

Frameworks & Methods II:
Project Final Report

- Group 6 -
Elyan Palladino
Woon Sup Kim
Shivam Bathija
Laurensia Charlene Gono
Yushan Lin

Columbia University
APAN 5205 | Professor Kitty Kay Chan
April 19, 2022

I. Statement of the Problem

Technological advancements such as credit cards and contactless payments have taken over the finance industry and shaped how, when, and where people make purchases today. This digital transformation, along with COVID-19-related concerns of virus transmission has resulted in many merchants becoming cash-free (Filipiak, 2020). The frequency of cash-related crimes is also a driving force in the adoption of technology and credit cards. The Federal Trade Commission (FTC) found that only 3% of all payments made in the United States in 2020 were accredited to cash payments (FTC, 2022). Bloomberg estimates that in the United States alone, the number of truly cashless vendors jumped from 8% to 31% between March and April 2020 (Poon, 2020). Not only are credit cards more accepted, but their total volume of transactions is rising again to pre-covid levels, as consumers have resumed their pre-pandemic purchasing patterns (Marte, 2021). Yet the FTC reported over 390,000 cases of credit card fraud in 2020 – an increase of 44.6% from 2019 (Daly, 2021). Moreover, credit card fraud was the leading type of identity theft in 4 out of the last 5 years, making it the second most common type of identity theft overall - with an increase of 48% from 2019 to 2020 (Mint, 2021). Thus, digital payments are, now more than ever, vulnerable to fraudulent activities.

In the interests of their customers and to combat credit card fraud, most banks offer certain services that notify their clients of any suspicious transactions on their cards. Based on their selected preferences, users immediately get notified about these transactions via email, phone, text, or mobile application. Yet users still can't self-audit these transactions. They rely on the efficacy of the banks in detecting fraudulent activity. By providing a self-audit service, banks could gain an edge over their competitors and create a robust security system that allows users to be more aware of their past activities, thereby providing clients with a feeling of security. This is in addition to the humongous losses that the bank could prevent by consumers identifying these frauds in advance - an amount that is estimated to hit a total of \$40bn by 2027 ("Credit Card", 2020). This service would also decrease the dependency of clients on their banks for transaction auditing. The self-auditing feature could also be crucial for clients that share their accounts with multiple people (i.e. joint accounts), as tracking all transactions can be a tedious and time-consuming process. The new self-auditing process outlined in this project could be accessed with the click of a mouse on the bank's online portal. Then, customers can review their past activity and see predicted transactions that are suspected to be fraudulent based on dollar amount, transaction location, and vendor.

II. Research Questions

1. ***Are there any specific geographical locations with more fraudulent activities?*** Research shows that the states with the most identity theft and fraud are Washington, Colorado, and Kansas (WalletHub, 2021). However, there are mixed findings on geographical data. Chargebacks911 revealed that the states with the highest risk for credit card fraud are Idaho and Massachusetts (Chargebacks911, 2021). This research question will be explored, therefore, to confirm and expand on existing published data.
2. ***Are there any specific merchant categories with more fraudulent activities?*** It has been found that there are three main categories of vendor fraud: billing schemes, check tampering schemes, and bribery or extortion schemes (RSM). Yet there is little research on what merchant categories cause the most fraudulent transactions. Examining this question will expand on current research about vendor fraud by specifying what categories are at the highest risk for credit card fraudulent activities.

3. ***Can suspicious transactions and purchase patterns be determined without data mining individual transactions?*** Detecting each fraudulent transaction is becoming increasingly challenging as the volume of transaction data becomes large with a growing number of customers. The traditional methods of data mining individual transactions to discern fraudulent activities will not be sufficient. Supervised machine learning methods are feasible but not ideal as they are computationally expensive, thus impractical. By using unsupervised learning methods, banks will be able to detect purchase patterns and identify suspicious activity without having to verify individual transactions. Recent research on fraud detection using unsupervised machine learning algorithms (Bodepudi, 2021) proposes a novel unsupervised method to detect anomalies. This project will dive deeper into the findings from Bodepudi's study and apply them to further optimize fraud detection in credit card transactions.

II. Description of the Data

The raw dataset for the self-audit system contains credit card transactions from January 1st, 2019 through December 31st, 2020. It consists of transactions from 999 unique credit cards, with 1,852,394 observations and 23 variables with no missing values. All corresponding figures and tables discussed in this report can be found in the appendix, including the breakdown of all variables in Table 1.

Figure 1 describes the number of transactions conducted by each credit card. The large variance in the histogram suggests that the dataset represents different types of credit card users. Figure 1 also shows that most credit card users have made between 500 and 3,000 transactions, with a mean of 1,854 transactions. The code for Figure 1 shows that the data was grouped by "cc_num" and can also be found in the appendix. Figure 2, instead, describes the number of purchases in US dollars. The histogram on the left is the total amount of purchases combining all transactions made by each credit card user, showing a range from \$1,348 to \$411,669 spent. The right-skewed distribution shows that a large majority of credit card users spend on smaller amounts. The histogram in the middle is the average amount of purchases on a transaction made by each credit card user, which ranges from \$46 to \$949. The data also reveals that a large majority of the subjects are using their credit cards to purchase small, cheaper items. This is congruent with the outliers as shown on the box plot to the right of Figure 2. Lastly, Table 2 shows the proportions of the number of transactions in each of the merchant categories and the total amount of dollars that have been spent in those categories.

Basic data cleaning and data transformation were conducted to transform the data into one that is more appropriate for developing a self-audit system. A detailed breakdown of the transformed data can be found in Table 3 in the appendix. After this was done, the dataset became more suitable for developing a mock self-audit system and answering the stated research questions. It provides comprehensive credit card information that records all transactions that have occurred in the past. The variables include transaction timestamp, name of the merchant, transaction amount, name of cardholder, credit card number, and location of transaction (longitude and latitude). With this information, analyses can be conducted on an individual's spending data to predict fraudulent transactions.

III. Summary of Analytical Techniques Used

There were two main analytical techniques used to explore fraudulent activities in the dataset and answer the stated research questions – Cluster Analysis and Spatial Analysis.

1. **Cluster Analysis:** Two separate clustering analyses were conducted to identify suspicious transactions: one on the dollar amount spent by customers and the other on the merchant categories. The intention is to ensemble the results of these clustering techniques to successfully detect fraudulent transactions.
 - a. **Cluster analysis on dollar amount:** This analytical technique helped understand the users' purchasing behavior without analyzing individual transactions, answering the third research question. Large clusters indicate regular purchase behavior. Conversely, small clusters indicate irregular purchase behavior. The transactions in the small clusters can be flagged as possible fraudulent transactions.
 - b. **Cluster analysis on merchant categories:** With a similar logic applied to dollar amounts, an independent clustering analysis was applied to merchant categories. This technique helped identify merchant categories that the customers do not regularly visit and flag them as probable fraudulent transactions.
2. **Spatial Analysis:** This technique was used to determine whether there are visible patterns in geographical locations related to fraudulent activities. This information can be used by users to know if their location is a high-risk fraud area. It can also be used to highlight transactions in these locations.

IV. Analysis

The 999 users in the dataset were grouped into personas with similar purchasing habits. This process was done through k-means clustering analysis, categorizing users by similarity based on the number of transactions each user has made. The analysis led to seven personas, following the cluster suggestion from the silhouette method shown in Figure 3. Table 4 shows a summary of the persona profiles identified. A representative from each of the profiles was chosen to conduct techniques for fraud detection analysis. These representatives were the candidates that were closest to the centroid of the clusters.


Hierarchical clustering, k-means clustering, and model-based clustering on dollar spend were conducted, and k-modes clustering on merchant category was conducted for each of the selected candidates. This was done to determine any suspicious transactions that are outside of the users' regular purchase behavior (i.e. small clusters). It is worth noting that the number of clusters for hierarchical clustering and k-means clustering was decided using a silhouette method. The number of clusters for model-based clustering was chosen automatically by the Mclust function, which allows for determining the number of clusters that produces the highest BIC. Table 5 summarizes each technique's performance (specificity) for each persona profile. Table 6, instead, summarizes the number of false positives for each profile. Figure 4 shows a plot for both specificity and false positives (previously summarized in Tables 5 and 6). The best-performing models for each of the profiles were selected based on the technique that produced the highest specificity with the lowest false positives. Lastly, Table 7 shows a summary of the models that performed the best for each persona profile, revealing that the clustering models work well for profiles 3, 4, 5, 6, and 7 but don't work as effectively for profiles 1 and 2.

V. Results

The analysis above was performed on one user to test the fraud-detecting model and yielded the following self-audit system results. All figures can be found in this section and the appendix.

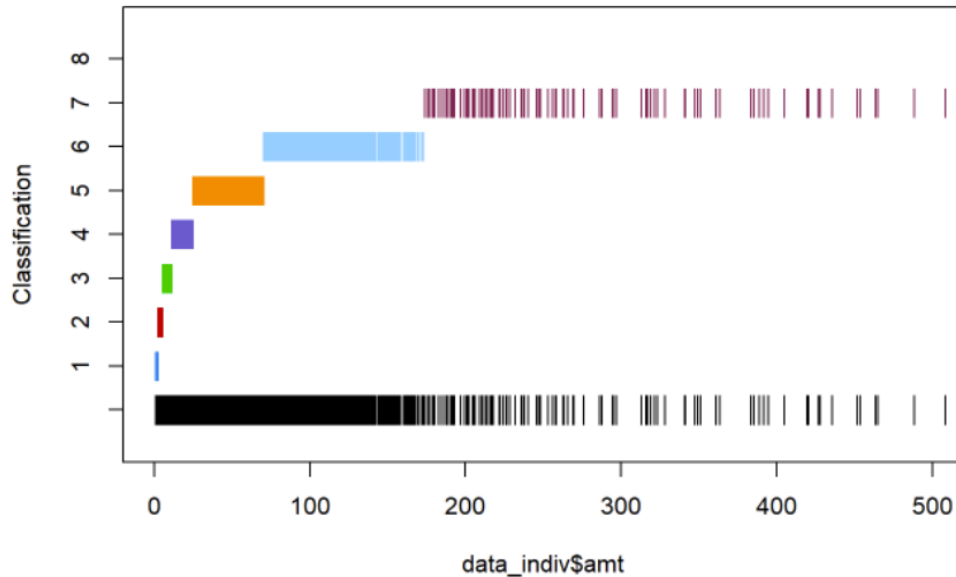
1. The user is prompted to enter their credit card number as shown in Figure 5 below.

Figure 5. Credit-Card Prompt Box

A screenshot of an R prompt dialog box. The title bar says "R prompt". Inside the dialog, there is a text label "Enter credit card number" above a single-line text input field. At the bottom right of the dialog, there are two buttons: "OK" and "Cancel".

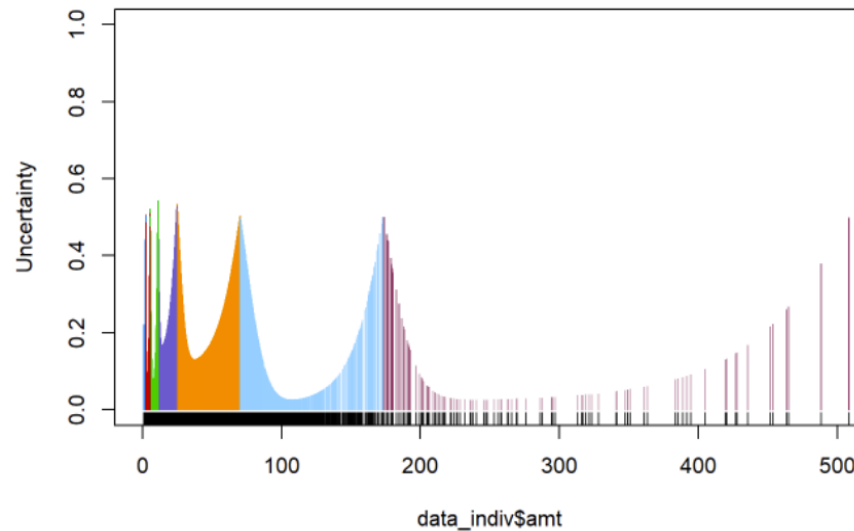
2. The user is categorized into one of the 7 profiles based on the number of their transactions. The selected user for this example was extracted from profile 1. The technique for this user to detect fraudulent transactions was model-based clustering, as it performed the best for Profile 1.
3. The model-based clustering method chose 8 as the number of clusters. Figure 6 shows the number of clusters and how the data points were assigned to the clusters (the figure is clipped to show only the main clusters).

Figure 6. Number and Classification of Clusters



4. Figure 7 shows the uncertainty levels of the data points that are assigned to the clusters, only showing the most important clusters like Figure 6. Figure 7 shows high levels of uncertainties for the data points in the small cluster suggesting that we should be skeptical of the results.

Figure 7. Clusters Uncertainty



5. The metrics shown in Figure 8 below summarize how the model-based clustering method performed for this user. The key metric for evaluation is specificity. The 0.4 Specificity suggests that the model detected 40% of the actual fraudulent transactions.

Figure 8. Summary of Model-Based Clustering for Profile 1 User

```
Confusion Matrix and Statistics

      Reference
Prediction  0    1
0      4240    9
1       137    6

      Accuracy : 0.9668
      95% CI   : (0.961, 0.9719)
      No Information Rate : 0.9966
      P-Value [Acc > NIR] : 1

      Kappa : 0.0702

      Mcnemar's Test P-Value : <2e-16

      Sensitivity : 0.96870
      Specificity : 0.40000
      Pos Pred Value : 0.99788
      Neg Pred Value : 0.04196
      Prevalence : 0.99658
      Detection Rate : 0.96539
      Detection Prevalence : 0.96744
      Balanced Accuracy : 0.68435
```

6. Although Figure 8 does not show an encouraging number for this model, Figure 9 compares the performance of this model against a simple logistic regression model. This shows that the model-based analysis performs better than logistic regression.

Figure 9. Comparison of Model-Based to Logistic Regression.

	Model Based	Logistic Regression
Actual Fraud Count	15	15
Predicted Fraud Count	143	1
Predicted Correct	6	0
% Detected	40.0%	0
% Correct	4.2%	0

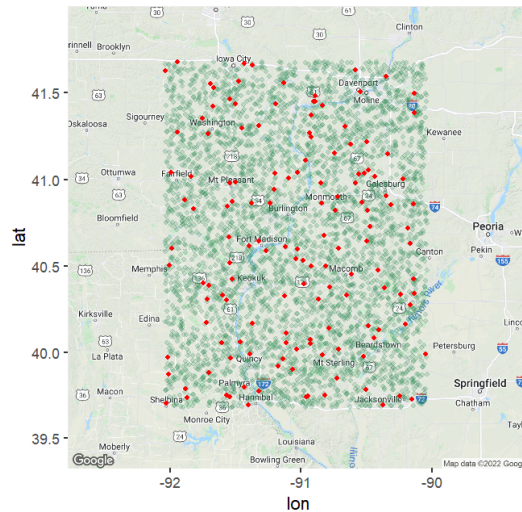
A deeper analysis was conducted on transactions that were predicted to be fraud to extract their merchants. Figure 10 below shows the results, revealing that the merchant category most likely to result in fraudulent activities is misc_net.

Figure 10. Cluster Analysis for Predicted Fraudulent Merchant Categories

Merchant <chr>	Proportions (%) <dbl>	Sum of Purchases (\$) <dbl>
misc_net	18.18	13463.53
misc_pos	17.48	10992.80
shopping_pos	17.48	25191.93
shopping_net	10.49	10125.14
home	9.09	2810.56
personal_care	6.29	1870.20
entertainment	4.90	1436.96
food_dining	4.90	2117.86
kids_pets	4.90	1746.85
health_fitness	3.50	1136.30
grocery_pos	1.40	701.42
travel	1.40	1593.92
gas_transport	0.00	NA
grocery_net	0.00	NA

Spatial analysis was used to visualize fraudulent merchant locations after conducting cluster analysis on dollars spent. This technique serves to look for visible patterns in geographical locations that may be at high risk for fraud. The expectation was a lump of fraudulent merchants in common geographic locations. However, as seen in figure 11 below, the spatial analysis did not show any visible patterns.

Figure 11. Spatial Analysis (Predicted Fraudulent Locations) for the Selected User from Profile 1.



VI. Conclusion & Recommendations

Users are in need of a self-auditing system due to the increase in credit card fraudulent activities. This will allow them to review their transactions and pinpoint any highlighted payments that may be fraudulent on their own. Given the size of the selected dataset - nearly 2 million data points - and likely much more in real-world financial transactions, using supervised techniques would prove to be computationally expensive and impractical. Moreover, the model runs the risk of overfitting, which would reduce the accuracy of fraudulent predictions made on new transactions. These limitations provide the scope of the third research question: determining fraudulent transactions with unsupervised data mining techniques.

Seven different persona profiles were developed to categorize the users based on their purchase behavior. This helped to evaluate them separately to determine which technique performs the best on each of the profiles. Then, four unsupervised clustering methods and one ensemble method (model-based and categorical clustering) were applied. Using Specificity and the number of False Positives as the key parameters of evaluation, the best performing techniques were identified for each persona profile.

Although the selected techniques performed better than a simple logistic regression model for all of the profiles, none of the methods performed sufficiently well for profiles 1 and 2. Therefore, this solution is not recommended for users belonging to profiles 1 and 2 and further development is necessary. Furthermore, all methods for all personas yielded a high number of false positives. Consequently, customers should be encouraged to verify which transactions are actually fraud as many of them could be false positives. Yet all methods performed better for users with smaller numbers of transactions. This implies that reducing the number of transactions may improve the accuracy of the models, as fraudulent transactions would stand out more in smaller datasets. This can be done by allowing the user to specify a time range when using the self-audit system, such as displaying transactions that occurred only in the past month. In terms of merchant categories, clustering analysis showed that the top three categories most likely to be fraudulent are `misc_net`, `misc_pos`, and `shopping_pos`.

Lastly, no visible patterns were identified through spatial analysis. However, this may be due to the fact that the selected dataset is a simulated dataset. The dataset contains fraudulent transactions created with defined proportions of merchants. A dataset suitable for both clustering and spatial analysis collected by credit card transactions may work better with the stated models in this report. Further investigations in this area are recommended.

References

- Credit Card Fraud Detection: How Machine Learning Can Protect Your Business From Scams. (2020, August 21). *Altexsoft*.
- Filipiak, P. (2020). COVID-19: the viral spread of a cashless society? *The Financial Times*.
<https://www.ft.com/partnercontent/comarch/covid-19-the-viral-spread-of-cashless-society.html>
- Marte, J. (2021, November 9). U.S credit card use returning to pre-pandemic patterns, NY Fed report finds. *Reuters*.
<https://www.reuters.com/business/us-credit-card-use-returning-pre-pandemic-patterns-ny-fed-report-finds-2021-11-09/>
- Poon, L. (2020, July 14). Coronavirus Hastens the Rise of the Cashless Economy. *Bloomberg Citylab*.
<https://www.bloomberg.com/news/articles/2020-07-14/the-costs-of-an-increasingly-cashless-economy>

Appendix

Figure 1. Histogram of Number of Transactions for Each Credit Card

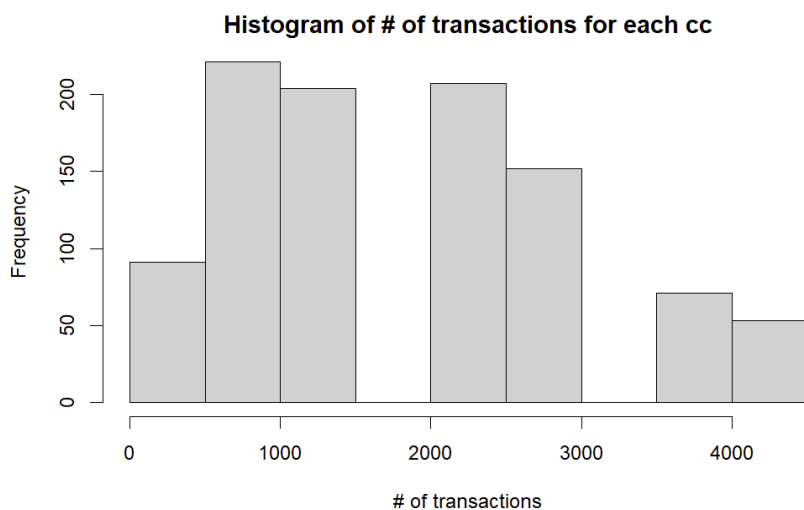


Figure 1 Code. Grouping data by cc_num.

```
data2 = data %>%  
  group_by(cc_num) %>%  
  summarize('count' = n(), 'mean_amt' = mean(amt), 'sum_amt' = sum(amt))  
hist(data2$count, main = "Histogram of # of transactions for each cc", xlab = "# of transactions")
```

Figure 2. Amount of Purchases in USD

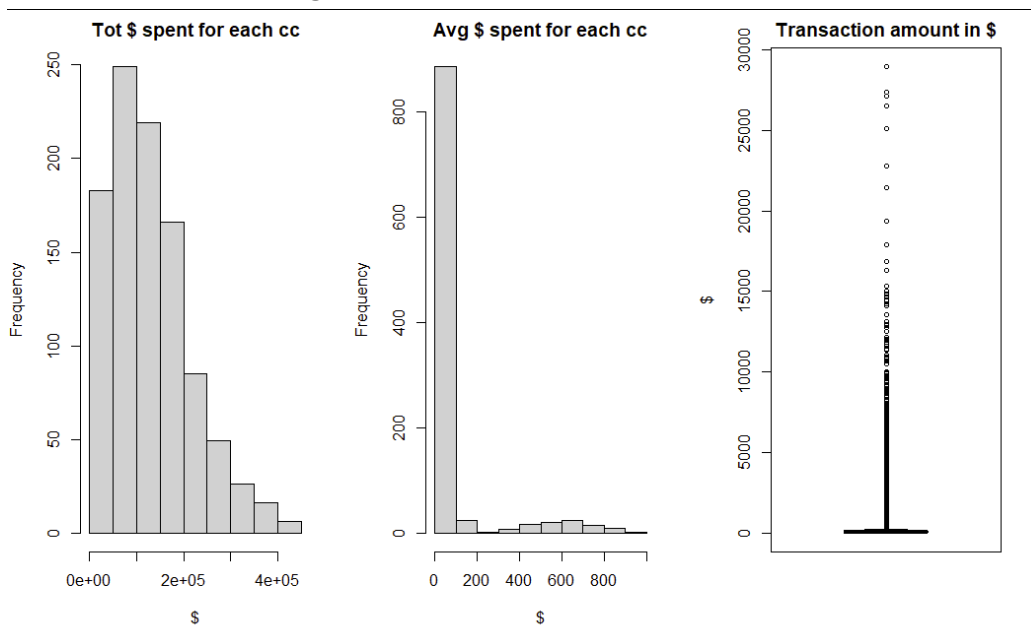


Figure 3. Silhouette Method

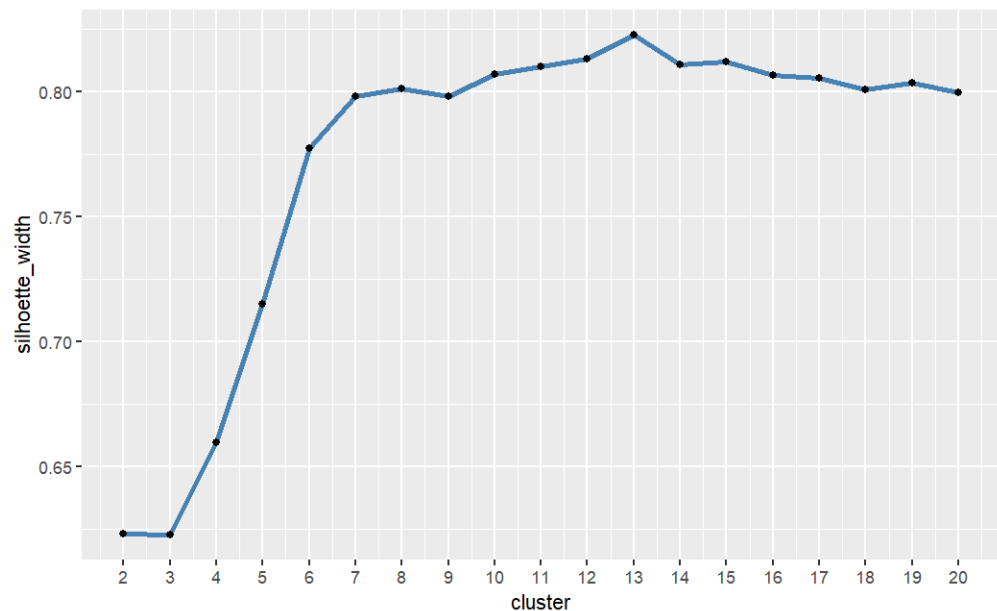


Figure 4. A plot of Tables 5 and 6 (Specificity and False Positives)

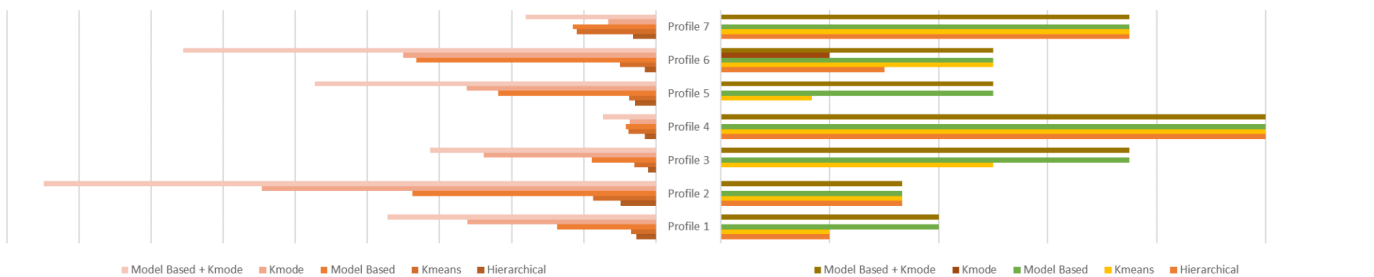


Figure 5. Credit-Card Prompt Box

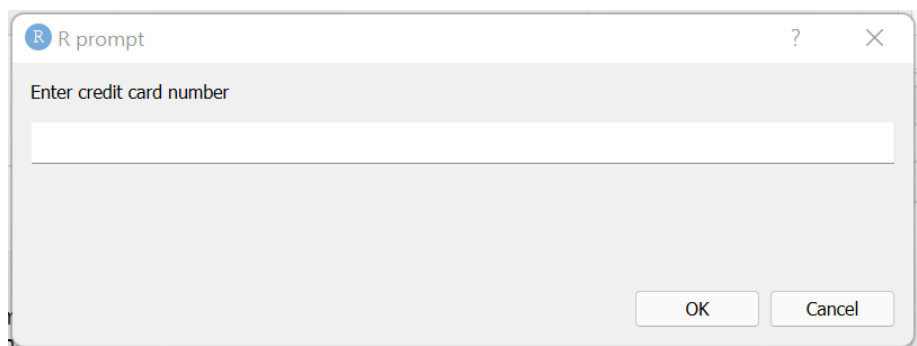


Figure 6. Number and Classification of Clusters

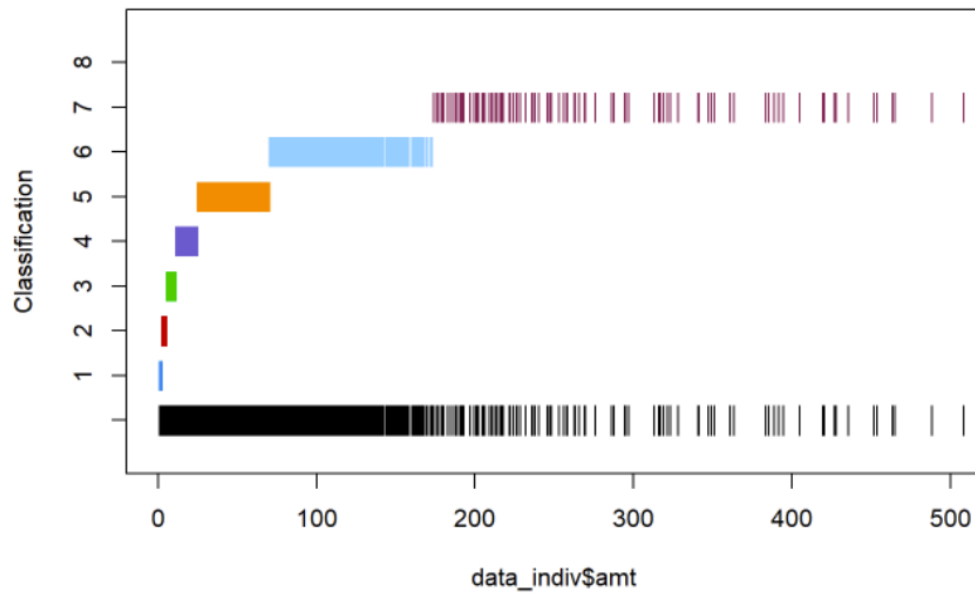


Figure 7. Clusters Uncertainty

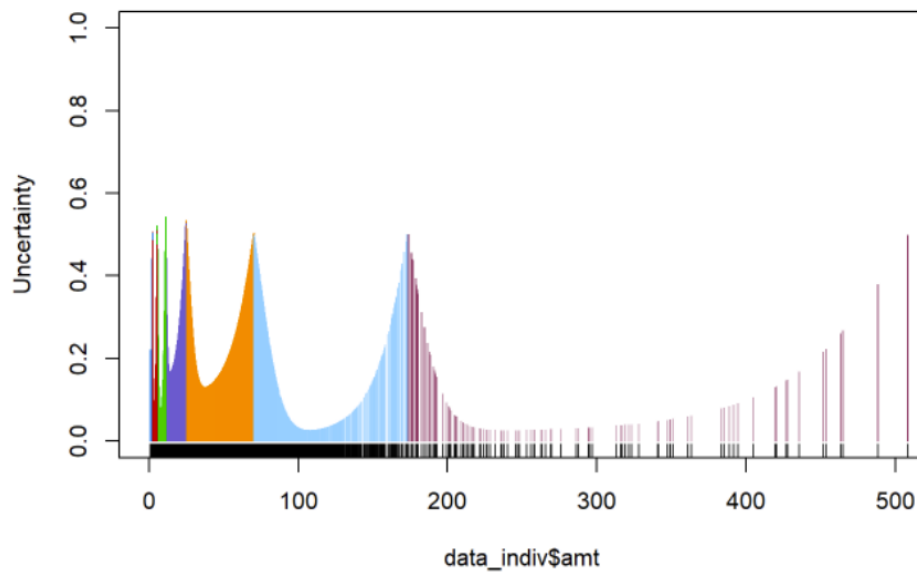


Figure 8. Summary of Model-Based Clustering for Profile 1 User

Confusion Matrix and Statistics

```

      Reference
Prediction 0    1
0  4240    9
1   137    6

Accuracy : 0.9668
95% CI : (0.961, 0.9719)
No Information Rate : 0.9966
P-Value [Acc > NIR] : 1

Kappa : 0.0702

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.96870
Specificity : 0.40000
Pos Pred Value : 0.99788
Neg Pred Value : 0.04196
Prevalence : 0.99658
Detection Rate : 0.96539
Detection Prevalence : 0.96744
Balanced Accuracy : 0.68435

```

Figure 9. Comparison of Model-Based to Logistic Regression.

	Model Based	Logistic Regression
Actual Fraud Count	15	15
Predicted Fraud Count	143	1
Predicted Correct	6	0
% Detected	40.0%	0
% Correct	4.2%	0

Figure 10. Cluster Analysis for Predicted Fraudulent Merchant Categories

Merchant <chr>	Proportions (%) <dbl>	Sum of Purchases (\$) <dbl>
misc_net	18.18	13463.53
misc_pos	17.48	10992.80
shopping_pos	17.48	25191.93
shopping_net	10.49	10125.14
home	9.09	2810.56
personal_care	6.29	1870.20
entertainment	4.90	1436.96
food_dining	4.90	2117.86
kids_pets	4.90	1746.85
health_fitness	3.50	1136.30
grocery_pos	1.40	701.42
travel	1.40	1593.92
gas_transport	0.00	NA
grocery_net	0.00	NA

Figure 11. Spatial Analysis (Predicted Fraudulent Locations) for the Selected User from Profile 1.

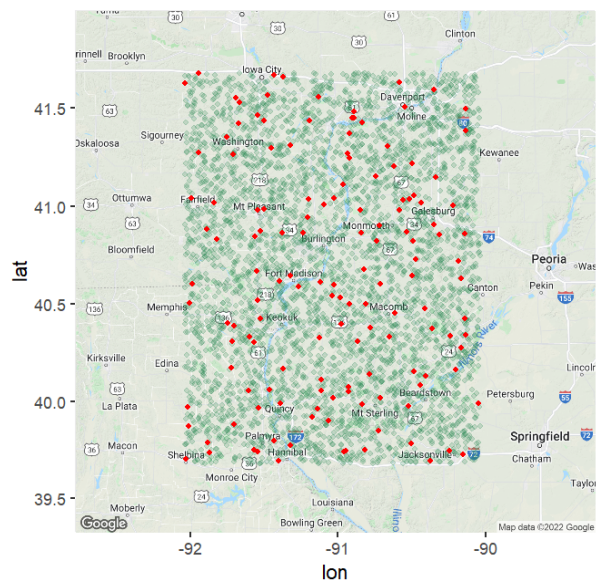


Table 1. Raw Dataset Variables Breakdown

Name	Type	Description
X	integer	Row ID
trans_date_trans_time	character	Transaction Timestamp
cc_num	numeric	Credit Card Number
merchant	character	Name of the Merchant
category	character	Type of Merchant
amt	numeric	Transaction Amount in \$
first	character	First Name of the Cardholder
last	character	Last Name of the Cardholder
gender	character	Gender of the Cardholder
street	character	Address of the Cardholder - Street
city	character	Address of the Cardholder - City
state	character	Address of the Cardholder - State
zip	integer	Address of the Cardholder - Zip Code
lat	numeric	Address of the Cardholder - latitude

		coordinate
long	numeric	Address of the Cardholder - longitude coordinate
city_pop	numeric	Population of the City that the Cardholder lives in
job	character	Cardholder's profession
dob	character	Cardholder's Date of Birth
trans_num	character	Transaction Number
unix_time	integer	Transaction Time in Unix Code
merch_lat	numeric	Address of the Merchant - latitude coordinate
merch_long	numeric	Address of the Merchant - longitude coordinate
is_fraud	integer	Response Variable

Table 2. Proportions of Transactions & Total Spent by Merchant Category.

category <chr>	proportions (%) <dbl>	Sum of Purchases (\$) <dbl>
entertainment	7.24	8602727
food_dining	7.06	6666408
gas_transport	10.15	11935568
grocery_net	3.50	3483204
grocery_pos	9.51	20550944
health_fitness	6.62	6628642
home	9.47	10209698
kids_pets	8.73	9303807
misc_net	4.89	7268762
misc_pos	6.17	7159471
personal_care	7.02	6250311
shopping_net	7.52	12112930
shopping_pos	8.99	13135052
travel	3.13	6477809

Table 3 . Cleaned Dataset Variables Breakdown

Name	Type	Description
trans_date_trans_time	Date/Time	Transaction Timestamp
cc_num	Factor	Credit Card Number
merchant	Factor	Name of the Merchant

category	Factor	Type of Merchant
amt	Numerical	Transaction Amount in \$
first	Factor	First Name of the Cardholder
last	Factor	Last Name of the Cardholder
gender	Factor	Gender of the Cardholder
street	Factor	Address of the Cardholder - Street
city	Factor	Address of the Cardholder - City
state	Factor	Address of the Cardholder - State
zip	Factor	Address of the Cardholder - Zip Code
lat	Factor	Address of the Cardholder - latitude coordinate
long	Factor	Address of the Cardholder - longitude coordinate
merch_lat	Factor	Address of the Merchant - latitude coordinate
merch_long	Factor	Address of the Merchant - longitude coordinate
zip2	Factor	First 3 digits of the cardholder's zip code
merch_country	Factor	Location of the Merchant - country
merch_state	Factor	Location of the Merchant - state
merch_county	Factor	Location of the Merchant - county

Table 4. Summary of Personas Profiles

Profile	Number of Transactions in the past 2 years
Profile 1	~4400
Profile 2	~3800
Profile 3	~3000
Profile 4	~2500
Profile 5	~1700
Profile 6	~1000
Profile 7	~300

Table 5. Specificity of Each Technique for Each Profile.

Profile	Hierarchical	Kmeans	Model Based	Kmode	Model Based + Kmode
Profile 1	0.2	0.2	0.4	0	0.4
Profile 2	0.3333	0.3333	0.3333	0	0.3333
Profile 3	0	0.5	0.75	0	0.75
Profile 4	1	1	1	0	1
Profile 5	0	0.1667	0.5	0	0.5
Profile 6	0.3	0.5	0.5	0.2	0.5
Profile 7	0.75	0.75	0.75	0	0.75

Table 6. False Positives for Each Profile.

Profile	Hierarchical	Kmeans	Model Based	Kmode	Model Based + Kmode
Profile 1	27	34	137	261	372
Profile 2	49	87	338	547	849
Profile 3	11	30	89	239	313
Profile 4	15	38	42	36	73
Profile 5	29	37	219	262	473
Profile 6	15	50	332	350	656
Profile 7	32	110	115	66	181

Table 7. Summary of best-performing models for each profile.

	Best Model	Specificity	True Positives	False Positives
Profile 1	Model Based	40%	6	137
Profile 2	Hierarchical	33%	1	49
Profile 3	Model Based	75%	3	89
Profile 4	Hierarchical	100%	2	15
Profile 5	Model Based	50%	3	219
Profile 6	Hierarchical	50%	3	15
Profile 7	Hierarchical	75%	3	32