Master in Data Science at Utica College
Course: DSC-503-Z1 Statistical Methods
Professor Name: Dr. Ho, Jing-Mao
Student Name: Henry J. Hu
Title: The Most Effective Way to Detect Fraudulent Activities in Banking Data

Introduction

According to the American Bankers Association's Deposit Account Fraud Survey, fraud against bank deposit accounts amounted to $25.1 billion in 2018. As banking businesses implemented new technologies to combat financial fraudulent activities, the criminals became more sophisticated in their techniques of carrying out these activities. Therefore, as data scientists, we must keep researching new ways of detecting financial fraudulent activities in banking data.

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?".

During the exploratory data analysis of this data set from Kaggle, it was found that logistic regression was the appropriate predictive model for this data because the values captured by the dependent variables are dichotomous. It was found that due to homogeneity in the data values of one of the two dependent variables, only one dependent variable could be selected for this study. Also, it was found that only five out of nine independent variables could be the drivers of the selected dependent variable. Lastly, it was found that after several iterations of the regression model, due to multicollinearity, the final resulting stepwise iteration of the regression model could have only four independent variables. Thus, the final most parsimonious regression model was identified to have only one dependent variable and four independent variables.

Literature Review

Article No. 1

Article Information:

“Detection of Fraud in Financial Statements: French Companies as a Case Study” by Ines Amara, Anis Ben Amar, and Anis Jarboui, published in July of 2013. These researchers are faculties and Ph.D. students of the University of Sfax in Tunisia. The article was published in the International Journal of Academic Research in Accounting, Finance and Management Sciences. It has a total of 12 pages.

Article Summary:

The purpose of the research conducted in this article was to test the impact of the "Fraud Triangle" elements on the detection of fraud in the financial statements. A Fraud Triangle is a term used in financial accounting to explain the motivation behind the individual who committed the fraud. Fraud in financial statements is a common phenomenon in large corporations where high-level managements are very disconnected from the rest of the organizational hierarchy. Often, these high-level managers who committed these fraudulent activities do not leave much of an audit trail. Therefore, it is very difficult for the internal auditor to detect these types of fraudulent activities. The best way to detect these types of accounting frauds is by relying on powerful computers to quickly sort through all electronic financial statements, calculate all the numbers, and ingest them into machine learning models such as logistic regression for fraud analysis.

The researcher in this article built a logistic regression model to predict the possibility that a firm is a victim of fraud in financial statements. The model has a total of one dependent variable and five independent variables. This logistic regression model has an R-squared value of 0.165, and only one of the five independent variables is statistically significant with a p-value of 0.049. Comparing the model (M1) statistics from this article with the model (M2) statistics fitted with Kaggle's financial fraud data for this paper, M1 is statistically very insignificant.

Research Project Relationship:

Fraud of financial statements is very similar to the fraud of real-time customer financial transactions. The only distinct difference is the criminal's status while carrying out the crime. One was the head of the bank and the other was the customer of the bank. For the same intention and purpose, both types of criminals want to steal money. Therefore, the same types of predictive models could be fitted to both types of financial transactions.

Article Critique:

The article perfectly explained all the details of fitting a logistic regression model onto the data of company financial statements. Also, both the statistical analysis and the mathematic used for evaluating the experiment are adequate.

<u>Article No. 2</u>

Article Information:

"Risk Analysis in Money Laundering A Case Study" by Joana Filipe Oliveira Marques, published by Instituto Superior Técnico in November of 2015. This research article has a total of 10 pages.

Article Summary:

The purpose of the research conducted in this article was to build logistic regression models for assessing bank client's risk levels. The "Risk Analysis in Money Laundering A Case Study" research explored two logistic regression models. One is the Private Entities logistic regression model (PELRM), and the other one is the Corporate Entities logistic regression model (CELRM). Their equations are Risk Level = 0.3286* Nationality +0.1995*Country of Residence +0.5004*job and Risk Level = 0.2729*Country of residence +0.6664*CAE+0.0943*CAE2+0.1624*Headquarters, respectively. These two logistic regression models do closely resemble the logistic regression model (LGM) fitted with Kaggle's financial fraud data for this research paper. The difference is that both PELRM and CELRM have only dichotomous independent variables while LGM has a combination of both categorical and continuous variables. PELRM has an R^2 value of 0.9908 and all independent variables have a confidence interval of 95% or p-values of 0.05. CELRM has an R^2 value of 0.8957 and all independent variables have a confidence interval of 95% or p-values of 0.05. Given that the R-Squared value cut-off point is 0.5 and the p-value cut-off point is 0.05, these two models are statistically significant. While the last iteration of the logistic regression model (LGM) for this research resulted in an R^2 value of 0.97 and an AIC value of 13.267. Therefore, by comparing just the R^2 values alone, the LGM model is statistically more significant than the CELRM model and is statistically less significant than the PELRM model.

Research Project Relationship:

This article was selected to be reviewed for possible statistical models to answer the research question. Also, the only difference between money laundering and financial transaction fraud is the former does not steal the money while the latter does steal the money. But both types of criminal activities could be detected by the movement of money. Therefore, the same types of predictive models could be fitted to both types of financial transactions.

Article Critique:

The article did not mention either the level of collinearity or multicollinearity between the independent variables. Also, there was no analysis of confounding variables. Therefore, it is difficult to assess whether the regression models' levels of significances were caused by multicollinearity and confounding variables.

Article No. 3

Article Information:

"An Analytical Approach to Detecting Insurance Fraud Using Logistic Regression" by Wilson, J. Holton, published in August of 2009. The researcher is an economics professor at Central Michigan University. The article was published in the Journal of Finance and Accountancy. It has a total of 15 pages.

Article Summary:

The purpose of the research conducted in this article was to find a predictive model that could precisely detect fraudulent activities associated with auto insurance claims. The patterns of information submitted to the insurance claim database could be used to train the predictive model. Because the outcome variable is dichotomous, the logistic regression model was selected to be fitted with the data. The stepwise iteration started with six independent variables. However, after it was evaluated for collinearity and multicollinearity, only two independent variables could be included with the model. The final model was able to predict 81.6% of the legitimate claims and 59.2% of the fraudulent claims, which is statistically not very significant. But given the nature of the insurance business, the large amount of data bias could skew the model's predictability power.

The article did mention the level of collinearity and multicollinearity between the independent variables. However, nowhere in the article that these calculations are mentioned. Also, there was no analysis of confounding variables. Therefore, it is difficult to assess whether the regression models' levels of significances were caused by multicollinearity and confounding variables.

Research Project Relationship:

Auto insurance claim fraud is very similar to the fraud of real-time customer financial transactions. The only distinct difference is the falsify information of the former is claim filing information and the latter is the transaction amount. Both types of information patterns lead to the indication of either fraud or no fraud. Therefore, the same types of predictive models could be fitted to both types of financial transactions.

Article Critique:

The article did not go into detail about the statistical significance of each of the relationships that lead to the determination of whether the model as a whole is significant or not. For example, it did mention the use of Chi-Square statistic and Pseudo R-squared values to determine the significance of relationships, but nowhere in the article, the results of the calculations are mentioned.

Data Source

Description

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

Context

Due to the highly sensitive nature of financial data and the legality associated it, there is a lack of publicly available secondary financial data sets to be used for educational purposes. Therefore, researchers have to use synthetic data sets such as this one to conduct their researches.

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 carrying out the data analysis, only 10,000 out of 1,048,575 data records were randomly selected for this research.

Table 1: Dependent Variables (DV)

Field Name	Variable Name	Data Type	Field Description	Unit	Conversion
isFraud	Is Fraud	Ordinal	A flag that indicates whether the transaction activity is 0 = Not fraudulent activity, 1 = fraudulent activity. These are the transactions made by the fraudulent agents inside the simulation. In this specific dataset, the fraudulent behavior of the agents aims to profit by taking control or customer accounts and try to empty the funds by transferring to another account and then cashing out of the system.	A binary number of either 0 or 1.	No conversion was made.
isFlaggedFraud	Is Flagged Fraud	Ordinal	A flag that indicates whether the transaction activity is 0 = Not fraudulent activity, 1 = fraudulent activity. The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction. **Since only zero values were recorded for this variable, it could not be used for regression modeling.**	A binary number of either 0 or 1.	No conversion was made.

Table 2: Independent Variables (IV)

Field Name	Variable Name	Data Type	Field Description	Unit	Conversion
step	Step	Interval	Maps a unit of time in the real world. In this case, 1 step is 1 hour. Total steps 744 (30 days of the simulation).	A number from 1 to 95.	No conversion was made.
type	Transaction Type	Nominal	The type of transactions. For example, CASH-IN, CASH-OUT, DEBIT, PAYMENT, and TRANSFER.	Text	No conversion was made.
amount	Transaction Amount	Ratio	Amount of the transaction in local currency.	Amount in US Dollars	No conversion was made.
nameOrig	Origination Account Number	Nominal	Customer who started the transaction.	Text	No conversion was made.
oldbalanceOrg	Old Origination Account Balance	Ratio	Initial balance before the transaction.	Amount in US Dollars	No conversion was made.
newbalanceOrig	New Origination Account Balance	Ratio	The new balance after the transaction.	Amount in US Dollars	No conversion was made.
nameDest	Destination Account Number	Nominal	Customer who is the recipient of the transaction.	Text	No conversion was made.
oldbalanceDest	Old Destination Account Balance	Ratio	Initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).	Amount in US Dollars	No conversion was made.
newbalanceDest	New Destination Account Balance	Ratio	New balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).	Amount in US Dollars	No conversion was made.

Exploratory Data Analysis

Central Tendency

Several conclusions could be drawn from observing the descriptive statistics depicted in Table 3 below. First, it appears that all the mean values are much closer to the minimum values than the maximum values. Second, it appears that all the mode values are zeros. The second observation provides the reason for the first observation. The mean values are skewed due to the huge number of zeros that are present in each of the variables. The median values are also much closer to the minimum values as well. This is a clear indication of the existence of outliers in the data. However, a better way to detect outliers is to calculate the z-score of each variable. Grubbs's Test utilizes Z-Scores.

From observing the box plots between Figure 1 and Figure 5 below, it appears that the largest average Transaction Amount value belongs to Transaction Type TRANSFER, largest average Old Origination Account Balance value belongs to Transaction Type CASH_IN, largest average Old Destination Account Balance value belongs to Transaction Type TRANSFER, and largest average New Destination Account Balance value belongs to Transaction Type TRANSFER. Also, it appears that the box plots were not able to capture any of the averages for the Transaction Type PAYMENT. Also, these box plots reveal that there are lots of outliers in the data.

Table 3: Descriptive Statistics Table

Variable Name	**Minimum**	**Maximum**	**Range**	**Mean**	**Median**	**Mode**	**Variance**	**Standard Deviation**	**Median Absolute Deviation**
Transaction Amount	3	9,887,819	3.21-9887819	155,905	74,261	NA	73.9+9	271,895	101,702
Old Origination Account Balance	0	30,000,000	0-30000000	942,470	15,562	0	1.E+13	3,162,776	23,072
New Origination Account Balance	0	30,300,000	0-30300000	960,685	0	0	1.E+13	3,199,075	0
Old Destination Account Balance	0	34,500,000	0-34500000	934,378	124,280	0	5.E+12	2,241,551	184,257
New Destination Account Balance	0	35,200,000	0-35200000	1,071,953	211,726	0	6.E+12	2,356,684	313,905

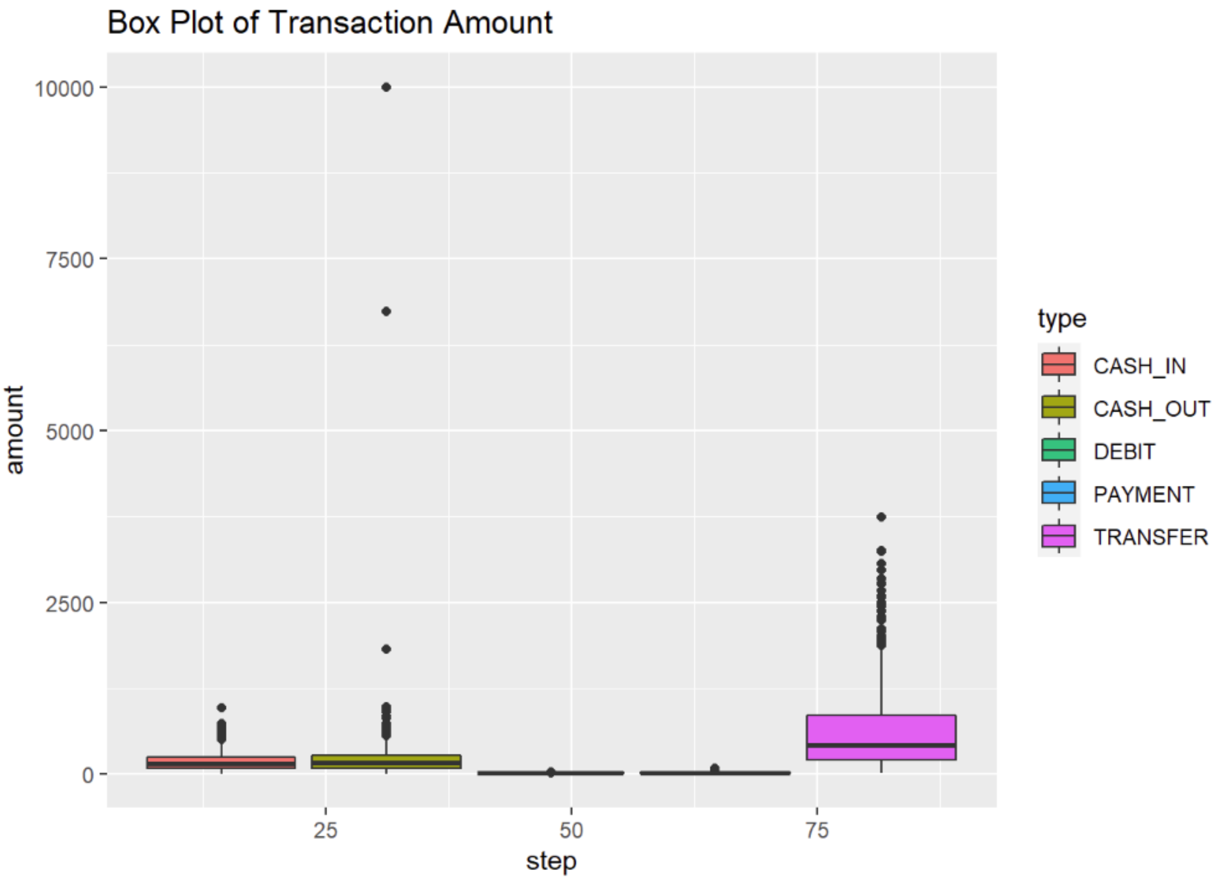

Figure 1: Box Plot

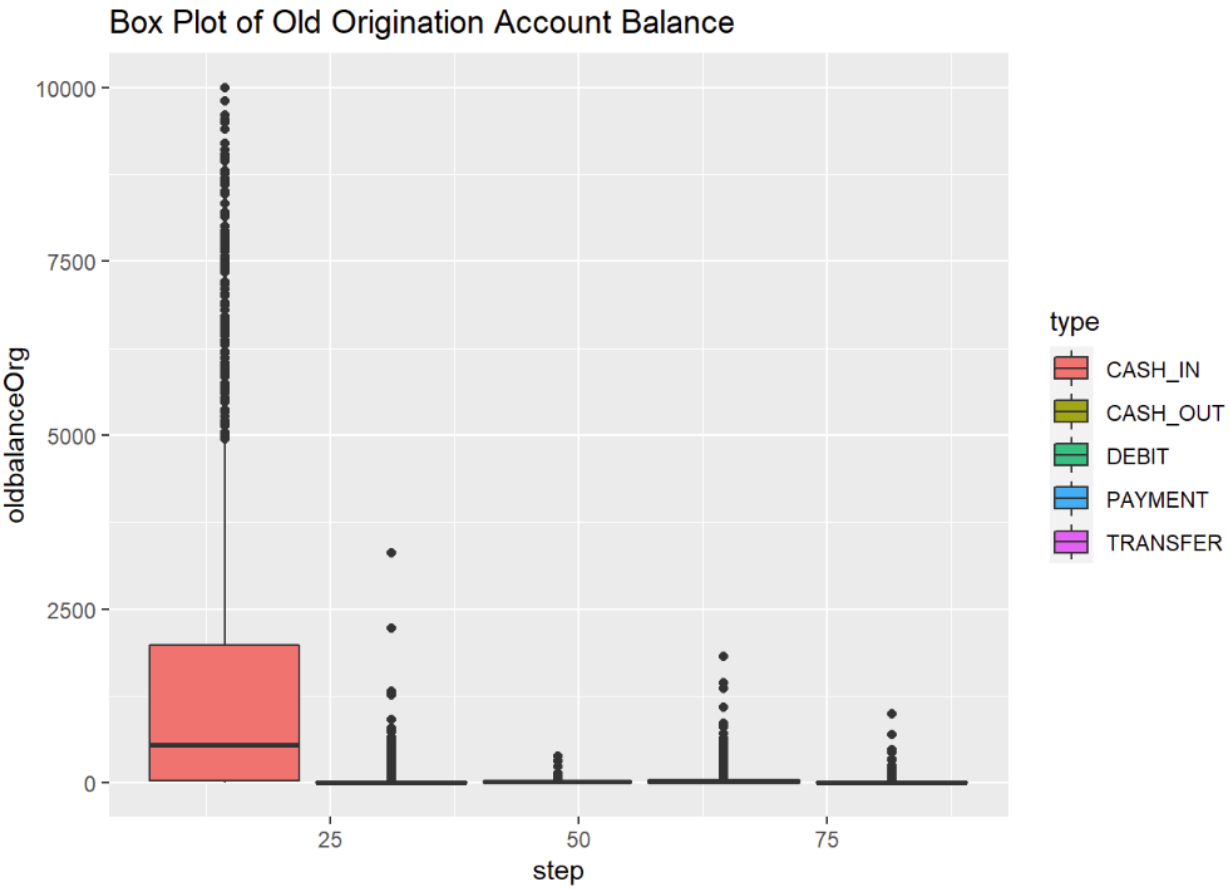

Figure 2: Box Plot

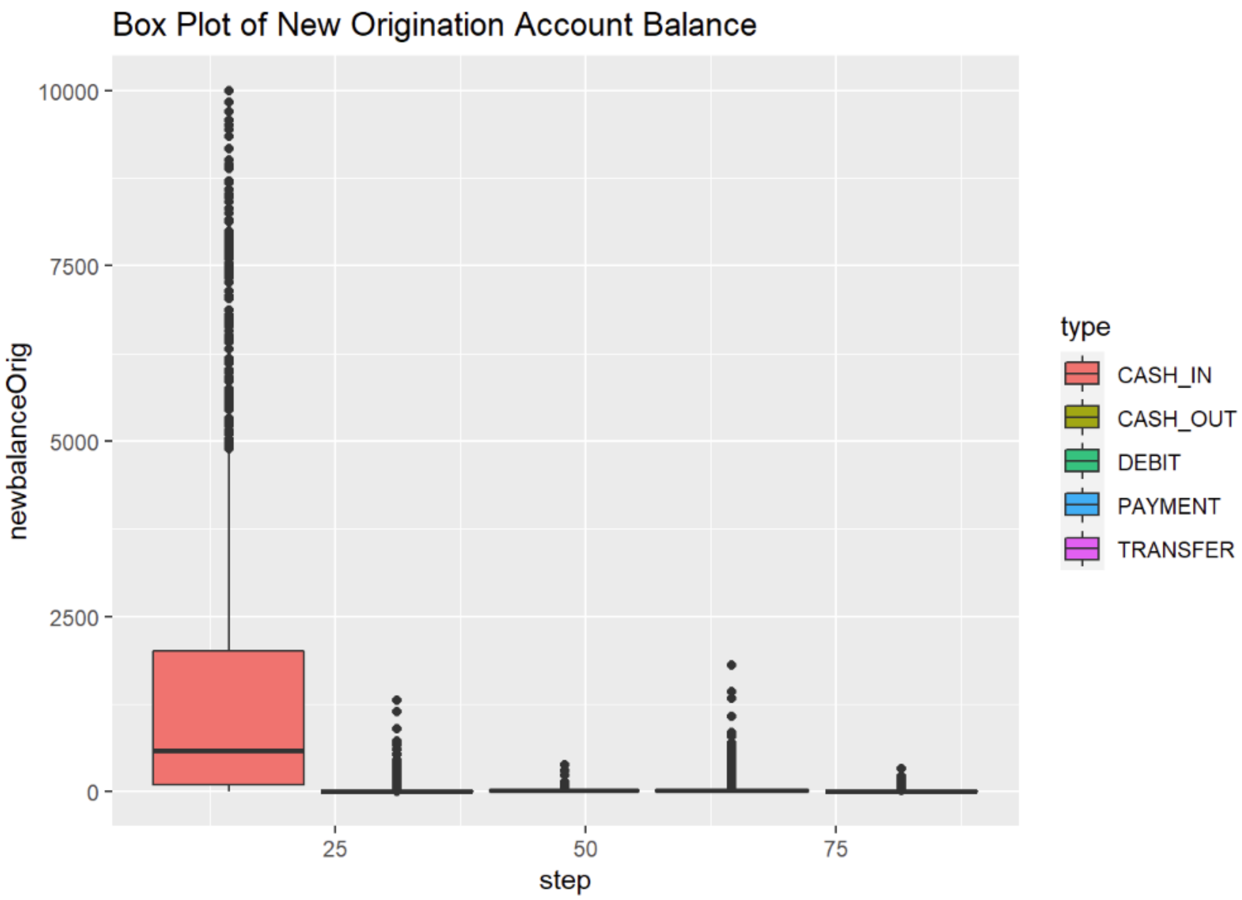

Figure 3: Box Plot

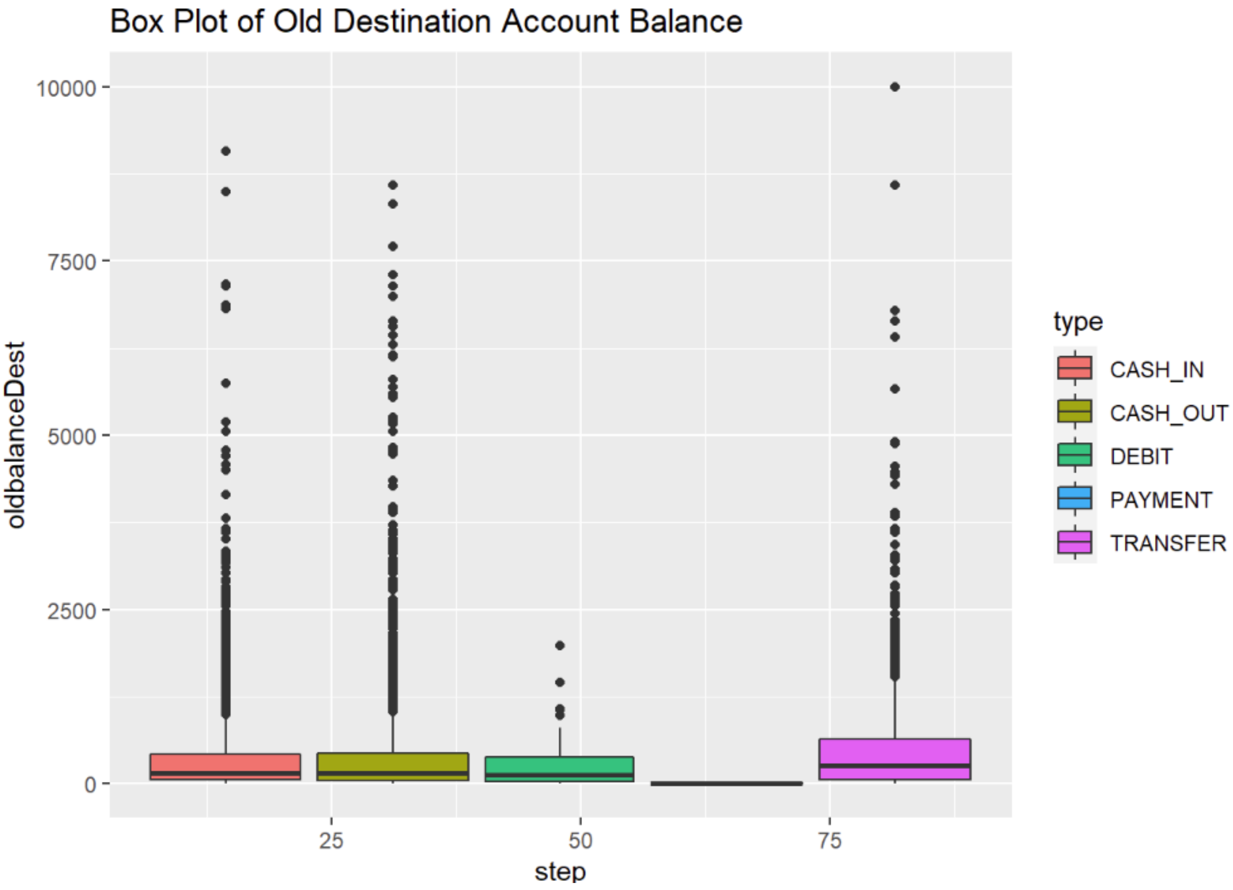

Figure 4: Box Plot

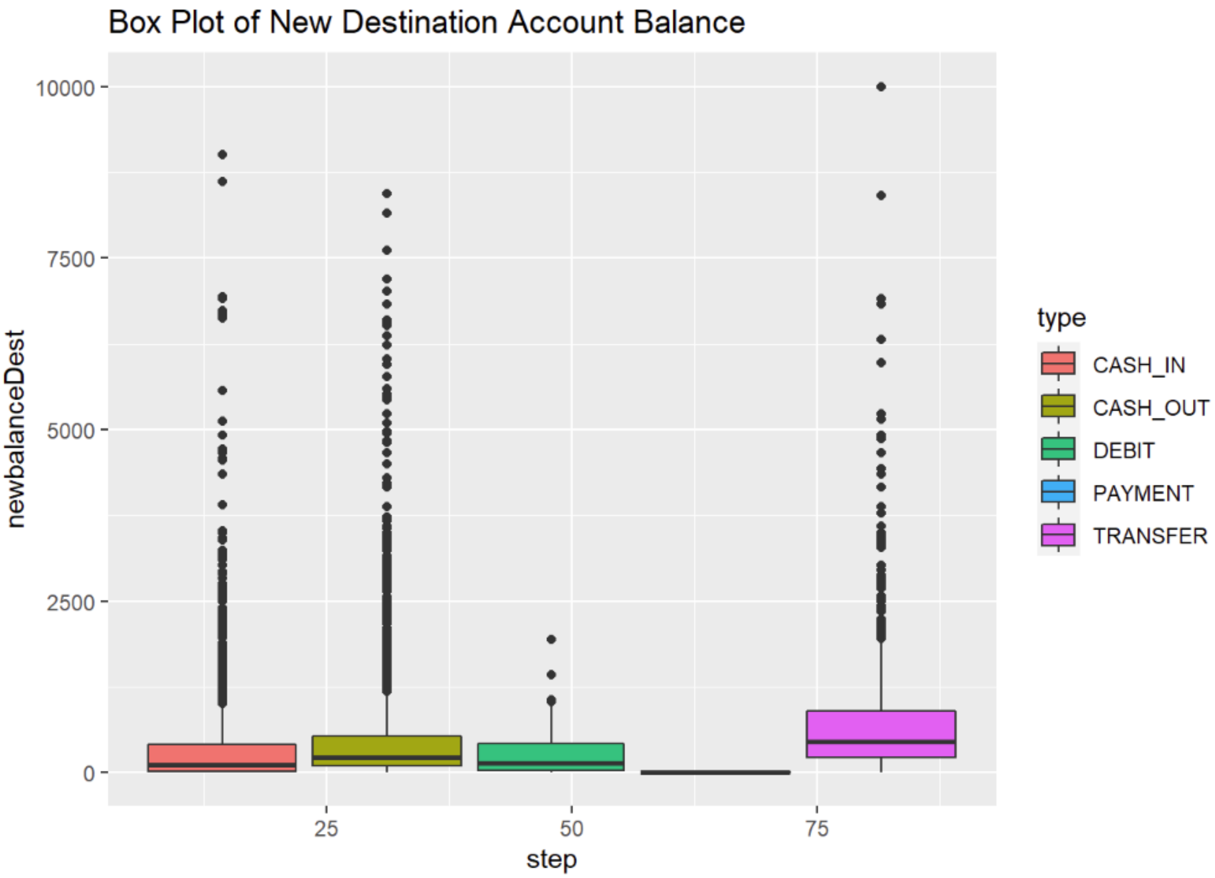

Figure 5: Box Plot

Dispersion Tendency

Several conclusions could be drawn from observing the descriptive statistics depicted in Table 3 above. Except for Transaction Amount, both the Standard Deviation values and the Median Absolute Deviation values are very far away from the mean values. The Standard Deviation values are larger than the Median Absolute Deviation values is because the calculation of Standard Deviation put more weight on the outliers. The average of the Standard Deviation values for variables Old Origination Account Balance, New Origination Account Balance, Old Destination Account Balance, and New Destination Account Balance is 2,740,022. While the average of the maximum values for these four variables is 32,500,000. This tells us that on average, the data points of these four variables are located at either about 29,759,978, or 35,240,022 on the y-axis. With the largest maximum value at 35,200,000, the Standard Deviation values tell us that the average data points are very close to the high end of the y-axis. The average of the Median Absolute Deviation values for these four variables is 130,309. According to the Median Absolute Deviation values, on average the data points of these four variables are located at either about 32,369,691 or 32,630,309 on the y-axis. From these calculations, it is evidenced that the dispersion of these data points is very extreme.

From observing the scatter charts between Figure 6 and Figure 10 below, it appears that on the plot of Transaction Amount there are outliers between 6500 and 10000 on the y-axis. It appears that the transaction activities are concentrated in blocks of steps 0-25, 30-50, and 90-100. Also, it appears that the majority of the transaction activities belong to transaction types CASH_IN, CASH_OUT, and TRANSFER.

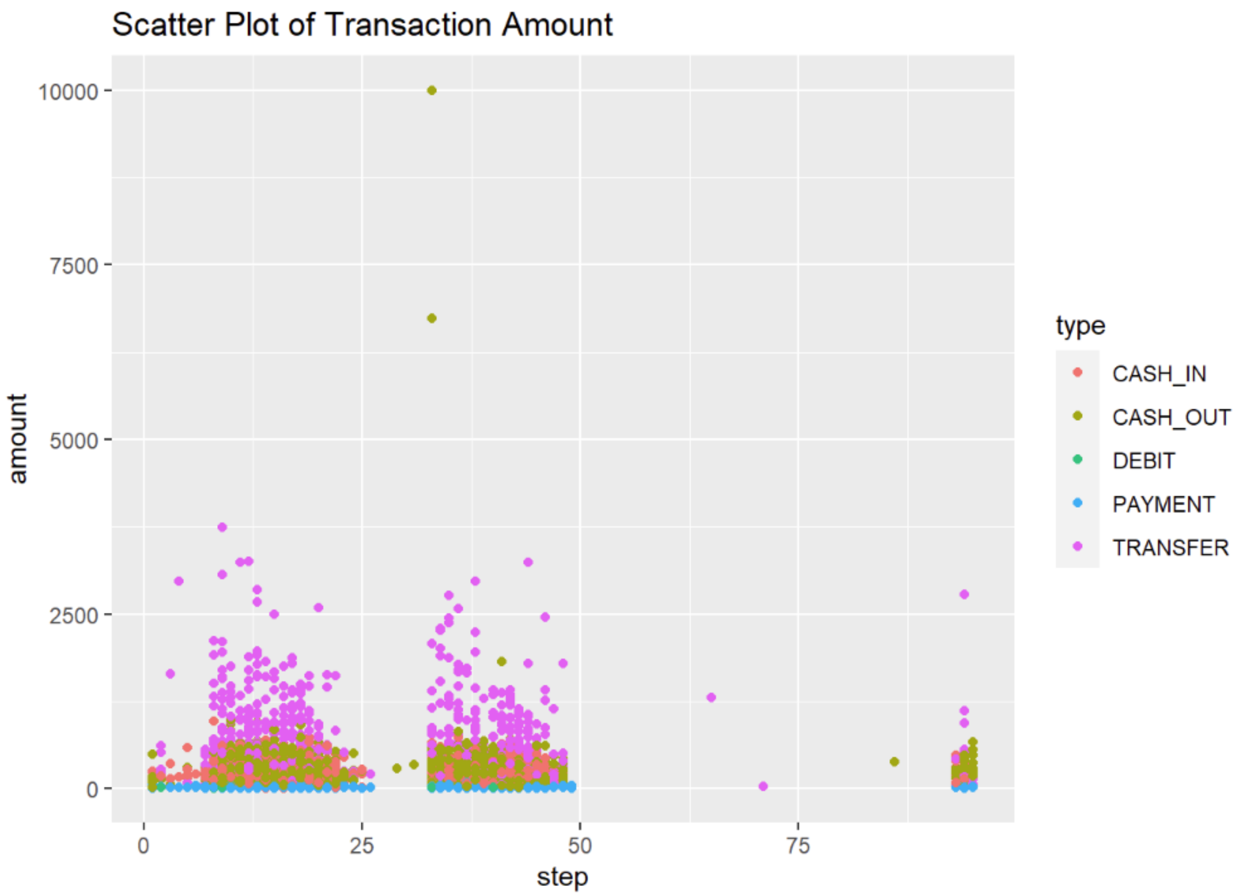

Figure 6: Scatter Chart

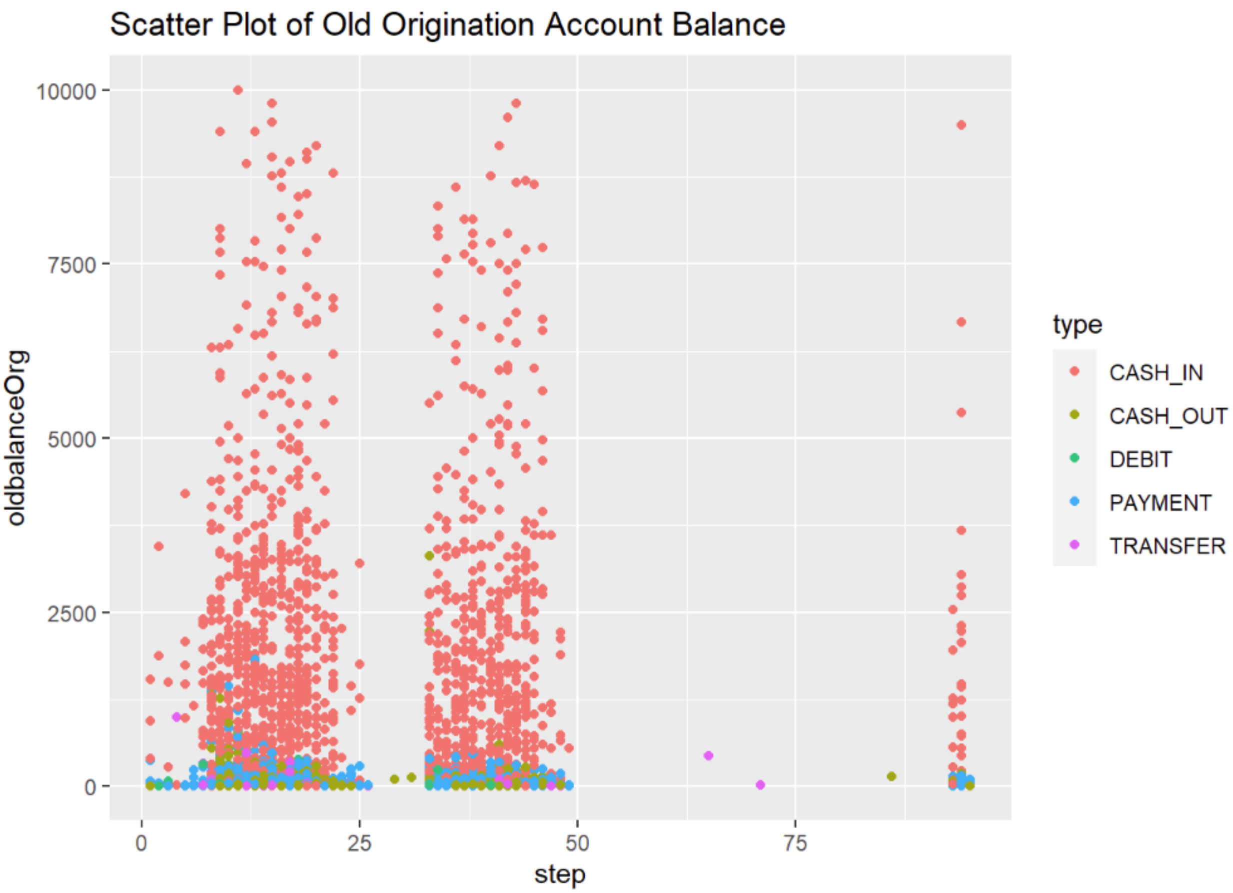

Figure 7: Scatter Chart

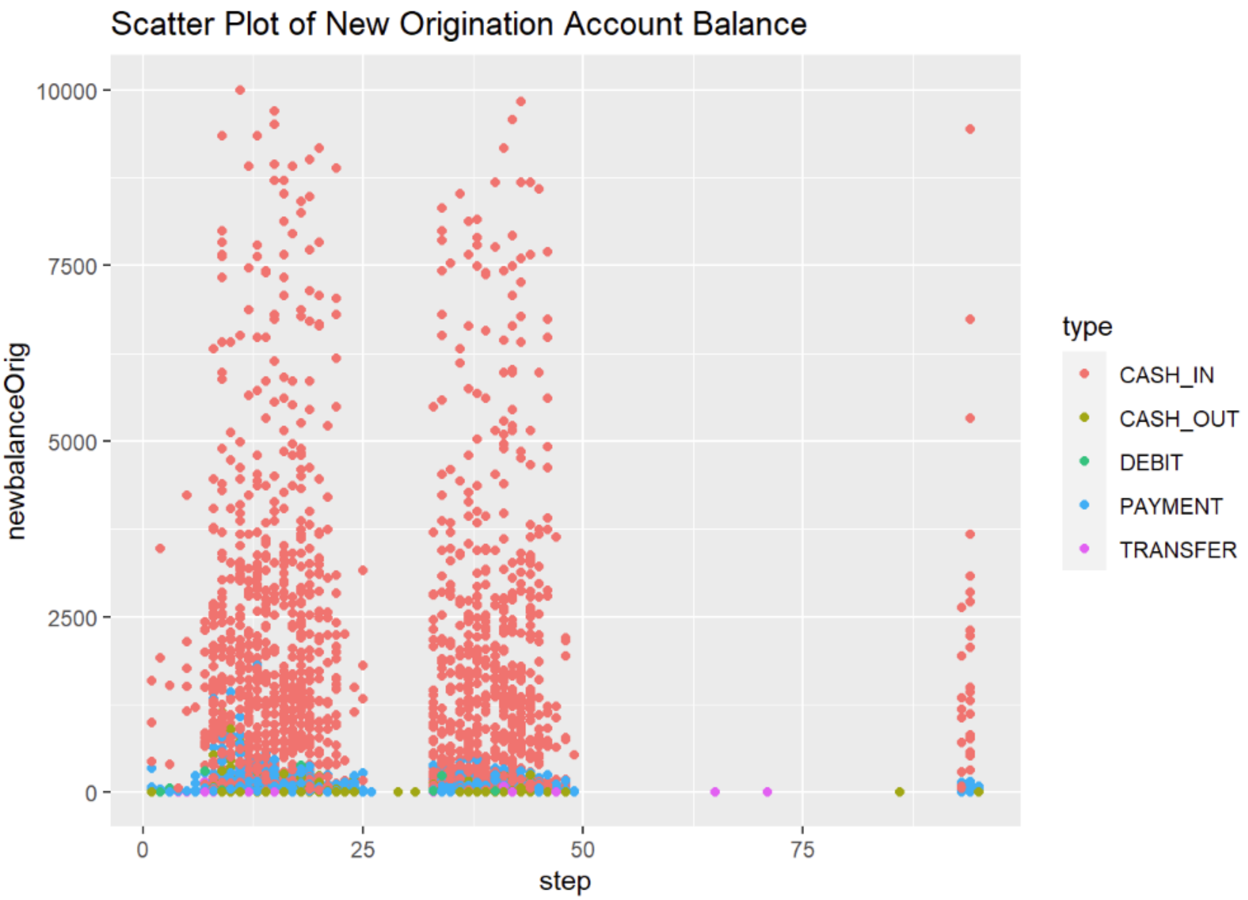

Figure 8: Scatter Chart

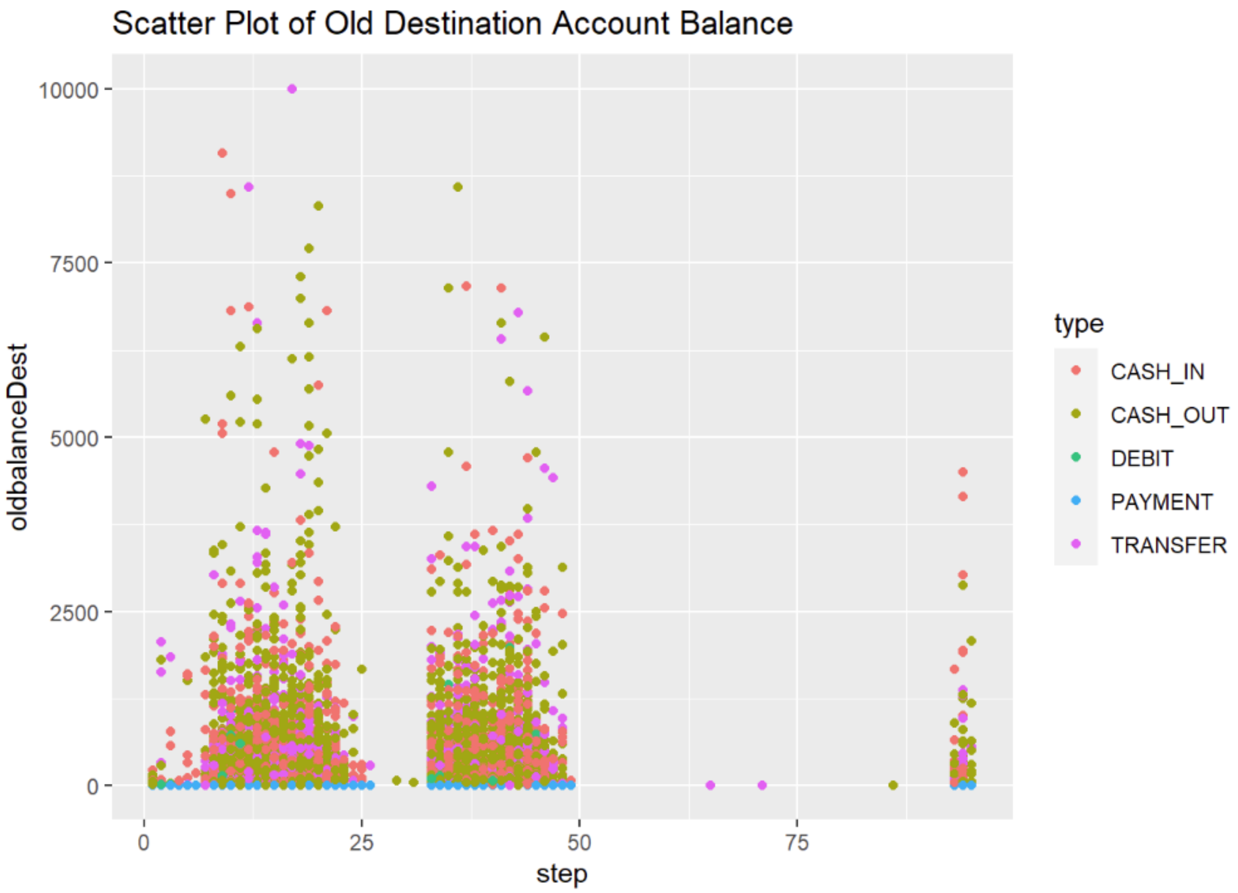

Figure 9: Scatter Chart

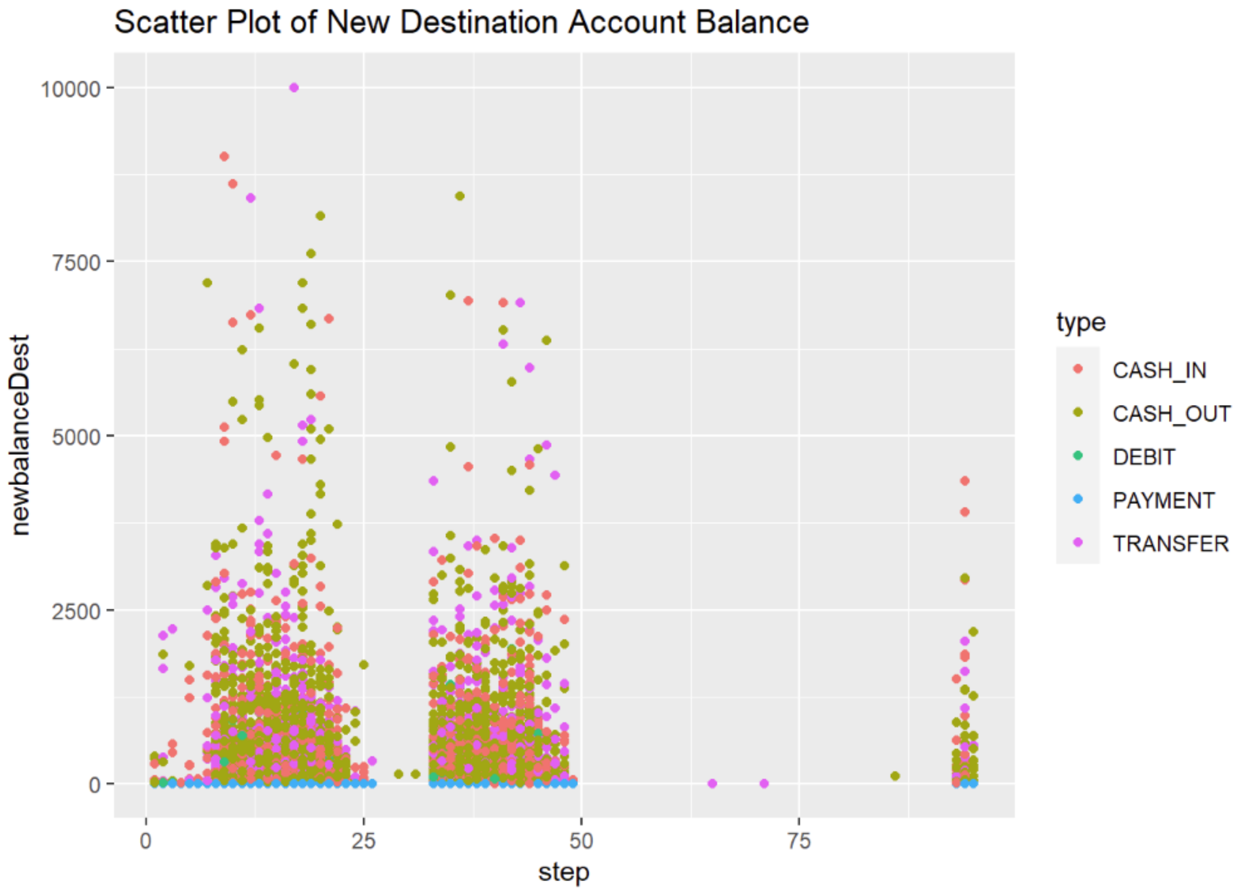

Figure 10: Scatter Chart

Outlier Calculation

As mentioned previously above, besides observing plots and charts, another way to identify outliers is to carry out the Grubbs Test. This is a test to check for the existence of outliers associated with each independent variable in the data frame. This test is based on Z-Scores. The function's null hypothesis is that there are no outliers. If the p-value is smaller than 0.05, then the null hypothesis could be rejected, and the alternative hypothesis that there is at least one outlier could be accepted. The two-tail test is carried out for this data frame. As a result of the test, all variables have p-values smaller than 0.05. Given a significant cut-off point of 0.05, all these variables have outliers.

Correlation Statistics

From observing the correlation coefficients depicted in Table 4 below, it appears that there are two pairs of variables that have correlation coefficients greater than 0.5 in the positive direction. They are newbalanceOrig and oldbalanceOrg, and newbalanceDest and oldbalanceDest. It is also evidenced by their p-values of smaller than 0.05 in Table 5 below, which given the significant cut-off point of 0.05 indicate the existence of strong correlations within themselves. Strong correlations between independent variables is an indication of multicollinearity.

Table 4: Correlation Coefficients

	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest
amount	1	0.01618856	0.00057682	0.20474725	0.306024317
oldbalanceOrg	0.016188561	1	**0.998690398**	0.07701329	0.04734309
newbalanceOrig	0.00057682	**0.9986904**	1	0.07829642	0.045078138
oldbalanceDest	0.204747252	0.07701329	0.078296415	1	**0.97911197**
newbalanceDest	0.306024317	0.04734309	0.045078138	**0.97911197**	1

Table 5: Correlation P-Values

	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest
amount	1	0.1055	0.954	2.20E-16	2.20E-16
oldbalanceOrg	0.1055	1	**2.20E-16**	2.20E-16	2.18E-06
newbalanceOrig	0.954	**2.20E-16**	1	4.48E-15	6.50E-06
oldbalanceDest	2.20E-16	2.20E-16	4.48E-15	1	**2.20E-16**
newbalanceDest	2.20E-16	2.18E-06	6.50E-06	**2.20E-16**	1

Zeros and Nulls

This synthetic data set from Kaggle doesn't have any null value in it. However, there are lots of 0's in both the fields that hold the new and old account balances.

Confounding Variable Analysis

Confounding happens when the collinearity between two or more independent variables creates a false causal relationship between the independent variable(s) and the dependent variable. It is important to eliminate these false causations because they could cause the regression model to make bias predictions.

From the Correlation Statistics depicted in table 4 and Table 5 above, two pairs of independent variables have correlation coefficients greater than 0.5 which indicate statistically significant relationships with the dependent variable. These pairs are newbalanceOrig and oldbalanceOrg with R-value of 0.99869040, and newbalanceDest and oldbalanceDest with R-value of 0.97911197. If one of the independent variables from any of these two pairs has a strong association with the dependent variable, then the existence of a confounding variable is a possibility. From examining the T-Test Statistics from R Studio output below, the only independent variable which has a strong association with the dependent variable is newbalanceOrig with a p-value of 2.2e-16. Therefore, the set (newbalanceOrig,oldbalanceOrg, isFraud) should be evaluated for the possibility of having a confounding variable. Because confounding is a matter of causation and not a matter of correlation, stepwise regression has to be carried out to determine if the confounding variable does exist.

The stepwise iteration result from R Studio output below reveals that variable newbalanceOrig has a strong effect on both the independent variable oldbalanceOrg and the y-intercept, which means the dependent variable isFraud is also strongly affected as well. However, independent variable oldbalanceOrg has very little effect on either variable newbalanceOrig or the y-intercept, which means the dependent variable isFraud is not very much affected by this independent variable. Therefore, the independent variable newbalanceOrig is the confounding variable. Since this variable causes the false causal relationship between independent variable oldbalanceOrg and the dependent variable isFraud, it should be evaluated for possible removal from the final regression equation. A likelihood ratio test between the models with and without this independent variable is necessary.

T-Test Statistics (R Studio output)

```
t.test(amount ~ isFraud, data = input_data)
##
##  Welch Two Sample t-test
##
## data:  amount by isFraud
## t = -2.0471, df = 11, p-value = 0.0653
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3876.4524   140.3762
## sample estimates:
## mean in group 0 mean in group 1
##        165.2713       2033.3094
```

```
t.test(oldbalanceOrg ~ isFraud, data = input_data)
##
##  Welch Two Sample t-test
##
## data:  oldbalanceOrg by isFraud
## t = -1.1745, df = 11.027, p-value = 0.2649
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1015.621   308.715
## sample estimates:
## mean in group 0 mean in group 1
##        323.4184        676.8714
```

```
t.test(newbalanceOrig ~ isFraud, data = input_data)
##
##  Welch Two Sample t-test
##
## data:  newbalanceOrig by isFraud
## t = 30.032, df = 9987, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  296.4224 337.8201
## sample estimates:
## mean in group 0 mean in group 1
##        327.1212         10.0000
```

```
t.test(oldbalanceDest ~ isFraud, data = input_data)
##
##  Welch Two Sample t-test
##
## data:  oldbalanceDest by isFraud
## t = 0.55923, df = 11.098, p-value = 0.5871
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -159.7707  268.7751
## sample estimates:
## mean in group 0 mean in group 1
##        280.6288        226.1266
```

```
t.test(newbalanceDest ~ isFraud, data = input_data)
##
##  Welch Two Sample t-test
##
## data:  newbalanceDest by isFraud
## t = -1.1897, df = 11.011, p-value = 0.2592
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1023.5674   305.2258
## sample estimates:
## mean in group 0 mean in group 1
##        313.7966        672.9674
```

Stepwise Iteration to Check Confounding (R Studio output)

```
#(newbalanceOrig,oldbalanceOrg, isFraud)
coef(lg_out1)
##    (Intercept) newbalanceOrig  oldbalanceOrg
##    65.60482430    -7.34751469     0.03620983
coef(lg_out2)
##    (Intercept) newbalanceOrig
##      130.36452      -13.65033
coef(lg_out3)
##   (Intercept) oldbalanceOrg
## -6.8110634295  0.0001846006
```

Conclusion of Exploratory Data Analysis

From this Exploratory Data Analysis exercise, we learned that it is very important to carry out this exercise before fitting any predictive model to the data. This exercise allows the data scientist to have a good understanding of the data, know the limitations present in the data, be aware of the biases present in the data, and the possible outcomes if the predictive model is applied, and if the data is suitable for fitting the model at all.

Methodology/Findings

Hypotheses for The Selected Regression Model

Null Hypothesis:

Changes in any of the independent variables will not affect the change in the dependent variable.

Alternative Hypothesis:

Changes in at least one of the independent variables will affect the change in the dependent variable.

Binary Logistic Regression Assumptions

Linearity and Log odds:

There must be a linear relationship between the log odds of the dependent variable and the independent variables. However, there should be no linear relationship between the independent variables and the odd-ratio or the outcome probability.

No Multicollinearity:

Independent variables must not be too highly correlated with each other.

Collinearity:

There must be some degree of collinearity between variables.

Variable Type:

The dependent variable must be categorical. The independent variables could be either continuous or categorical.

Outcome Value:

The outcome values of the dependent variable must be dichotomous.

Independent Observations:

The observations must be independent of each other. The observations should not come from repeated measurements or matched data. If observations are related to one another, then the model will tend to overweight the significance of those observations.

Model Iteration

During the first stepwise iteration of the regression model, the null model is used. This iteration will reveal the model statistics when there is no independent variable. The statistics will include the R-squared value, the Akaike information criterion (AIC), and the P-value of the logistic regression. The R-squared value indicates the percentage of the actual data points that could be explained by the regression. The AIC value is a maximum likelihood estimator for the logistic regression model. The smaller the AIC value the better fitted the model. The p-value of each independent variable indicates the pairwise significance of the association between each independent variable and the dependent variable. The iteration result is depicted in the 1st Stepwise Regression Iteration section below. The result reveals that none of the model statistics is statistically significant. A Model P-Value of 1.0 indicates a regression model with no significance.

During the second stepwise iteration of the regression model, the independent variable amount was added to the null model. The iteration result is depicted in the 2nd Stepwise Regression Iteration section below. Given a significant cut-off point of 0.05, the independent variable amount with a p-value of 0.0 indicates a very significant relationship with the dependent variable. A McFadden R-squared value of 0.2277059 means that the model can explain only 22.77% of the real observed data points. A Model P-Value of 8.173884e-11 indicates a moderately significant regression model.

During the third stepwise iteration of the regression model, both the independent variables amount and newbalanceDest were added to the null model. The iteration result is depicted in the 3rd Stepwise Regression Iteration section below. Given a significant cut-off point of 0.05, the independent variable amount has a p-value of 0.0, which indicates a highly significant relationship with the dependent variable. Independent variable newbalanceDest has a p-value of 0.001, which indicates a moderately significant relationship with the dependent variable. A McFadden R-squared value of 0.2645303 means that the model can explain 26.45% of the real observed data points. A Model P-Value of 2.241185e-11 indicates a moderately significant regression model. All independent variables have variance inflation factors smaller than 2.25, which given a cut-off point of 5.0 means that multicollinearity is moderately weak.

During the fourth stepwise iteration of the regression model, all three independent variables, amount, newbalanceOrig, and oldbalanceOrg were added to the null model. The iteration result is depicted in the 4th Stepwise Regression Iteration section below. Given a significant cut-off point of 0.05, none of the independent variables is significant. However, the A McFadden R-Squared value is 0.971051, which means that the model can explain 97% of the real observed data points. A Model P-Value of 0.0 indicates a highly significant regression model. However, all independent variables have variance inflation factors greater than 5.0, which given a cut-off point of 5.0 means that the independent variables are highly correlated.

The purpose of the fifth stepwise iteration is to figure out which independent variables are the control variables and which are the variables of interest. As the article posted by Paul Allison on Statistical Horizons explained, “The variables with high VIFs are control variables, and the variables of interest do not have high VIFs”. The iteration result is depicted in the 5th Stepwise Regression Iteration section below. Given a significant cut-off point of 0.05, none of the independent variables is significant. However, the A McFadden R-Squared value is 0.9694564, which means that the model can explain 97% of the real observed data points. A

Model P-Value of 0.0 indicates a highly significant regression model. However, all independent variables except the newbalanceDest have variance inflation factors greater than 5.0, which given a cut-off point of 5.0 means that the independent variables are highly correlated.

1st Stepwise Regression Iteration (R Studio output)

Stepwise

```
oldw <- getOption("warn")
options(warn = -1)
lg_out <- glm(isFraud ~ 1, data=input_data, family=binomial(link="logit"))
summary(lg_out)
##
## Call:
## glm(formula = isFraud ~ 1, family = binomial(link = "logit"),
##     data = input_data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.049  -0.049  -0.049  -0.049   3.667
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.7242     0.2888  -23.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 185.4  on 9999  degrees of freedom
## Residual deviance: 185.4  on 9999  degrees of freedom
## AIC: 187.4
##
## Number of Fisher Scoring iterations: 9
```

McFadden R-Squared

```
library(DescTools)
##
## Attaching package: 'DescTools'
## The following objects are masked from 'package:Hmisc':
##
##    %nin%, Label, Mean, Quantile
r2 <- PseudoR2(lg_out, which = "McFadden")
r2_adj <- PseudoR2(lg_out, which = "McFaddenAdj")
print(r2)
## McFadden
##        0
print(r2_adj)
## McFaddenAdj
## -0.01078772
```

Model P-Value

```
# log-likelihood of the null model
ll.null <- lg_out$null.deviance/-2
# log-likelihood of the fancy model
ll.proposed <- lg_out$deviance/-2
pv <- 1-pchisq(2*(ll.proposed-ll.null),df=(length(lg_out$coefficients)-1))
print(pv)
## [1] 0
```

2nd Stepwise Regression Iteration (R Studio output)

Stepwise

```
oldw <- getOption("warn")
options(warn = -1)
lg_out <- glm(isFraud ~ amount, data=input_data, family=binomial(link="logit"))
summary(lg_out)
##
## Call:
## glm(formula = isFraud ~ amount, family = binomial(link = "logit"),
##     data = input_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7964  -0.0409  -0.0362  -0.0343   3.8567
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.479162   0.388782  -19.24  < 2e-16 ***
## amount       0.001739   0.000302    5.76 8.42e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 185.40  on 9999  degrees of freedom
## Residual deviance: 143.18  on 9998  degrees of freedom
## AIC: 147.18
##
## Number of Fisher Scoring iterations: 10
# Coefficient Confident Intervals
confint.default(lg_out)
##                    2.5 %       97.5 %
## (Intercept) -8.241160491 -6.717163959
## amount       0.001147438  0.002331135
# Odd Ratios
exp(coef(lg_out))
##  (Intercept)       amount
## 0.0005647303 1.0017407998
options(warn = oldw)
```

McFadden R-Squared

```
library(DescTools)
r2 <- PseudoR2(lg_out, which = "McFadden")
r2_adj <- PseudoR2(lg_out, which = "McFaddenAdj")
print(r2)
##  McFadden
## 0.2277059
print(r2_adj)
## McFaddenAdj
##   0.2061304
```

Model P-Value

```
# log-likelihood of the null model
ll.null <- lg_out$null.deviance/-2
# log-likelihood of the fancy model
ll.proposed <- lg_out$deviance/-2
pv <- 1-pchisq(2*(ll.proposed-ll.null),df=(length(lg_out$coefficients)-1))
print(pv)
## [1] 8.173884e-11
```

3rd Stepwise Regression Iteration (R Studio output)

Stepwise

```
oldw <- getOption("warn")
options(warn = -1)
lg_out <- glm(isFraud ~ amount + newbalanceDest, data=input_data,
family=binomial(link="logit"))
summary(lg_out)
##
## Call:
## glm(formula = isFraud ~ amount + newbalanceDest, family = binomial(link = "logit"),
##     data = input_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8526  -0.0410  -0.0389  -0.0347   4.2873
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.2103469  0.3992348 -18.060  < 2e-16 ***
## amount          0.0026718  0.0004997   5.347 8.95e-08 ***
## newbalanceDest -0.0023694  0.0012464  -1.901   0.0573 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 185.40  on 9999  degrees of freedom
## Residual deviance: 136.35  on 9997  degrees of freedom
## AIC: 142.35
##
## Number of Fisher Scoring iterations: 11
# Coefficient Confident Intervals
confint.default(lg_out)
##                     2.5 %       97.5 %
## (Intercept)    -7.992832648 -6.427861e+00
## amount          0.001692401  3.651197e-03
## newbalanceDest -0.004812312  7.341906e-05
# Odd Ratios
exp(coef(lg_out))
##    (Intercept)        amount newbalanceDest
##   0.0007389008   1.0026753714   0.9976333585
options(warn = oldw)
```

McFadden R-Squared

```
library(DescTools)
r2 <- PseudoR2(lg_out, which = "McFadden")
r2_adj <- PseudoR2(lg_out, which = "McFaddenAdj")
print(r2)
##  McFadden
## 0.2645303
print(r2_adj)
## McFaddenAdj
##   0.2321672
```

Model P-Value

```
# log-likelihood of the null model
ll.null <- lg_out$null.deviance/-2
# log-likelihood of the fancy model
ll.proposed <- lg_out$deviance/-2
pv <- 1-pchisq(2*(ll.proposed-ll.null),df=(length(lg_out$coefficients)-1))
print(pv)
## [1] 2.241185e-11
```

Multicollinearity

```
library(car)
# Variance Inflation Factors (>5?)
vif(lg_out)
##        amount newbalanceDest
##      2.213578       2.213578
```

4th Stepwise Regression Iteration (R Studio output)

Stepwise

```
oldw <- getOption("warn")
options(warn = -1)
lg_out <- glm(isFraud ~ amount + newbalanceOrig + oldbalanceOrg, data=input_data,
family=binomial(link="logit"))
summary(lg_out)
##
## Call:
## glm(formula = isFraud ~ amount + newbalanceOrig + oldbalanceOrg,
##     family = binomial(link = "logit"), data = input_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6307   0.0000   0.0000   0.0000   1.7204
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1672.50     815.34   2.051   0.0402 *
## amount           -74.13      64.10  -1.156   0.2475
## newbalanceOrig  -318.33     205.29  -1.551   0.1210
## oldbalanceOrg    225.01     194.49   1.157   0.2473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 185.396  on 9999  degrees of freedom
## Residual deviance:   5.367  on 9996  degrees of freedom
## AIC: 13.367
##
## Number of Fisher Scoring iterations: 25
# Coefficient Confident Intervals
confint.default(lg_out)
##                    2.5 %     97.5 %
## (Intercept)      74.45535 3270.53526
## amount        -199.76973   51.51659
## newbalanceOrig -720.69041   84.02664
## oldbalanceOrg  -156.17933  606.20116
# Odd Ratios
exp(coef(lg_out))
##    (Intercept)        amount newbalanceOrig  oldbalanceOrg
##            Inf   6.415646e-33  5.626259e-139   5.260161e+97
options(warn = oldw)
```

McFadden R-Squared

```
library(DescTools)
r2 <- PseudoR2(lg_out, which = "McFadden")
r2_adj <- PseudoR2(lg_out, which = "McFaddenAdj")
print(r2)
## McFadden
## 0.971051
print(r2_adj)
## McFaddenAdj
##   0.9279002
```

Model P-Value

```
# log-likelihood of the null model
ll.null <- lg_out$null.deviance/-2
# log-likelihood of the fancy model
ll.proposed <- lg_out$deviance/-2
pv <- 1-pchisq(2*(ll.proposed-ll.null),df=(length(lg_out$coefficients)-1))
print(pv)
## [1] 0
```

Multicollinearity

```
library(car)
# Variance Inflation Factors (>5?)
vif(lg_out)
##        amount newbalanceOrig  oldbalanceOrg
##   2.477457e+07   3.197941e+01   2.479807e+07
```

5th Stepwise Regression Iteration (R Studio output)

Stepwise

```
oldw <- getOption("warn")
options(warn = -1)
lg_out_2 <- glm(isFraud ~ amount + newbalanceOrig + oldbalanceOrg + newbalanceDest,
data=input_data, family=binomial(link="logit"))
summary(lg_out_2)
##
## Call:
## glm(formula = isFraud ~ amount + newbalanceOrig + oldbalanceOrg +
##     newbalanceDest, family = binomial(link = "logit"), data = input_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5915   0.0000   0.0000   0.0000   1.7901
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.489e+03  7.675e+02   1.940   0.0524 .
## amount         -6.903e+01  5.949e+01  -1.160   0.2459
## newbalanceOrig -2.896e+02  1.918e+02  -1.510   0.1310
## oldbalanceOrg   2.095e+02  1.805e+02   1.161   0.2457
## newbalanceDest  1.562e-03  4.282e-03   0.365   0.7153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 185.3960  on 9999  degrees of freedom
## Residual deviance:   5.6627  on 9995  degrees of freedom
## AIC: 15.663
##
## Number of Fisher Scoring iterations: 25
# Coefficient Confident Intervals
confint.default(lg_out)
##                    2.5 %     97.5 %
## (Intercept)      74.45535 3270.53526
## amount         -199.76973   51.51659
## newbalanceOrig -720.69041   84.02664
## oldbalanceOrg  -156.17933  606.20116
# Odd Ratios
exp(coef(lg_out))
##    (Intercept)         amount newbalanceOrig  oldbalanceOrg
##          Inf   6.415646e-33  5.626259e-139   5.260161e+97
# exp(cbind(OR = coef(lg_out_2), confint(lg_out_2)))
options(warn = oldw)
```

McFadden R-Squared

```
library(DescTools)
r2 <- PseudoR2(lg_out_2, which = "McFadden")
r2_adj <- PseudoR2(lg_out_2, which = "McFaddenAdj")
print(r2)
##  McFadden
## 0.9694564
print(r2_adj)
## McFaddenAdj
##   0.9155178
```

Model P-Value

```
# log-likelihood of the null model
ll.null <- lg_out$null.deviance/-2
# log-likelihood of the fancy model
ll.proposed <- lg_out$deviance/-2
pv <- 1-pchisq(2*(ll.proposed-ll.null),df=(length(lg_out$coefficients)-1))
print(pv)
## [1] 0
```

Multicollinearity

```
library(car)
# Variance Inflation Factors (>5?)
vif(lg_out_2)
##        amount newbalanceOrig  oldbalanceOrg newbalanceDest
##   2.529981e+07   3.051996e+01   2.531792e+07   1.155344e+00
```

Regression Models Comparison

1st stepwise iteration is insignificant. Starting with 2nd stepwise iteration, both the AIC value and Model P-Value are significantly larger than those of the model in the 3rd stepwise iteration. There is only a minor difference in the McFadden R-squared values between the models in the 2nd stepwise iteration and 3rd stepwise iteration. This means that the model in the 3rd stepwise iteration is statistically more significant.

The AIC value and the McFadden R-squared value of the model in the 3rd stepwise iteration significantly much larger and much smaller than those of the model in the 4th stepwise iteration, respectively. However, no independent variable of the model in the 3rd stepwise iteration has a VIF value greater than 5.0, while all the independent variables of the model in the 4th stepwise iteration have VIF values greater than 5.0. If comparing only the AIC value and the McFadden R-squared value alone, the model in the 4th stepwise iteration seems to be statistically more significant than the model in the 3rd stepwise iteration. However, that significant could be caused by the high level of multicollinearity.

The AIC value and the McFadden R-squared value of the model in the 4th stepwise iteration significantly much larger and much smaller than those of the model in the 5th stepwise iteration, respectively. However, while all independent variables of the model in the 4th stepwise iteration have VIF values greater than 5.0, only three out of four independent variables of the model in the 5th stepwise iteration have VIF values greater than 5.0. The only independent variable of the model in the 5th stepwise iteration that has a VIF value smaller than 5.0 is the newbalanceDest. As the article posted by Paul Allison on Statistical Horizons explained, “The variables with high VIFs are control variables, and the variables of interest do not have high VIFs". Therefore, the variable of interest, in this case, is newbalanceDest and the other three independent variables are control variables.

To make sure that there is a significant difference between the models in the 4th stepwise iteration and 5th stepwise iteration, a model comparison test was carried out. This test is depicted in the "Models Comparison Using Likelihood Ratio Test" right below. The result of the test showed that the two models are significantly different. Therefore, the final most parsimonious predictive model for this data set is concluded to be the model in the 5th stepwise iteration.

Models Comparison Using Likelihood Ratio Test (R Studio output)

```
anova (lg_out,lg_out_2,test="LRT")
## Analysis of Deviance Table
##
## Model 1: isFraud ~ amount + newbalanceOrig + oldbalanceOrg
## Model 2: isFraud ~ amount + newbalanceOrig + oldbalanceOrg + newbalanceDest
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     9996     5.3670
## 2     9995     5.6627  1 -0.29564
```

Regression Equation Identified

Y = ln(p/(1-p)) = b0 + b1(X1) + b2(X2) + b3(X3) + b4(X4)

Y = 1489 - 69.03(x1) - 289.6 (X2) + 209.5(X3) + 0.001562(X4)

p = isFraud = $(e^{bo+b1*x1+b2*x2+b3*x3+b4*x4})/(e^{bo+b1*x1+b2*x2+b3*x3+b4*x4}+1)$

b0 = Y-intercept = 1489

b1 = Coefficient 1 = -69.03

b2 = Coefficient 2 = -289.6

b3 = Coefficient 3 = 209.5

b4 = Coefficient 4 = 0.001562

X1 = amount = Transaction Amount

X2 = newbalanceOrig = New Origination Account Balance

X3 = oldbalanceOrg = Old Origination Account Balance

X4 = newbalanceDest = New Destination Balance

Regression Model Interpretation

The y-intercept or coefficient b0 for this equation is 1489. If independent variables X1, X2, X3, and X4 are all equal to zeros, then the log-odds Y is equal to 1489. Coefficients b1 and b2 both have negative signs, which mean if X3 and X4 are constant, any increase in the value of either X1 or X2, there will be a corresponding decrease in the value of log-odds Y. Coefficient b3 has a positive sign, which means that if all X1, X2, and X4 are constant, any increase in the value of X3 will correspond to an increase in the value of log-odds Y. Coefficient b4 has a positive sign, which means that if all X1, X2, and X3 are constant, any increase in the value of X4 will correspond to an increase in the value of log-odds Y. Coefficient b2 is 4.2 times the magnitude of coefficient b1, and coefficient b3 is 134,123 times the magnitude of coefficient b4.

In terms of the probability p, if b1(X1) + b2(X2) + b3(X3) + b4(X4) is greater than 0, then the probability value p = (e^(b0 + b1(X1) + b2(X2) + b3(X3) + b4(X4)) / (e^(b0 + b1(X1) + b2(X2) + b3(X3) + b4(X4))+1) will be above 0.5. A probability value above 0.5 indicates the present of fraudulent activity. For example, if X1=0, X2=0, X4=0 and X3=0.001, then probability p = e^(1489+(-69.03)(0)+(-289.6)(0)+(209.5)(0.001)+(0.001562)(0))/((1489+(-69.03)(0)+(-289.6)(0)+(209.5)(0.001)+(0.001562)(0))+1) = 0.552. Because the outcome variable is dichotomous, the probability value above or equal to 0.5 is evaluated to indicate a present of fraudulent activity, and below 0.5 is evaluated to indicate an absence of fraudulent activity. Meaning, even if all the Transaction Amount, New Origination Account Balance and New Destination Balance are zeros, any increase in the Old Origination Account Balance will result in the present of fraudulent activity. An example of where there is an absence of fraudulent activity would be X1=1, X2=0, X3=0, and X4=0. For this example, the probability value p =

e^(1489+(-69.03)(1)+(-289.6)(0)+(209.5)(0)+(0.001562)(0))/((1489+(-69.03)(1)+(-289.6)(0)+(209.5)(0)+(0.001562)(0))+1) = 0 indicates an absence of fraudulent activity. Meaning, while all the Old Origination Account Balance, New Origination Account Balance and New Destination Account Balance are zeros, an increase in the Transaction Amount will lower the probability that the transaction will be detected as fraudulent activity.

<u>Hypothesis Evaluation</u>

Type I and Type II Error:

Type I error is committed when the hypothesis is wrongly rejected. Type II error is committed when the hypothesis is wrongly accepted. For this research, neither Type I nor Type II error has been committed. It is because of the model statistics from the 4th stepwise iteration is undisputable. A McFadden R-squared of 0.971051 and an AIC value of 13.367, clearly indicates a highly significant regression model. Therefore, the null hypothesis that changes in any of the independent variables will not affect the change in the dependent could be confidently rejected. Hence, the alternative hypothesis of changes in at least one of the independent variables will affect the change in the is accepted.

Data Bias:

As depicted in Figure 11 below, fraudulent transaction activities are found only with transaction type CASH_OUT and TRANSFER. However, in reality, fraudulent transaction activities could exist with any transaction type. Therefore, it is a good decision to exclude the independent variable Transaction Type from the regression model.

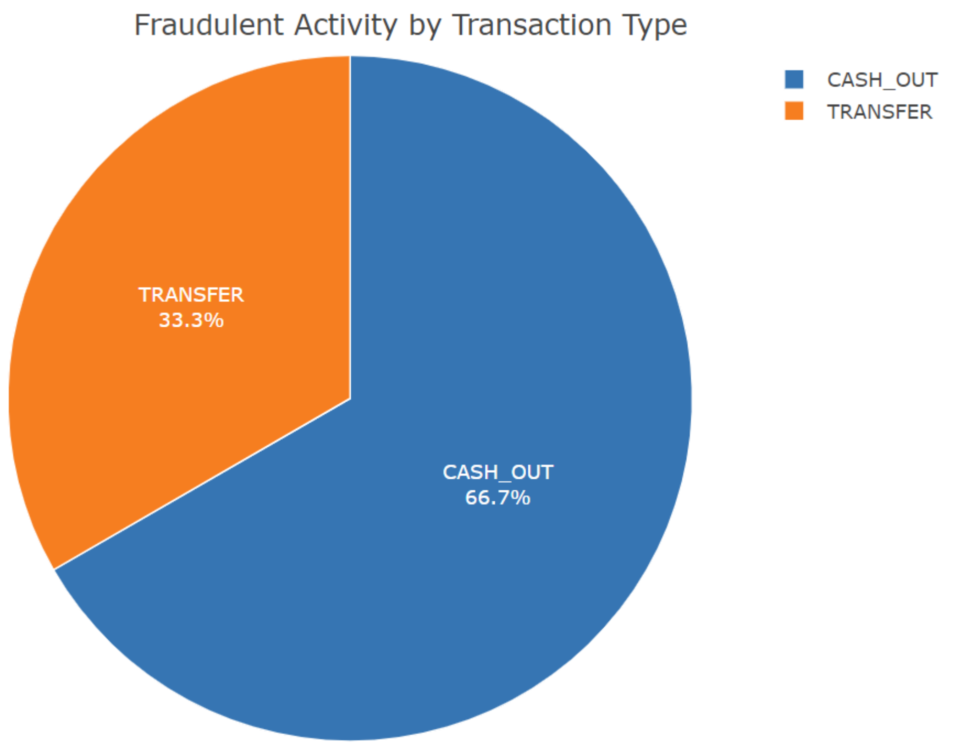

Figure 11: Pie Chart

How the research result answers class project question

The project research question is "What is the best data science model for detecting fraudulent activities in banking data?". The final most parsimonious regression iteration in this research shows that logistic regression is appropriate for detecting fraudulent activities in financial data. The model has a McFadden R-squared of 0.971051 and an AIC value of 13.367, which indicates a highly significant regression model. Therefore, the class project question has been answered.

Limitations of the analysis

The financial data set used for this research is synthetic. It was generated by a data simulator named PaySim. Therefore, the predictive model resulted from this analysis might not be able to make predictions with statistically significant precision if it is used with real financial data set. Also, due to the existence of a confounding variable, the resulting regression model might be either overestimating or underestimating the existence of fraudulent activities. This confounding variable prevents the model from making predictions with a high degree of precision. Also, this predictive model could not be used with financial fraud data that has more than one dependent variable or dependent variables that have more than two levels. Last, this predictive model cannot be used to make future predictions of financial fraudulent activities, because it is not a time series model.

Suggestions for a subsequent analysis

There are still many other regression models and machine learning algorithms to be explored for new possible ways of detecting fraudulent activities in financial data. However, to find the most parsimonious predictive model, the data that the model is trained on must first be the source of primary data. This data must be collected specifically for the intended research. Control experiments must be carried out to eliminate all confounding variables. The data must have a high degree of randomization, very minimal to no bias, very minimal to no outliers, and a lack of ethical issues. The possible statistical models to be explored next are K-Nearest Neighbor, Decision Tree, Neural Networks, Naïve Bayes classifier, and Benford's Law. The next steps are to explore these possibilities.

Discussion of social responsibility

This research could not have a negative social impact on individuals or institutions if the findings are not improperly used. It is because the secondary data used in this research is synthetic. The origin of the data source has been erased by the simulator and is not traceable. However, the data science model discovered from this research should not be used with real financial data from any financial institution. It is because the result of this research is based completely on a set of synthetic financial data. No one should be improperly using the findings from this research.

Discussion of ethics

This is the most ethical financial crime research. It is because neither the subjects of financial crime investigations nor the financial institutions are involved. It is because the secondary data set used in this research is synthetic. The data does not represent any individuals, institutions, or organizations. However, no one should improperly be using the findings from this research. Because it is unethical to have flawed predictions that could cause unnecessary investigations of individual account activities and unnecessary interrogations of individuals by authorities.

Conclusion

Several lessons were learned here. To find the most robust and parsimonious predictive model, the data scientist must follow these steps. First, make sure the sample data has minimal multicollinearity, a high degree of randomization, and no missing data value. Second, during model selection, make sure the data meet all the basic model assumptions. Third, make sure to carefully select a regression model that will not make too many biased predictions. Fourth, always use stepwise regression to find the best and the most parsimonious model. Last, do not include too many insignificant independent variables, because it could cause spurious causation. As Charles Wheelan has mentioned in chapter 12 of his Naked Statistics book, “If you put enough junk variables in a regression equation, one of them is bound to meet the threshold for statistical significance just by chance.”

Work Cited/References:

Allison, P. (September 10, 2012) - When Can You Safely Ignore Multicollinearity?. Retrieved from https://statisticalhorizons.com/multicollinearity (Accessed May 1st, 2020)

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

Detection of Fraud in Financial Statements: French Companies as a Case Study. Retrieved from https://www.researchgate.net/publication/272958533_Detection_of_Fraud_in_Financial_Statements_French_Companies_as_a_Case_Study (Accessed May 1st, 2020)

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

Marques, J. F. Oliveira (2015). Risk Analysis in Money Laundering A Case Study. Retrieved from https://fenix.tecnico.ulisboa.pt/downloadFile/563345090414324/resumo.pdf (Accessed May 1st, 2020)

Wheelan, C. J. (2013). Naked statistics: stripping the dread from the data (First Edition).
New York: W.W. Norton & Company, Inc.,
www.wwnorton.com

Wilson, J. Holton (August 2009). An Analytical Approach to Detecting Insurance Fraud Using Logistic Regression. Retrieved from http://connection.ebscohost.com/c/articles/46793982/analytical-approach-detecting-insurance-fraud-using-logistic-regression (Accessed May 1st, 2020)