Credit Card Transactions Fraud Detection Using R

Group 10

Final Project Report

Duc Hoang

Khushi Shah

Naman Atul Shah Rohit

Kumar

Tanay Parag Shah

GUIDED BY
PROFESSOR ZHE ZHANG
MIS 6356.006 - BUSINESS ANALYTICS WITH R
Fall 2022

THE UNIVERSITY
OF TEXAS AT DALLAS

December 8, 2022

Table of Contents

Executive Summary 3

Project Motivation.....4

Data Description.....5

Data Exploration/Analysis and Interpretation.....6

Business Intelligence Models 9

Conclusions.....11

Project results.....11

Project challenges 11

References.....11

Executive Summary

In this report, we have explored a credit-card transactions dataset to derive various BIbased conclusions and findings. This dataset from Kaggle had enough data for us to create histograms, perform linear regression and create a decision tree to derive some findings like category of merchants was more prone to frauds or how expensive purchases lead to more frauds. We were also able to put together a confusion matrix for the decision tree model. Lastly, we have listed down the conclusions, which includes project results as well as challenges we faced while curating the project.

Project Motivation

Nowadays, online payment gaining popularity because of easy and convenience use of ecommerce. It became very easy mode of payment. People choose online payment and eshopping; because of time convenience, transport convenience, etc. As the result of huge amount of e-commerce use, there is a vast increment in credit card fraud also. Fraud detection in credit card is a big problem, it becomes challenging due to two major reasons–first, the profiles of normal and fraudulent behaviors change frequently and secondly due to reason that credit card fraud data sets are highly skewed. Hence, we decided to find some conclusions on this very domain.

Data Description

For the group project, we are using second-hand data from Kaggle. The dataset is “Credit Card Transactions Fraud Detection Dataset” (https://www.kaggle.com/datasets/kartik2112/fraudetection?select=fraudTrain.csv), which is a simulated dataset containing legitimate and fraudulent credit card transactions from 2019 to 2020. The data contains transactions of 1000 customers and 800 merchants. Each transaction includes customer details, the merchant and category of purchase, and whether the transaction was fraudulent.

The dataset contains 23 columns. Some of the properties include transaction date, customer name, merchant name, amount of transaction, location of purchase, etc. By analyzing the data, we found the types of purchases that are most likely to be instances of fraud and whether older customers are more likely to be victims of credit card fraud.

Data Exploration/Analysis and Interpretation

After downloading and importing the CSV file into RStudio, we installed a few packages to assist the visualization. Upon a quick check, no missing value was found. We then formatted the data to only keep the important columns. The next thing we did was to analyze the relations among merchant category, transaction amount, and fraud count. R code can be accessed on GitHub.

We will create a few plots to have a general idea about the data that we are working with.

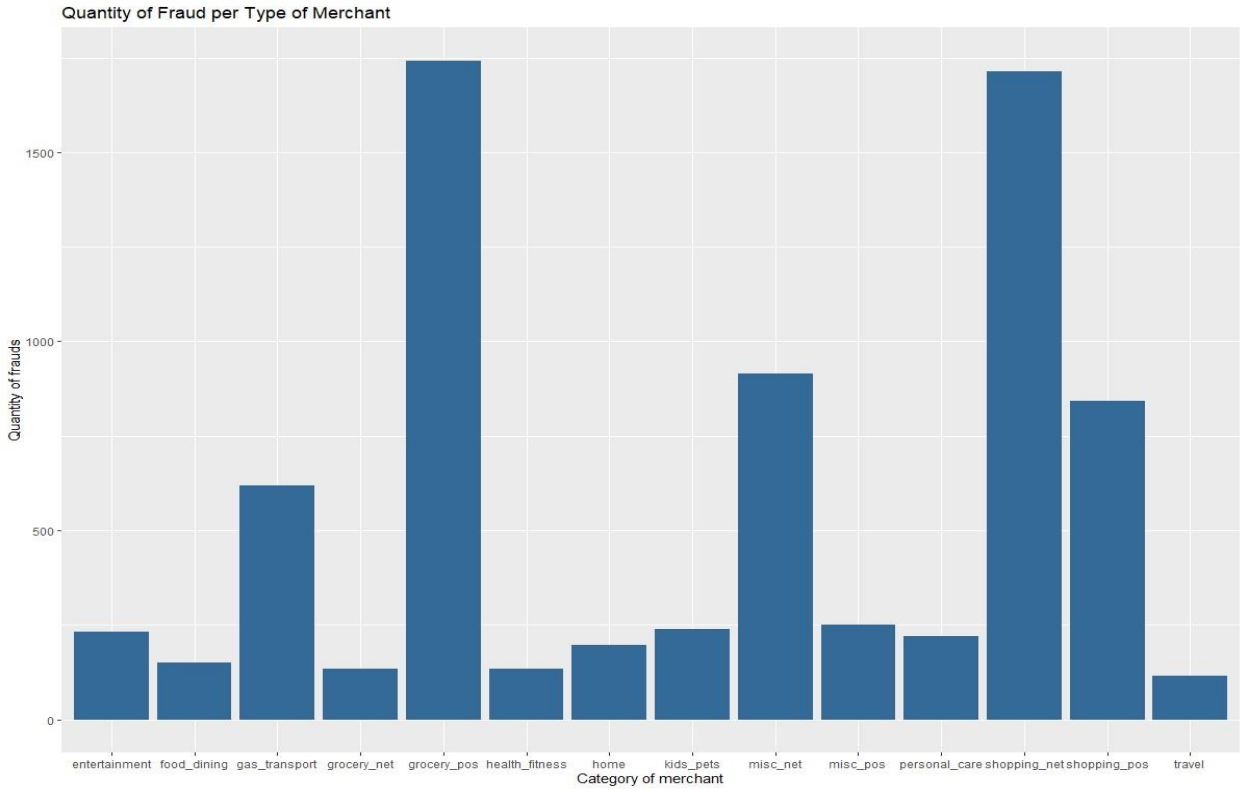

Figure 1: Plot of category of merchants and Quantity of frauds

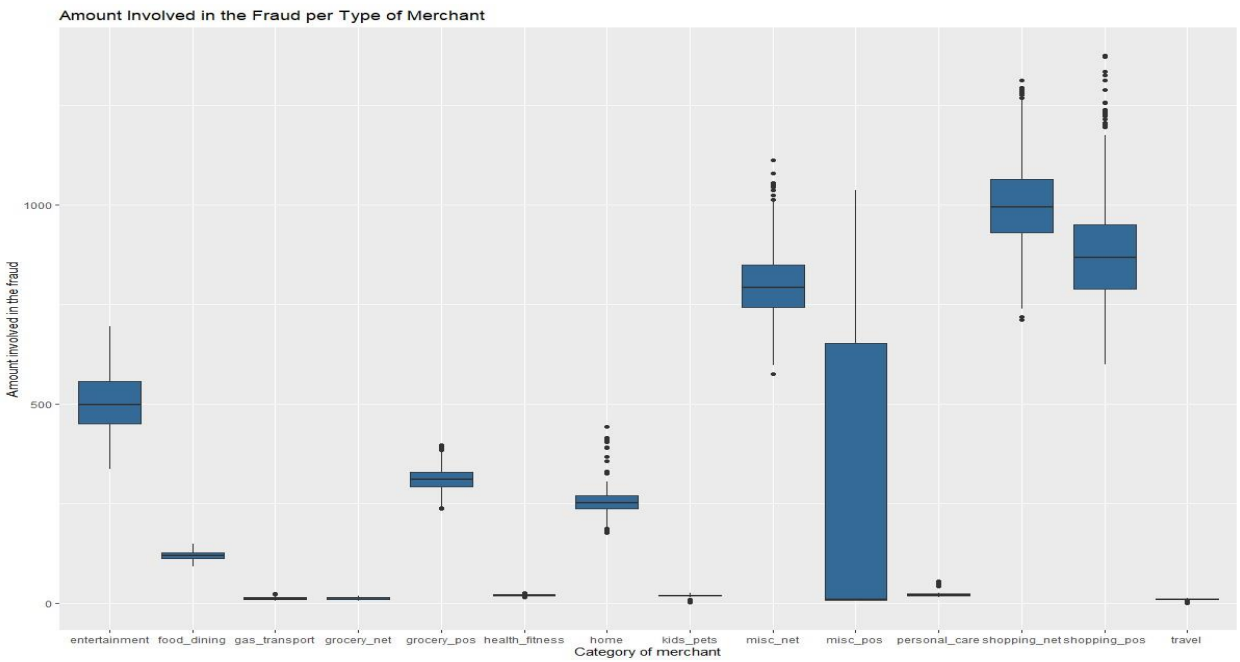

Figure 2: Plot of category of merchant and amount involved in the fraud

Based on figure 1, shopping_net and grocery_pos are the category of merchant are more likely to instances of fraud, however, figure 2 shows there were more instances of frauds with purchases more expensive in shopping_net and shopping_pos.

Next, we will analyze whether older customers are significantly more likely to be victims of credit card fraud. This is done with a bar-chart and a scatterplot.

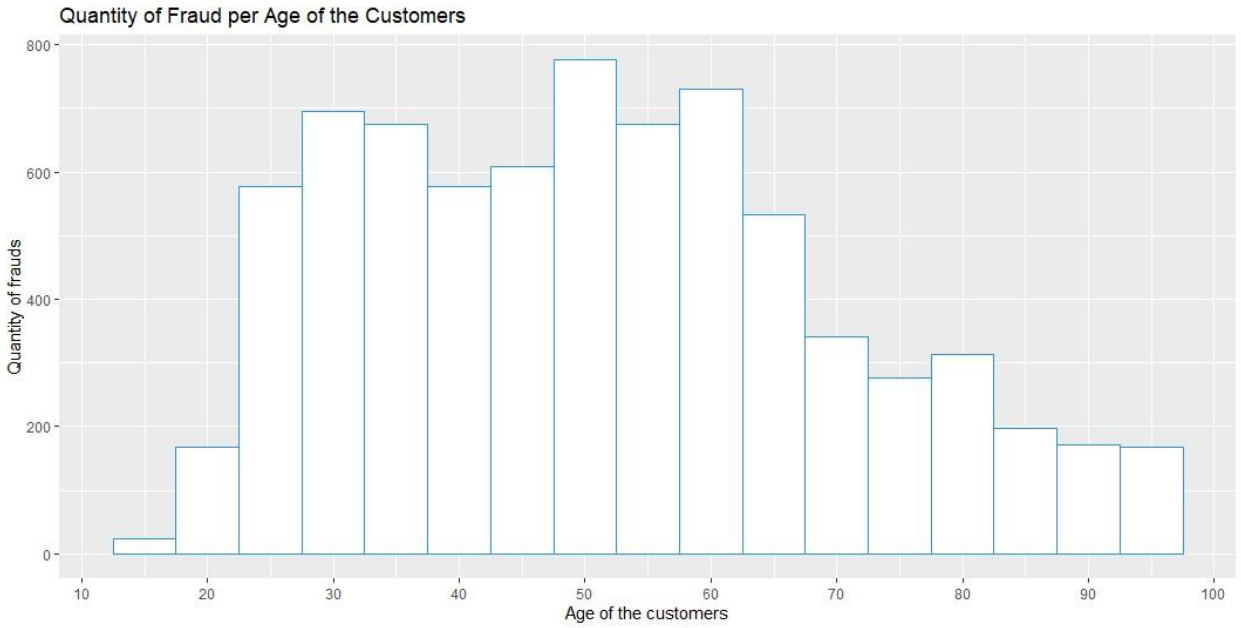

Figure 3: Plot of Age of customers vs Quantity of frauds

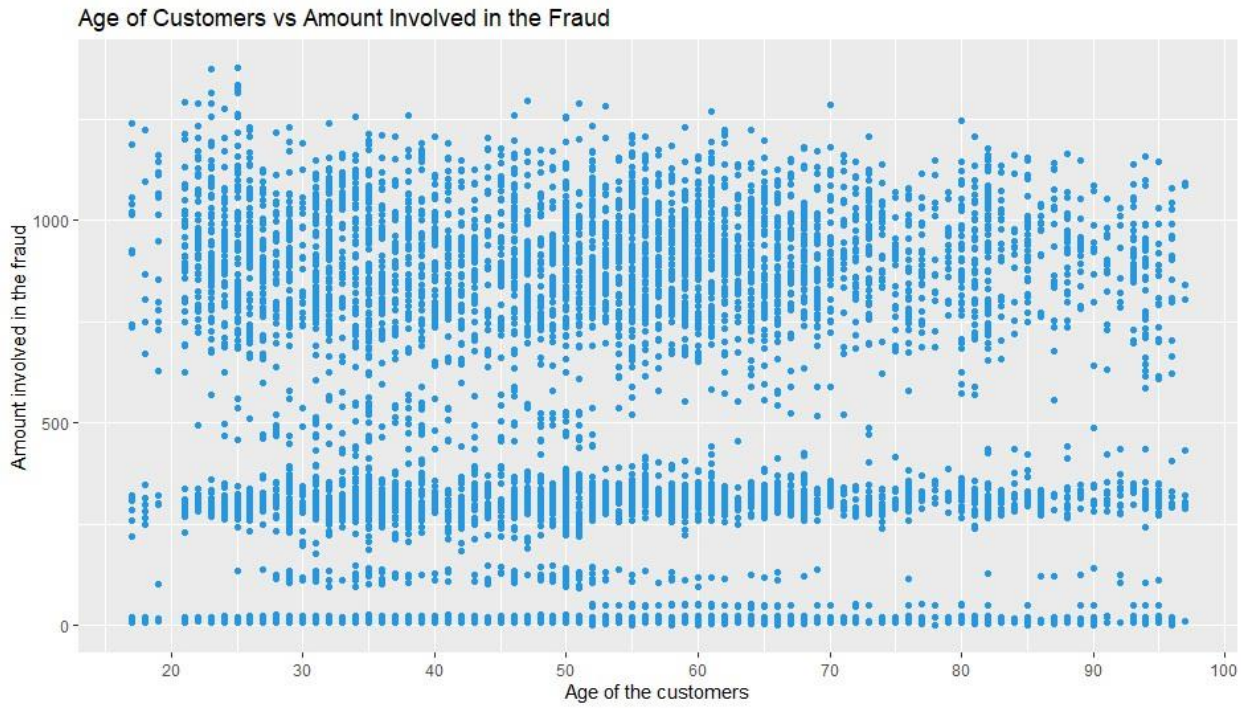

Figure 4: Plot of Age of customers vs Amount involved in the fraud

From figure 3 and 4, we can conclude that there is no evidence that older customers are more likely to be victims of credit card fraud. The histogram shows people are between 45 and 60 years old are more likely to be victims, moreover, the amount of money involving older people seems to be the same than youngers.

Machine Learning Model

Logistic Regression

Predictive analytics and categorization frequently make use of this kind of statistical model, also referred to as a logit model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. In logistic regression, the odds—that is, the probability of success divided by the probability of failure—are transformed using the logit formula. The following formulas are used to represent this logistic function, which is also referred to as the log odds or the natural logarithm of odds:

$$\text{Logit}(pi) = 1/(1+ \exp(-pi))$$
$$\ln(pi/(1-pi)) = \text{Beta}_0 + \text{Beta}_1*X_1 + \ldots + B_k*K_k$$

As we need to specify the cut of value for the prediction of the prediction to convert it into the dummy variables like 0,1. It might not be the best approach as based on the dataset the cutoff value to increase the predict the values for the new data might change. Thus, we would try to use Decision tree algorithm on that.

Decision Tree:

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

For the last part of our project, we are using both **Decision tree and Logistic regression method** to predict whether a transaction is fraudulent or non-fraudulent. Since both the provided training and testing dataset are too big, we must clear unnecessary columns as well as limiting the number of imported data (the whole dataset for training and 3,000 rows for testing set).

Decision tree plot:

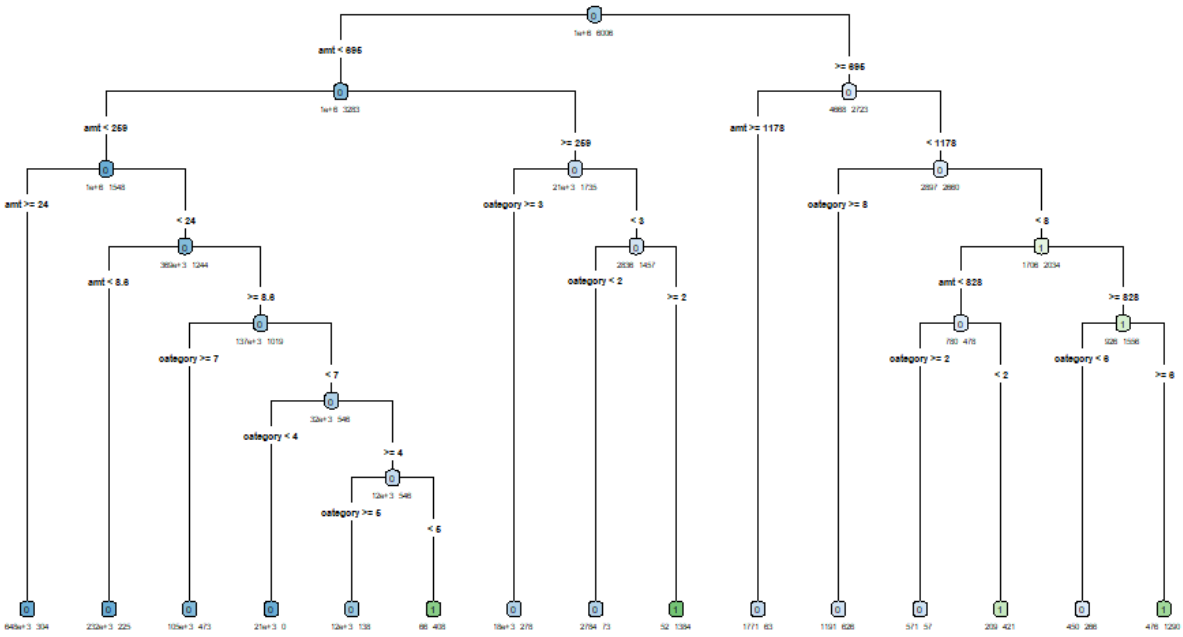

Figure 5: Plot of decision tree

	Reference	
Prediction	0	1
0	2976	15
1	6	3

Figure 6: Confusion matrix for decision tree model

Conclusions

Project results

- Shopping_net and grocery_pos are the category of merchant are more likely to instances of fraud, however, the are more instances of frauds with purchases more expensive in shopping_net and shopping_pos.
- There is no evidence that older customers are more likely to be victims of credit card fraud.
- The decision tree model turns out to be better, achieving 99.3% accuracy rate. On the other hand, the logistic regression model classified too many transactions (2,730) as fraudulent. This may have been caused by the limited amount of training data.

Project challenges

- Limited training data due to size, leading to possible too high/too low accuracy rate.
- Deciding which variables worth keeping and how to format those variables to be able to use decision tree and logistic regression in R. If we have to redo the ML model, we would select the trans_date_trans_time, category, amt, state, city_pop variables - How to create a confusion matrix for logistic regression model.

References

1. Kaggle training and testing data set https://www.kaggle.com/datasets/kartik2112/fraud-detection?select=fraudTrain.csv
2. R code https://github.com/shahnamana/Business-Analytics-in-R-Assignment/tree/main/Project 3. http://www.ijirset.com/upload/2020/september/104_Tejas_NC.PDF