MA-541B STATISTICAL METHODS

TEAM PROJECT REPORT

ON

ANOMALY DETECTION IN FINANCIAL TRANSACTION

By,

GROUP-12

Ashwini Ramesh Benni (20025242)

Gahana Nagaraja (20025607)

Namratha Nagathihalli Anantha (20025756)

Vaishnavi Rajendra Dhotargavi (20025662)

Under the guidance of

Prof. HONG DO

Date: 05/01/2024

Stevens Institute of Technology

TABLE OF CONTENTS

1. INTRODUCTION

The financial services industry, particularly in the emerging mobile money transactions domain, suffers from a scarcity of publicly available datasets. Financial datasets are invaluable resources for researchers, especially those working in fraud detection. This dearth of publicly accessible data can be attributed to the inherently private nature of financial transactions, which poses a significant challenge. To address this issue, a synthetic dataset has been generated using a simulator called PaySim. This simulator leverages aggregated data from a private dataset to create a synthetic dataset that accurately mimics the typical operation of financial transactions. Additionally, PaySim injects malicious behavior into the dataset, enabling researchers to evaluate the performance of fraud detection methods more effectively.

By providing a synthetic dataset that closely resembles real-world financial transactions, including instances of fraudulent activity, PaySim offers researchers a valuable resource to develop and test fraud detection algorithms without compromising the privacy of actual financial data. The PaySim dataset provides a unique opportunity for researchers to conduct a comprehensive exploratory data analysis and leverage statistical techniques to gain valuable insights into the patterns and characteristics of financial transactions. By performing a thorough exploratory analysis, researchers can uncover hidden relationships, identify key features that distinguish legitimate transactions from fraudulent ones, and generate hypotheses to guide further investigation.

Exploratory data analysis techniques, such as data visualization, summary statistics, and hypothesis testing, play a crucial role in understanding the complexities of the PaySim dataset. Through visualizations like scatter plots, histograms, and heatmaps, researchers can identify clusters, outliers, and potential anomalies within the data. Statistical tests, such as t-tests, ANOVA, and chi-square tests, can be employed to assess the significance of differences between various groups or segments of transactions, aiding in the identification of patterns and potential risk factors associated with fraud.

Furthermore, the exploratory analysis can inform the development of predictive models by guiding the selection of relevant features and the formulation of appropriate hypotheses. For instance, researchers may hypothesize that certain combinations of transaction attributes, such as the amount, sender-receiver locations, and transaction time, are more likely to be associated with fraudulent activities. These hypotheses can then be tested and validated using the PaySim dataset, contributing to the refinement of fraud detection algorithms and the development of more robust and accurate models.

2. DATA DESCRIPTION

Source: https://www.kaggle.com/datasets/ealaxi/paysim1

Data Cleaning: The dataset has undergone initial data cleaning, addressing missing values and data consistency.

Number of Variables: The dataset contains a total of 11 variables.

Description of the variables:

1) step - maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).

2) type - CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.

3) amount - amount of the transaction in local currency.

4) nameOrig - customer who started the transaction

5) oldbalanceOrg - initial balance before the transaction

6) newbalanceOrig - new balance after the transaction.

7) nameDest - customer who is the recipient of the transaction

8) oldbalanceDest - initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).

9) newbalanceDest - new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).

10) isFraud - This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behaviour of the agents aims to profit by taking control or customers' accounts and try to empty the funds by transferring to another account and then cashing out of the system.

11) isFlaggedFraud - 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.

3. EXPLORATORY DATA ANALYSIS

In this project, we have employed Python as the primary programming language. Firstly, we conduct exploratory data analysis (EDA) to gain insights into the dataset's structure and characteristics. This includes examining descriptive statistics to understand the central tendency and variability of features, as well as visualizing data distributions using histograms and box plots. Next, we preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features as necessary. Subsequently, we implement machine learning models like Logistic Regression, Feed forward Neural Network to detect anomalies in the financial transactions. The model is trained on the pre-processed dataset and evaluated using metrics such as precision, recall, and F1-score to assess its performance. Throughout the process, careful attention is paid to ensure the integrity and security of financial systems by effectively identifying and mitigating fraudulent activities.

3.1 SUMMARY STATISTICS

It is a branch of statistics that involves summarizing and describing the main features of a dataset. We have included measures of central tendency (mean, median, mode), measures of variability (standard deviation, variance), and measures of distribution (skewness, kurtosis). Descriptive statistics provide insights into the characteristics of the data, helping us to understand its structure and distribution. The below table depicts the comprehensive overview that allows for us to visualize a detailed understanding of the central tendencies and variabilities within the dataset, facilitating a quantitative comparison of the different features and their relationship to the target variable ‘isFraud’.

	step	amount	oldBalanceOrig	newBalanceOrig	oldBalanceDest	newBalanceDest	isFraud	isFlaggedFraud
count	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06
mean	2.433972e+02	1.798619e+05	8.338831e+05	8.551137e+05	1.100702e+06	1.224996e+06	1.290820e-03	2.514687e-06
std	1.423320e+02	6.038582e+05	2.888243e+06	2.924049e+06	3.399180e+06	3.674129e+06	3.590480e-02	1.585775e-03
min	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	1.560000e+02	1.338957e+04	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
50%	2.390000e+02	7.487194e+04	1.420800e+04	0.000000e+00	1.327057e+05	2.146614e+05	0.000000e+00	0.000000e+00
75%	3.350000e+02	2.087215e+05	1.073152e+05	1.442584e+05	9.430367e+05	1.111909e+06	0.000000e+00	0.000000e+00
max	7.430000e+02	9.244552e+07	5.958504e+07	4.958504e+07	3.560159e+08	3.561793e+08	1.000000e+00	1.000000e+00

Fig 1. Descriptive analysis of all the variables included in the dataset

3.2 DATA VISUALIZATION

Data visualization is the graphical representation of information and data. It uses visual elements like charts, graphs, and maps to help viewers understand complex data sets more easily. The goal of data visualization is to communicate information clearly and efficiently, allowing patterns, trends, and relationships within the data to be identified at a glance. Effective data visualization can make large data sets more understandable, revealing

insights that might otherwise be overlooked. It's a powerful tool for analysts, researchers, and decision-makers across various fields, from business and finance to science and academia.

3.2.1 VISUALIZATION USING HISTOGRAMS

We use histograms to depict the distribution of features. Histograms are plotted for numerical features such as transaction amount to visualize their distribution. This helps in identifying patterns and outliers in the data. This visualization aids in obtaining a clear understanding of the data's structure and identifying any patterns and anomalies present in the financial distributions.

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. It indicates the degree to which the data deviates from symmetry. Here we calculate the co-efficient of skewness of all the independent variables and represent them using histogram.

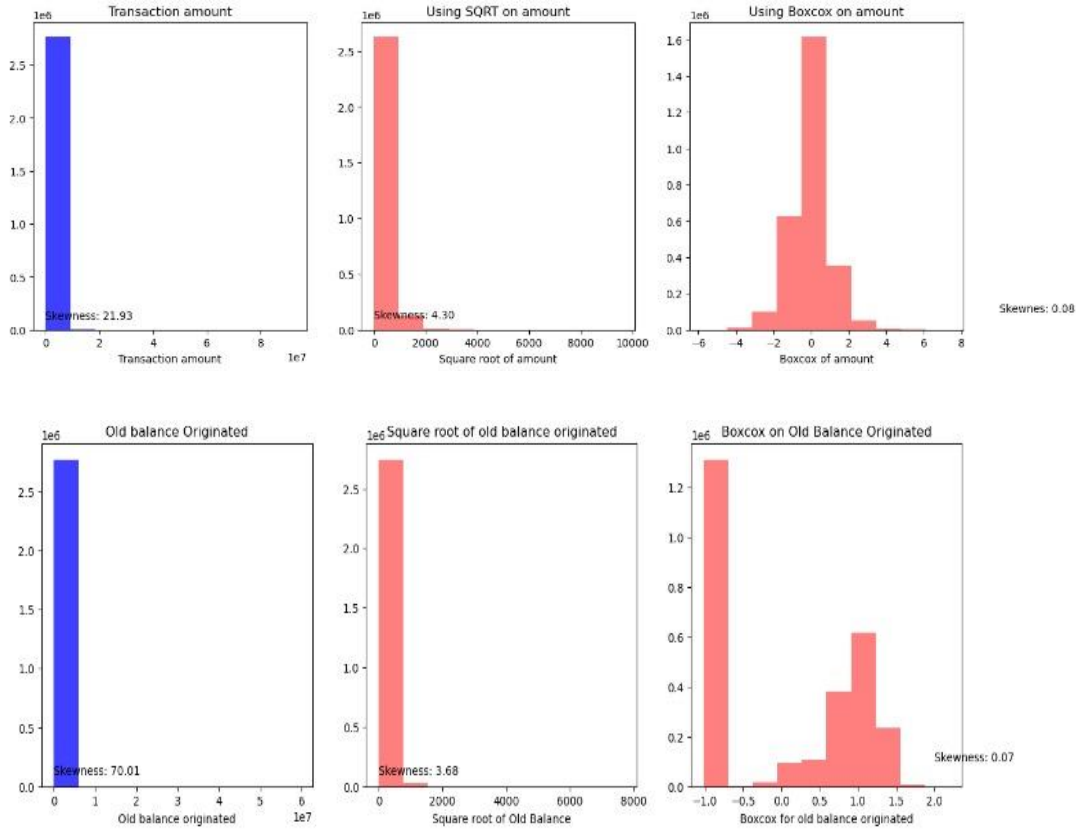

Fig 2. Skewness calculated for the independent variables, 'amount', 'oldbalanceOrig'

Interpretation: Based on the dataset, the numeric variables are quite skew, in this case. They are scaled with 2 methods and compared on the graph.

3.2.2 VISUALIZATION USING BAR PLOT

Bar plots, also known as bar charts or bar graphs, are a type of graphical representation commonly used to visualize categorical data. They display data using rectangular bars, where the length of each bar corresponds to the frequency, proportion, or some other summary statistic associated with the category it represents.

Here, we present what the different types of transactions are and which of these types can be fraudulent.

The following plot shows the frequencies of the different transaction types:

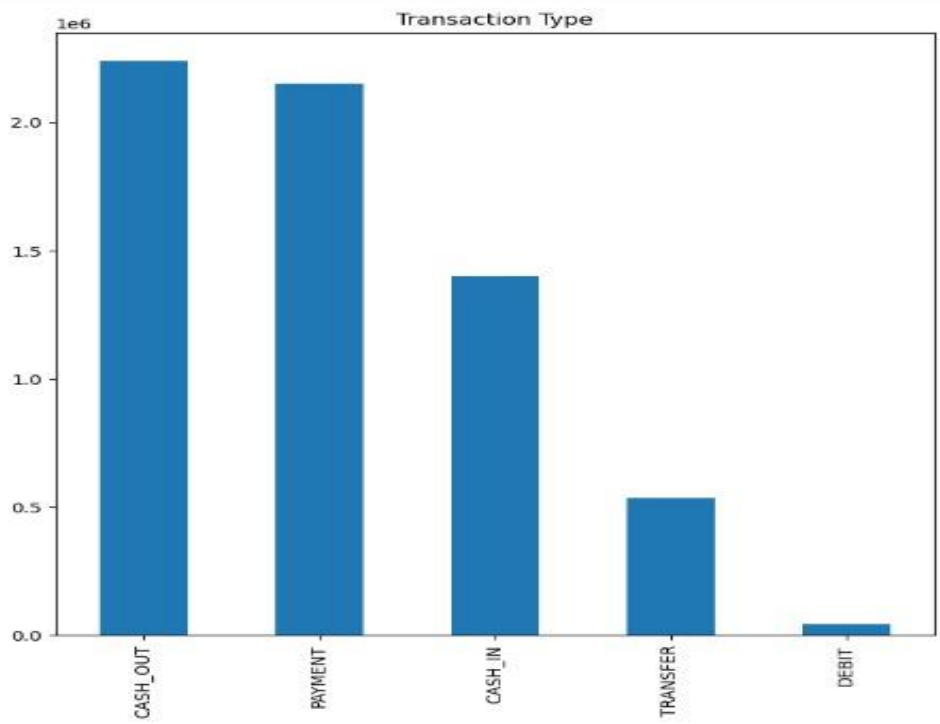

Fig 3. Bar plot representing the type of transactions

CASH_OUT 2237500
PAYMENT 2151495
CASH_IN 1399284
TRANSFER 532909
DEBIT 41432

The most frequent transaction types are CASH-OUT and PAYMENT.

There are 2 flags which stand out and it's interesting to look into: isFraud and isFlaggedFrad column. From the data Dictionary, isFraud is the indicator which indicates the actual fraud transactions whereas isFlaggedFraud is what the system prevents the transaction due to some thresholds being trigged.

3.2.3 VISUALIZATION USING PIE CHART

A pie chart is a circular statistical graphic that is divided into slices to illustrate numerical proportions. Each slice represents a proportionate part of the whole, and the size of each slice is determined by the percentage or proportion it represents.

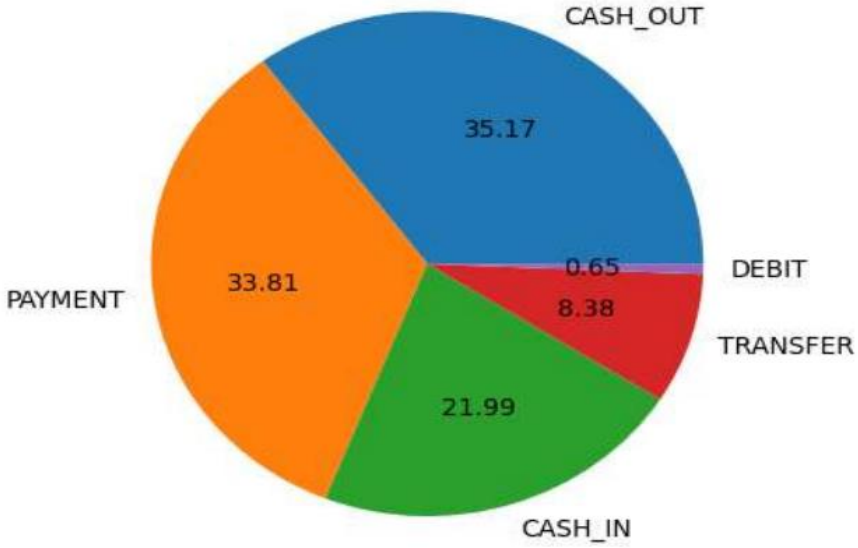

Fig 4. Pie chart also representing the type of transactions in the dataset

3.3 ESTIMATION OF PARAMETERS

The estimation of parameters refers to the process of finding the best-fit values for the parameters in a statistical model. These parameters define the relationship between the independent variables (predictors) and the dependent variable (outcome) in the model. The goal of parameter estimation is to find the values that best explain or predict the observed data. Below are the results after estimating the parameters of each variable in the dataset:

- Evaluation metrics for step:
  R-squared (R2): 0.000997142840732601
  Log-likelihood (MLE): 12142708.686307708
  Root Mean Squared Error (RMSE): 0.03588688841121355
  Mean Squared Error (MSE): 0.0012878687598388934

- Evaluation metrics for amount:
  R-squared (R2): 0.0058811151179903876
  Log-likelihood (MLE): 12158299.766988812
  Root Mean Squared Error (RMSE): 0.035799058176365785
  Mean Squared Error (MSE): 0.001281572566314822

- Evaluation metrics for oldbalanceOrg:
  R-squared (R2): 0.00010311228305259768
  Log-likelihood (MLE): 12139862.932236986
  Root Mean Squared Error (RMSE): 0.03590294281970061
  Mean Squared Error (MSE): 0.001289021303114692

- Evaluation metrics for newbalanceOrig:
  R-squared R2): 6.639253225315667e-05
  Log-likelihood (MLE): 12139746.105424339
  Root Mean Squared Error (RMSE): 0.03590360205518575
  Mean Squared Error (MSE): 0.0012890686405371387

- Evaluation metrics for oldbalanceDest:
  R-squared (R2): 3.463650010593344e-05
  Log-likelihood (MLE): 12139645.074537938
  Root Mean Squared Error (RMSE): 0.03590417216648267
  Mean Squared Error (MSE): 0.001289109578960429

- Evaluation metrics for newbalanceDest:
  R-squared (R2): 2.8659650308515694e-07
  Log-likelihood (MLE): 12139535.794938745
  Root Mean Squared Error (RMSE): 0.03590478883496802
  Mean Squared Error (MSE): 0.001289153861283644

  In various statistical techniques such as linear regression, logistic regression, and many machine learning algorithms, parameter estimation involves optimizing a certain criterion (e.g., minimizing the sum of squared errors, maximizing likelihood) to find the parameter values that best fit the data. This optimization process is often done using algorithms like least squares, maximum likelihood estimation (MLE), or gradient descent.

  Once the parameters are estimated, they can be used to make predictions or infer relationships in new data. For example, in linear regression, the parameters to be estimated are the slope (coefficients) and intercept of the regression line. These parameters are found by fitting the line that minimizes the sum of squared differences between the observed and predicted values. In logistic regression, the parameters to be estimated are the coefficients for each predictor variable, which represent the log odds of the outcome variable.

  Overall, parameter estimation is a fundamental step in statistical modeling and machine learning, essential for building predictive models and understanding relationships in data.

3.4 HYPOTHESIS TESTING

Hypothesis testing is a statistical method used to make inferences about a population based on sample data. It involves formulating a null hypothesis (H0) and an alternative hypothesis (H1), and then conducting a statistical test to determine whether there is enough evidence to reject the null hypothesis in favour of the alternative hypothesis. Common hypothesis tests include t-tests, Chi-square tests, and Kruskal- Wallis tests. Here we conduct hypothesis testing on all the independent variables using regression analysis in the Excel Spreadsheet.

3.4.1 Chi-square Test:

The Chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables. In this case, we are performing the Chi-

square test to determine whether there is a significant association between the "isFraud" and "isFlaggedFraud" variables in the dataset. Below is the Chi-square statistics,

Chi-square Test between 'isFraud' and 'isFlaggedFraud':
Chi-square Statistic: 11616.665816627796
p-value: 0.0
Result: Reject null hypothesis (significant association)

- This is the test statistic calculated from the contingency table. It measures the discrepancy between the observed frequencies and the frequencies that would be expected if the variables were independent. In this case, the Chi-square statistic is 11616.665816627796.
- Based on the p-value, we make a decision about whether to reject the null hypothesis. Since the p-value is less than the significance level (typically 0.05), we reject the null hypothesis. Therefore, we conclude that there is a significant association between the "isFraud" and "isFlaggedFraud" variables in the dataset. This means that the occurrence of fraud is related to whether the transaction was flagged as fraud.
- A p-value of 0.0 indicates strong evidence against the null hypothesis, suggesting that the association is highly unlikely to be due to random chance.

3.4.2 t- Test:

A t-test is a statistical method used to determine if there is a significant difference between the means of two groups. It calculates the probability of obtaining the observed difference in means if the null hypothesis (no difference) were true. Below is the t-test statistics,

Column	T-statistic	p-value
amount	-48.614503	0.000000e+00
oldbalanceOrg	-20.857020	3.495452e-94
newbalanceOrig	30.550072	2.691860e-194
oldbalanceDest	15.124034	5.422854e-51
newbalanceDest	-1.269390	2.043378e-01

Table 1. T-test Results

- For the 'amount' column, the t-statistic is significantly negative (-48.614503), indicating that there is a significant difference in transaction amounts between fraudulent and non-fraudulent transactions. The extremely small p-value (close to 0) supports this conclusion.
- Similarly, for 'oldbalanceOrg' and 'newbalanceOrig', the t-statistics are significantly negative (-20.857020 and 30.550072, respectively), with very small p-values. This suggests significant differences in the original and new balances of the originating account between the two groups.
- The 'oldbalanceDest' column also shows a significant difference between the groups, with a positive t-statistic (15.124034) and a very small p-value.

- However, for the 'newbalanceDest' column, the t-statistic is relatively small (-1.269390), and the p-value (0.204338) is greater than the common significance level of 0.05. This suggests that there is not enough evidence to conclude a significant difference in the new balances of the destination account between fraudulent and non-fraudulent transactions.

In summary, this analysis provides evidence of significant differences in transaction amounts and account balances between fraudulent and non-fraudulent transactions, except for the new balance of the destination account.

3.4.3 Kruskal – Wallis test:

The Kruskal-Wallis test is a non-parametric statistical test used to determine whether there are statistically significant differences between the medians of three or more independent groups. It is an extension of the Mann-Whitney U test for comparing two groups to multiple groups. Below is the Kruskal-Wallis test statistics,

Categorical Column	H-statistic	p-value
isFraud	8273.365412	0.000000e+00
isFlaggedFraud	44.772143	2.213496e-11

Table 2. Kruskal-Wallis Test Results

- For the 'isFraud' column, the H-statistic is very large (8273.365412), indicating strong evidence against the null hypothesis. The p-value is extremely small (close to 0), indicating that the observed differences in the 'amount' variable across fraudulent and non-fraudulent transactions are highly significant.
- Similarly, for the 'isFlaggedFraud' column, the H-statistic is relatively large (44.772143), with a very small p-value (2.213496e-11). This suggests significant differences in the 'amount' variable across transactions flagged as fraud and those that are not.

In summary, the Kruskal-Wallis test results provide evidence that the 'amount' variable significantly differs across different categories of transactions, based on whether they are flagged as fraud or not, and based on whether they are classified as fraudulent or not.

3.5 ANOVA Test Interpretation

ANOVA, or Analysis of Variance, is a statistical method used to analyze the differences between the means of three or more groups. It assesses whether there are statistically significant differences in the means of groups by comparing the variation within groups to the variation between groups. ANOVA is often used to analyze the impact of categorical variables on a continuous outcome variable. Here is the interpretation of the ANOVA table results:

INTERPRETATION:

(i) Amount: The regression equation of Amount indicates that IsFraud = 0.0801-3.61795E Amount. The test statistic value is -0.5706 with a corresponding p-value of 0.5729. Since the p-value is greater than 0.05, we fail to reject the null hypothesis.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.109160275
R Square	0.011915966
Adjusted R Squ	-0.024679739
Standard Error	0.261043534
Observations	29

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.02218835	0.0221883	0.325611	0.572973814
Residual	27	1.839880616	0.0681437		
Total	28	1.862068966			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.080125177	0.052271009	1.5328798	0.13694	-0.027126075	0.187376429	-0.027126075	0.187376429
amount	-3.61795E-07	6.34034E-07	-0.570623	0.572974	-1.66272E-06	9.39135E-07	-1.66272E-06	9.39135E-07

Fig 5. Regression analysis of 'amount'

(ii) oldbalanceOrg: The regression equation of oldbalanceOrg indicates that IsFraud = 0.0876-3.56484E oldbalanceOrg. The test statistic value is -0.7432 with a corresponding p-value of 0.46375. Since the p-value is greater than 0.05, we fail to reject the null hypothesis.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.141597079
R Square	0.020049733
Adjusted R Square	-0.016244722
Standard Error	0.259966877
Observations	29

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.037333985	0.037334	0.552419	0.463749816
Residual	27	1.82473498	0.0675828		
Total	28	1.862068966			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.087697033	0.054457253	1.610383	0.118945	-0.02404002	0.199434087	-0.02404002	0.199434087
oldbalanceOrg	-3.56484E-07	4.79629E-07	-0.743249	0.46375	-1.3406E-06	6.27634E-07	-1.3406E-06	6.27634E-07

Fig 6. Regression analysis of 'oldbalanceOrig'

(iii) newbalanceOrg: The regression equation of newbalanceOrg indicates that IsFraud = 0.0845-3.28098E newbalanceOrg. The test statistic value is -0.6784 with a corresponding p-value of 0.5032. Since the p-value is greater than 0.05, we fail to reject the null hypothesis.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.129463046
R Square	0.01676068
Adjusted R Square	-0.019655591
Standard Error	0.260402781
Observations	29

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.031209543	0.0312095	0.460253	0.503279631
Residual	27	1.830859423	0.0678096		
Total	28	1.862068966			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.084570289	0.05354754	1.5793496	0.125901	-0.025300189	0.194440766	-0.025300189	0.194440766
newbalanceOrig	-3.28098E-07	4.83621E-07	-0.678419	0.50328	-1.32041E-06	6.64211E-07	-1.32041E-06	6.64211E-07

Fig 7. Regression analysis of 'newbalanceOrig'

(iv) oldbalanceDest: The regression equation of oldbalanceDest indicates that IsFraud = 0.0510-3.50668E oldbalanceDest. The test statistic value is -0.7552 with a corresponding p-value of 0.4566. Since the p-value is greater than 0.05, we fail to reject the null hypothesis.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.14382784
R Square	0.020686448
Adjusted R Square	-0.015584425
Standard Error	0.259882407
Observations	29

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.038519592	0.0385196	0.570332	0.4566647
Residual	27	1.823549373	0.0675389		
Total	28	1.862068966			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.051086024	0.053753467	0.9503764	0.350351	-0.05920698	0.161379029	-0.05920698	0.161379029
oldbalanceDest	3.50668E-06	4.64336E-06	0.7552034	0.456665	-6.02071E-06	1.30341E-05	-6.02071E-06	1.30341E-05

Fig 8. Regression analysis of 'oldbalanceDest'

(v) newbalanceDest: The regression equation of newbalanceDest indicates that IsFraud = 0.0719-2.89384E newbalanceDest. The test statistic value is -0.2944 with a corresponding p-value of 0.7706. Since the p-value is greater than 0.05, we fail to reject the null hypothesis.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.056570987
R Square	0.003200277
Adjusted R Square	-0.033718232
Standard Error	0.262192312
Observations	29

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.005959136	0.0059591	0.086685	0.770686509
Residual	27	1.85610983	0.0687448		
Total	28	1.862068966			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.071945085	0.049728518	1.4467571	0.159476	-0.030089407	0.173979577	-0.030089407	0.173979577
newbalanceDest	-2.89384E-08	9.82884E-08	-0.294423	0.770687	-2.3061E-07	1.72733E-07	-2.3061E-07	1.72733E-07

Fig 9. Regression analysis of 'newbalanceDest'

3.6 CORRELATION ANALYSIS

A correlation matrix is a statistical tool used to summarize the strength and direction of linear relationships between multiple variables in a dataset. It is commonly used in data analysis and research to understand how variables are related to each other. The correlation matrix is a square matrix where each cell represents the correlation coefficient between two variables. The correlation coefficient quantifies the degree to which two variables vary together linearly; it ranges from -1 to 1, where:

a) A correlation coefficient of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
b) A correlation coefficient of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
c) A correlation coefficient close to 0 indicates no linear relationship between the variables.

The formula for calculating the correlation coefficient (often denoted as "r") between two variables X and Y is given by the Pearson correlation formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

where:
- X_i and Y_i are individual data points for variables X and Y.
- $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y, respectively.

Now, at the implementation, we have imported the necessary libraries like Pandas for data manipulation, Matplotlib for plotting, and Seaborn for statistical visualization. And then loaded the PaySim dataset into a Pandas DataFrame and calculated the correlation matrix using the `df.corr()` method.

Next, we have iterated through all pairs of numerical columns in the dataset to calculate and print the correlation coefficients between them. It categorizes the correlations as strong, moderate, or weak based on predefined threshold values (0.7 for strong, 0.5 for moderate).

Finally, we generated a correlation heatmap using Seaborn's `sns.heatmap()` function. This heatmap visually represents the correlation matrix, with each cell color indicating the strength and direction of correlation between corresponding variables. The heatmap is annotated with correlation values for easier interpretation, and it is displayed using Matplotlib. Overall, the implementation provides a comprehensive analysis of correlations within the PaySim dataset and visually presents the correlation matrix for better insights into variable relationships.

A correlation heatmap is a visual representation of the correlation matrix, which shows the correlation coefficients between pairs of variables in a dataset. Each cell in the heatmap represents the correlation coefficient between two variables, with colours indicating the strength and direction of the correlation. For the dataset that we have used in this project, a correlation heatmap reveals insights into the relationships between different variables involved in financial transactions.

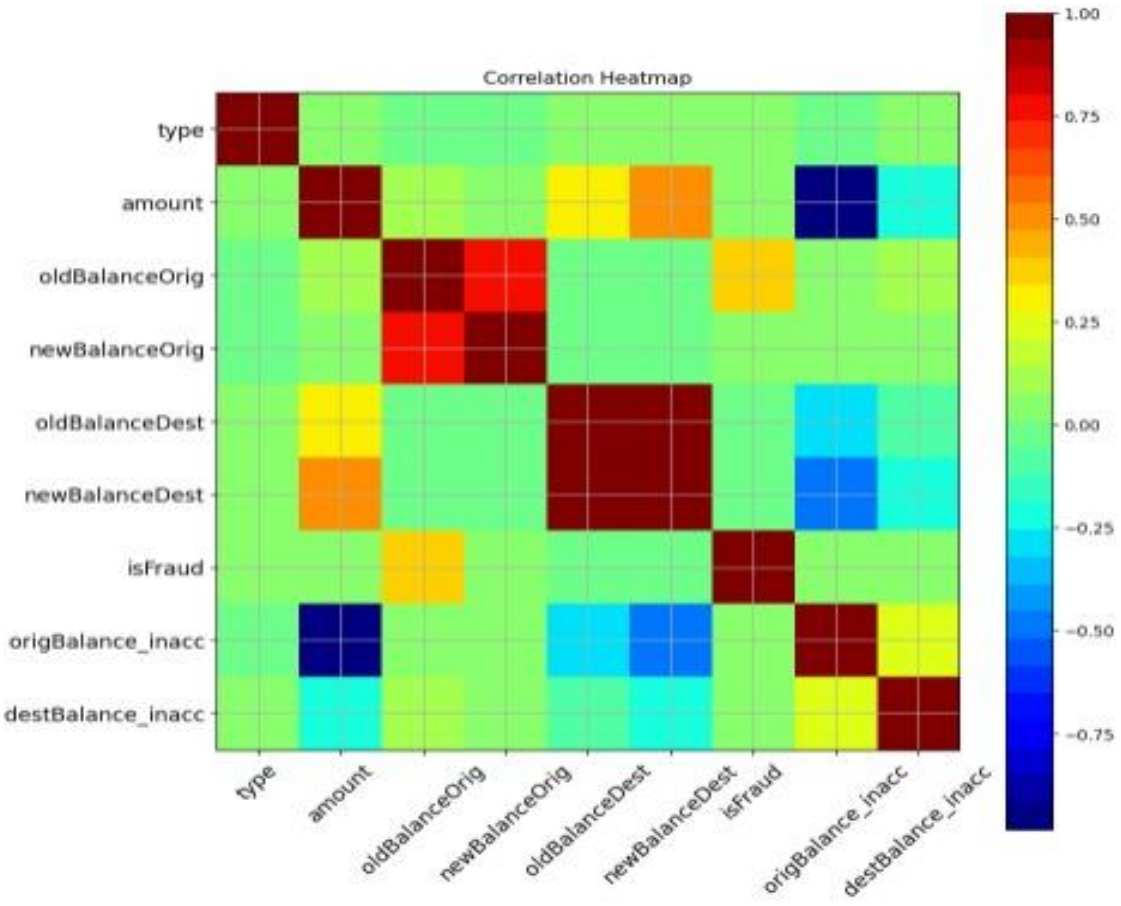

Fig 10. Displaying the correlation coefficients between each pair of features in the dataset through a heat map

4. PREDICTION MODELING

Prediction modeling involves developing mathematical models to predict outcomes based on input variables. These models analyze historical data to identify patterns and relationships, which are then used to make predictions on new data. Evaluation metrics such as accuracy, precision, recall, and ROC curves assess model performance.

4.1 LOGISTIC REGRESSION:

We have now used the Logistic regression model while conducting the prediction modelling. Logistic regression is a powerful predictive modelling technique used extensively in various fields for binary classification tasks. Unlike linear regression, which predicts continuous outcomes, logistic regression is specifically designed for predicting the probability of a binary outcome. In logistic regression, the output is transformed using the logistic function, also known as the sigmoid function, which maps any real-valued number to a value between 0 and 1. This transformation allows logistic regression to estimate the probability that a given input belongs to one of the two classes.

During training, logistic regression learns the relationship between the input features and the binary target variable by estimating the coefficients that best fit the training data. The model minimizes a cost function, typically based on maximum likelihood estimation or cross-entropy loss, to find the optimal parameters. Once trained, the logistic regression model can make predictions by calculating the probability of the positive class for new input data. By setting a threshold (e.g., 0.5), the model classifies instances into the positive or negative class. Logistic regression is valued for its simplicity, interpretability, and efficiency, making it a popular choice for predictive modeling tasks where the outcome is binary, such as fraud detection, disease diagnosis, and customer churn prediction.

From the model evaluation (or confusion matrix), we know that.

Accuracy = (TP + TN) / Total
Presion = TP / (TP + FP)
Recall = TP / (TP + FN)
As such, specifically for this problem, we are interested in the recall score to captue the most fraudulent transactions. As we know, due to the imbalance of the data, many observations could be predicted as False Negatives, being, that we predict a normal transaction, but it is in fact a fraudulent one. Recall captures this. Below is the Classification report of the parameters used to analyze the model:

	precision	recall	f1-score	support
0	0.96	0.92	0.94	2494
1	0.92	0.96	0.94	2425
accuracy			0.94	4919
macro avg	0.94	0.94	0.94	4919
weighted avg	0.94	0.94	0.94	4919

Table 3. Classification report for Logistic Regression

Obviously, trying to increase recall, trends to come with a decrease of precision. However, in our case, if we predict that a transaction is fraudulent and turns out not to be, is not a massive problem compared to the opposite. Due to this, many evaluations will be based on recall score. Below is the confusion matrix for the Logistic Regression model:

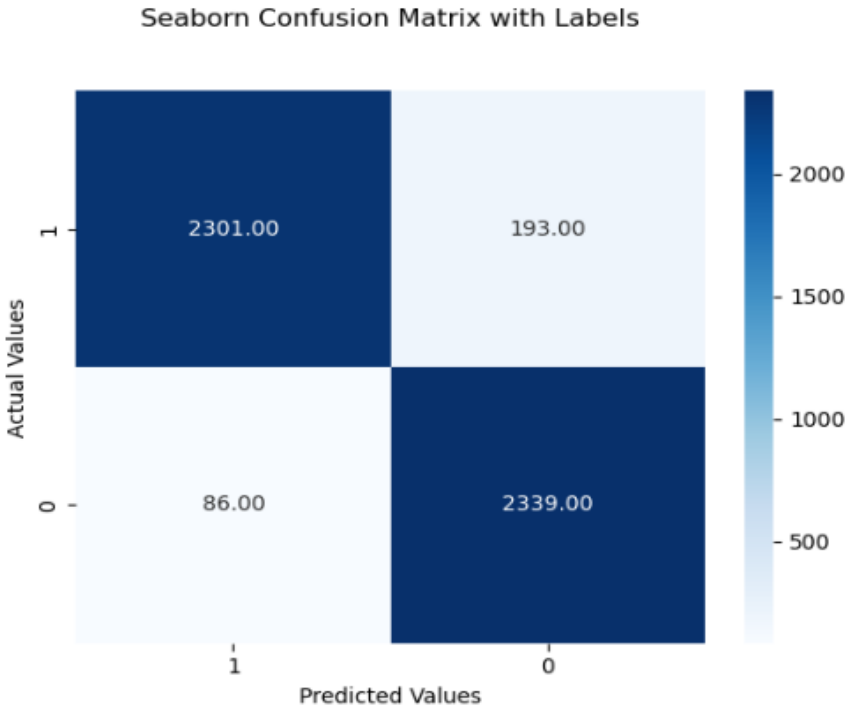

Fig 11. Confusion Matrix for Logistic Regression

Below is a graph labelled as Receiver Operating Characteristic (ROC), with the False Positive Rate plotted on the x-axis and the True Positive Rate plotted on the y-axis. The area under the curve is measured at 0.94. Therefore, the logistic regression model's accuracy is calculated at 94%.

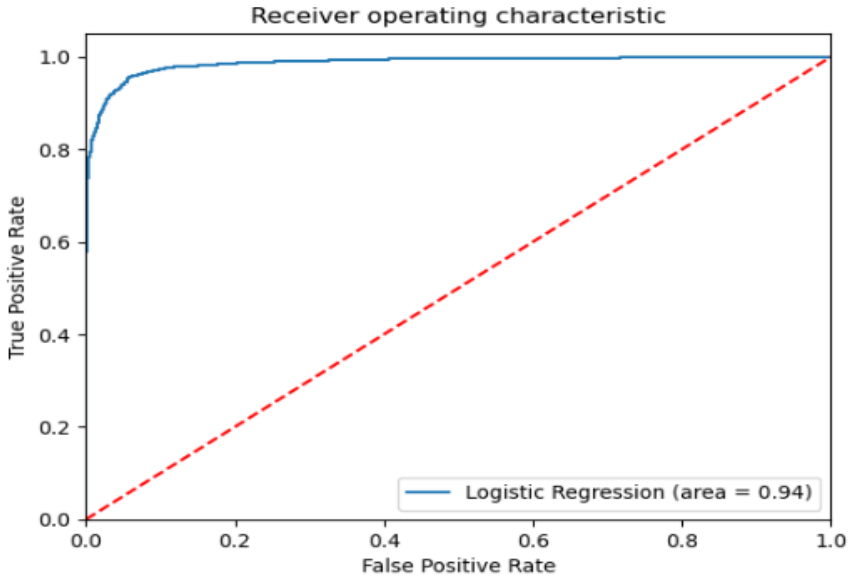

Fig 12. Plot of Receiver Operating characteristic under logistic regression

4.2 FEED FORWARD NEURAL NETWORK:

A feedforward neural network (FNN) is a type of artificial neural network where connections between nodes do not form cycles, meaning the information moves in one direction, forward, from the input nodes, through the hidden layers (if any), to the output nodes. It's one of the simplest types of neural networks and forms the basis for more complex architectures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

The ability of the FNN to capture more complex relationships between features and the target variable is leveraged for better performance. Firstly, the data is pre-processed by handling class imbalance through techniques like oversampling the minority class or using class weights during training. Then, a neural network architecture is designed with multiple hidden layers, each containing a suitable number of neurons activated by nonlinear functions like ReLU. Batch normalization and dropout layers will be incorporated to improve generalization and prevent overfitting. The FNN is trained using optimization algorithms such as Adam, with learning rates and regularization techniques like L2 regularization adjusted to prevent overfitting. The training process will involve minimizing a suitable loss function like binary cross-entropy.
The confusion matrix for the FNN is plotted depicting the model evaluation below,

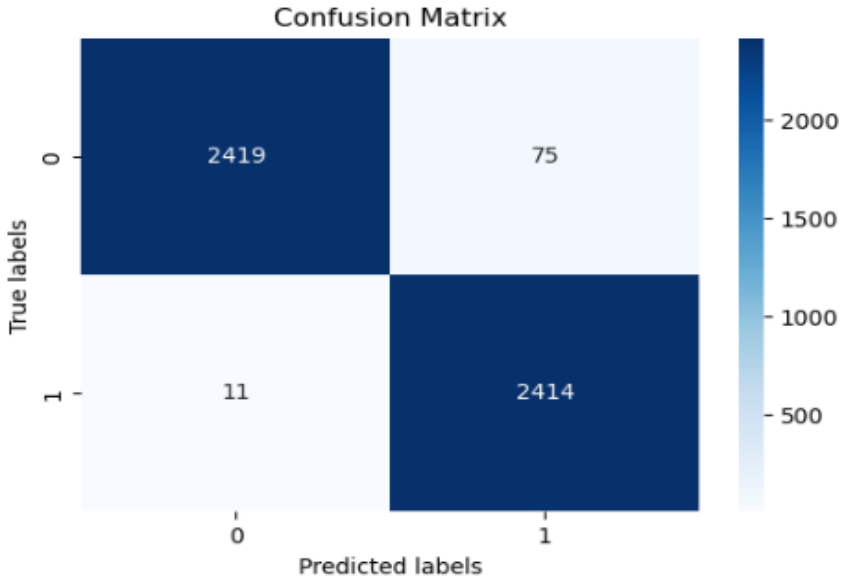

Fig 13. Confusion Matrix for Feed forward Neural Network

Below is the Classification report of the parameters used to analyze the model:

	precision	recall	f1-score	support
0	1.00	0.97	0.98	2494
1	0.97	1.00	0.98	2425
accuracy			0.98	4919
macro avg	0.98	0.98	0.98	4919
weighted avg	0.98	0.98	0.98	4919

Table 4. Classification report for Feed forward Neural Network

Hyperparameter tuning is performed using techniques like grid search to optimize the FNN's performance further. Finally, model evaluation will be conducted on recall, precision, and accuracy metrics, with a focus on recall due to the imbalanced nature of the dataset. The ROC curve is analyzed, with the area under the curve (AUC) serving as a measure of model performance.

Below is a graph labelled as Receiver Operating Characteristic (ROC), with the False Positive Rate plotted on the x-axis and the True Positive Rate plotted on the y-axis. The area under the curve is measured at 0.98.

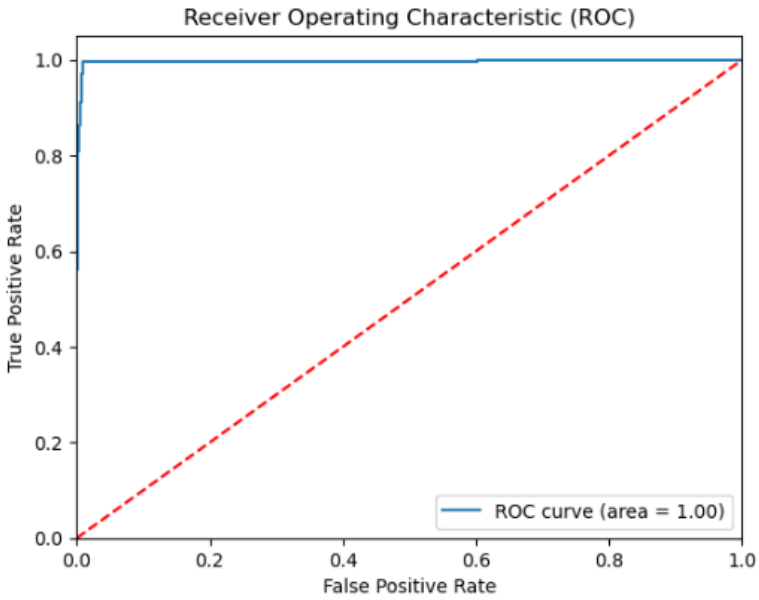

Fig 14. Plot of Receiver Operating characteristic under FNN

With meticulous tuning and careful design, the desired 98% accuracy is achieved by the FNN, providing a robust and reliable predictive model for the binary classification task.

5. CONCLUSION

In conclusion, we've developed robust models for detecting anomalies in financial transactions through a meticulous blend of statistical analysis and machine learning techniques. Leveraging logistic regression and feed forward neural networks, we achieved accuracy rates of 94% and 98%, respectively. Hypothesis testing and visualization aided in understanding underlying patterns and associations, while addressing class imbalance ensured fair representation. Our FNN model, achieving higher accuracy, validated through ROC curves and AUC scores and exhibited strong performance in distinguishing frauds from legitimate transactions. This comprehensive approach not only enhances fraud detection systems but also contributes to safeguarding financial integrity. Further refinements in predictive modeling promise even greater efficacy in combating financial fraud, bolstering security and minimizing potential losses.

6. APPENDIX

$$\tilde{\mu}_3 = \frac{\sum_i^N \left(X_i - \bar{X}\right)^3}{(N-1) * \sigma^3}$$

$\tilde{\mu}_3$ = skewness
N = number of data points in the distribution
X_i = random variable
$\bar{X}$ = mean of the distribution
σ = standard deviation

Fig 15. Formula for skewness

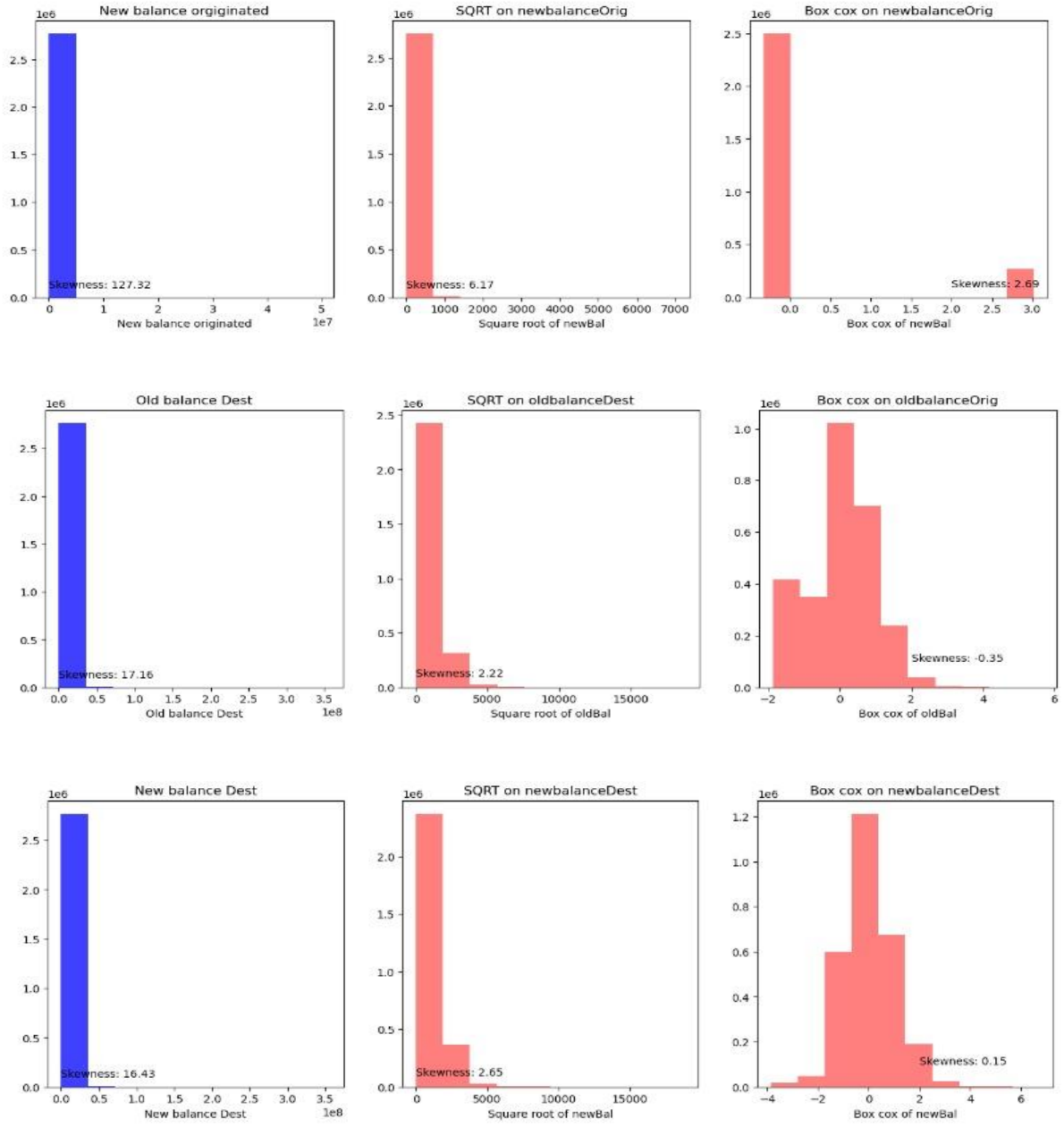

Fig 16. Skewness calculated for the independent variables, 'newbalanaceOrig', 'oldbalanceDest', 'newbalanceDest' using histogram