ALY 6040: Data mining

Online Fraud Detection

ALY 6040 – Data Mining

College of Professional Studies

Northeastern University - Vancouver

REPRESENTATIVES

Murtaza Vora.
Shamim Sherafati.
Milan Prajapati.
Abhilash Dikshit.
Minesh Patil.

Online Fraud Detection

1] Abstract

Online fraud has become a significant threat in the digital era as people increasingly rely on online transactions. This project focuses on analyzing a dataset on online fraud to gain insights into fraudulent behaviour and develop strategies for fraud detection and prevention. The dataset contains variables such as transaction type, amount, customer details, and transaction outcomes. The project includes exploratory data analysis (EDA) to understand the statistical measures, payment types, and fraud transaction distributions. Various data mining techniques and algorithms are applied, including data scaling using robust scaling and dimensionality reduction using Principal Component Analysis (PCA). Predictive algorithms such as decision trees, random forests, and logistic regression are utilized to classify fraudulent transactions. The performance of these algorithms is evaluated using accuracy metrics and ROC curves. The results show exceptional accuracy rates but highlight challenges such as false positives and false negatives. Addressing these challenges is crucial to improve the overall performance of fraud detection models. The project emphasizes the significance of data mining and machine learning in combating online fraud and provides valuable insights for developing robust fraud detection systems. Further research can be conducted to refine the models and enhance their performance, contributing to a safer and more secure online environment.

2] Introduction

In the current digital era, online fraud is becoming a bigger threat. As people rely more and more on online transactions, thieves are coming up with creative and clever ways to take advantage of the system and defraud unwary victims of their money and personal information. Online fraud may come in a variety of shapes and sizes, making it challenging for people and businesses to defend themselves. These forms can range from phishing schemes and identity theft to credit card fraud and bogus online stores.

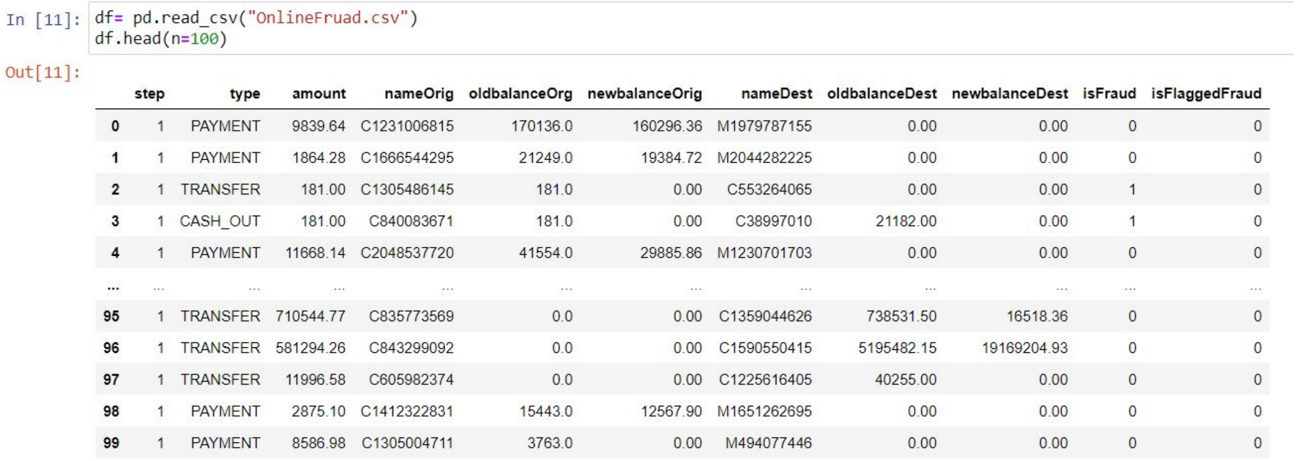
The rise of e-commerce and online transactions has led to a significant increase in payment fraud. According to a report by Nilson, global payment card losses reached $27.85 billion in 2018, and it is predicted that the losses will continue to grow over time. Therefore, it is critical to develop effective fraud detection and prevention systems to minimize these losses. Therefore, to protect yourself from these risks, it's crucial to grasp the characteristics of online fraud and the strategies that fraudsters employ.

In this situation, studying online fraud data might offer insightful information about the recurring themes, recurring trends, and recurring behaviours that underlie these illicit operations. we will explore the topic of online fraud data, its importance, and the tools and techniques that can be used to detect and prevent online fraud.

The variables defined in our dataset are listed below.

1. step: represents a unit of time where 1 step equals 1 hour
2. type: type of online transaction
3. amount: the amount of the transaction
4. nameOrig: customer starting the transaction
5. oldbalanceOrg: balance before the transaction
6. newbalanceOrig: balance after the transaction
7. nameDest: recipient of the transaction

8. oldbalanceDest: initial balance of the recipient before the transaction
9. newbalanceDest: the new balance of the recipient after the transaction
10. isFraud: fraud transaction



```
In [11]: df= pd.read_csv("OnlineFruad.csv")
         df.head(n=100)
```

Out[11]:

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.00	0.00	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.00	0.00	0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.00	0.00	1	0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.00	0.00	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.00	0.00	0	0
...	...	...	...	...	...	...	...	...	...	...	...
95	1	TRANSFER	710544.77	C835773569	0.0	0.00	C1359044626	738531.50	16518.36	0	0
96	1	TRANSFER	581294.26	C843299092	0.0	0.00	C1590550415	5195482.15	19169204.93	0	0
97	1	TRANSFER	11996.58	C605982374	0.0	0.00	C1225616405	40255.00	0.00	0	0
98	1	PAYMENT	2875.10	C1412322831	15443.0	12567.90	M1651262695	0.00	0.00	0	0
99	1	PAYMENT	8586.98	C1305004711	3763.0	0.00	M494077446	0.00	0.00	0	0

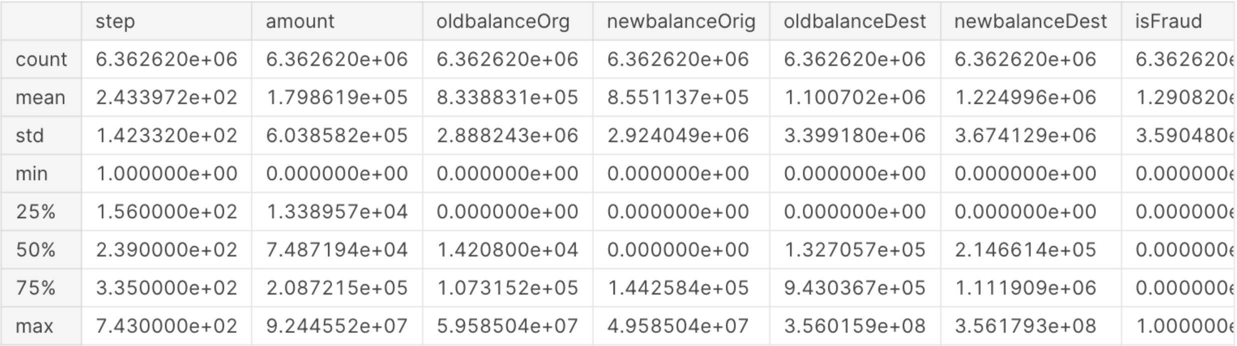
2.1 Variable table

3] Proposal

Have you ever questioned the methods used by online scammers to defraud unknowing victims out of their hard-earned money? As the number of online transactions has increased, fraudsters have developed creative new methods for committing their crimes. As a result, it is now more crucial than ever to analyze data on online fraud in order to comprehend the strategies and methods these criminals employ. The goal of this project proposal is to examine a dataset on online fraud to find patterns and trends that could aid in reducing fraud efforts in the future. We can get insights into fraudster behaviour, spot fraudulent conduct in real time, and ultimately stop financial loss for both individuals and companies by utilizing powerful data analytics and machine learning approaches.

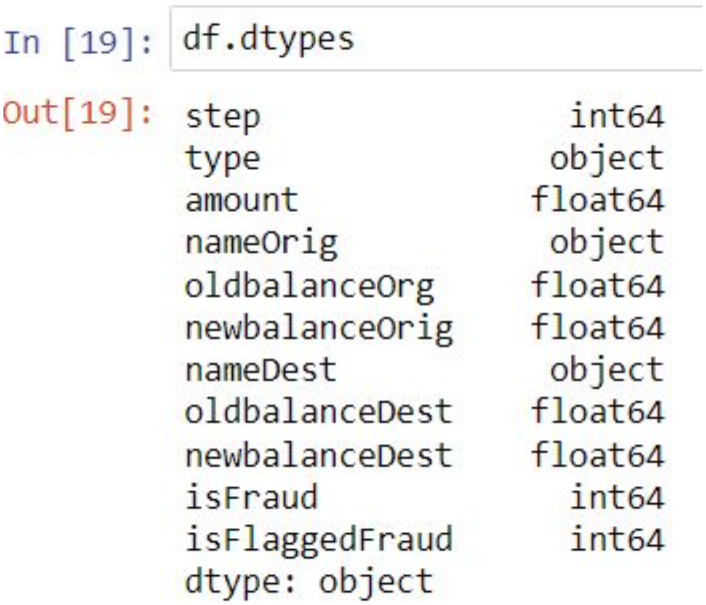
4] Exploratory Data Analysis

The chart below displays each variable's mean, median, mode and quantiles as well as other common statistical measures.



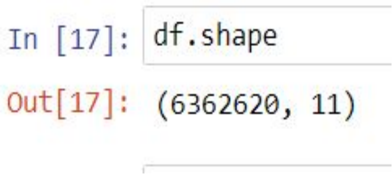
	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud
count	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e
mean	2.433972e+02	1.798619e+05	8.338831e+05	8.551137e+05	1.100702e+06	1.224996e+06	1.290820e
std	1.423320e+02	6.038582e+05	2.888243e+06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

4.1 Statistical measure



```
In [19]: df.dtypes

Out[19]: step                int64
         type               object
         amount            float64
         nameOrig           object
         oldbalanceOrg     float64
         newbalanceOrig    float64
         nameDest           object
         oldbalanceDest    float64
         newbalanceDest    float64
         isFraud             int64
         isFlaggedFraud      int64
         dtype: object
```

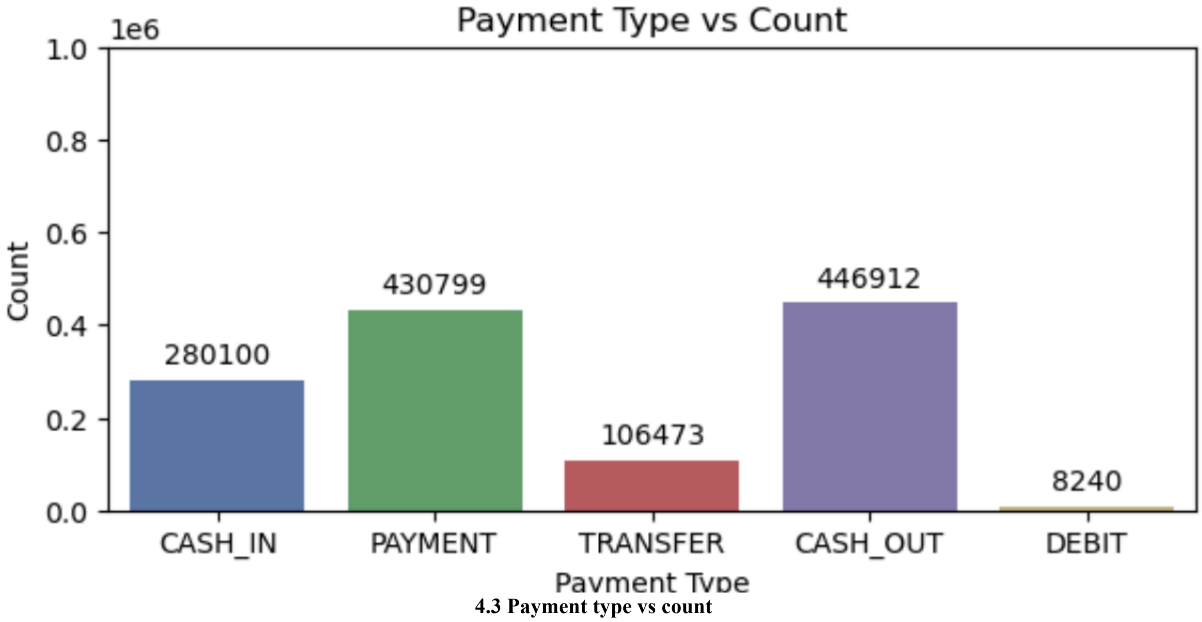


```
In [17]: df.shape

Out[17]: (6362620, 11)
```

4.2 descriptive measure

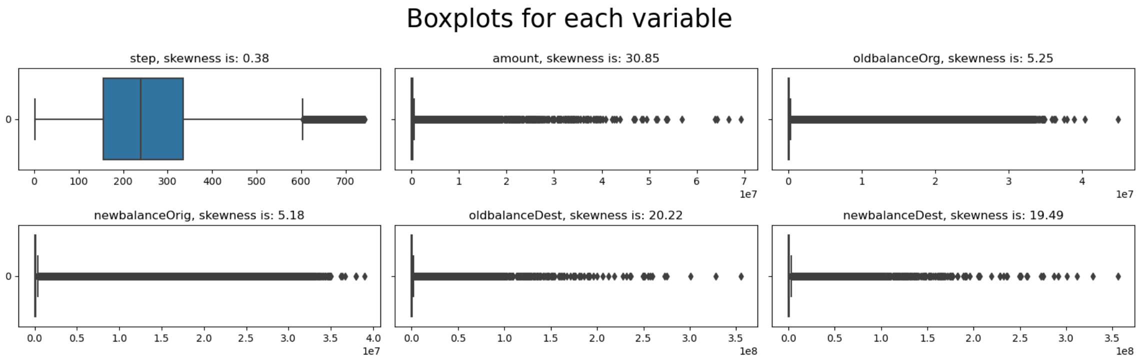
The chart's variable shapes and data types are listed below. We can see that the majority of our columns are numeric and contain the data types int and float. We also have 11 columns and 6362620 rows.



4.3 Payment type vs count

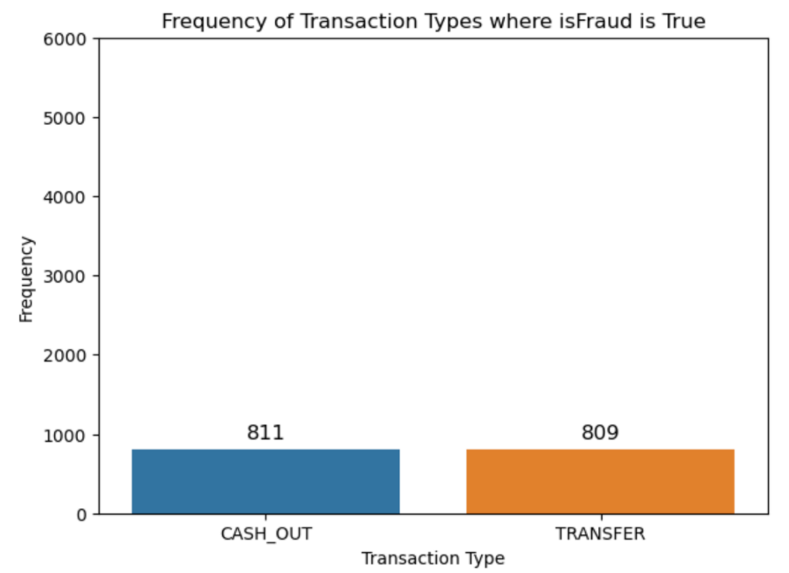
The "Payment Type vs Count" graph provides a visual representation of the different payment types used in each context, along with their respective frequencies or counts. This graph serves as a valuable tool to analyze the distribution and popularity of various payment methods among a specific population or within a particular domain.

By examining this graph, one can quickly identify the most prevalent payment types (cash and payments) as well as those that are less commonly used(debit). It offers insights into the preferences and habits of individuals or groups when it comes to making payments, aiding in the understanding of consumer behaviour and trends.



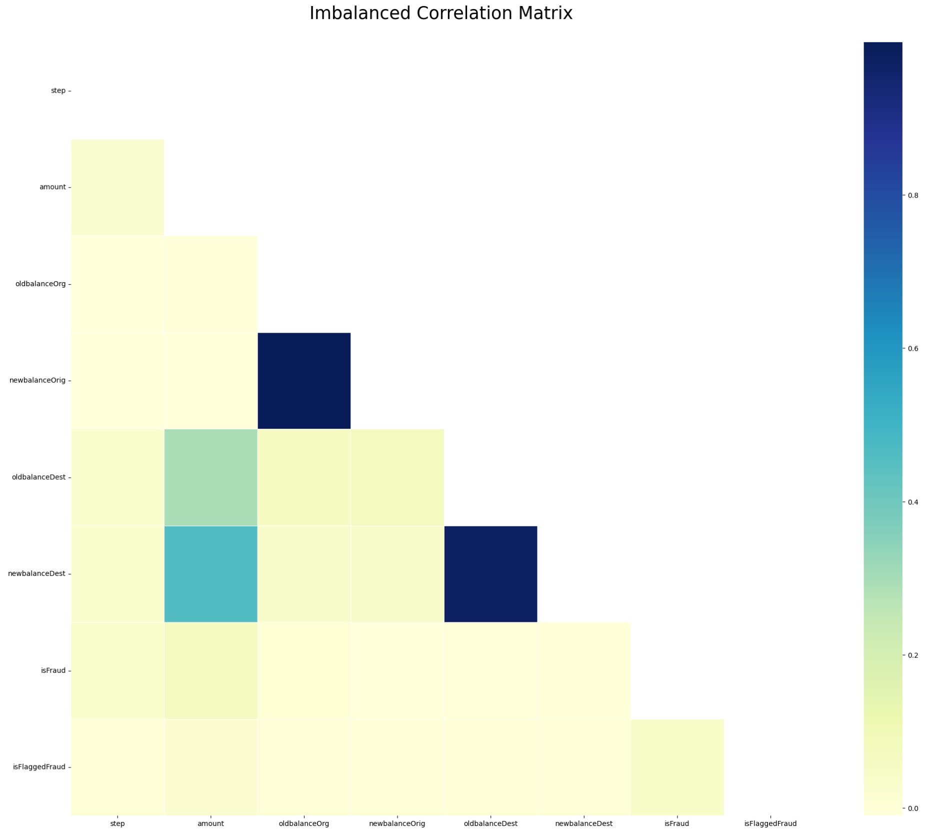
4.4 Boxplots

The boxplot is a useful visualization tool for understanding the central tendency, spread, and distribution of payment values. It allows viewers to quickly grasp the range of payments, identify any skewness or asymmetry in the data, and spot any potential outliers. By providing a visual summary of the payment dataset, enables comparisons across different groups or categories and facilitates the detection of anomalies or trends in payment behaviour.



4.5 Fraud transaction type

This graph displays the distribution of payment types with a focus on fraudulent transactions. It provides insights into the prevalence and impact of fraudulent activities within different payment methods. There is an equal distribution of fraudulent transactions between cash and transfer.



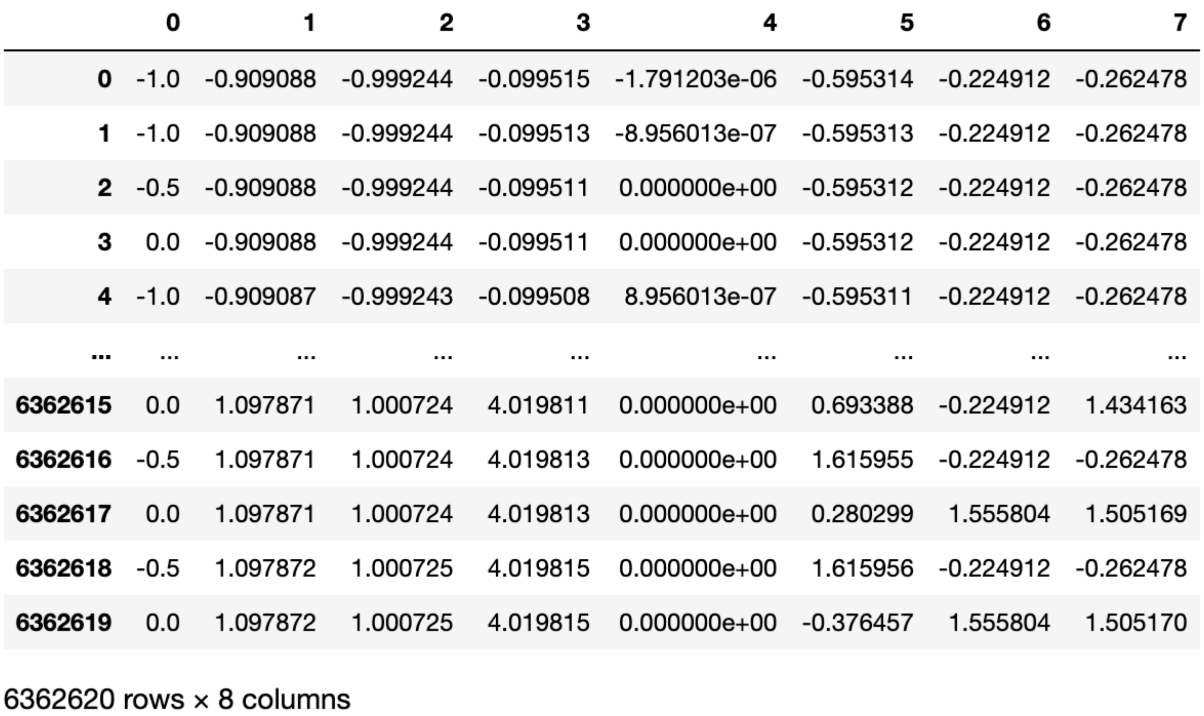
4.6 Correlation plot

A correlation matrix is a mathematical representation that provides valuable insights into the relationships between multiple variables within a dataset. It is commonly used in statistics and data analysis to explore the strength and direction of associations between pairs of variables. We can observe that there is a relatively low correlation between our variables. However, there is some degree of correlation between old balance, new balance, and amount.

5] Predictive algorithms

1] Data Scaling.

Robust Scaling: - As our dataset has approximately 6.3 million rows. It was important to scale the dataset as we had different ranges of values in the features. Robust scaling is a valuable preprocessing technique that provides a robust and resistant approach to feature scaling while mitigating the influence of outliers. Using robust statistical measures such as the median and interquartile range allows for more reliable and accurate data analysis and model building.

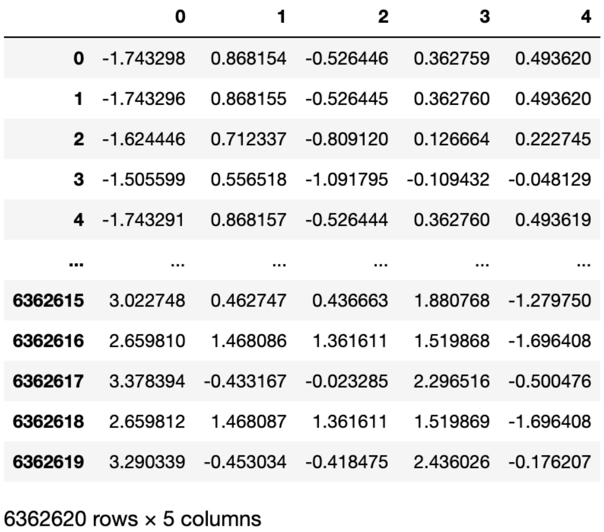


	0	1	2	3	4	5	6	7
0	-1.0	-0.909088	-0.999244	-0.099515	-1.791203e-06	-0.595314	-0.224912	-0.262478
1	-1.0	-0.909088	-0.999244	-0.099513	-8.956013e-07	-0.595313	-0.224912	-0.262478
2	-0.5	-0.909088	-0.999244	-0.099511	0.000000e+00	-0.595312	-0.224912	-0.262478
3	0.0	-0.909088	-0.999244	-0.099511	0.000000e+00	-0.595312	-0.224912	-0.262478
4	-1.0	-0.909087	-0.999243	-0.099508	8.956013e-07	-0.595311	-0.224912	-0.262478
...	...	...	...	...	...	...	...	...
6362615	0.0	1.097871	1.000724	4.019811	0.000000e+00	0.693388	-0.224912	1.434163
6362616	-0.5	1.097871	1.000724	4.019813	0.000000e+00	1.615955	-0.224912	-0.262478
6362617	0.0	1.097871	1.000724	4.019813	0.000000e+00	0.280299	1.555804	1.505169
6362618	-0.5	1.097872	1.000725	4.019815	0.000000e+00	1.615956	-0.224912	-0.262478
6362619	0.0	1.097872	1.000725	4.019815	0.000000e+00	-0.376457	1.555804	1.505170

6362620 rows × 8 columns

5.1.1 Robust Scaler

PCA:- Principal Component Analysis is a powerful technique for dimensionality reduction, feature extraction, and data exploration. By transforming high-dimensional data into a lower-dimensional representation, PCA enables easier analysis and visualization while preserving the essential information. Thus, after running PCA the algorithm generated 5 new features instead of the 8 we had.

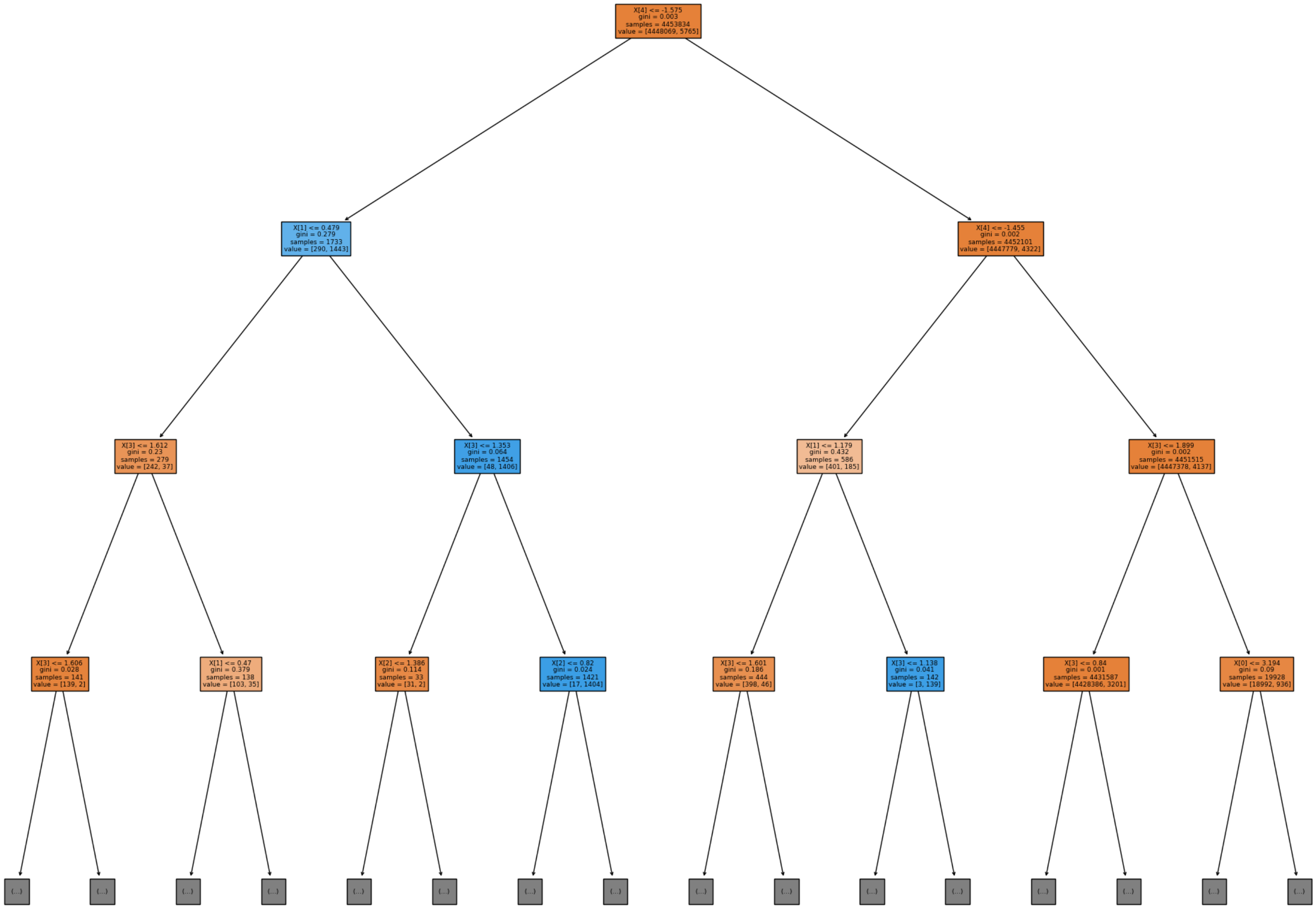


	0	1	2	3	4
0	-1.743298	0.868154	-0.526446	0.362759	0.493620
1	-1.743296	0.868155	-0.526445	0.362760	0.493620
2	-1.624446	0.712337	-0.809120	0.126664	0.222745
3	-1.505599	0.556518	-1.091795	-0.109432	-0.048129
4	-1.743291	0.868157	-0.526444	0.362760	0.493619
...	...	...	...	...	...
6362615	3.022748	0.462747	0.436663	1.880768	-1.279750
6362616	2.659810	1.468086	1.361611	1.519868	-1.696408
6362617	3.378394	-0.433167	-0.023285	2.296516	-0.500476
6362618	2.659812	1.468087	1.361611	1.519869	-1.696408
6362619	3.290339	-0.453034	-0.418475	2.436026	-0.176207

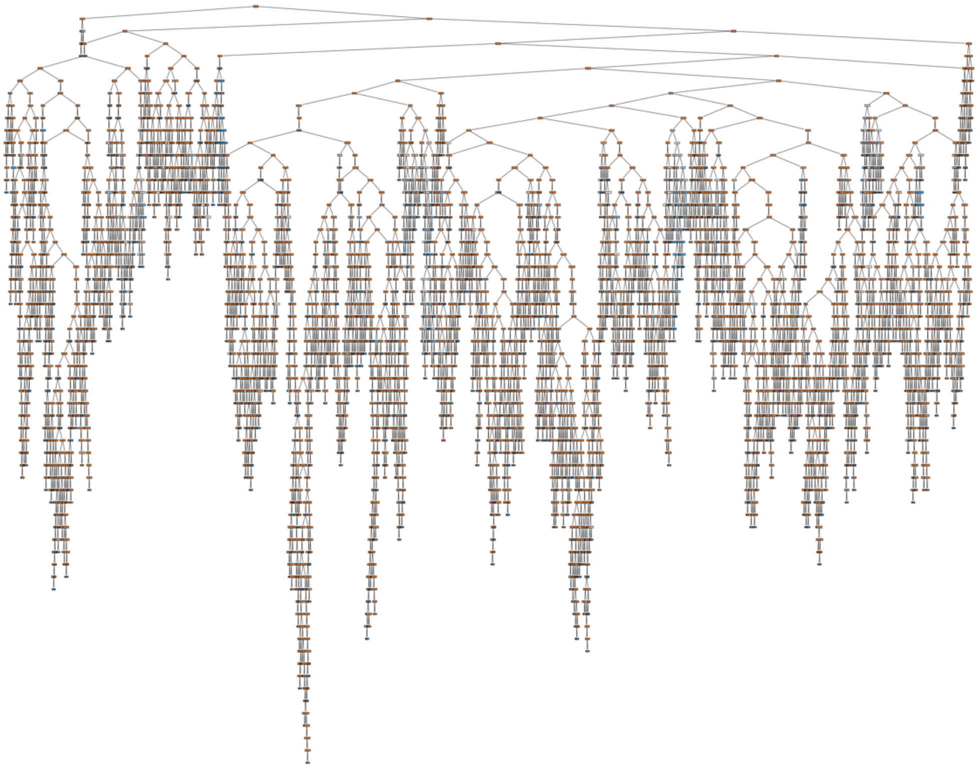
6362620 rows × 5 columns

5.1.2 PCA

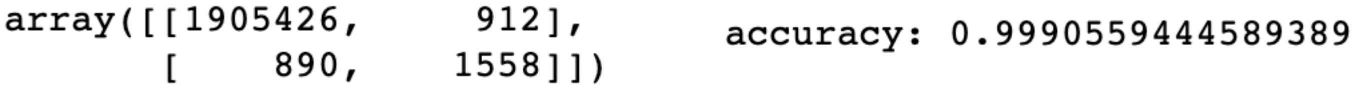
2] Decision Tree:- Decision trees are versatile and intuitive machine-learning algorithms used for both classification and regression tasks. They offer interpretability, handle nonlinear relationships, and provide feature importance rankings. However, decision trees are prone to overfitting and may not perform well on unseen data.



5.2.1: Subsection of Decision Tree



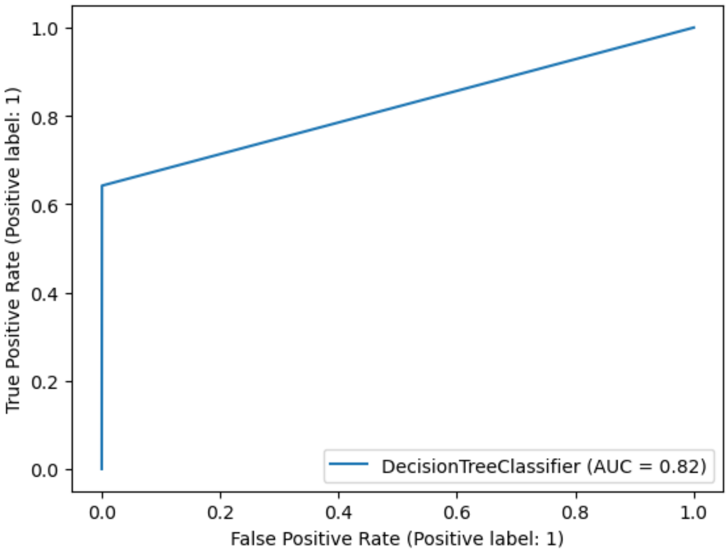
5.2.2 Full Decision Tree



```
array([[1905426,     912],
       [    890,    1558]])          accuracy: 0.9990559444589389
```

5.2.2 Accuracy Metrix.

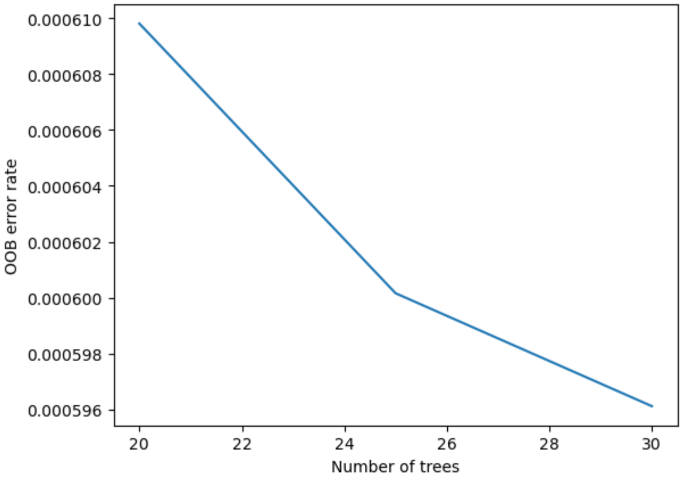
After executing the algorithm, we achieved an outstanding accuracy level of 99.9%. However, a minor issue arises from this result. We have encountered a significantly high number of false positives, specifically 912 cases, where our model incorrectly identifies negative values (fraudulent transactions) as positive values (valid transactions). It is crucial for us to address this problem and explore methods to minimize the occurrence of such errors.



5.2.3 ROC

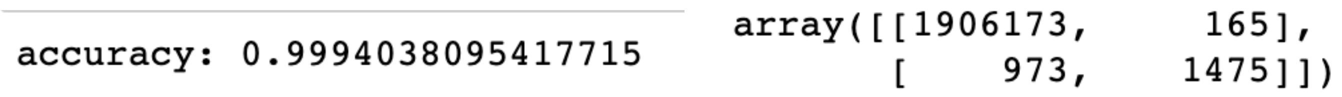
The Receiver Operating Characteristic (ROC) graph is a graphical representation that illustrates the performance of a binary classification model at various classification thresholds. It plots the true positive rate (TPR), also known as sensitivity or recall, on the y-axis against the false positive rate (FPR) on the x-axis. For the decision tree, our AUC is only 0.82.

3] Random forest:- Random Forest is a powerful machine learning algorithm that combines the strength of decision trees and ensemble learning. It provides robust predictions, handles high-dimensional data, and offers insights into feature importance. Random Forest has numerous applications across various domains and is a go-to choice for many data scientists and machine learning practitioners.



5.3.1 OOB

Out-of-bag (OOB) is a concept used in Random Forest and other ensemble learning methods, such as Bagging. It provides a way to estimate the performance of the model without the need for a separate validation dataset. In our algorithm, we have used to OOB to determine the optimal number of trees.

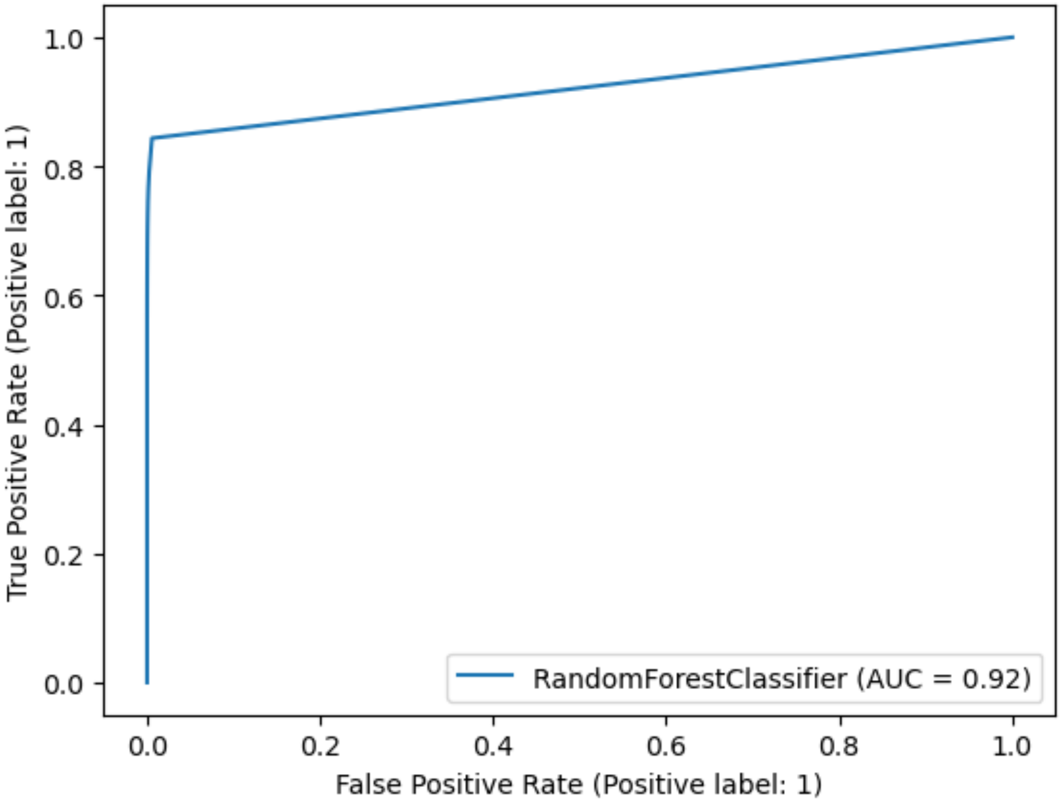


```
accuracy: 0.9994038095417715
```

```
array([[1906173,     165],
       [    973,    1475]])
```

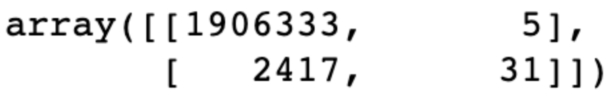
5.3.2 Accuracy Metrix

Upon running the algorithm, we obtained an exceptional accuracy rate of 99.9%. Nevertheless, there is a slight drawback associated with this outcome. We have come across a notable issue concerning false positives, specifically 165 instances, where our model mistakenly classifies negative values (fraudulent transactions) as positive values (valid transactions). It is imperative that we tackle this problem promptly and investigate strategies to further mitigate the occurrence of these errors.

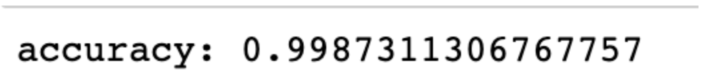


5.3.3 ROC

4] Logistic Regression: - logistic regression is a widely used algorithm for binary classification tasks. It models the relationship between features and the probability of the positive class using the logistic function. Logistic regression provides interpretable coefficients, enables understanding of the impact of features, and offers flexibility in adjusting the decision boundary. It is a valuable tool in predictive modelling and understanding the factors influencing binary outcomes.

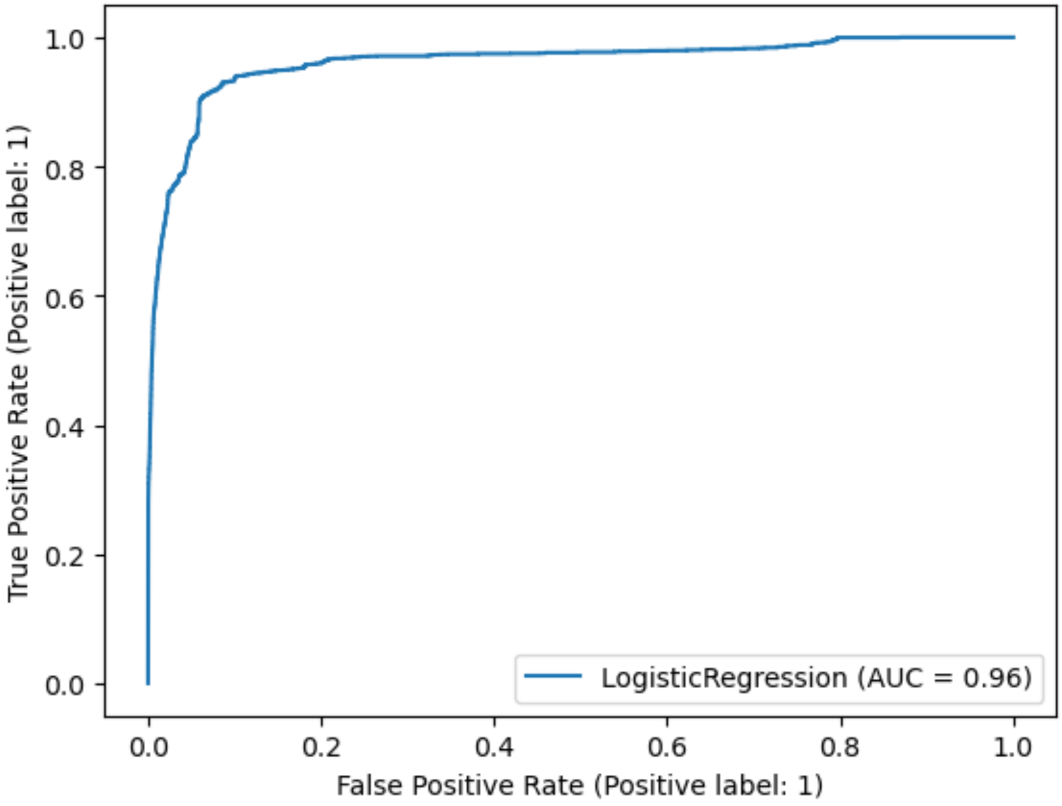


```
array([[1906333,       5],
       [   2417,      31]])
```



```
accuracy: 0.9987311306767757
```

5.4.1 Accuracy Metrix.



5.4.2 : ROC Graph

After executing the algorithm, we achieved an outstanding accuracy level of 99.9%. Additionally, we have successfully reduced the occurrence of false positives. However, we have noticed a relatively higher number of false negatives, where the model incorrectly classifies positive instances (valid transactions) as negative (fraudulent transactions). While this situation is tolerable, it is important to focus on analyzing false negatives and true negatives to achieve our objectives. Further investigation and analysis can help us refine the model and improve its performance in accurately identifying fraudulent transactions.

Conclusion

In this project, we have explored the dataset on online fraud to understand the characteristics of fraudulent transactions and develop strategies to detect and prevent such activities. Our analysis involved various data mining techniques and algorithms, including exploratory data analysis (EDA), data scaling, principal component analysis (PCA), decision trees, random forests, and logistic regression.

Through our analysis, we have gained valuable insights into the dataset and the nature of online fraud. We have observed patterns and trends in fraudulent behaviour, identified key features that contribute to fraud detection, and evaluated the performance of different algorithms in classifying fraudulent transactions.

Our findings indicate that the algorithms used, including decision trees, random forests, and logistic regression, achieved exceptional accuracy rates of 99.9%. However, we have also encountered challenges such as a relatively high number of false positives and false negatives. It is crucial to address these issues to improve the model's overall performance and minimize fraud detection errors.

Overall, this project highlights the importance of data mining and machine learning techniques in combating online fraud. By leveraging the power of data analysis, we can develop effective strategies and models to identify fraudulent transactions and protect individuals and businesses from financial loss. Further research and analysis can be conducted to refine the models, investigate the causes of false positives and false negatives, and enhance the overall performance of fraud detection systems. This project contributes to the ongoing efforts in developing robust fraud detection systems and provides valuable insights for individuals and organizations involved in online transactions. By understanding the underlying patterns and behaviours of fraudsters, we can work towards creating a safer and more secure online environment.

Reference

1. Roy, R. (2022, April 17). *Online payments fraud detection dataset*. Kaggle. Retrieved April 21, 2023, from https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset