Classification of Credit Card Fraud Cases

Meichan Huang
School of Information Studies
Syracuse University
Syracuse, NY, 13210

Abstract

In recent years, the incidence of credit card fraud has surged, necessitating the development of more effective detection strategies. Our study assesses the performance of three machine learning models—logistic regression, decision trees, and random forest—in identifying and predicting fraudulent credit card transactions. We aim to compare their effectiveness in accurately detecting fraud. Additionally, various data preprocessing were investigated (SMOTE sampling, PCA transformation, Target Encoding etc.) to determine their impact on improving the models' predictive accuracy. The results showed that random forest generally performed well, followed by logistic regression, and decision tree.

1. Introduction

Credit card fraud detection has become increasingly necessary due to the transition from traditional physical payment methods to digital platforms. This shift has been accelerated by advancements in digital technologies, which have revolutionized how financial transactions are conducted. With the convenience of internet banking and online purchases, credit cards have become a primary tool for consumers to manage financial activities remotely, from the comfort of their homes or offices.

However, the rise in the volume of digital transactions has been paralleled by an increase in the incidence of credit card fraud. Fraudulent activities can lead to significant financial losses for both consumers and financial institutions, erode trust in digital platforms, and have broader implications for economic stability. Effective fraud detection methods are therefore essential to identify and prevent unauthorized transactions, safeguard consumer information, and maintain the integrity of the financial system.

Studies in this field have been plethora; leveraging various analytical and machine learning algorithms to detect patterns indicative of fraud. Traditional statistical methods, such as logistic regression, have been employed due to their interpretability and effectiveness in binary classification problems. However, the complexity of fraud patterns has led to the exploration of more sophisticated algorithms like decision trees and random forests, which enhance prediction accuracy by combining multiple decision trees to address overfitting issues.

Therefore, in this project, there were major business questions as follows:

1. Which factors are most strongly associated with fraudulent transactions? For example:
   - In which transaction categories does fraud occur most frequently?
   - Are certain customer demographics (like age or gender) more susceptible to fraud?
   - Which geographic locations have higher instances of fraud?
   - During what times (months, days, and specific hours) do fraudulent transactions occur most often?

2. Which machine learning model is the most effective in identifying and preventing credit card fraud through the analysis of transaction features, Logistic Regression, Decision Tree, or Random Forest?

2. Method

2.1 Data Set and General Consideration

The dataset employed for this project was sourced from Kaggle.com and includes simulated credit card transactions, both legitimate and fraudulent, spanning from January 1, 2019, to December 31, 2020. It encompasses transactions from 1,000 customers across a network of 800 merchants. The training dataset comprises 1,296,675 transaction records, while the testing dataset includes 555,719 records. Each dataset is pre-labeled to indicate fraudulent and non-fraudulent transactions (denoted by 1 and 0, respectively) and features 23 variables. Among these, quantitative variables include city population and transaction-related identifiers such as card numbers and transaction numbers. Additionally, there are 12 qualitative variables, including city, state, zip code, merchant details, and names." Given the presence of numerous categorical variables and the significant variance in the scales of the numeric data, data transformation is essential. Standardizing the format of categorical data and normalizing the range of numeric values is necessary to enhance the performance of machine learning models.

It is worth noting that the dataset has an unbalanced distribution of fraud vs. non-fraud classes (Figure. 1). Specifically, there were 1,289,169 transactions that have been classified as non-fraudulent, making up 99.42% of the dataset; 7,506 transactions were classified as fraudulent, which constitutes only 0.58% of the dataset.

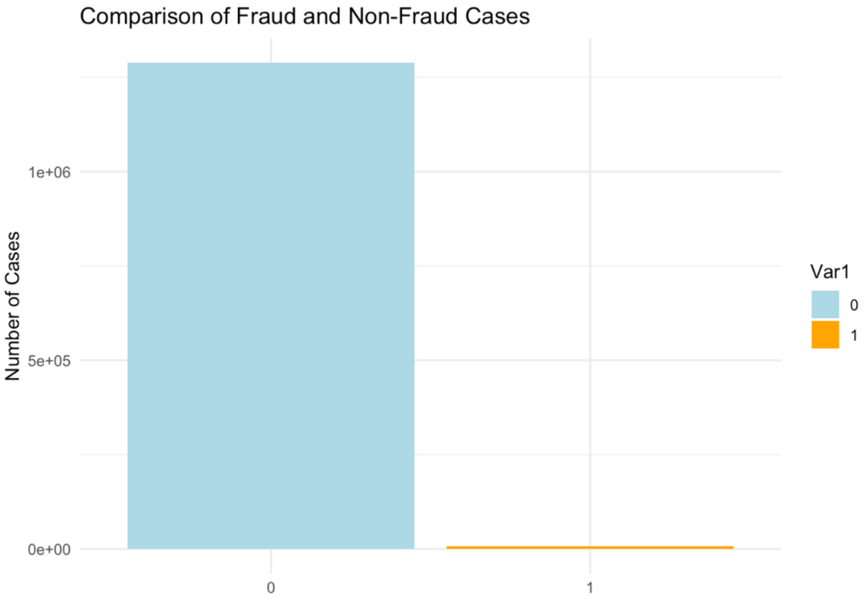


Figure 1. Number of non-fraud vs. fraud in the train set

Addressing imbalanced datasets is a well-documented challenge in data science, as highlighted by Wu (2017). There are multiple strategies for handling such imbalances. One option is to proceed without addressing the imbalance. Alternatively, resampling techniques can be employed, such as under-sampling the majority class, over-sampling the minority class, or applying advanced methods like SMOTE (Synthetic Minority Over-sampling Technique) and ROSE to create a more balanced dataset. In this project, two methods were evaluated: one approach did not address the imbalance during training, and the other utilized SMOTE.

Moreover, considering the dataset's characteristics, standard metrics such as accuracy may not be sufficient for model evaluation. Therefore, alternative metrics, including precision (also known as specificity), recall (also known as sensitivity), F1 scores, and the Area Under the Receiver Operating Characteristic Curve (AUC), were employed to provide a more comprehensive assessment of model performance.

2.2 Data Preprocessing

2.2.1. Data Type Conversion

Upon brief inspection, the dataset contains two distinct attribute types: char (such as city and state) and numeric (such as amount and transaction time). Prior to building predictive models, it is necessary to transform all categorical attributes into a numeric format to facilitate processing by machine learning algorithms. Several data conversion steps were conducted:

(1) Convert char data to factor; and encode highly cardinal categorical data: Character data were initially transformed into categorical factors in R using the as.factor() function. However, some categories, such as "city," had a high number of unique values (849 levels), significantly increasing the analysis's dimensionality. To address this, Target Encoding was applied to these variables to condense their dimensionality by replacing each category with a value based on its association with the target variable, calculated as the ratio of the count of the target outcome to the total occurrences of the category. Target Encoding is particularly useful for tree-based methods like decision trees and random forests, as it helps circumvent the "curse of dimensionality" associated with one-hot encoding. Nevertheless, it has its drawbacks, such as the risk of target leakage, which occurs when the encoded features indirectly contain information about the target, potentially misleading the model during training. Due to technical challenges in performing this encoding within R, the procedure was carried out using Python, with the details provided in Appendix C.

(2) Binary data encoding: In this dataset, the 'gender' variable was dummy encoded for analytical purposes, with '0' representing Female and '1' representing Male.

(3) Create new attributes using the time variable: In the dataset, the date of birth (DOB) of customers and the time of transactions were used to derive new variables. Specifically, the transaction month, day of the week, hour of the transaction, and the customer's age were extracted to enhance the analysis.

(4) Log transform atm variables: When visualizing the distribution of atm v. fraud cases, the "amt" variable exhibits minimal intervals between smaller values and wider gaps between larger ones (Figure 2).

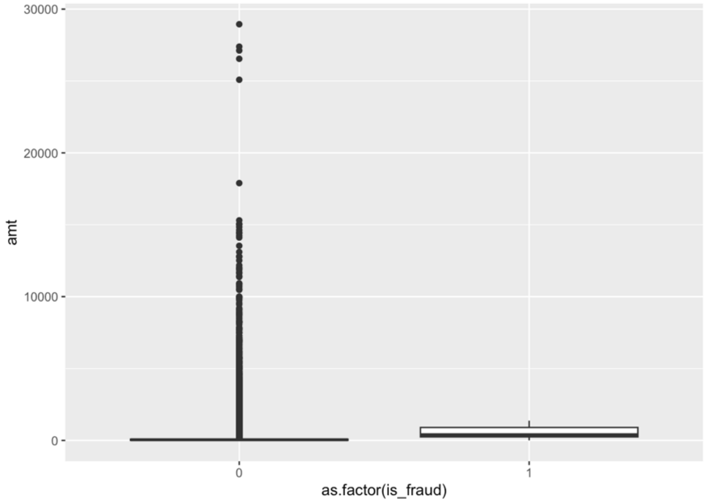


Figure 2: Distribution of atm on non-fraud and fraud cases

(5) Geolocation (lat and long) transformation: Since the distance between the customers' long and lat and merchant's long and lat could be potential indication of fraud. Therefore, transformation were performed on these columns. The absolute value of the differences of lat and long between the customer and merchant were calculated and transformed into the variable "displacement" measured in miles.

To address this, the variable was transformed using logarithmic scaling, which narrows the range between larger numbers while expanding the intervals between smaller figures (Figure 3)

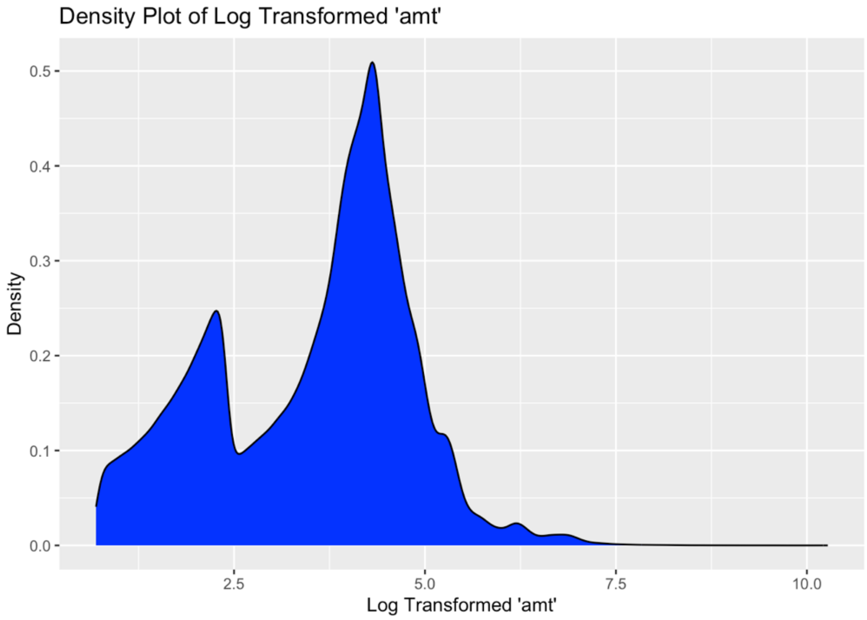


Figure 3: Distribution of Log Transformed amt

After the data cleaning, redundant and irrelevant columns were removed. The new dataset is consist of 13 columns, including one predictor variable – is_fraud, and 12 explanatory variables, namely city_pop, trans_hour, month, age, displacement, amt_log, category_target_sklearn, state_target_sklearn, city_target_sklearn, job_target_sklearn, weekofday_target_sklearn, and gender_binary.

2.2.2. Data Transformation and Standardization

PCA was first conducted in an attempt to reduce the dimensionality of the data; however, it did not significantly help to reduce the number of clusters. Cluster 10 was able to explain 99% of the variation, however, there were only 12 explanatory variables to be used after the data cleaning.

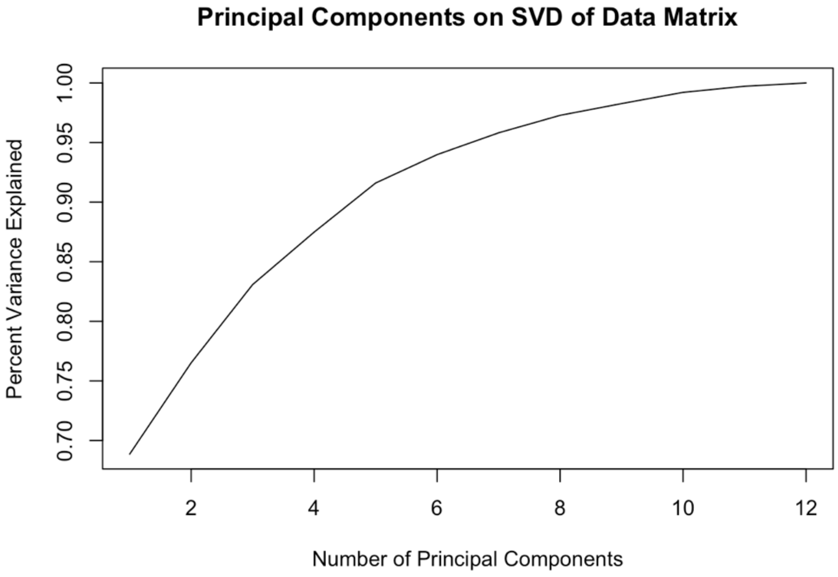


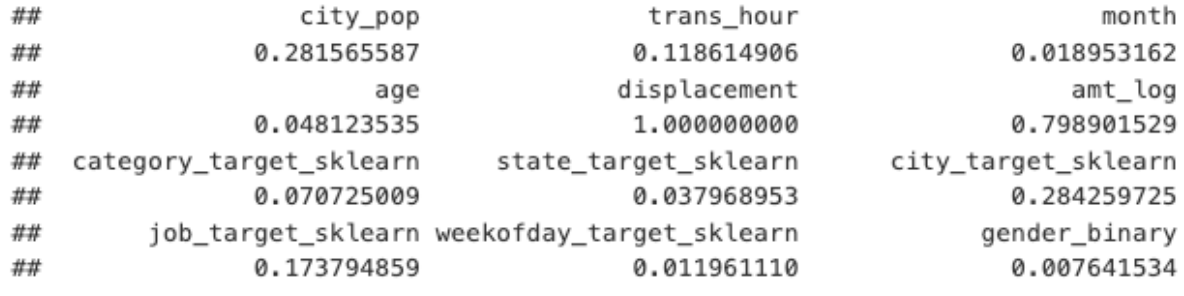
Figure 4: Plot of PCA Analysis

Therefore, all the transformed data were scaled to have mean of 0 and a standard deviation of 1 using the purr package.

2.2.3. Feature Selection

After the data transformation and standardization, it is essential to conduct the correlation study to see the relationship between single indicator variable with the target variable. Since it is a binary classification, linear correlation is not approparaite, instead, cramer's V correlation was used to compute the individual correlation.

Table 1 **Cramers V. correlation is_fraud ~ .**



```
##                 city_pop              trans_hour                   month
##              0.281565587             0.118614906             0.018953162
##                      age            displacement                 amt_log
##              0.048123535             1.000000000             0.798901529
##  category_target_sklearn    state_target_sklearn     city_target_sklearn
##              0.070725009             0.037968953             0.284259725
##       job_target_sklearn weekofday_target_sklearn           gender_binary
##              0.173794859             0.011961110             0.007641534
```

The results showed that the log of transaction amount (amt_log) demonstrated the strongest association with the target variable, indicating its potential importance in predicting fraud. On the other hand, month, weekday, and gender showed negligible correlations, suggesting they contribute little to the model's predictive capability and were removed from the subsequent model building. Worth noticing, the perfect

2.2.4 Dealing with imbalanced data: Default v.s. SMOTE sampling

In the study, two data conditions were evaluated: one in its original state with imbalanced classes of non-fraud and fraud, and the other after applying the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE, which generates synthetic samples rather than just replicating the minority class, was used to create a more balanced dataset. By interpolating between a rare observation and its nearest neighbors, SMOTE constructs artificial samples, effectively equalizing the distribution of non-fraudulent and fraudulent cases for training purposes. To implement this technique, the DMwR package, now archived in CRAN, was utilized. The SMOTE function from this package was employed, resulting in a modified dataset with 30,024 non-fraud and 22,518 fraud cases ready for algorithmic analysis.

2.3 Classification Algorithms

Three distinct algorithms were employed in the analysis: logistic regression, decision tree, and random forest. Logistic regression, which is well-suited for binary outcomes such as fraud versus non-fraud, has demonstrated commendable accuracy levels. The decision tree, a non-linear classifier, works by partitioning the dataset into progressively smaller subsets. While advantageous for its simplicity and interpretability, a decision tree's main drawback is its propensity to overfit the training data. The random forest algorithm mitigates this by constructing multiple decision trees and averaging their predictions, thus enhancing the model's generalizability and robustness against overfitting.

The three algorithms—logistic regression, decision tree, and random forest—were evaluated on both the original imbalanced dataset and the balanced dataset created using SMOTE. Their performance was assessed using metrics such as Accuracy, Precision, Recall (or Sensitivity), F1 Score, and the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). These metrics provided a comprehensive view of each model's predictive ability.

Accuracy measures the ratio of the total number of correct predictions of fraud to the total number of predictions (fraud and non-fraud) made by the model, which is calculated as:

Sensitivity/Recall measures the proportion of actual positives correctly identified (out of all actual positives).

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

Precision measures the proportion of positive identifications that are actually correct (out of all positive identifications made by the model).

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

3 Results

In the subsequent section, a comparative analysis of the performance of three algorithms on two datasets—one imbalanced and one balanced through SMOTE—is presented. For detailed results, refer to the accompanying HTML outputs of the code provided in Appendix A for the

SMOTE-adjusted data and Appendix B for the imbalanced data. correlation of displacement is suspect and warrants further investigation to determine if it should be retained.

Also, there were no strong correlation between all the independent variables, therefore, there is not a huge concern of collinearity.

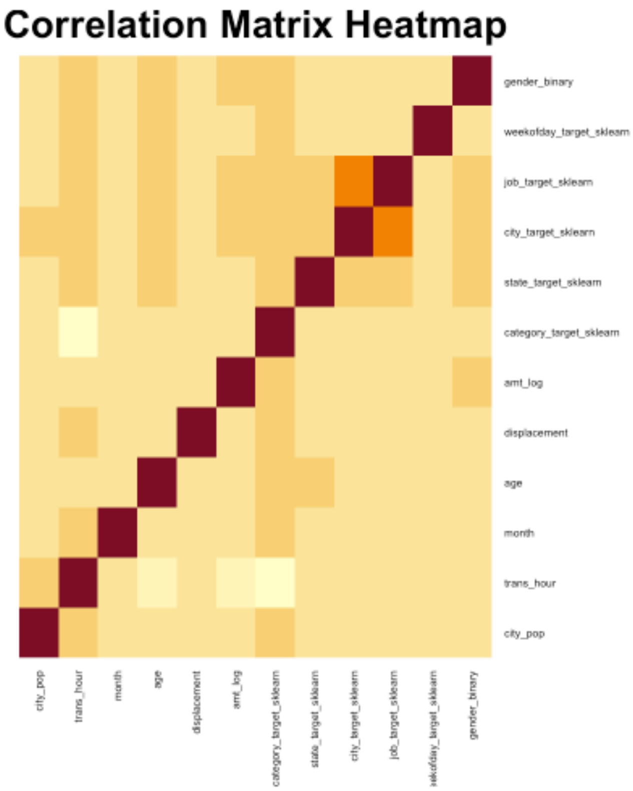


Figure 5: Correlation Heatmap of Independent Variables

Therefore, the final data that is used for the modeling is as follows:

```
## 'data.frame':    1296675 obs. of  8 variables:
##  $ city_pop                : int  3495 149 4154 1939 99 21
##  $ is_fraud                : Factor w/ 2 levels "0","1": 1
##  $ trans_hour              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ displacement            : num  60.2 18.8 67.3 63.5 59.9
##  $ amt_log                 : num  1.79 4.68 5.4 3.83 3.76
##  $ category_target_sklearn: num  0.01446 0.0141 0.00248 0
##  $ city_target_sklearn     : num  0 0 0 0.0304 0 ...
##  $ job_target_sklearn      : num  0.00169 0.00216 0.01566
```

3.1 Exploratory Analysis

The original data was rich in details, therefore, some exploratory analysis to conducted to understand the characteristics of non-fraudulent and fraudulent cases. Specifically, investigations were conducted to determine if any attributes exhibited a significant inclination towards distinguishing between fraud and non-fraud.

First, The analysis identified several transaction categories that exhibited a higher likelihood of fraud. Specifically, four categories—shopping_pos, shopping_net, misc_net, and grocery_pos—were found to be particularly prone to fraudulent activities.

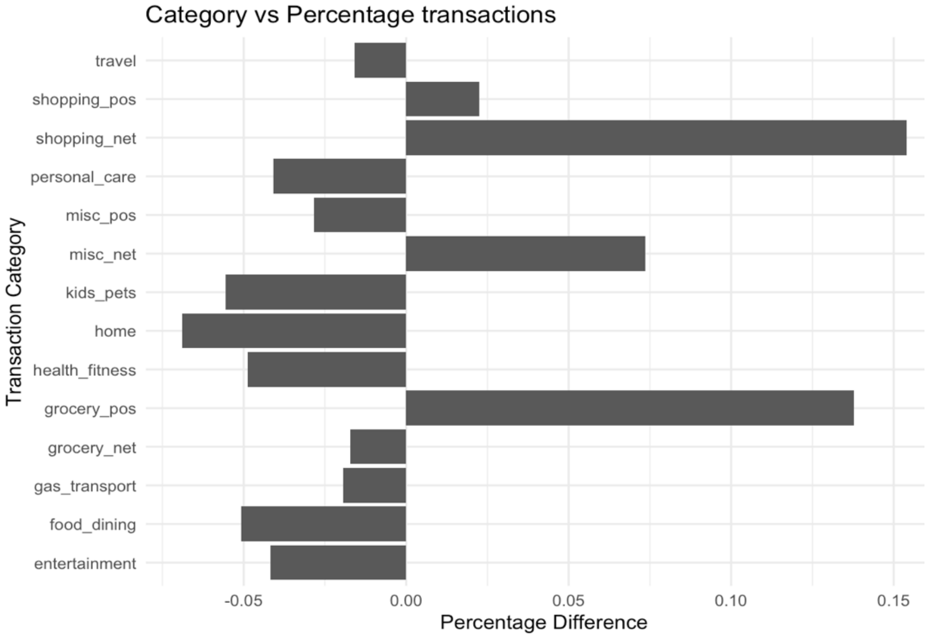
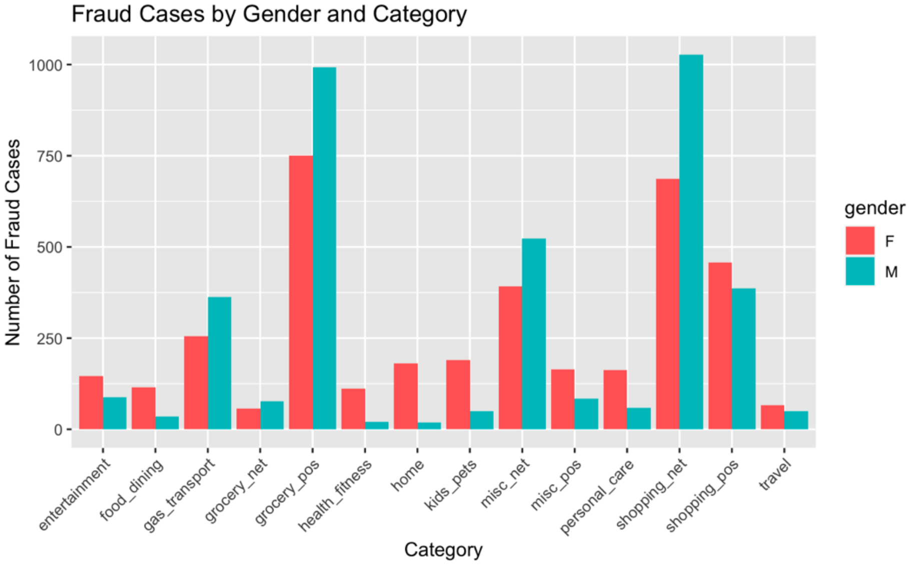


Figure 6: Percentage differences among categories in fraud cases

The analysis also revealed that fraudulent transactions varied between genders across different transaction categories. Male cardholders were more frequently subjected to fraud in categories such as gas_transport, grocery_pos, and misc_net, which are notably prevalent in fraudulent activities. Conversely, female cardholders were more often the targets of fraud in categories like home, kids_pets, and personal care.



Age has proven to be a variable in distinguishing between fraudulent and non-fraudulent transactions. Analysis of the age distribution of cardholders (Figure 7) reveals that non-fraudulent transactions predominantly involve cardholders aged 35 and 40, with another significant occurrence at around 50 years. Conversely, fraudulent transactions display a more uniform age distribution, with peaks at 35 years and a notable frequency in the 50 to 55-year age range. The results suggest an age-related variation in susceptibility to credit card fraud, with individuals over 50 being slightly vulnerable to fraudulent activities, which is anticipated.

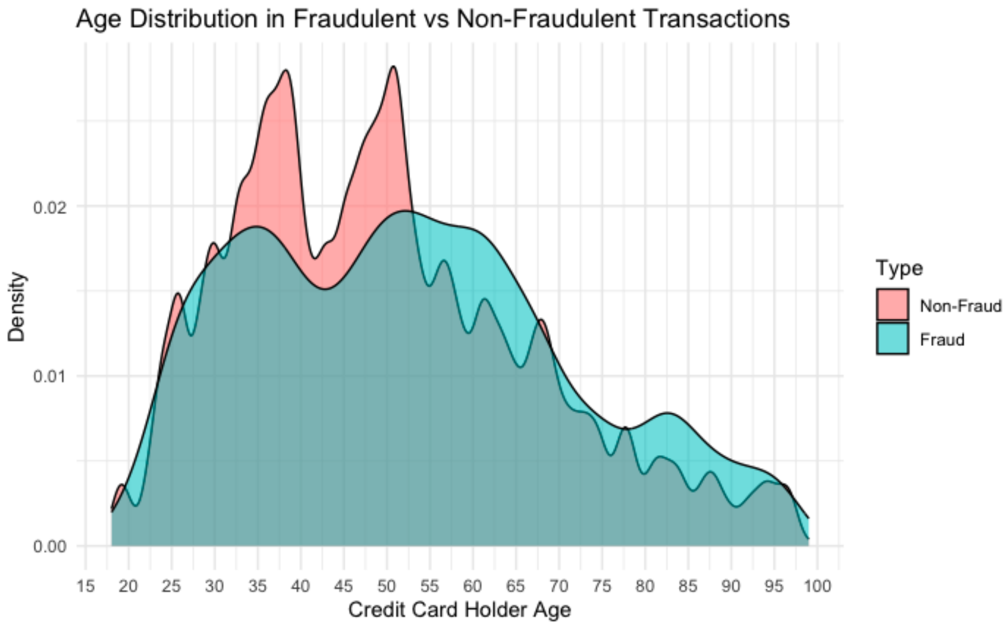


Figure 7: Credit card holder age distribution between fraud and non-fraud events

The analysis indicates that while there is no discernible difference in the distribution of fraud cases between genders, females exhibit a wider age range in fraudulent cases compared to males, as shown in Figure 8.

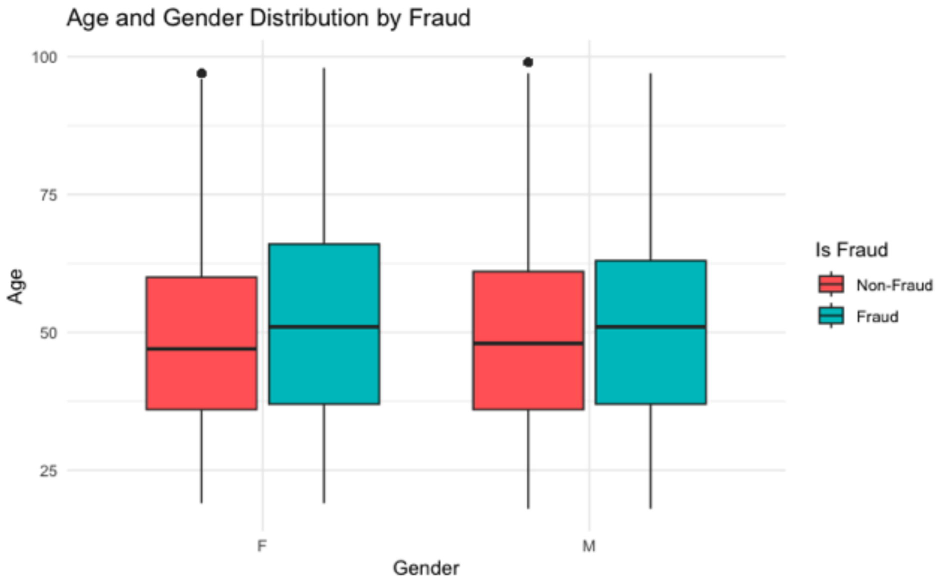


Figure 8. Age and Gender Distribution of Cardholders by Fraud

Regarding the geographical distribution of fraudulent cases, the data indicates that certain states have higher occurrences. New York (NY), Pennsylvania (PA), and Texas (TX) top the list, with Ohio (OH) and California (CA) also showing significant numbers. Potential reasons for this include their dense populations, status as major tourist destinations, presence of international ports, and higher average incomes. However, these hypotheses remain unverified without further data to substantiate the correlations.

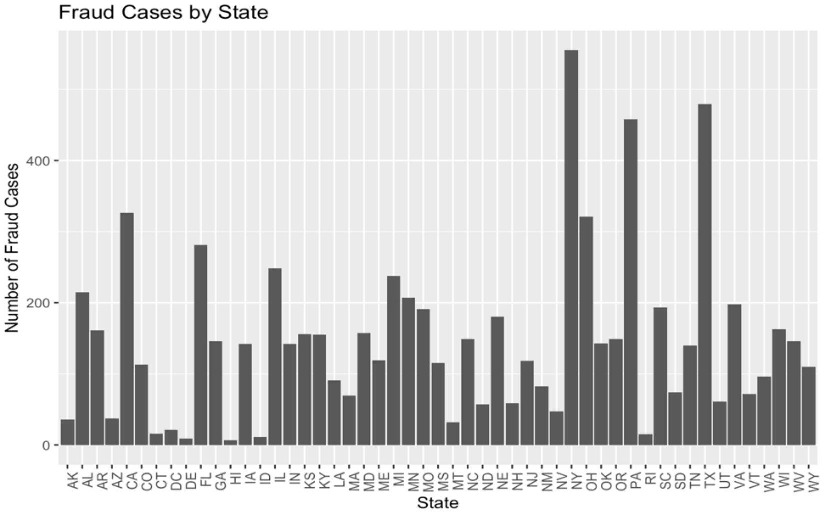
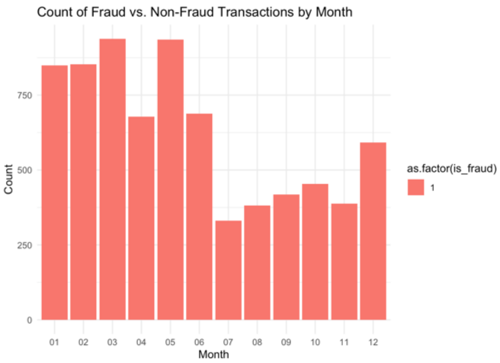
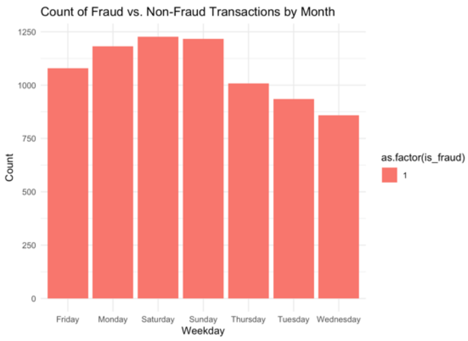
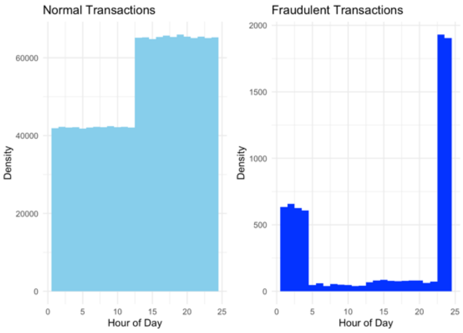


Figure 9. Distribution of fraud cases in the states

Lastly, the timing of transactions was found to be a significant indicator of fraud, as revealed by exploratory data analysis. The impact of the month, day of the week, and hour on the likelihood of a transaction being fraudulent varied (refer to Figures 10 to 12). The trends suggest that fraudulent activities tend to occur more frequently during January to March and in May, are more common on Fridays and Saturdays, and peak from late evening to midnight.







Therefore, to summarize, credit card fraud is influenced by transaction categories, with shopping_pos, shopping_net, misc_net, and grocery_pos being particularly vulnerable. Age plays a role, with a more uniform age distribution in fraudulent cases, especially noticeable among individuals aged 35 and those between 50 to 55 years. Fraud occurrences show temporal patterns, tending to happen more often from January to March, in May, on Fridays and Saturdays, and during late-night hours, with the effect on the transaction hours particularly strong. Geographically, states like New York, Pennsylvania, Texas, Ohio, and California report more fraud, likely due to their high population densities, statuses as travel hubs, and economic affluence.

Gender analysis reveals no significant difference in fraud distribution, although females show a broader age range in fraud cases. Therefore, this factor was not considered in the modeling.

3.2 Modeling to Predict Fraud Cases

This section reports on the outcomes of applying three machine learning algorithms to two datasets: one imbalanced and the other balanced using SMOTE. Although additional models were developed, not all yielded on par results. The models discussed here were selected for presentation based on their better predictive performance within their respective classes.

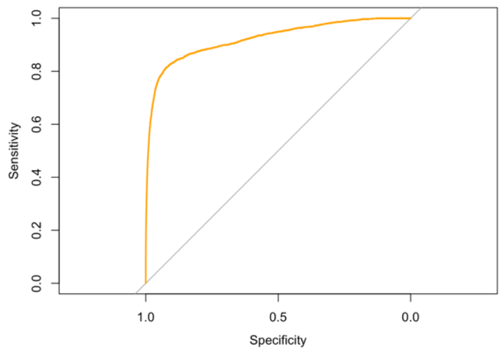
3.2.1 Logistic Regression

Table 2 demonstrated the performance of logistic regression ML algorithm using imbalanced dataset and the SMOTE data respectively.

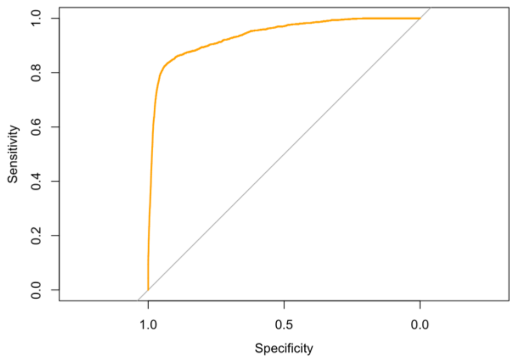
Table 2: Comparative performance matrices of imbalanced and SMOTE data using Logistic Regression

Logistic Regression	Imbalanced data	SMOTE data
Accuracy	0.975	0.887
Precision	0.097	0.0289
Recall (sensitivity)	0.65	0.88
F1 score	0.169	0.056

AUC curves also demonstrated that balanced data performed slightly better.



For imbalanced data, AUC is 0.9227



For SMOTE data, AUC is 0.9375

Neither model proved ideal in performance based on the performance matrices. The Logistic Regression model trained on imbalanced data achieved a high accuracy of 97.5%, yet its recall of 65% suggests it fails to detect a substantial portion of fraud cases. The accuracy does not seem to tell the true performance of the model. The precision for this model is notably low at 9.7%, indicating many false positives.

Conversely, the model trained on SMOTE data, while less accurate at 88.7%, has a high recall of 88%, indicating it is good at identifying most fraud cases. However, this comes at the cost of precision, which is only 2.89%, resulting in a high rate of false positive. If the priority is to detect fraud, the SMOTE model would be more effective, despite the increase in false positives. If reducing false positives is crucial, even at the expense of missing more fraudulent transactions, the model using imbalanced data would be more appropriate, despite its limitations.

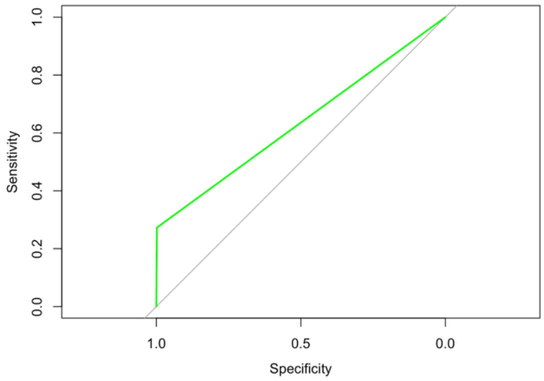
Another index that should be considered is the AUC value. The AUC represents the degree to which the model is capable of distinguishing between classes, which in this context are the fraudulent and non-fraudulent transactions. An AUC of 0.9227 for the model trained on imbalanced data indicates a very good ability to distinguish between the two classes, which means that there is a 92.27% chance that the model will be able to differentiate a random fraudulent transaction from a non-fraudulent one. An AUC of 0.9375 for the model trained on SMOTE data suggests an even better discriminative ability, with a 93.75% chance of distinguishing between a fraudulent and a non-fraudulent transaction.

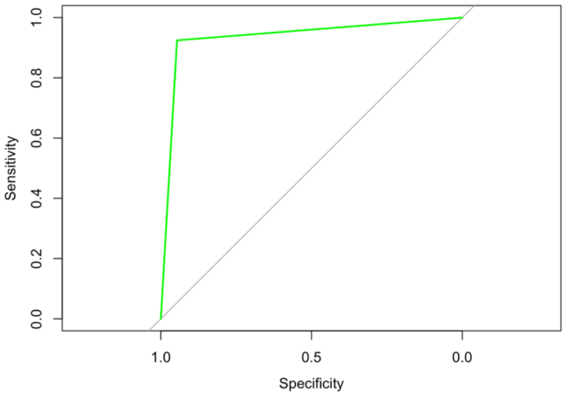
3.2.2 Decision Tree

The next model discussed here used Decision Tree to examine both the imbalanced and SMOTE data. Table 3 demonstrated the performance of decision tree on these two datasets respectively. The parameter control setting on this model is (cp = 0, minsplit = 2, maxdepth = 5)

Table 3. Comparative performance matrices of imbalanced and SMOTE data using Decision Tree

Decision Tree	Imbalanced data	SMOTE data
Accuracy	0.995	0.887
Precision	0.272	0.06
Recall (sensitivity)	0.356	0.94
F1 score	0.308	0.1128





For imbalanced data, AUC is 0.63

For SMOTE data, AUC is 0.935

The decision tree model trained on imbalanced data boasts a high accuracy of 99.5%. However, this model only identifies 35.6% of actual fraud cases (recall) and has a precision of 27.2%, meaning that less than a third of its fraud predictions are correct. The combined measure of effectiveness, the F1 score, is relatively low at 30.8%.

On the contrary, the decision tree model trained on SMOTE data has a lower accuracy of 88.7% but a much higher recall of 94%, indicating it captures the vast majority of fraud cases. The trade-off is its precision, which is only 6%, indicating a high rate of false positives. The F1 score for this model remains low at 11%.

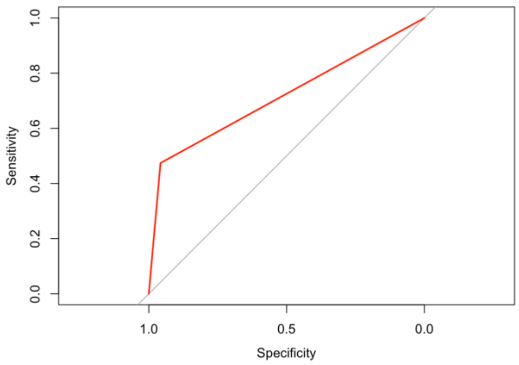
When considering the Area Under the Curve (AUC) as a performance metric, the model using SMOTE data outperforms the one trained on imbalanced data. The higher AUC value for the SMOTE-trained model suggests a significantly better ability to differentiate between the classes, reinforcing the effectiveness of SMOTE data in training decision tree models for fraud detection.

3.2.3 Random Forest

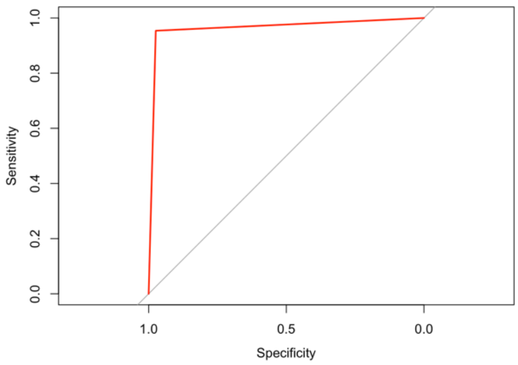
The last model discussed here used Random Forest to examine both the imbalanced and SMOTE data. Table 4 demonstrated the performance of decision tree on these two datasets respectively. Different parameter settings were required due to computational constraints. For the imbalanced data, which is considerably larger and more computationally demanding, causing full memory usage, the model parameters were set to a smaller number of trees (ntree = 50) and a limited number of variables tried at each split (mtry = 8). In contrast, the reduced size of the SMOTE dataset allowed for a more robust model with a higher number of trees (ntree = 1000), facilitating a more thorough analysis without overextending the system's resources.

Table 4. Comparative performance matrices of imbalanced and SMOTE data using Decision Tree

Random Forest	Imbalanced data with ntree = 50	SMOTE data with ntree = 1000
Accuracy	0.995	0.97
Precision	0.02	0.12
Recall (sensitivity)	0.47	0.97
F1 score	0.04	0.22



For imbalanced data, AUC is 0.71



For SMOTE data, AUC is 0.96

The Random Forest model, when applied to imbalanced data using 50 trees, shows a high accuracy of 99.5% but falls short in precision at a mere 2%, indicating a large number of false positives. Its recall is moderate at 47%, catching less than half of the actual fraud cases, and the F1 score is low at 0.04, reflecting the imbalance between precision and recall.

On the other hand, the Random Forest model trained on SMOTE data with 1000 trees shows slightly lower accuracy at 97% but significantly improves in recall, achieving a high rate of 97%, which suggests it is very effective at identifying fraud cases. The precision improves to 12%, and the F1 score increases to 0.22, which is better than the imbalanced model but still indicates room for improvement in balancing the rate of false positives and correctly identified fraud cases.

4. Discussion and Conclusions

To summarize, comparing all six models, the Random Forest model using SMOTE data stands out as the most effective in fraud detection among the three tested algorithms, with a good balance of accuracy, recall, and a reasonable rate of false positives.

There is a notable trend across the algorithms when using SMOTE data; they tend to produce more false positives, incorrectly labeling non-fraudulent transactions as fraudulent. Depending on a company's tolerance for false positives, the utility of models, particularly the Decision Tree, may vary. Companies seeking to minimize false alerts might consider alternative approaches, while those prioritizing the detection of fraud may find the increased false positives an acceptable trade-off.

As for which features have a strong indication to fraud, credit card fraud tends to concentrate in certain transaction categories such as shopping (both at physical locations and online) and grocery purchases. Patterns in fraudulent activities also vary by age, with spikes among those aged 35 and 50-55, and timing, with a higher incidence in early months of the year, particularly in May, and during weekends and late nights. Moreover, states with large populations and economic significance, like New York and California, report higher levels of fraud, possibly reflecting their dense and affluent environments. However, gender is not a strong indicator of fraud in this dataset.

This project has several limitations that warrant additional investigation. Addressing the imbalance in the data is crucial for accurately representing the true characteristics of fraud. The techniques employed here, including the use of SMOTE, did not yield optimal results, suggesting that alternative methods of sampling could be explored.
Furthermore, the selection and treatment of features require closer scrutiny. For instance, the variable representing the distance between the merchant and the customer was anticipated to be significant in identifying fraud. However, it showed a 100% correlation with fraud cases in the data preprocessing, indicating a potential issue with how the feature was calculated or used, which could be misleading. This anomaly necessitates a more thorough examination to ensure the reliability of feature selection and to improve model performance.

References

[1] Trevisan, V. (2022). Target-encoding Categorical Variables. Retrieved from https://towardsdatascience.com/dealing-with-categorical-variables-by-using-target-encoder-a0f1733a4c69

[2] Wu, Y. and Radewagen, R. (2017). 7 Techniques to Handle Imbalanced Data. Retrieved from https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html.