Student Name: Henry J. Hu
Course: DSC-611-Z1 Data Visualization
Final Project

Introduction

According to the American Bankers Association's Deposit Account Fraud Survey, published in January of 2020, fraud against bank deposit accounts amounted to $25.1 billion in the year 2018. The prevalence of financial fraud is not only monetary damaging the banking industry but also it is destroying the reputation of the industry as a whole. As banking businesses implemented new technologies to combat financial fraudulent activities, the criminals became more sophisticated in their techniques of carrying out these criminal activities. Therefore, as data scientists, we must keep researching new ways of detecting these types of financial fraudulent activities in banking data.

Research Question

The purpose of this class research project is to answer the question "What is the best data science model for detecting fraudulent activities in banking data?".

Data Source

The data used in this research is a synthetic mobile banking data set from Kaggle. Due to the highly sensitive nature of banking data and the legal issues surrounding it, no publicly available real banking data set with fraud indicators could be found. It has been known that a none disclosure statement has to be signed by many parties to obtain real banking data set. Not unless this is a paid research, then it is sensible to invest money into the legal protection of obtaining the real data set and the policy of its usage.

Description

Synthetic financial data set from Kaggle to be used for financial fraud data analysis.

Method of Data Collection

Real financial logs data of more than 14 countries from a mobile financial service system implemented in an African country was fed into a simulator named PaySim to produce this synthetic data, which also removed all traces of the original data source. Therefore, the original data source is not traceable.

Data Temporal

30 days.

Data Size

1,048,575 observations and 11 variables.

Data Geographic Information

Synthetically worldwide.

Data Sampling

Due to limitations with the computer used for training the individual machine learners, only 200,000 out of 1,048,575 data records were randomly selected for this research.

Dependent Variables

Field Name	Data Type	Field Description	Unit
isFraud	Binary	This is a flag that indicates whether the transaction is fraudulent or not.	A binary number of either 0 or 1.

Independent Variables

Field Name	Data Type	Field Description	Unit
type	Nominal	CASH-IN, CASH-OUT, DEBIT, PAYMENT, and TRANSFER.	Text
amount	Ratio	Amount of the transaction in local currency.	Amount in US Dollars
nameOrig	Nominal	Customer who started the transaction	Text
oldbalanceOrg	Ratio	Initial balance before the transaction	Amount in US Dollars
newbalanceOrig	Ratio	The new balance after the transaction	Amount in US Dollars
nameDest	Nominal	Customer who is the recipient of the transaction	Text
oldbalanceDest	Ratio	Initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).	Amount in US Dollars
newbalanceDest	Ratio	New balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).	Amount in US Dollars

Analytical approach

The exploratory data analysis of this data set confirmed the existence of both confounding and severe multicollinearity in the data. Also, the majority of outliers are associated with activities identified as fraudulent in the actual data. Therefore, we have decided that it is better to fit this data with nonlinear classification models. It is because confounding, multicollinearity, and outliers do not affect the performance of nonlinear classification models. These classification models include Decision Tree, Random Forest, and Artificial Neural Networks. Decision Tree algorithm was used for blending the Random Forest algorithm together with the Artificial Neural Networks algorithm to form the Ensemble Learner. Also, due to technological limitations, a small sample of the data was used to both train and test these models instead. Since we wanted to ensure that the sample taken repented the large population, we used random sampling. This random sampling technique is depicted in Figure 14 below. Last, since for classification algorithms to yield the correct and accurate results, both the training and testing data must be normalized on the same scale. This data normalization technique is depicted in Figure 15 below.

Findings

If compared with the Standalone Random Forest model, the resulting Ensemble Learner was identified to be the most robust and parsimonious classification model. It was able to classify the transaction activities with a Kappa statistic of 0.76 and a sensitivity percentage of 78.83%. Therefore, the result has answered the research question. But further research is necessary to determine if it is the best model in today’s industry.

Narrative framework

Ensemble Learner Is Better at Detecting Frauds Than Standalone Random Forest.

Target audience

Financial institutions which are in need of implementing new technology for detecting fraudulent activities within their customer transactional data.

Visualization Theories and Techniques

The visual techniques we used were ISO-Measure and humor, Frame Narrative, color contrasting, enhanced plot labeling, plot annotation, plot overlaying, and interactive plotting and charting. The reason we used ISO-Measure and humor is that financial fraudulent activities are a serious matter which all financial institutions want to minimize. We used humor to break the ice with the executives and make them feel at ease that there is a solution to their problem. We used ISO-Measure to relate to these executives on a personal level. That way they could quickly and correctly grasp the exact message which we were trying to convey. The before and after of these visual improvements are depicted between Figure 1 and Figure 4 below. As one could see that there was a drastic improvement in the visuals between the before and after visual enhancements. Also, we used Frame Narrative to convince the executives that the Ensemble Learner was better than the Standalone Random Forest classifier. This technique is depicted in the visual of Figure 5 below. We used color contrasting in all of our visual enhancements to highlights the important aspects of the visuals and to ease the audience's cognitive effort. The best example of this is depicted in the pie chart of Figure 12 below. We used enhanced plot labeling in all of our visual enhancements to both ease the audience's cognitive effort and capture their attention. All of our enhanced plots and charts had clearly labeled axes with axis lines, tick marks, grid lines, and titles. The best example is depicted in Figure 13 below. We used plot annotation to enhance the audience's cognitive ability to quickly grasp the message we were trying to convey. The annotation of lines and bars on the plot made it easier for the audience to quickly spot the magnitude or degree of each line or bar on the plot. The best example of this is depicted in the overlayed plot of Figure 13 below. We used plot overlaying to combine different views and perspectives into one single plot. For example, as depicted in Figure 13 below, we combined the visual of transaction count vs. transaction amount with the visual of fraud indicator to convey to the audience a narrative that both the positive and negative fraudulent activities were captured by the transactions with the amount of 10M. Last, we used Tableau's interactive plotting features to allow the audience to further explore the detail of the plot or chart. Features such as dynamic plot annotation, full screen, download, and share allowed the company executives to quickly explore and have a discussion about the details. This helped improved the audience's understanding of the concepts being presented.

Ethical considerations

Our only purpose for enhancing the visuals of these plots and charts was so that the audience could quickly, easily, and correctly grasp the concepts, stories, messages, and narratives that we were trying to convey. We worked very hard to ensure that the enhanced visuals will not cause the audience to have the wrong perception. Therefore, our intention for visual enhancement has always been completely ethical.

Biases

We don’t have any of the data, selection or story, narrative, optical, memory, cultural, or confirmation biases in our newly enhanced visuals. We avoid data bias by utilizing the technique of random sampling. We avoid selection bias by including all the possible stories and data for a particular narrative. We avoid narrative bias by making sure that there is a direct link between the facts found in our model results and the narrative. We avoid optical bias by test viewing our visuals from all angles to make sure that they convey the same message. We avoid memory bias by not repeating any sentence or phrase in the visual more than once. We avoid cultural bias by excluding anything that is culturally specific from the visuals. Last, we avoid confirmation bias by not basing our stories and narratives on popular ideas or concepts that have been around for years. Therefore, our visuals are bias-free.

Social responsibility

Our visuals do not specifically target any financial institution nor any specific individual. Also, our narrative does not convey that the predictions from these models are of absolute certainty. After all, nothing in data science is of absolute certainty. Therefore, we are completely socially responsible.

Hyperlinks of Visuals

R Code, Old Plots and Charts

https://aaacomply.com/data_science/DSC609/Henry_Hu_Moduel_7_Programming.htm

Figure 10: New Transaction Activity by Fraudulent Type Chart

https://public.tableau.com/views/DSC_611_Module_8_Final_Project/TransactionActivitybyFraudulentType?:language=en&:display_count=y&:origin=viz_share_link

Figure 11: New Fraudulent Activity by Transaction Type Chart

https://public.tableau.com/views/DSC_611_Module_8_Final_Project/FraudulentActivitybyTransactionType_1?:language=en&:display_count=y&:origin=viz_share_link

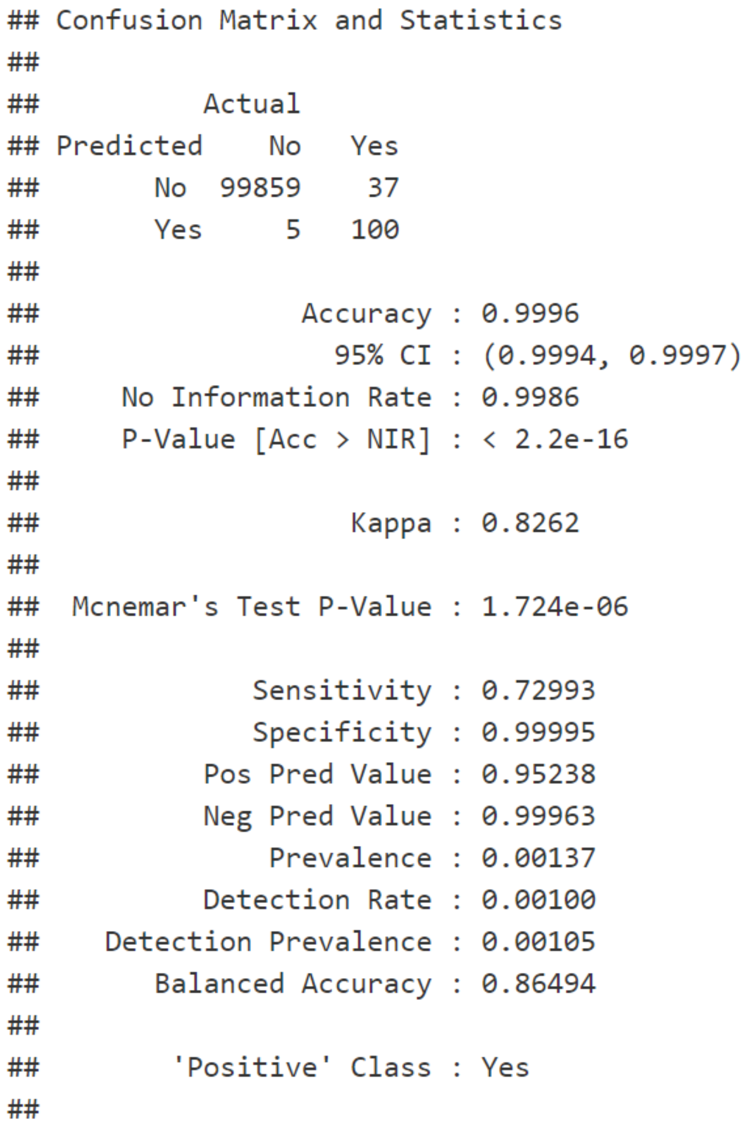
Figure 12: New None Fraudulent Activity by Transaction Type Chart

https://public.tableau.com/views/DSC_611_Module_8_Final_Project/NoneFraudulentActivitybyTransactionType?:language=en&:display_count=y&:origin=viz_share_link

Figure 13: Both Good and Bad Transactions at 10M

https://public.tableau.com/views/DSC_611_Module_8_Final_Project/BothGoodandBadTransactionsat10M?:language=en&:display_count=y&publish=yes&:origin=viz_share_link

Standalone Random Forest Accuracy Matrix



```
## Confusion Matrix and Statistics
##
##          Actual
## Predicted    No   Yes
##       No  99859    37
##       Yes     5   100
##
##                Accuracy : 0.9996
##                  95% CI : (0.9994, 0.9997)
##     No Information Rate : 0.9986
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8262
##
##  Mcnemar's Test P-Value : 1.724e-06
##
##             Sensitivity : 0.72993
##             Specificity : 0.99995
##          Pos Pred Value : 0.95238
##          Neg Pred Value : 0.99963
##              Prevalence : 0.00137
##          Detection Rate : 0.00100
##    Detection Prevalence : 0.00105
##       Balanced Accuracy : 0.86494
##
##        'Positive' Class : Yes
##
```

Figure 1: Standalone Random Forest Accuracy Matrix

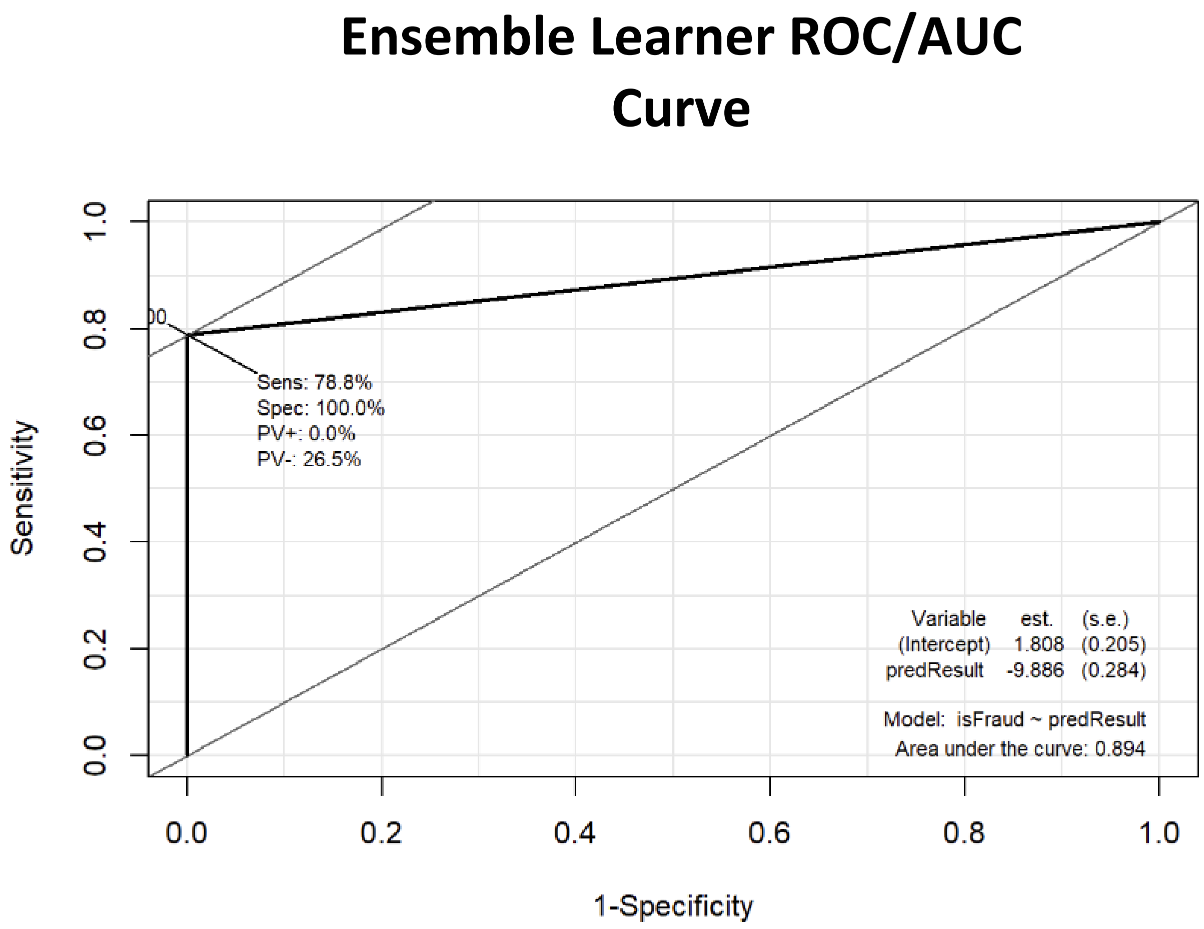


Figure 2: Old Ensemble Learner ROC/AUC Curve

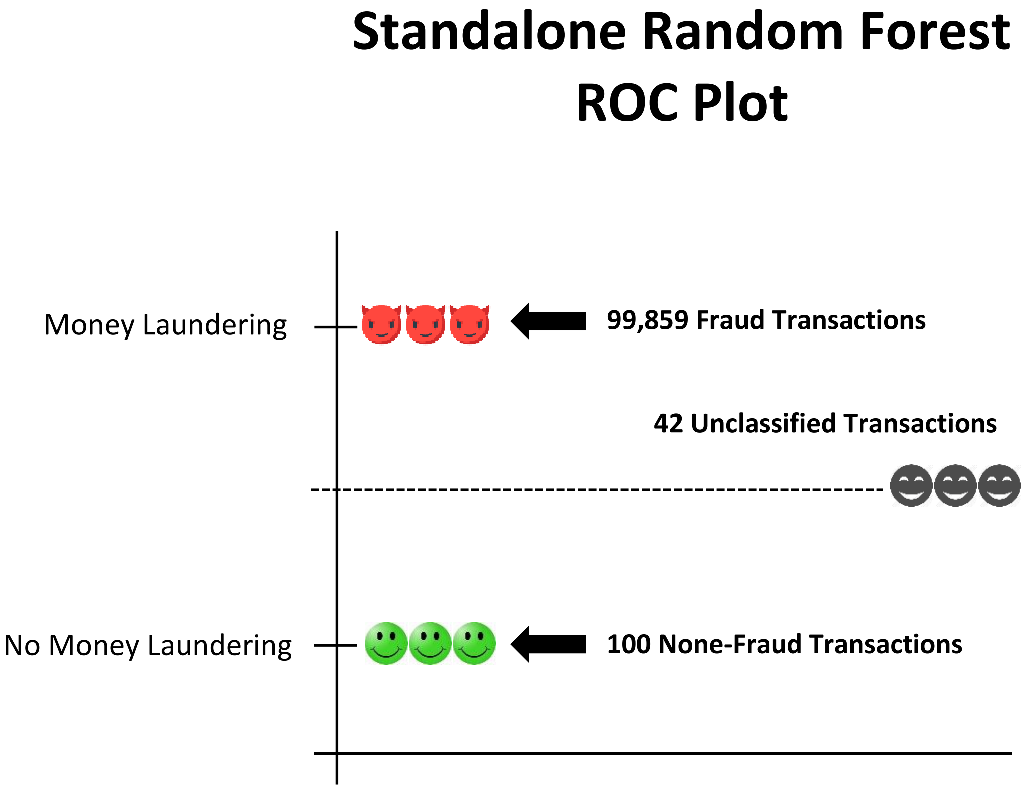

Standalone Random Forest
ROC Plot
Money Laundering
99,859 Fraud Transactions
42 Unclassified Transactions
No Money Laundering
100 None-Fraud Transactions

Figure 3: Standalone Random Forest ROC Plot

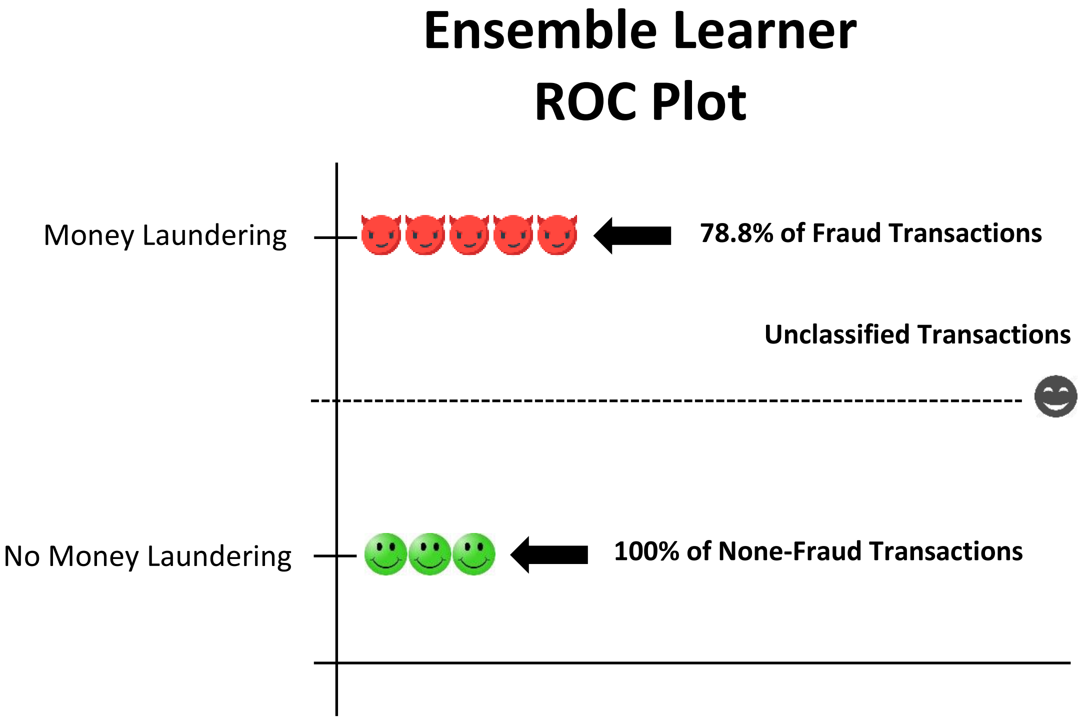
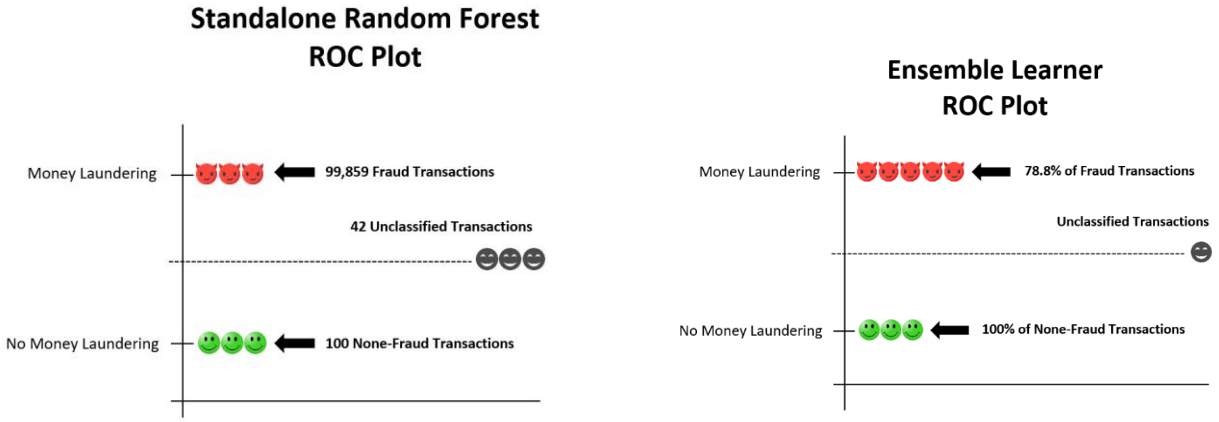

Ensemble Learner
ROC Plot
Money Laundering
78.8% of Fraud Transactions
Unclassified Transactions
No Money Laundering
100% of None-Fraud Transactions

Figure 4: New Ensemble Learner ROC Plot

Ensemble Learner is Superior



Standalone Random Forest ROC Plot

Money Laundering
99,859 Fraud Transactions
42 Unclassified Transactions
No Money Laundering
100 None-Fraud Transactions

Ensemble Learner ROC Plot

Money Laundering
78.8% of Fraud Transactions
Unclassified Transactions
No Money Laundering
100% of None-Fraud Transactions

Figure 5: Frame Narrative

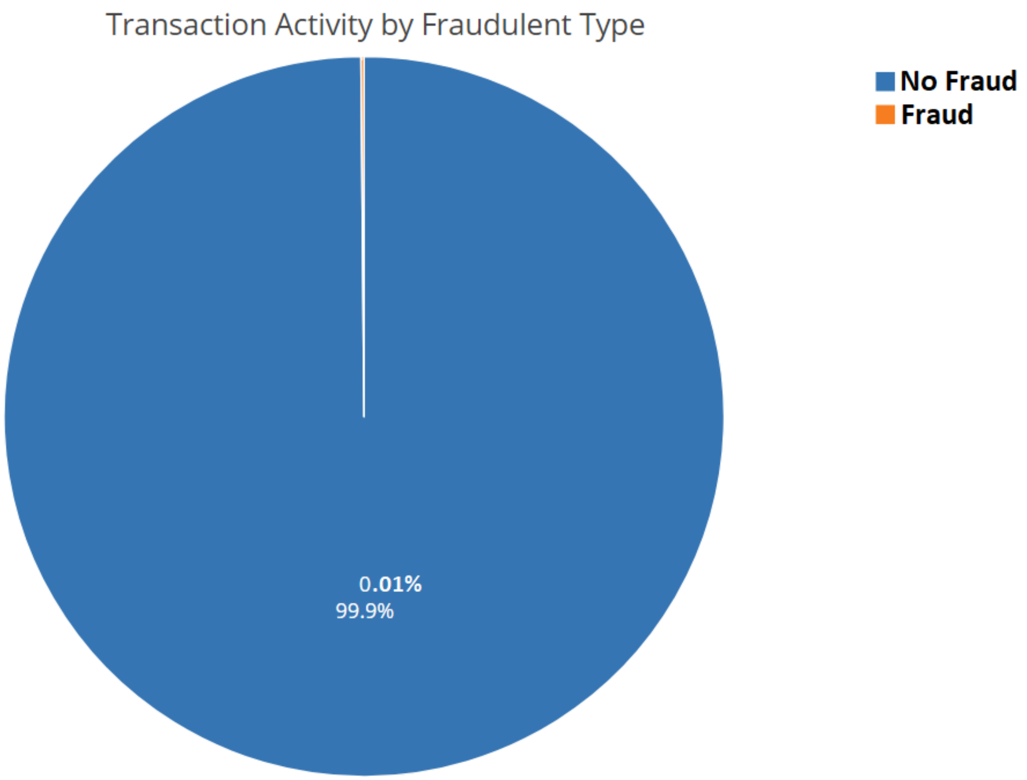


Figure 6: Old Transaction Activity by Fraudulent Type Chart

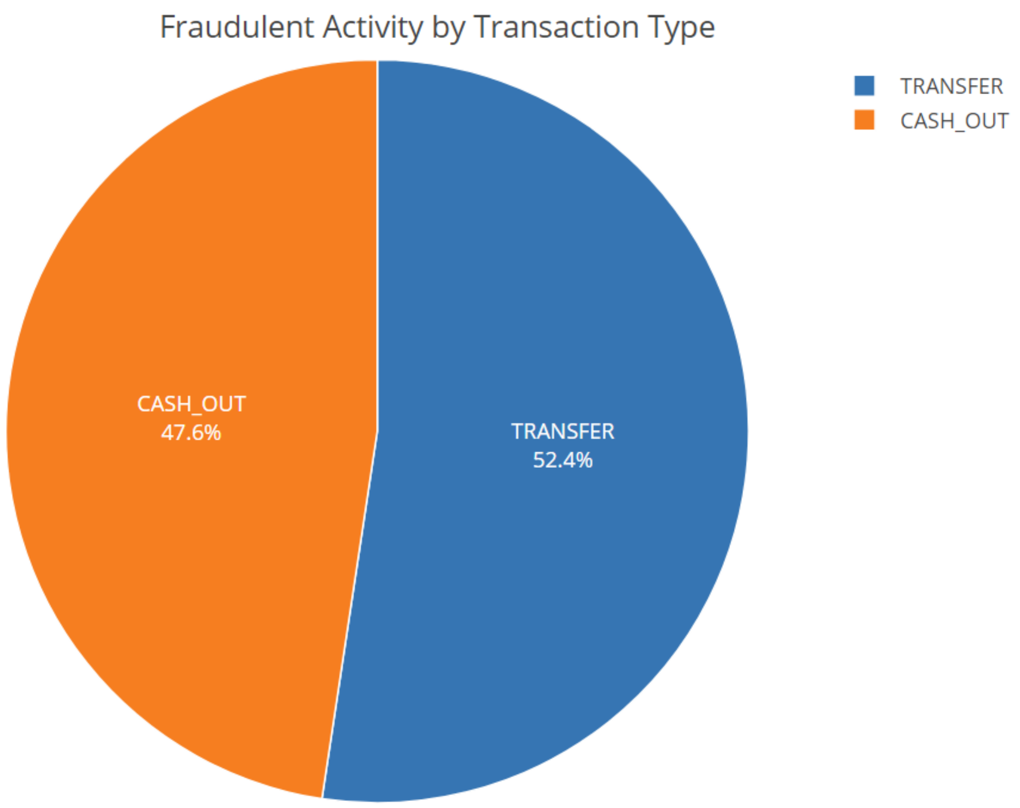


Figure 7: Old Fraudulent Activity by Transaction Type Chart

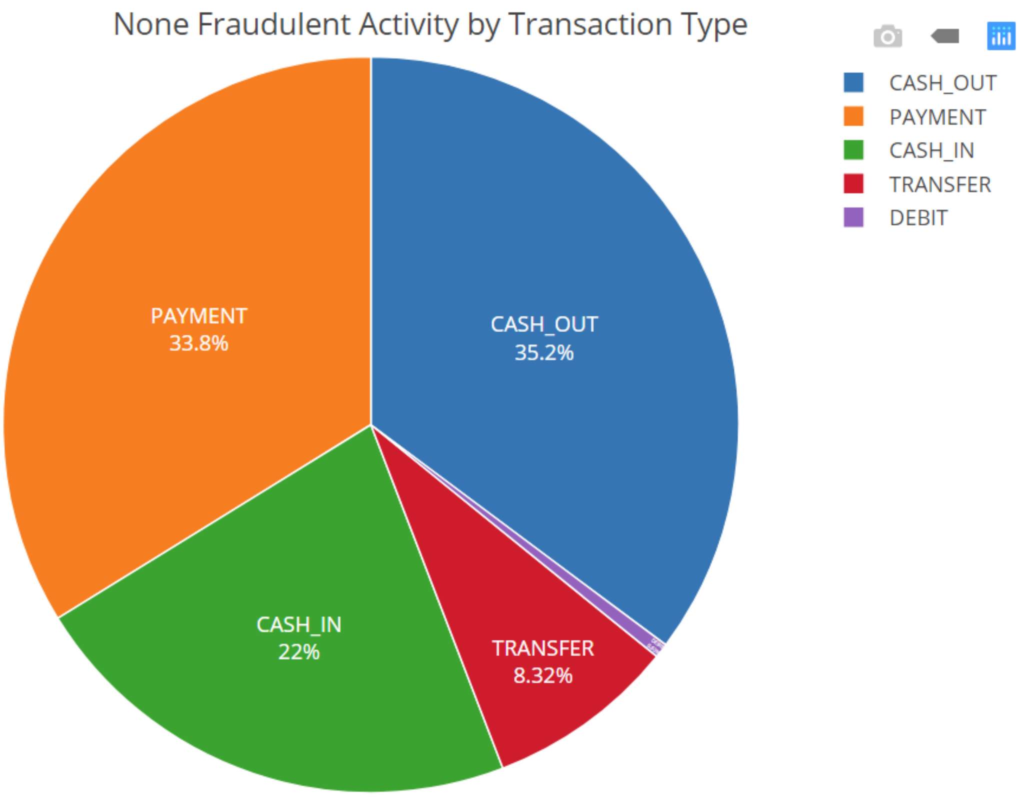


Figure 8: Old None Fraudulent Activity by Transaction Type Chart

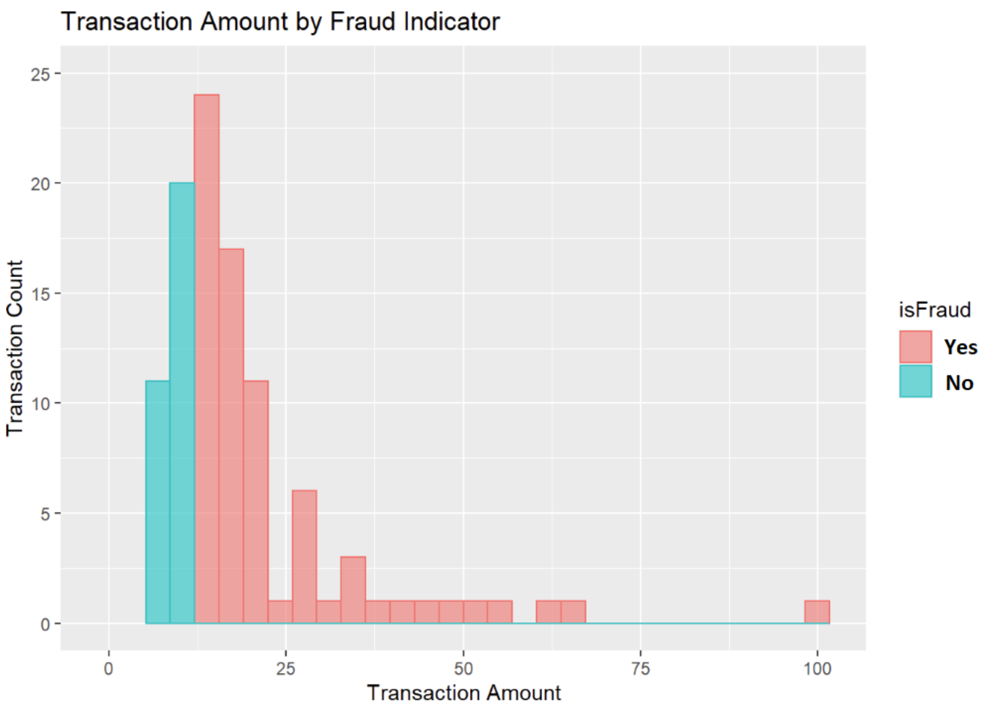


Figure 9: Old Transaction Amount by Fraud Indicator Plot

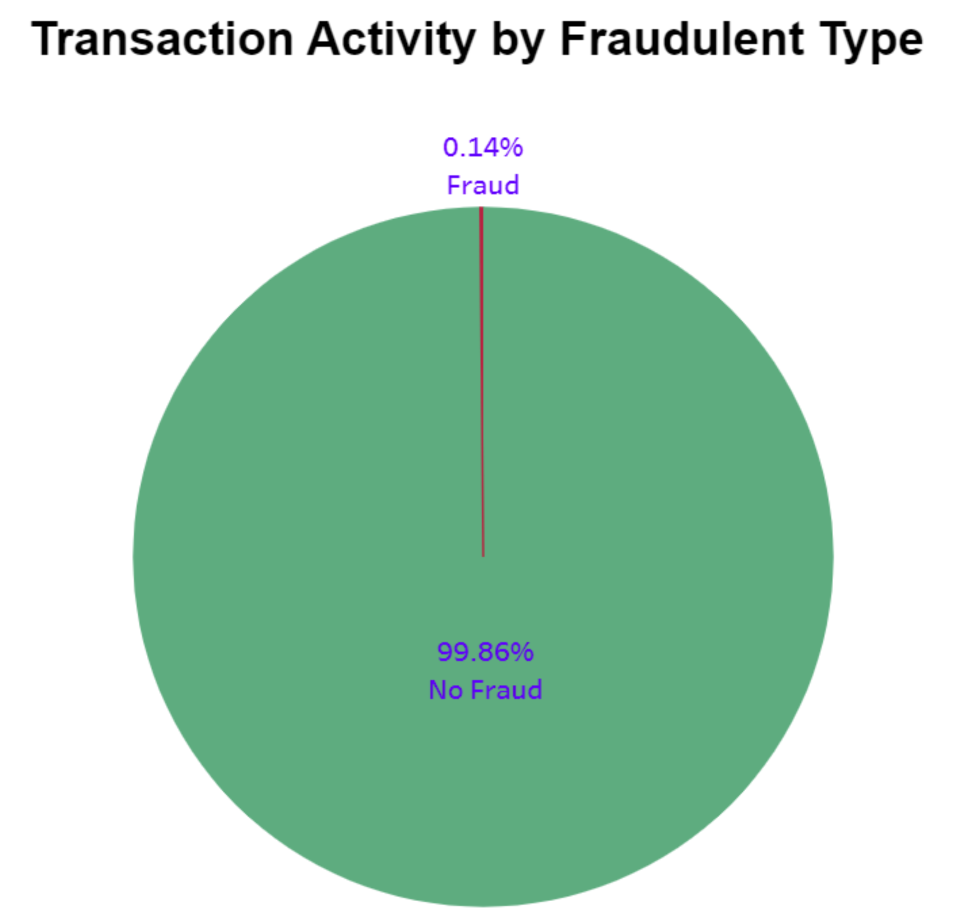


Figure 10: New Transaction Activity by Fraudulent Type Chart

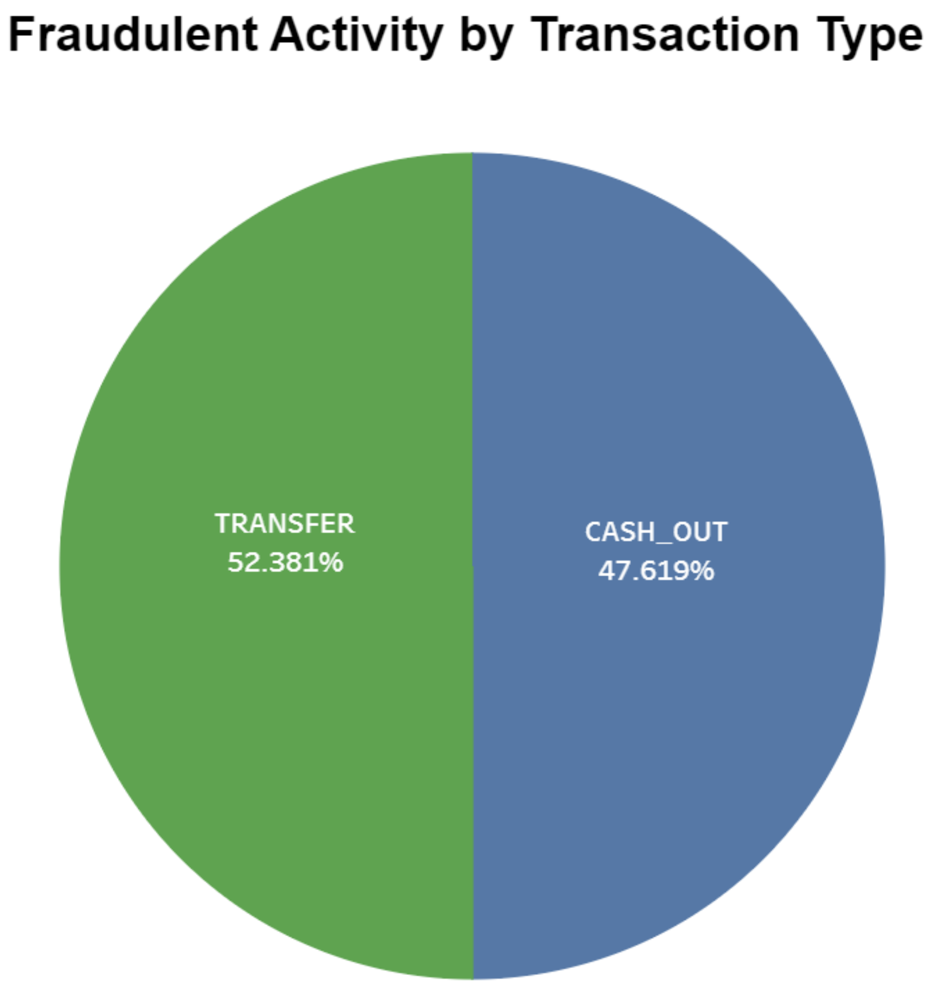


Figure 11: New Fraudulent Activity by Transaction Type Chart

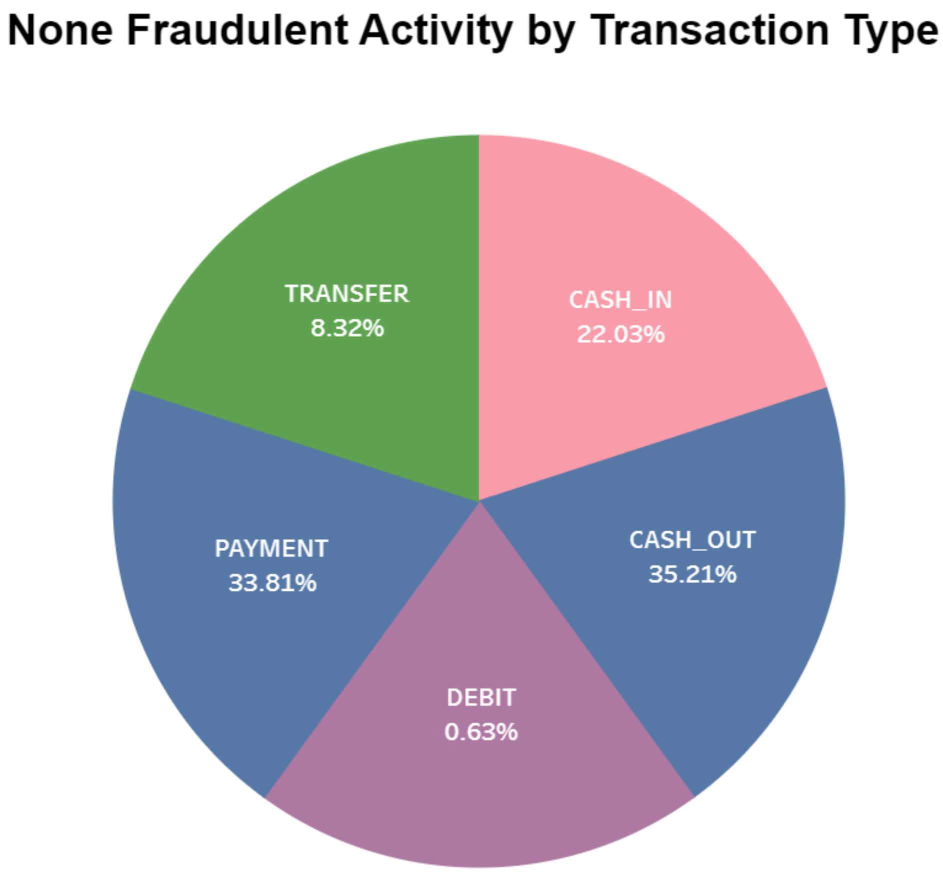


Figure 12: New None Fraudulent Activity by Transaction Type Chart

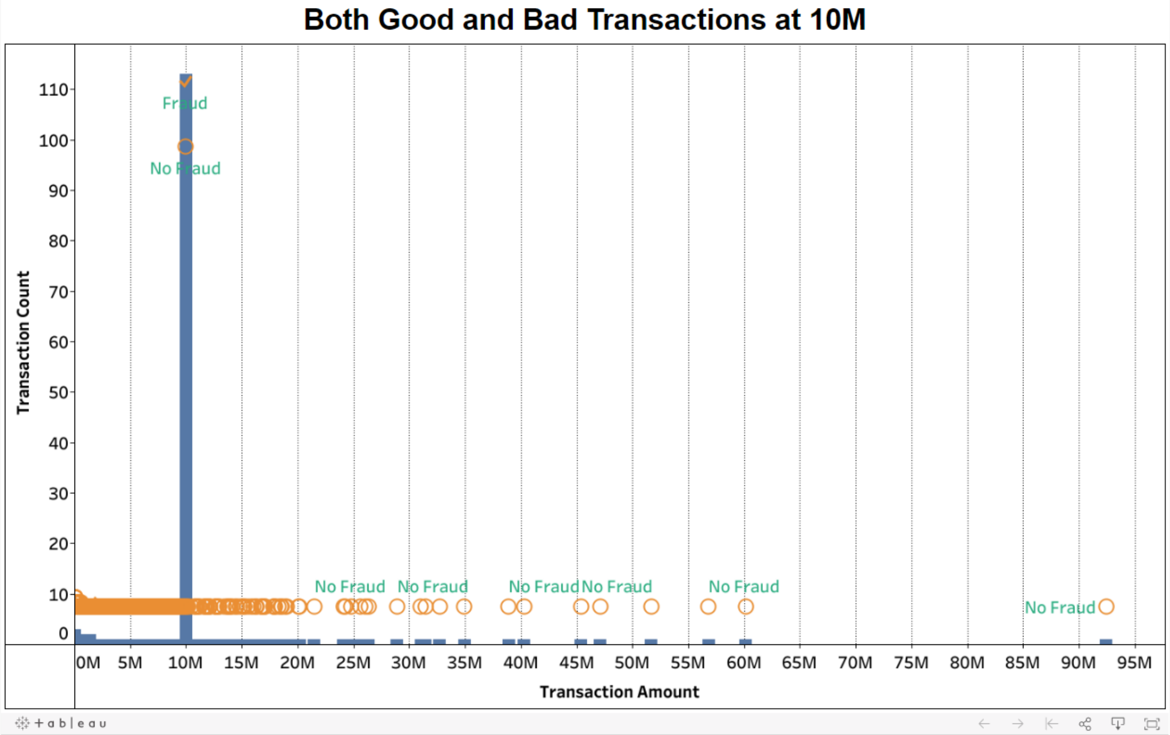
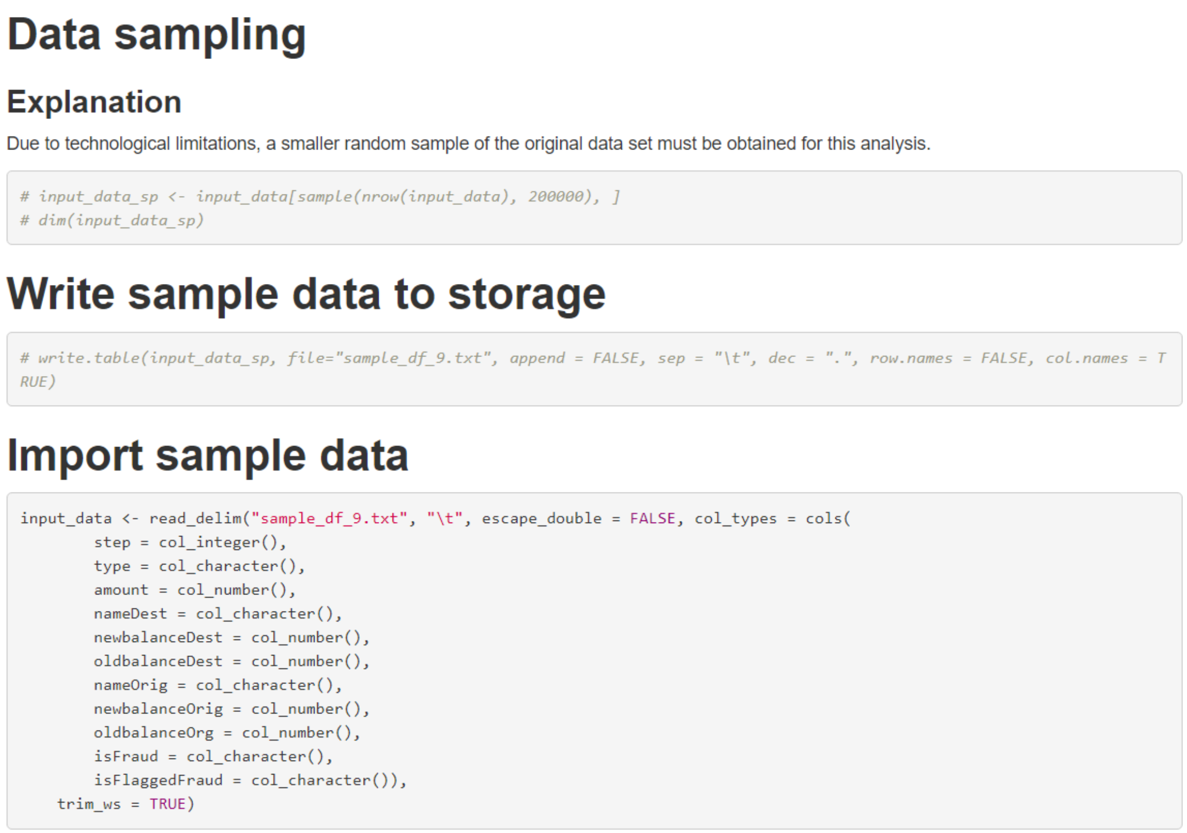


Figure 13: Both Good and Bad Transactions at 10M



Data sampling

Explanation

Due to technological limitations, a smaller random sample of the original data set must be obtained for this analysis.

```
# input_data_sp <- input_data[sample(nrow(input_data), 200000), ]
# dim(input_data_sp)
```

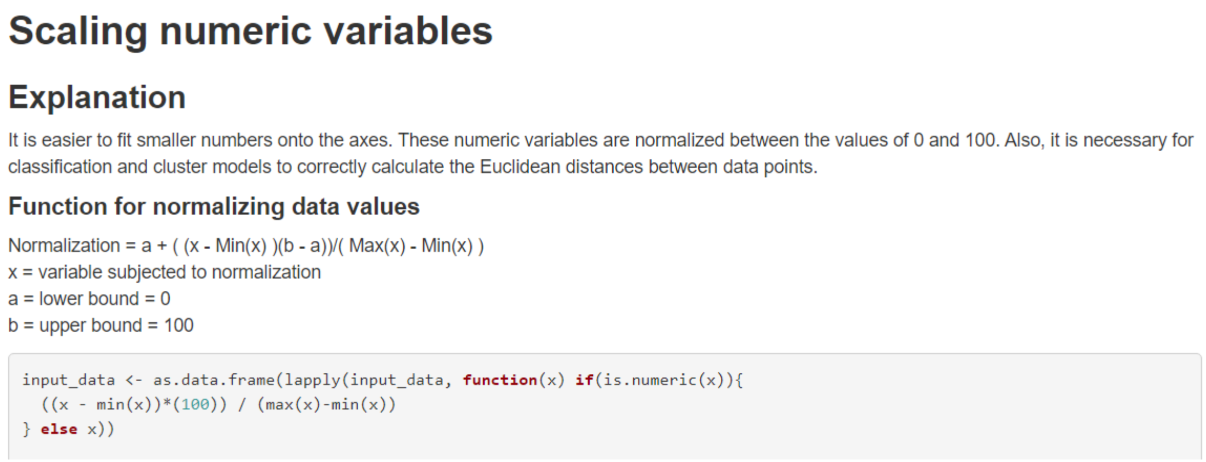
Write sample data to storage

```
# write.table(input_data_sp, file="sample_df_9.txt", append = FALSE, sep = "\t", dec = ".", row.names = FALSE, col.names = T
RUE)
```

Import sample data

```
input_data <- read_delim("sample_df_9.txt", "\t", escape_double = FALSE, col_types = cols(
        step = col_integer(),
        type = col_character(),
        amount = col_number(),
        nameDest = col_character(),
        newbalanceDest = col_number(),
        oldbalanceDest = col_number(),
        nameOrig = col_character(),
        newbalanceOrig = col_number(),
        oldbalanceOrg = col_number(),
        isFraud = col_character(),
        isFlaggedFraud = col_character()),
    trim_ws = TRUE)
```

Figure 14: Data Sampling



Scaling numeric variables

Explanation

It is easier to fit smaller numbers onto the axes. These numeric variables are normalized between the values of 0 and 100. Also, it is necessary for classification and cluster models to correctly calculate the Euclidean distances between data points.

Function for normalizing data values

Normalization = a + ((x - Min(x))(b - a))/(Max(x) - Min(x))
x = variable subjected to normalization
a = lower bound = 0
b = upper bound = 100

```
input_data <- as.data.frame(lapply(input_data, function(x) if(is.numeric(x)){
  ((x - min(x))*(100)) / (max(x)-min(x))
} else x))
```

Figure 15: Data Normalization R Code

Work Cited/References

American Bankers Association. Deposit Account Fraud Survey. Retrieved from https://www.aba.com/news-research/research-analysis/deposit-account-fraud-survey-report (Accessed May 1st, 2020)

Berengueres, J., Fenwick, A., Sandell, M. (2019). Introduction to Data Visualization & Storytelling: A Guide for The Data Scientist (First Edition). Independently published

Kaggle Inc. Synthetic Financial Data sets for Fraud Detection. Retrieved from https://www.kaggle.com/ntnu-testimon/paysim1/data# (Accessed May 1st, 2020)