1. Introduction

December 10, 2021

|  |  |
| --- | --- |
| Predict credit card frauds  Credit card fruad detecton | Tinglei Ruan, Yutong Chen, Fanru Zhou, Preetika Babu.  STAT4600 – Intermediate Statistics Model Analytics ----taught by Dr. JinFang |

Due to their simplicity of use and convenience, credit cards have become increasingly popular as a method of payment in recent years. Fraudsters are altering their malevolent operations to take advantage of the circumstances as a result of this problem. According to the Nilson Report, a magazine covering worldwide payment systems, overall losses due to fraud totaled USD 27.85 billion in 2018, and are expected to reach USD 35.67 billion in 2023.

## **1.1 Business Analytics problem**

When a transaction is carried out by an unauthorized party without the knowledge of the rightful owner and/or relevant institution, it is said to be fraudulent. Our research problem is finding to predict if the transaction is a fraudulent one or not.

## 1.2 Why is this research problem important

The American economy relies heavily on credit cards. We have become a "cashless society" as a result of the widespread usage of credit cards and checks.  As a result of the rapid growth of credit card usage, it has its own drawbacks. The more the usage of the credit card more are frauds.

* The banks would be benefited by this as they can decrease the fraudulent happening and can increase their profits by showing the less fraudulent rate to attract their new customers. It also helps in reducing the loses.
* Our research problem will also help in increasing the economy of the country as credit cards have a major role in it.
* When a credit card fraud happens, it effects the customer credit score, oue problem aims to reduce situations like these because the credit score plays a very important role in future loans and financial aids.

## 1.3 How does it relate to the STAT4600 class?

This project will be using statistical knowledge that been taught during the STAT4600 course, range from calculating central tendency, dispersion, sampling dataset, visualizing dataset, and finally building a logistic regression model. All the methods that are applied during this project help to strengthen team members’ statistical competence.

# 2. Methods

## 2.1 Dataset Overall Description

The team members got and downloaded the dataset from Kaggle. The dataset is a population dataset and has over 1.2 million observations along with 25 variables. The dataset contains all the information surrounding a credit card transaction, including the credit card number, transaction data, specific location, and demographical information of credit card holder.

## 2.2 Description of Analysis and Modeling Approach

As the business problem declared above is identifying which factor(s) result in a fraudulent transaction, this project will adopt calculation, visualization, and logistic regression to find the answers. The team members will do a lot of Factors-wised visualizations, and the visualizations will present a straightforward view of the trend. After that, logistic regression will further justify which factor(s) really impact the result.

## 2.3 Methods

### 2.3.1 Chapter 1: Introduction

* P.13: Elements in this dataset are transaction numbers, which serve as primary keys.
* P.14: Variable: amt; Observation: NC is the value of first row and column ‘state’; Dataset: fraudTrain.
* P.15: Quantitative variable: amt, city\_pop, unix\_time, Age.

Qualitative variable: trans\_date\_trans\_time, cc\_num, merchant, category, first, last, gender, street, city, state, zip, lat, long, job, dob, trans\_num, merch\_lat, merch\_long, is\_fruad, Time\_Category (added).

* P.17: Discrete variable: Age, city\_pop, unix\_time.

Continuous variable: amt.

* P.19: The cross-section of the data set should be different columns at one specific time, then the time-series is a data column recorded in time sequence by the same unified index.
* P.20: Population is the dataset fraudTrain, which reserves all the records of transactions. Sample is a portion of records that the team members will select from the population.
* P.21: Census implies if the team members use the whole dataset to conduct analysis to get a result; survey means that the team members will only use a portion of the dataset to analyze.
* P.22: Filtering the dataset by specifying the value of state is ‘MA’ produces a representative sample that reflects transactions happened in Massachusetts.
* P.23: sample(df\_train[“gender”], 1000, replace = TRUE)

Sample(df\_train[“gender”], 1000, replace = FALSE)

* P.26: Fig.1 shows a random sampling, and Fig.2 shows nonrandom sampling.
* P.29: In this data set, sampling error refers to the difference between the selected variable or element and the true value. For example, the average age in this data set is 46 years old, but the selected samples are 42,34 and 43 years old, which obviously has errors. Non-sampling errors are often caused by human factors. For example, we used the wrong calculation method, or the respondents were hiding something in their answers.
* P.31: The selection error occurred when only participants who were interested in the survey answered the questions, such as someone who had experienced credit card fraud but didn't answer the questionnaire.
* P.33: The nonresponse error occurs when the respondents did not answer the questionnaire.
* P.34: The response error occurs when participants answer the questions incorrectly or misunderstand the quiz.
* P.35: Because the members of the voluntary response sample are self-selected volunteers, they tend to take positions at extreme of the topic. This can lead to biased and unreliable results.
* P.36: Fig.1 is a good example of simple random sampling.
* P.45: The dataset is an observational study because all the transactions are happened naturally by real credit card holders, not designed by scientists.

### 2.3.2 Chapter 2: Organizing and graphing data

* P.10: Fig.3 is a frequency distribution table for the qualitative variable “Category”.
* P.11: Fig.4 is a relative frequency distribution.
* P.15: Fig.5 is a Pareto Chart applying to variable category, and Fig.6 presents the same attribute by Barchart.
* P.19-23: Fig.7 shows the distribution of quantitative variable ‘amt’
* P.29: Fig.8 shows the histogram of column Age.
* P.30:

### 2.3.3 Chapter 3: Numerical Descriptive Measures

First of all, according to the dataset, the team reorganized the category and state variables to create category of merchant count and state count for this portion of project. The attribute, category, in the dataset been analyzed, means the category of merchant, for example, online shopping, in-store shopping, and food dining. Table.1 explains the count of category, which shows the distribution of each category. Besides, Table.2 tells the distribution of how many transactions happened in each state.

* P.7: Table.3 lists all the means of quantitative variables in the dataset.

|  |  |
| --- | --- |
| Mean of Variables | |
| Variables | Mean value |
| Category of merchant count | 92619.64 |
| Amount | 70.35 |
| City population | 88824 |
| State count | 25425 |

*Table.3 Means of variables*

* P.8 & 13: Taking state count outliers to do k% trimmed mean. Because of the limitation of count, around 12% trimmed mean is taken. Then DE:9, RI:550, AK:2120, PA:79847, NY:83501, TX:94876 are cut. So the 12% trimmed mean of state count is 23017.16.
* P.9: Table.4 prepares the medians of all the quantitative variables in the dataset.
* P.10: Difference between mean and median, the median which is resistant is less affected by outliers, while mean which is nonresistant is affected more by outliers.
* P.11: Mode of each variable in the dataset is appear in Table.5. According to the table, the team first mix the first name and last name to get mode of a person Scott Martin, who has 3697 records.Then the team mix the street, city, state, zip code, latitude and longitude to get a popular address in the dataset: 864 Reynolds Plains Uledi PA, 15484,（39.8936，-79.7856）, which has 2507 records.
* P.12: All the variables are unimodal.
* P.13: The team uses 10 amount values to create a weighted mean as shown by Table.6. The Weighted mean of 10 amount=∑xw/∑w=708.05/10=$ 70.805
* P.19-21: The first variable is amount, according to the amount data, the mean of amount is 70.35, median is 47.52, and the mode is 1.14. So, the mode is the smallest value, and the mean is the largest one. Mean>median>mode, for a histogram and a frequency distribution curve skewed to the right. The second variable is city population, according to the city population data, the mean of it is 88824, the median is 2456, and the mode is 606. Hence, the mode is the smallest value and the mean is the largest one. The relationship of mean, mode and median of city population is as same as amount data. Mean>median>mode, for a histogram and a frequency distribution curve skewed to the right.
* P.23: Table.7 gives range of each quantitative variable in the dataset.
* P.24: Table.8 exhibits variance and standard deviations in the dataset. Standard deviations of all four variables are larger values which means that all the four variables are spread over a relatively large range around mean.
* P.28-31: The category of merchant count’s coefficient of variation is 30.62%.

The state count’s coefficient of variation is 78.98%.

* P.47: The average count of category for customers is 92619.64 with a standard deviation of 28361.55, Using Chebyshev’s theorem, find the minimum percentage of customer between 35896.54 and 149342.74
* P.49-50: Because all of data in the dataset are not normal distribution, so no bell-shaped distribution, no empirical rule.
* P.52: The team use whole state data in the dataset instead of only using 20 observations to calculate percentile. Percentile of 20 for amount: P20= (20 \* 1296675) / 100 = 259335
* P.53 The team use whole state count data in the dataset instead of only using 20 observations to calculate percentile. Percentile rank of 1 for amount: X1 = ( 9/1296675)\* 100% =0.00069%
* P.54: The team use whole state count data and category of merchant count in the dataset instead of only using 20 observations to calculate the quartiles and IQR.

State count: Q2=22996, Q1=11768, Q3=30266, IQR=Q3-Q1=25425

Category of merchant count: Q2 = 92737.5, Q1 = 79655, Q3 = 116672, IQR = 37017

* P.56-61: According to Fig.11, the team use category of merchant count in the dataset instead of only using 20 observations to create the box-and-whisker plot.

图表, 箱线图

描述已自动生成

*Figure 11Box and Whisker plot of category*

### 2.3.4 Chapter 4: Probability

* P.9: Experiment: selecting a customer and identify the gender. Outcomes: Female, Male

Sample space: the collection of all outcomes for the experiment {Female, Male}

* P.10: Event: the outcomes of selecting customers. Simple event: selecting two customers whether the customer selected each time is a male or a female. Each of the four outcomes (E1=FM, E2=FF, E3=MF and E4=MM) for this experiment is a simple event.
* P.12: Compound event: selecting three customers and observing whether the customer selected each time is from TX or not. Let W be the event that at most one customer from TX is selected.
* P.25: Mutually exclusive event: a customer is fraud or not. Table.9 is an example of the two-dimensions contingency table. Independent event: A selected customer is a female or made a fraudulent transaction. Dependent event: The probability that a person who is fraud is a female.

P (Female| is fraud) = (3735/ 7506) = 0.4976. Table.10 is an example of marginal probabilities.

* P.26: Complementary event: Selecting a customer is a female or a male.
* P.29: Intersection of events: selecting a customer is a female and fraud.

P ( Female and is fraud) = 54.74% \* 0.58% = 31.75%

* P.33: Union of events: a customer is fraud or not.

P (is fraud or nonfraud) = 1

### 2.3.5 Chapter 7: Sampling Distribution

7-2: The mean amount of credit fraud in our data set fraudtest.csv is 69.3928 and the standard deviation is 156.7459. The team tried to find the mean and standard deviation for different sample sizes:

**50000** :- mean : 69.3928 Standard deviation: 150.5911

**100000 :-** mean : 69.3928 Standard deviation: 152.6439

**350000 :-** mean : 69.3928 Standard deviation: 149.6920

7-5: The team tried to find out what is the probability of the mean amount to be in between 100 and 120 dollars and the probability is 0.0492

The most frauds happen from the merchant company “fraud\_Abbott-Rogahn” so now we try to find the mean and standard deviation of the population that fall under this category.

Mean: 66.1983

Standard deviation: 68.9141

**7-10**

Given that, population proportion ( p ) = 0.75

Selecting random sample size ( n ) = 1400 from our dataset

The team found p(0.765<p<0.78)

=p((0.765-0.75)/sqrt((0.75\*(1-0.75)/1400))<p<((0.78-0.75)/sqrt((0.75\*(1-0.75)/1400)))

=p(1.3<z<2.59)

=p(z<2.59) – p(z<1.3)

= 0.9952-0.9032

= 0.092

### 2.3.6 Chapter 8: Estimations of the mean and proportion

Assume Samples = 100

Mean: 69.3928

Standard deviation: 156.7459

At 90%CI

CI = (69.3928 – 1.65\*156.7459/sqrt(100), 69.3928 + 1.65\*156.7459/sqrt(100))

CI lies in (43.5297,95.2558)

## 2.4 Choice of Model

As the business problem the team was focusing on is identify whether a transaction is a fraud or not, which is a two-level classification problem, the team members decided to apply binomial logistic regression to explore influenceable factors. Variables like Time\_Category, state, Age, amt, category, and so on will be gathered to build the model. After building the model and playing with arguments and variables to get the best accuracy, team members will interpret estimates and corresponding significance from the result.

## 2.5 Model development and analysis details

### 2.5.1 Data Engineering

In order to transform values of categorical variables to numeric, the team members adopt various of encoding techniques, including one-hot encoding, and ordinal encoding. They divided a day into four time-categories, involving morning, afternoon, evening, and midnight. Based on the economic situation of a state, values of variable state were coded into 1 to 5, and the higher number means the higher GDP per capita of the state. In addition, the original value of column gender, ‘F’ and ‘M’, were transformed into 0 and 1. Finally, they applied one-hot encoding on variable category and Time\_Category, which resulted in 18 more variables in the dataset.

### 2.5.2 Modeling Process

At first, the team built the model with the listed variables: state, amt, gender, city\_pop, Age, and all the encoded variables of category and Time\_Category. The result, according to fig.9, revealed

A screenshot of a computer

Description automatically generated with low confidence

Figure 10 Coefficient of first modeling

that variable gender, city\_pop, and state are not statistically significant because their p-value are 0.11, 0.74, and 0.22, respectively. Therefore, the team eliminated the three variables in further modeling. After reducing the three variables, the team remodeled the dataset and got the information as presented in fig.10. Interpretations from the table are variables including amount of purchase, age, categorygas\_transport, Time\_CategoryEvening, and Time\_CategoryMidnight have a positive relationship with the response variable, while variables like categoryentertainment, categoryhealth\_fitness, and categoryshopping\_net had a negative relationship against the target variable. The positive relationship means that if the value of the variable goes high, then the record tends to be a fraud, and the negative relationship goes to an opposite way.

Text

Description automatically generated

Figure 10 Logistic Regression Result

# 3. Results & Discussion

## 3.1 Results: Presentation of results

The modeling process turns out some results. Elder credit card holders have a relatively high potential to make a fraudulent transaction; transactions that have a high amount of value tend to be a fraud; Transactions that happen during evening and midnight are substantially problematic than transactions that are made in the daytime. In contrast, sex of transaction holders, variation of the 51 states, and population of city of the credit-card holder don’t determine whether a transaction is a fraud or not.

## 3.2 Visualizations & Analysis

Fig.12, the state-wise fraudulent rate, conveys that fraud rate varies little amongst the 51 states except Alaska. The reason that Alaska shows a compellingly high fraud rate is because there are a lot of homeless people, and many of them vanish, most likely into gang-run sex trafficking. Oil and natural gas are no longer the state's biggest moneymakers. Prisons are. The state receives funding from the federal government to house, clothe, and feed inmates. However, because the reformation agenda has been a failure, the recidivism rate is likely to be in the top 5% of the country. Alaska is also one of the most popular destinations for extended holidays, which draws attention to credit card fraud as one of the most widely utilized payment methods.

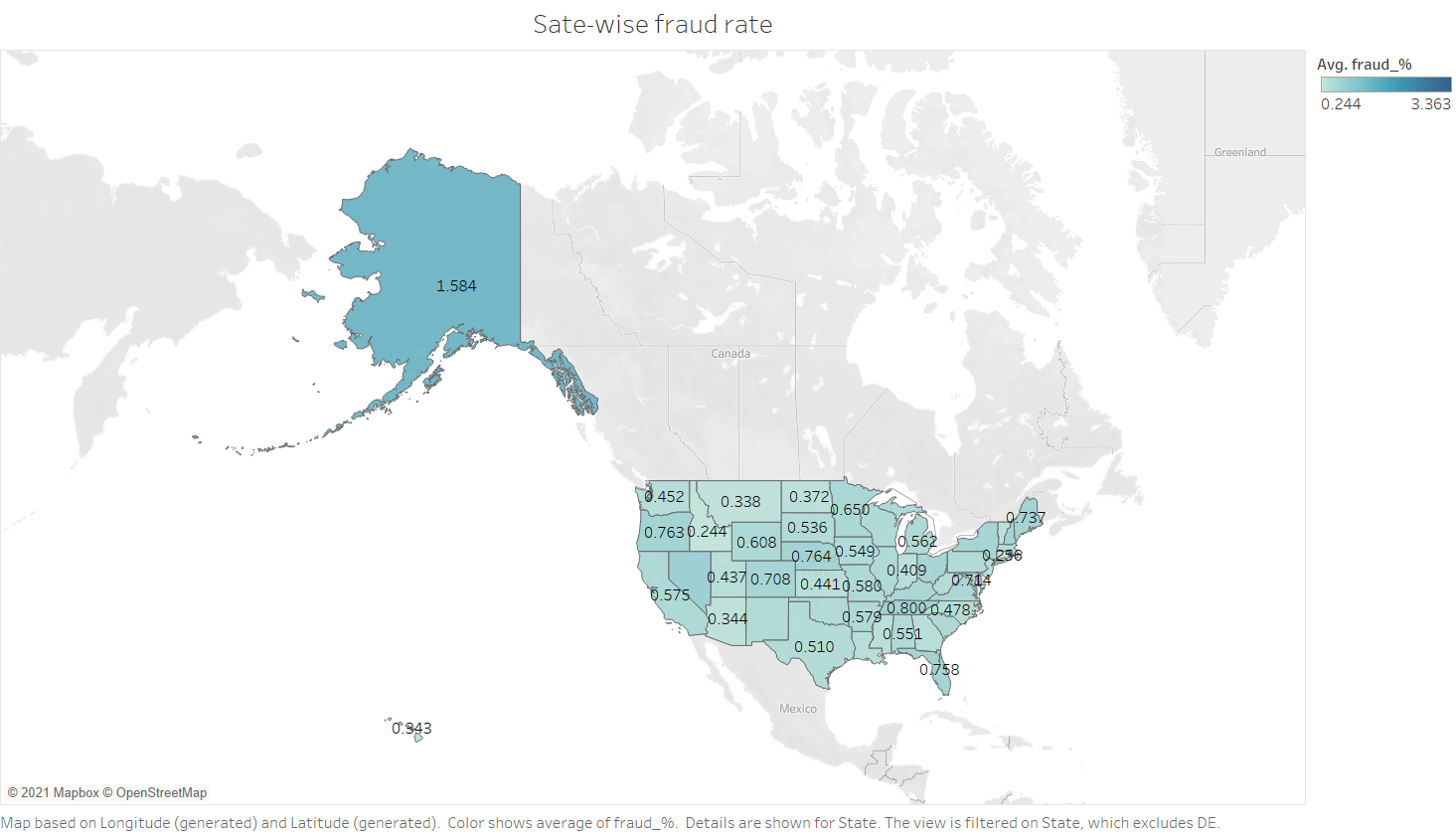


Figure 12 State\_Wise Fraud Rate

Fig.13 clearly presents the difference of fraud rate amongst different categories. The top four categories are online shopping, online miscellaneous transactions, store grocery, and store shopping, meaning that these four categories are much likely to cause frauds.

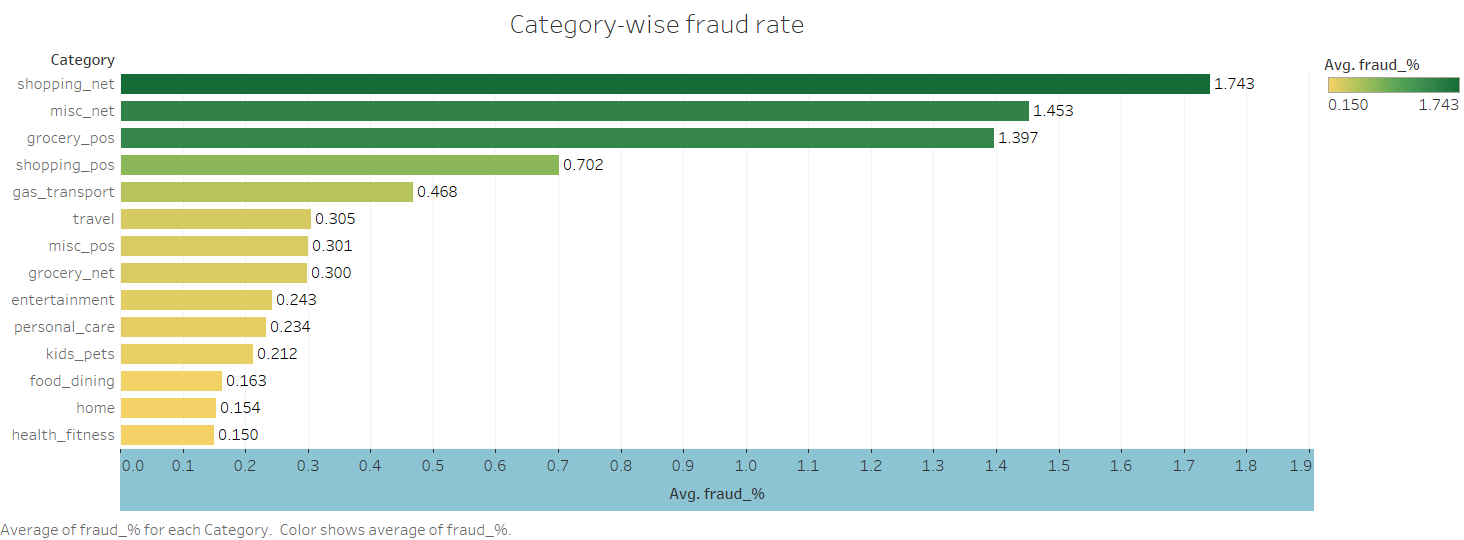


Figure 13 category-wise fraud rate

Fig.14 communicates that the gap of fraud rate between male and female is not significant, which justifies the finding of logistic regression.

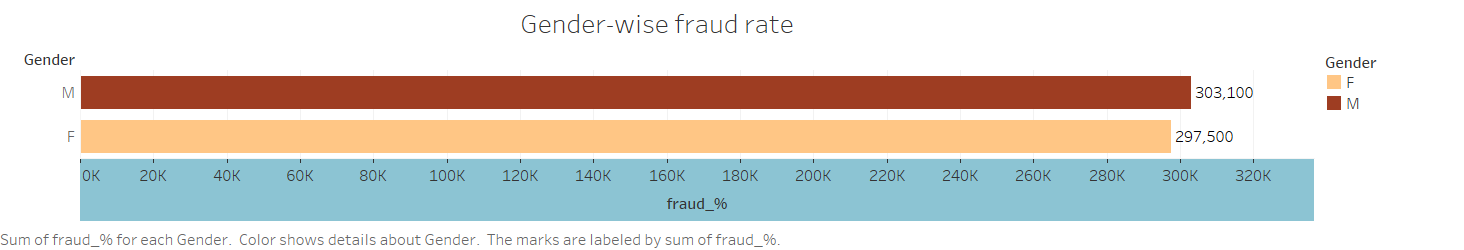


Figure 14 Fraud rate of different sex

Fig.15 is a monthly count of fraud transactions. The tendency tells that the number of monthly fraud transactions is almost stable except the winter days. A possible explanation to this scenario is that working opportunity during the cold months is less, hence resulting more criminals including credit card frauds.

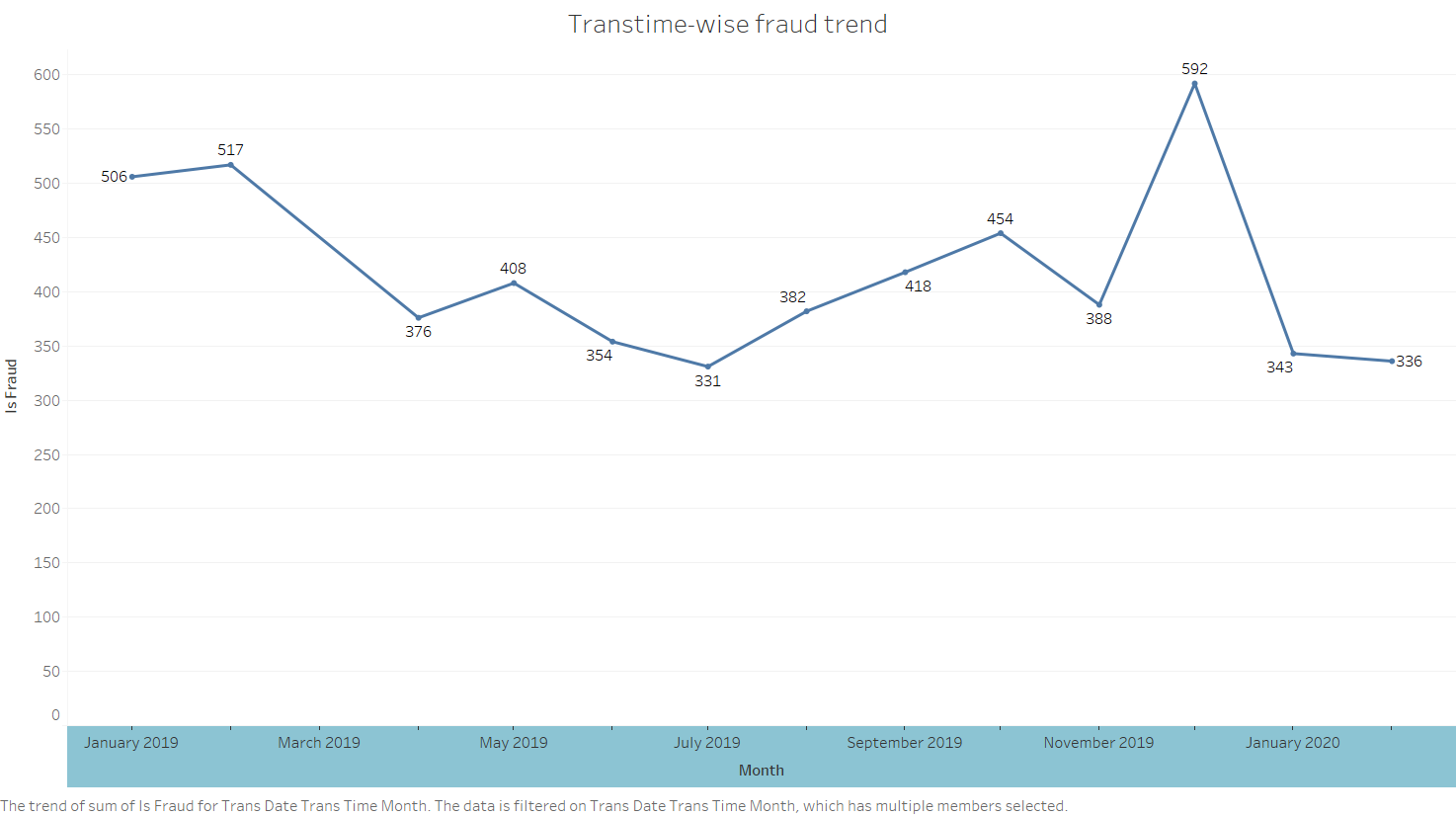


Figure 15 Fraud rate by month

Also, for better understanding how the population of a city affect the fraud rate, the team members categorize the population of a city into three categories, comprising of large city, small city, and medium city. The length of bar chart, as shown in Fig.16, testifies the random relationship between population of city and fraud rate, which were proposed by the regression model.

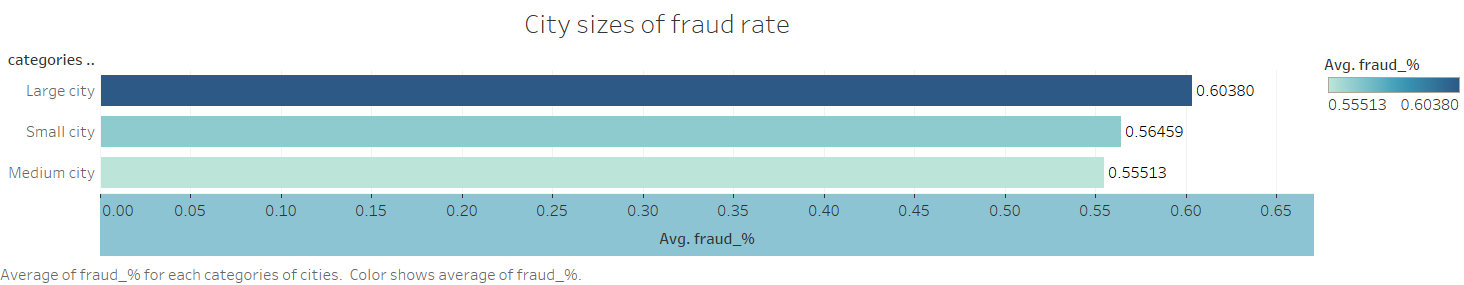


Figure 16 Fraud rate by city population

# 4. Conclusion

## 4.1 Key Findings

Midnight and evening produce a significant proportion of fraud transactions. Elder credit card holders and higher amount of money are problematic. Fields like healthcare, fitness, pet, and dining are much more safer than online/in-store shopping, gas, and in-store grocery.

## 4.2 Business Insights & Recommendations.

The Management of the credit card center could adopt the model to discern whether a transaction is a fraud a not. For example, they should put their effort on night-time transactions and certain categories. However, for a better accuracy, they should manually check ambiguous transactions by viewing other information of the card holder, e.g., recent transaction, monthly income etc.

# 5. Appendix.

## 5.1 Tables & Figures

Text, table

Description automatically generated

Figure 2 Random Sampling

Table

Description automatically generated

Figure 3 Nonrandom Sampling

A picture containing table

Description automatically generated

Figure 4 Categorical Table

A picture containing table

Description automatically generated

Figure 5 Relative Frequency

Chart, scatter chart

Description automatically generated

Figure 6 Pareto Chart

Chart, bar chart

Description automatically generated

Figure 7 Barchart

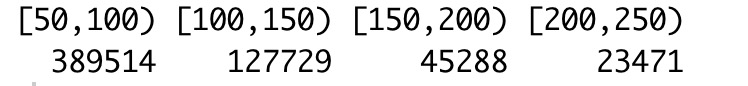


Figure 8 Quantitative Frequency Table

Chart, histogram

Description automatically generated

Figure 9 Histogram 'Age'

|  |  |  |
| --- | --- | --- |
| The Distribution of the Category of merchant count | | |
| Category | Count | Percentage(%) |
| travel | 40507 | 3.123913 |
| grocery\_net | 45452 | 3.505273 |
| misc\_net | 63287 | 4.880714 |
| misc\_pos | 79655 | 6.14302 |
| health\_fitness | 85879 | 6.623017 |
| personal\_care | 90758 | 6.999287 |
| food\_dining | 91461 | 7.053502 |
| entertainment | 94014 | 7.25039 |
| shopping\_net | 97543 | 7.522548 |
| kids\_pets | 113035 | 8.717296 |
| shopping\_pos | 116672 | 8.997783 |
| home | 123115 | 9.494669 |
| grocery\_pos | 123638 | 9.535003 |
| gas\_transport | 131659 | 10.15359 |
|  | Sum=1296675 | 100 |

|  |  |  |
| --- | --- | --- |
| The Distribution of the States | | |
| States | Count | Percentage(%) |
| DE | 9 | 0.000694 |
| RI | 550 | 0.042416 |
| AK | 2120 | 0.163495 |
| HI | 2559 | 0.197351 |
| DC | 3613 | 0.278636 |
| ID | 5545 | 0.427632 |
| NV | 5607 | 0.432414 |
| CT | 7702 | 0.593981 |
| NH | 8278 | 0.638402 |
| UT | 10699 | 0.82511 |
| AZ | 10770 | 0.830586 |
| MT | 11754 | 0.906472 |
| VT | 11768 | 0.907552 |
| SD | 12324 | 0.950431 |
| MA | 12376 | 0.954441 |
| CO | 13880 | 1.07043 |
| ND | 14786 | 1.140301 |
| NM | 16407 | 1.265313 |
| ME | 16505 | 1.272871 |
| TN | 17554 | 1.35377 |
| OR | 18597 | 1.434207 |
| WA | 18924 | 1.459425 |
| WY | 19322 | 1.490119 |
| LA | 20965 | 1.616828 |
| MS | 21188 | 1.634025 |
| KS | 22996 | 1.773459 |
| NE | 24168 | 1.863844 |
| NJ | 24603 | 1.897391 |
| WV | 25691 | 1.981298 |
| GA | 26063 | 2.009987 |
| MD | 26193 | 2.020013 |
| OK | 26671 | 2.056876 |
| IA | 26985 | 2.081092 |
| IN | 27580 | 2.126979 |
| KY | 28475 | 2.196001 |
| SC | 29190 | 2.251142 |
| VA | 29250 | 2.25577 |
| WI | 29368 | 2.26487 |
| NC | 30266 | 2.334124 |
| AR | 31127 | 2.400524 |
| MN | 31714 | 2.445794 |
| MO | 38403 | 2.961652 |
| AL | 40989 | 3.161085 |
| FL | 42671 | 3.290801 |
| IL | 43252 | 3.335608 |
| MI | 46154 | 3.559412 |
| OH | 46480 | 3.584553 |
| CA | 56360 | 4.346502 |
| PA | 79847 | 6.157827 |
| NY | 83501 | 6.439624 |
| TX | 94876 | 7.316868 |
|  | Sum=1296675 | 100 |

Table 2 Count of State

|  |  |
| --- | --- |
| Mean of Variables | |
| Variables | Mean value |
| Category of merchant count | 92619.64 |
| Amount | 70.35 |
| City population | 88824 |
| State count | 25425 |

Table 3 Mean of variables

|  |  |
| --- | --- |
| Median of Variables | |
| Variables | Median value |
| Category of merchant count | 94014 |
| Amount | 47.52 |
| City population | 2456 |
| State count | 22996 |

Table 4 Median of variables

|  |  |  |
| --- | --- | --- |
| Mode of Variables | | |
| Variables | Mode value | Records |
| Category of merchant count | gas\_transpot | 131659 |
| Transaction date and time | 2019-04-22 16:02: 01 | 7 |
| Merchant | fraud\_Kilback LLC | 3521 |
| Amount | 1.14 | 440 |
| First name | Christopher | 21518 |
| Last name | Smith, | 23394 |
| Gender | Female | 709863 |
| Street address | 864 Reynolds Plains | 2507 |
| City | Birmingham | 4499 |
| State count | TX, | 94876 |
| Zip code | 73754 | 2905 |
| Latitude | 36.385 | 2905 |
| Longitude | -98.0727 | 2905 |
| City population | 606 | 4376 |
| Job position | Film/video editor | 7940 |
| Date of birth | 1977-03-23 | 4552 |
| Unix\_time | 1335110521 | 4 |
| Merchant latitude | 41.014694 | 4 |
| Merchant longitude | -87.11641 | 3 |

Table 5 Mode of Variables

|  |  |  |
| --- | --- | --- |
| Amount  x | Count  w | xw |
| 161.61 | 1 | 161.61 |
| 84.94 | 1 | 84.94 |
| 73.60 | 1 | 73.60 |
| 105.84 | 1 | 105.84 |
| 65.49 | 1 | 65.49 |
| 107.22 | 1 | 107.22 |
| 9.21 | 1 | 9.21 |
| 10.68 | 1 | 10.68 |
| 8.05 | 1 | 8.05 |
| 81.41 | 1 | 81.41 |
|  | ∑w=10 | ∑xw=708.05 |

Table 6 Weighted mean of amount of transaction

|  |  |
| --- | --- |
| Range of variables | |
| Variable | Range |
| Amount | (1.0, 28948.9) |
| Zip code | (1257, 99783) |
| Latitude | (20.0271, 66.6933) |
| Longitude | (-165.6723, -67.9503) |
| City population | (23, 2906700) |
| Unix\_time | (1325376018, 1371816817) |
| Merchant latitude | (19.02779, 67.51027) |
| Merchant longitude | (-166.6712, -66.9509) |

Table 7 Range of variables

|  |  |  |
| --- | --- | --- |
| Variance and standard deviation of variables | | |
| Variable | Variance | Standard deviation |
| Category of merchant count | 804377498 | 28361.55 |
| Amount | 25701.23 | 160.316 |
| City population | 91177643760 | 301956.4 |
| State count | 403274177 | 20081.69 |

Table 8 Variance and Standard Deviation of Variables

## 5.2 R Codes

### 5.2.1 Chapter 1 & 2 Code

category1 <- table(df\_train$category)

sum\_c <- sum(summary(factor(df\_train$category)))

frequency <- summary(factor(df\_train$category)) / sum\_c

frequency

library(qcc)

pareto.chart(category1)

barchart(category1)

sample1 <- df\_train[df\_train$trans\_num %in% sample(df\_train$trans\_num,1000),]

for (i in 1:nrow(sample1)) {

if (sample1$amt[i] > 250) {

sample1$amt\_c[i] = 'high\_amt'

} else if(sample1$amt[i] > 200) {

sample1$amt\_c[i] = 'rela\_high\_amt'

} else if (sample1$amt[i] > 150) {

sample1$amt\_c[i] = 'middle\_amt'

} else if (sample1$amt[i] > 100) {

sample1$amt\_c[i] = 'rela\_low\_amt'

} else {

sample1$amt\_c[i] = 'low\_amt'

}

}

View(sample1)

table(sample1$amt\_c)

breaks = seq(50,250,50)

cut = cut(df\_train$amt,breaks, right = FALSE)

freq <- table(cut)

freq

hist(df\_train2$Age)

polygon(freq)

polygon(x= df\_train$Age[1:5],y = df\_train$amt[1:5])

plot.new()

polygon(1:9, c(2,1,2,1,1,2,1,2,1))

plot.new()

5.2.2 Chapter 3 & 4

df<-read.csv("C:\\Users\\49109\\OneDrive\\??????\\intermed stats model analytics\\fraudTrain.csv")

summary(df)

amt\_df<-df$amt

amt\_df[amt\_df>quantile(amt\_df, 0.1)|amt\_df<quantile(amt\_df,0.9)]

boxplot(amt\_df)

?mode

mode(df$trans\_date\_trans\_time)

getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

v<-df$trans\_date\_trans\_time

result<-getmode(v)

print(result)

merchant<-df$merchant

print(getmode(merchant))

category<-df$category

print(getmode(category))

amt<-df$amt

print(getmode(amt))

first<-df$first

print(getmode(first))

last<-df$last

print(getmode(last))

library(dplyr)

left\_join(df$first,df$last,by="name")

name<-paste(df$first, df$last)

print(getmode(name))

gender<-df$gender

print(getmode(gender))

street<-df$street

print(getmode(street))

city<-df$city

print(getmode(city))

state<-df$state

print(getmode(state))

address<-paste(df$street,df$city,df$state)

print(getmode(address))

zip<-df$zip

print(getmode(zip))

lat<-df$lat

print(getmode(lat))

city\_pop<-df$city\_pop

print(getmode(city\_pop))

job<-df$job

print(getmode(job))

dob<-df$dob

print(getmode(dob))

unitx<-df$unix\_time

print(getmode(unitx))

count(getmode(unitx))

cc<-df$cc\_num

print(getmode(cc))

getmode(df$cc\_num)

getmode(df$merch\_lat)

getmode(df$merch\_long)

median(df$trans\_date\_trans\_time)

median(df$zip)

median(df$lat)

median(df$long)

median(df$city\_pop)

median(df$dob)

median(df$unix\_time)

median(df$merch\_lat)

median(df$merch\_long)

mean(df$trans\_date\_trans\_time)

summary(df$city\_pop)

range(df$amt)

range(df$zip)

range(df$lat)

range(df$long)

range(df$city\_pop)

range(df$unix\_time)

range(df$merch\_lat)

range(df$merch\_long)

var(df$trans\_date\_trans\_time)

var(df$merchant)

var(df$amt)

sd(df$amt)

var(df$city\_pop)

sd(df$city\_pop)

mean(df$amt)

sample(df$amt,10,replace=T)

### 5.2.3 Modeling Code

# STAT 4600 Project

# Import Data

library(caret)

library(varhandle)

library(e1071)

library(readr)

library(dplyr)

df\_train <- read\_csv("~/Desktop/fraudTrain.csv")

View(df)

str(df)

df\_test <- df

# Data Engineering

# Adding Variable Age

library(lubridate)

df\_train$Age <- (year(Sys.time()) - year(df\_train$dob))

# Transforming Transaction time into a categorical variable

for (i in 1:nrow(df\_train)){

if (format(df\_train$trans\_date\_trans\_time[i],"%H") > "17"){

df\_train$Time\_Category[i] <- "Evening"

}

if (("11" < format(df\_train$trans\_date\_trans\_time[i],"%H")) && (format(df\_train$trans\_date\_trans\_time[i],"%H") <= "17")){

df\_train$Time\_Category[i] <- 'Afternoon'

}

if (("04" < format(df\_train$trans\_date\_trans\_time[i],"%H")) && (format(df\_train$trans\_date\_trans\_time[i],"%H") <= "11")){

df\_train$Time\_Category[i] <- 'Morning'

}

if (("00" <= format(df\_train$trans\_date\_trans\_time[i],"%H")) && (format(df\_train$trans\_date\_trans\_time[i],"%H") <= "04")){

df\_train$Time\_Category[i] <- 'Midnight'

}

}

View(df\_train)

df\_train <- read\_csv("fraudtrain.csv")

# One-hot Encoding for categogy

dmy <- dummyVars("~category+Time\_Category",data = df\_train)

df\_dmy <- data.frame(predict(dmy,newdata = df\_train))

df\_train1 <- cbind(df\_train,df\_dmy)

View(df\_train1)

df\_train2 <- df\_train1

#Transform Gender into Numerical Value

for (i in 1:nrow(df\_train2)) {

if (df\_train2$gender[i] == "F"){

df\_train2$gender[i] <- 0

} else {

df\_train2$gender[i] <- 1

}

}

# Transforming Function

gender\_trans <- function(vector,i){

if (vector[i] == "F"){

vector[i] <- 0

} else {

vector[i] <- 1

}

}

for (i in 1:nrow(df\_train)) {

gender\_trans(df\_train$gender,i)

}

df\_train2 <- fraudtrainf

# Data filtering

df\_train2s <- df\_train2 %>% select(-trans\_date\_trans\_time,-cc\_num,-merchant,-category,-first,-last,-street,-city,-Time\_Category,-zip,-lat,-long,-job,-dob,-trans\_num,-unix\_time,-merch\_lat,-merch\_long)

df\_train2s <- df\_train2s %>% select(-1)

df\_train2s <- df\_train2s %>% select(-gender\_n)

df\_train2sl <- df\_train2s %>% select(amt,gender,city\_pop,Age,is\_fraud)

df\_train2ss <- df\_train2s %>% select(-20,-21,-22,-23)

View(df\_train2s)

# Data Partitioning:

set.seed(71)

train\_ind <- sample(seq\_len(nrow(df\_train2s)), size = 0.90\*nrow(df\_train2s))

train <- df\_train2s[train\_ind, ]

train1 <- train[train$is\_fraud ==1,]

train1 <- train1[sample(nrow(train1),1000, TRUE),]

train2 <- train[train$is\_fraud ==0,]

traindex <- sample(nrow(train2),1000,TRUE)

train\_unf <- train[traindex,]

train\_final <- rbind(train1,train\_unf)

test <- df\_train2s[-train\_ind, ]

View(train\_final)

# Logistic Regression

logistic.fits <- glm(is\_fraud~ .-gender-state-city\_pop, data = train\_final, family = binomial)

summary(logistic.fits)

logistic.probs <- predict(logistic.fits, test, type = "response")

glm.pred <- rep(0,129668)

glm.pred[logistic.probs > .5] <- 1

table(glm.pred,test$is\_fraud)

accuracy\_of\_LR <- mean(glm.pred == test$is\_fraud)

accuracy\_of\_LR

df\_time <- df\_time %>%

arrange(desc(Time\_Category)) %>%

mutate(prop = mean / sum(mean) \*100) %>%

mutate(ypos = cumsum(prop)- 0.5\*prop )

ggplot(df\_time, aes(x="", y=mean, fill=Time\_Category)) +

geom\_bar(stat="identity", width=1, color="white") +

coord\_polar("y", start=0) +

theme\_void()

# Testifying State

sample1 <- df\_train2s[df\_train2s$amt %in% sample(df\_train2s$amt,400),]

high\_income <- c('DC','MA','NY','AK','ND','CA')

relatively\_high\_income <- c('CT','WA','WY','DE','NJ','MD','IL','TX','CO','MN')

mid\_income <- c('NE','HI','NH','VA','PA','Iowa','KS','SD','OR','OH','WI','RI','LA','UT','OK','GA','NV')

relatively\_low\_income <- c('IN','VT','NC','TN','MI','MO','NM','FL','AZ','MT','ME','KY','SC','AL','ID','WV')

low\_income <- c('AR','MS')

for (i in 1:nrow(sample1)){

if (sample1$state[i] %in% high\_income) {

sample1$State[i] <- 4

} else if(sample1$state[i] %in% relatively\_high\_income){

sample1$State[i] <- 3

} else if (sample1$state[i] %in% mid\_income) {

sample1$State[i] <- 2

} else if (sample1$state[i] %in% relatively\_low\_income){

sample1$State[i] <- 1

} else if (sample1$state[i] %in% low\_income) {

sample1$State[i] <- 0

}

}

View(sample1)

sample.fits <- glm(is\_fraud~ .-state, data = sample1, family = binomial)

summary(sample.fits)