ANA 515 Assignment 4

Gyu Hyun Byun

2022-10-15

# 1. Discuss the business problem/goal

The goal is to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

# 2. Identify where the dataset was retrieved from

The dataset was retrieved from <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?resource=download>

# 3. Identify the code that imported and saved your dataset in R

The dataset is saved as “creditcard\_raw”.

setwd("/Users/Qram/OneDrive/Documents/McDaniel/ANA 515 Fundamentals of Data Storage/Week 8")  
creditcard\_raw <- read.csv("creditcard.csv")

# 4. Describe your data set

The original data has 284807 rows and 31 columns. The names of the columns and a brief description of each are:

names(creditcard\_raw) #to show column names

## [1] "Time" "V1" "V2" "V3" "V4" "V5" "V6" "V7"   
## [9] "V8" "V9" "V10" "V11" "V12" "V13" "V14" "V15"   
## [17] "V16" "V17" "V18" "V19" "V20" "V21" "V22" "V23"   
## [25] "V24" "V25" "V26" "V27" "V28" "Amount" "Class"

#Time = Number of seconds elapsed between this transaction and the first transaction in the dataset  
#V1 ~ V28 = may be result of a PCA Dimensionality reduction to protect user identities and sensitive features(V1-V28)  
#Amount = Transaction amount  
#Class = 1 for fraudulent transactions, 0 otherwise

# 5-8. Discuss any data preparation, missing values, errors, the modeling, and the output. Provide explanation with any visuals.

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.173% of all transactions.

sum(creditcard\_raw$Class == "1") #to show the number of fraudulent transactions

## [1] 492

nrow(creditcard\_raw) #to show the number of total transactions

## [1] 284807

percent(sum(creditcard\_raw$Class == "1")/nrow(creditcard\_raw), accuracy = 0.001) #to show the % of fraudulent transactions

## [1] "0.173%"

creditcard\_fraud <- creditcard\_raw[creditcard\_raw$Class == "1",] #to create a subdataset for fraudulent transactions  
creditcard\_nofraud <- creditcard\_raw[creditcard\_raw$Class == "0",] #to create a subdataset for non-fraudulent transactions  
  
summary(creditcard\_fraud$Amount) #to show the summary of the Amount variable of the fraudulent subdataset

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 1.00 9.25 122.21 105.89 2125.87

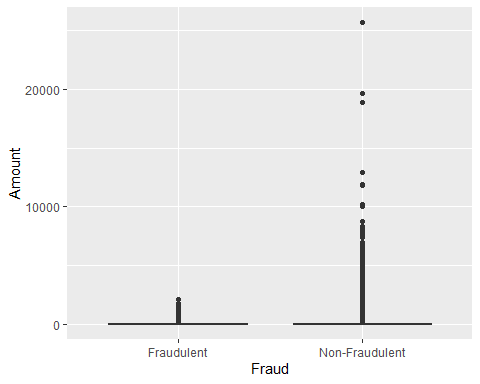
summary(creditcard\_nofraud$Amount) #to show the summary of the Amount variable of the non-fraudulent subdataset

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 5.65 22.00 88.29 77.05 25691.16

Per the summaries above, we cannot seem to identify a notable difference in the Amount variable between the two subdataset.

Let’s investigate further to determine if the Amount variable can help us tell fraudulent transactions from normal transactions.

creditcard\_raw$Fraud[creditcard\_raw$Class == "1"] <- "Fraudulent" #to create a new variable, Fraud, and mark it "Fraudulent" if the transaction is classified as fraudulent per the Class variable.  
creditcard\_raw$Fraud[creditcard\_raw$Class == "0"] <- "Non-Fraudulent" #to create a new variable, Fraud, and mark it "Non-Fraudulent" if the transaction is not classified as fraudulent per the Class variable.  
  
ggplot(creditcard\_raw, aes(x = Fraud, y = Amount)) + geom\_boxplot() #to draw a graph to compare the transaction amounts between the two groups.



t.test(Amount ~ Fraud, data = creditcard\_raw, var.equal = TRUE, paired = FALSE) #to perform a t-test to draw a conclusion.

##   
## Two Sample t-test  
##   
## data: Amount by Fraud  
## t = 3.0056, df = 284805, p-value = 0.002651  
## alternative hypothesis: true difference in means between group Fraudulent and group Non-Fraudulent is not equal to 0  
## 95 percent confidence interval:  
## 11.80029 56.04031  
## sample estimates:  
## mean in group Fraudulent mean in group Non-Fraudulent   
## 122.21132 88.29102

Based on the boxplot and t-test performed above, the mean of the variable Amount is a good indicator of identifying fraudulent transactions as the p-value is only 0.002651.