent transactions.

5 Final Technique

5.1 Decision Making

After assessing the three different modeling techniques that banking institutions most commonly use to detect credit fraud within customer purchases, we have decided to pursue and refine the decision tree (CART) model to help predict fraudulent transactions within our given dataset. Two key decisions went into making this final decision to critique an already solid decision tree model. The first factor in the decision was the zero False Negative (FN) predictions. Should we find ourselves in the situation of the bank, protecting our customers' financials and livelihood is much more important than spending money on sending out alerts to determine if the purchase is fraudulent or not. When seeing the zero in the FN portion of the confusion matrix, we knew that this model would be a phenomenal start to our final model. The second decision that helped us pick the decision tree model was the vast room for improvement within the precision and recall portions of the final output. The final indicator of how "good" a model is is the F-1 score, which gives a measurement between precision and recall, which was a low 18 percent. Overall, there are many avenues for improvement, and within this section, we will produce a model that has high precision, recall, and most importantly, a high F-1 Score.

5.2 Final Modeling

To improve our Decision Tree CART Model, we tuned multiple hyperparameters, focusing on reducing false negatives while maintaining accuracy. The complexity parameter (`cp`) was lowered from 0.005 to 0.002 to allow a deeper tree, capturing more patterns in fraudulent transactions. Additionally, fraud cases were weighted 5 times higher than non-fraudulent cases to improve

sensitivity within our model testing.

Further refinements included setting `minsplit` to 10 and `minbucket` to 5 (half of `minsplit`), ensuring meaningful splits in our decision bins and reducing non-meaningful bins as much as possible. The maximum tree depth (`maxdepth`) was capped at 10 to balance complexity and prevent overfitting. Finally, the classification threshold was adjusted from 0.5 to 0.35, increasing fraud detection sensitivity at the cost of some precision.

Parameter	**Tuned Value**
`cp`	$0.005 \rightarrow 0.002$
weights	Fraud: 5, Non-Fraud: 1
`minsplit`	10
`minbucket`	5
`maxdepth`	10
Classification Threshold	$0.5 \rightarrow 0.35$

Table 4: Summary of Model Tuning Adjustments

These optimizations of our hyperparameters led to a significant improvement in recall and an F1-score exceeding 0.55, which is considered adequate for a credit fraud detection model. While tuning all of these parameters and refining our model, we did sacrifice our ultimate goal of having 0 false negative results. The following matrix shows our final model results:

Predicted	**Actual 0**	**Actual 1**
0	386,444	1,183
1	385	991

Table 5: Confusion Matrix for Predicted vs Actual

While compromising our goal of having 0 False Negative transactions, we skyrocketed our recall and precision to an impressive 0.72 and 0.456, which is far better than we originally had on our Decision Tree (CART) Model. While the confusion matrix tells the story of the final

models ability to make a decision, another great plot is a variable importance plot, which tells the modeler what variables are most important when a machine learning model is trying to make unique decisions to help classify fraudulent and non-fraudulent transactions. Here is the final results for the most important variables in the dataset:

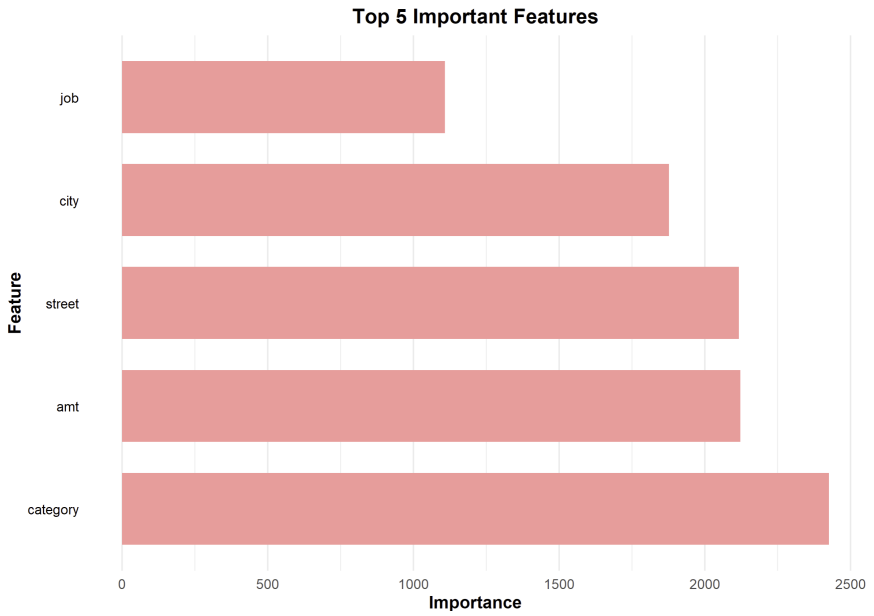


Figure 13: Variable Importance Plot

Diving into the importance plot, we are able to draw a lot of similarities from our exploratory data analysis. Like we saw before, variables such as Category, Amount, and Address are incredibly important to the machine learning model when trying to make the final decision. These 4 categories are more than twice as important as any of the other variables, so it is important to note and keep in mind that these respective variables are what the Decision Tree CART model is relying on the most when trying to make a decision.

When evaluating models for credit fraud, recall and precision are two of the biggest factors when determining how "Good" a model is, and these scores fall in the "Acceptable" category in terms of fraudulent detection models. While the final Decision Tree (CART) Model tuned with all listed hyperparameters ended up being the best model by far, there are some limitations. While using a personal computer with limited computing power, the machine learning algorithms can only be as good as the computer they are being ran by. It is important to note that a much more complex and detailed model could be produced on a machine with much more computing power, but we unfortunately have zero access to a machine like that.

6 Conclusion

The Decision Tree CART Model that was produced provided an F-1 Score of 55.83%, which is adequate for a credit fraud detection model in the financial industry. Like we had explored in our preliminary findings, variables such as Age (Figure 2), Amount (Figure 5), Time of Day (Figure 6), and Date (Figure 10) are incredibly important in deciding if a transaction is fraudulent. These findings were later confirmed by an importance plot produced with our final Decision Tree CART Model, which said these 4 variables were most important when determining the authenticity of a customer transaction.

Potential further improvements would be an expansion in computing power, as a model is eventually limited in terms of the hyperparameters being set and needs more computing power to analyze larger portions of the dataset. Randomly subsetting a dataset is acceptable with this many transactions, but the more unique transactions a predictive model can use to make a decision is what will ultimately provide the best final model for credit fraud detection.

References

[1] Adele Cutler, David Cutler, and John Stevens. *Random Forests*. In *Machine Learning - ML*, volume 45, pages 157–176, 2011. ISBN: 978-1-4419-9325-0. https://doi.org/10.1007/978-1-4419-9326-7_5.

[2] Ali Nassif, Manar Abu Talib, Qassim Nasir, and Fatima Dakalbab. *Machine Learning for Anomaly Detection: A Systematic Review*. *IEEE Access*, pages 1–1, 2021. doi:10.1109/ACCESS.2021.3083060.

[3] Ben-Dov, J., & Ben David, A. (n.d.) *Machine-Learning Techniques for Detecting New Account Fraud*. Transmit Security. https://transmitsecurity.com/blog/machine-learning-techniques-for-detecting-new-account-fraud Accessed February 22, 2025.

[4] Choksi, P. (n.d.). Credit card transactions dataset. Kaggle. https://www.kaggle.com/datasets/priyamchoksi/credit-card-transactions-dataset/data Accessed February 25, 2025.

[5] Content Team (2015) *Credit Card Fraud*. Legal Dictionary. https://legaldictionary.net/credit-card-fraud/ Accessed February 26, 2025.

[6] Jason M. Klusowski. *Analyzing CART*. arXiv preprint arXiv:1906.10086, 2020. https://arxiv.org/abs/1906.10086

[7] R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org Accessed February 25, 2025.

Andrew Marchant is a senior undergraduate student in Data Science and Mathematics at South Dakota State University. Andrew plans to apply his passion for the outdoors with his education by attending graduate school or entering a work environment that involves fisheries and wildlife research.

Gage Scholting is a senior undergraduate student in Data Science at South Dakota State University. Gage plans to attend South Dakota State University for Graduate School, where he will enhance his skills in data science and apply them in a financial field after completing his Master's Degree.