## Financial Payment Services Fraud Prediction

## Bargavi Kongara

## Group Number- 7 & Section- 01

## 4550 Analytics Programming

## Project Proposal

## Theyab Alhwiti

## **Abstract:**

Fraud detection is the process of finding illegal acts, such as phishing, credit card theft, and identity theft, that results in the acquisition of money or property under false pretenses. In our project, we will use logistic regression to try and gain a better detection rate. We aim to demonstrate how supervised ML techniques can be used to classify data with high class imbalance with high accuracy. We did a detailed data exploration and cleaning of data by choosing a machine-learning algorithm and logistic regression to find an optimal solution to identify fraudulent payments. This will not only help us to detect frauds easily but be more consistent and accurate with our results.

h

## **1. Introduction**

1. **Motivation of the paper in detail**

Our major motivation behind choosing fraud detection as our project is that this has become a significant issue in the world. With increasing value for money and the number of trades happening throughout the world every second, it is vital to ensure that every transaction is true and real. It is humanly not possible to keep a check on whether every transaction is genuine and is not a Fraud. Hence our goal is to provide a project which will help us in showing how to detect frauds more consistently and accurately giving as minimum errors as possible.

1. **Specific problem under study:**

Fraud negatively affects a company’s bottom line, its character and discourages possible prospects and current customers likewise to engage with it.

More frequently than not, for any fraud detected, the association ends up paying for the losses. Also, it takes the loyal customers down from them while attracting further fraudsters.

Given the scale and reach of the utmost of these vulnerable companies, it has become necessary for them to stop these frauds from passing or indeed always prognosticate all suspicious conduct beforehand.

Frauds can range from really small amounts like remitment for e-commerce orders to threatening (to association’s actuality) like public exposure of customers’ credit card details.

Machine learning comes to the deliverance then. On setting up automated data science processes with deep learning algorithms, associations can greatly reduce the threat of their exposure to utmost of similar frauds.

1. **Why studying problem is important**

For years, fraud has been a major issue in sectors like banking, medical, insurance and numerous others. Due to the increase in online services through different payment options, similar as credit/ debit cards, etc., fraudulent conditioning has also increased. Fraudsters or culprits have come veritably professed in finding escapes so that they cannot be traced further. Since no system is perfect and there's always a loophole between them, it has become a grueling task to make a secure system for authentication and protecting guests from fraud. So, Fraud detection algorithms are veritably useful for precluding frauds.

## **2. Method and Analysis**

1. **Data description and source:**

Financial data is typically private, thus there aren't many publicly accessible datasets that may be utilized for research.

We downloaded the data from kaggle link: <https://www.kaggle.com/code/arjunjoshua/predicting-fraud-in-financial-payment-services/notebook>

The fact that this dataset is now one of just four on Kaggle providing information on the increased danger of financial digital fraud highlights how challenging it is to get such information. The distribution of positive and negative classes is extremely unbalanced, which is the key technical difficulty it presents for forecasting fraud.

1. **Variable description**

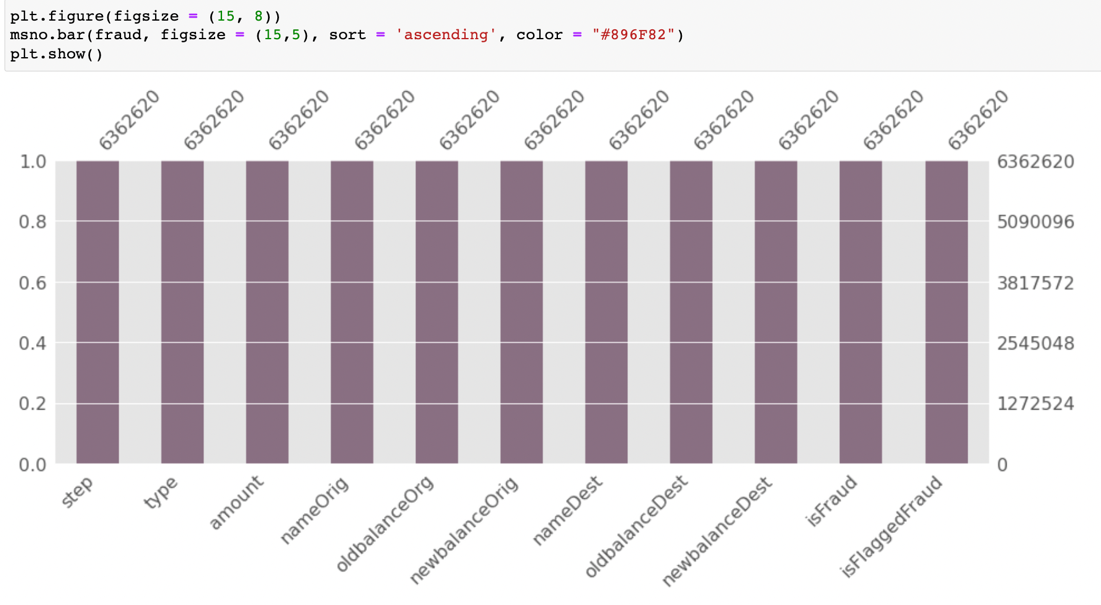
|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Variable Type** |
| step | Maps a unit of time in the real world. In this case 1 step is 1 hour of time. | Integer |
| type | CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER | String |
| amount | amount of the transaction in local currency | Integer |
| nameorig | Origin of the Transaction | String |
| oldbalanceOrg | initial balance before the transaction | Integer |
| newbalanceOrig | customer's balance after the transaction. | Integer |
| nameDest | Recipient of the transaction | String |
| oldbalanceDest | initial recipient balance before the transaction. | Integer |
| newbalanceDest | recipient's balance after the transaction. | Integer |
| isFraud | identifies a fraudulent transaction (1) and non-fraudulent (0) | Boolean |

1. **Data preparation:** (including the process if you combined the data)

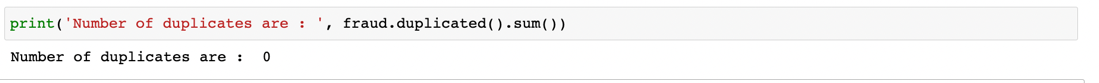
The data used for this analysis is a synthetically generated digital transactions dataset using a simulator called PaySim. PaySim simulates mobile money transactions based on a sample of real transactions extracted from one month of financial logs from a mobile money service implemented in an African country. It aggregates anonymized data from the private dataset to generate a synthetic dataset and then injects fraudulent transactions. The dataset has over 6 million transactions and 11 variables. There is a variable named ‘isFraud’ that indicates actual fraud status of the transaction. This is the class variable for our analysis.

1. **Data cleaning:**

In this phase, we also check if there are any missing values in the dataset. The following code and output indicate the total number of missing / NA values in all columns, which is zero.



We also have checked if there are duplicate values. The following code output indicates that there are no duplicate values.

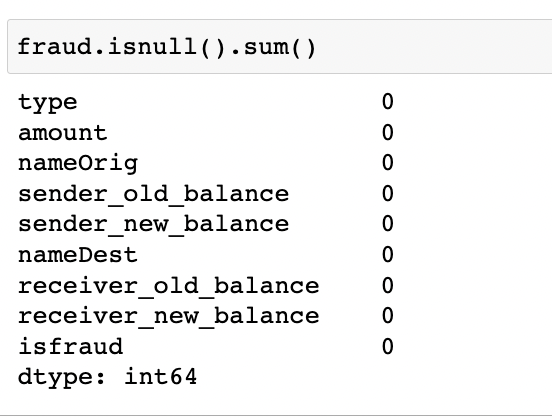


Later, we removed the irrelevant columns like Step, nameOrign, nameDest and isFalggedFraud.

Then, we renamed some of our columns for our comfort and better understanding.

1. **Identifying and handling missing values:**

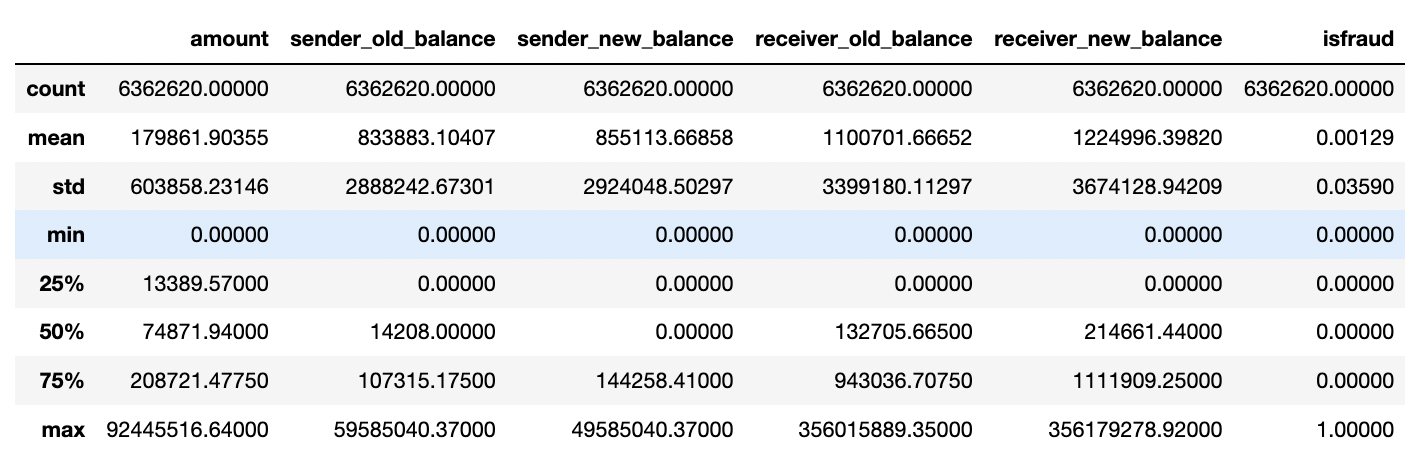
The below screenshot indicates that there are no missing values for the given data set.



## **3. Results**

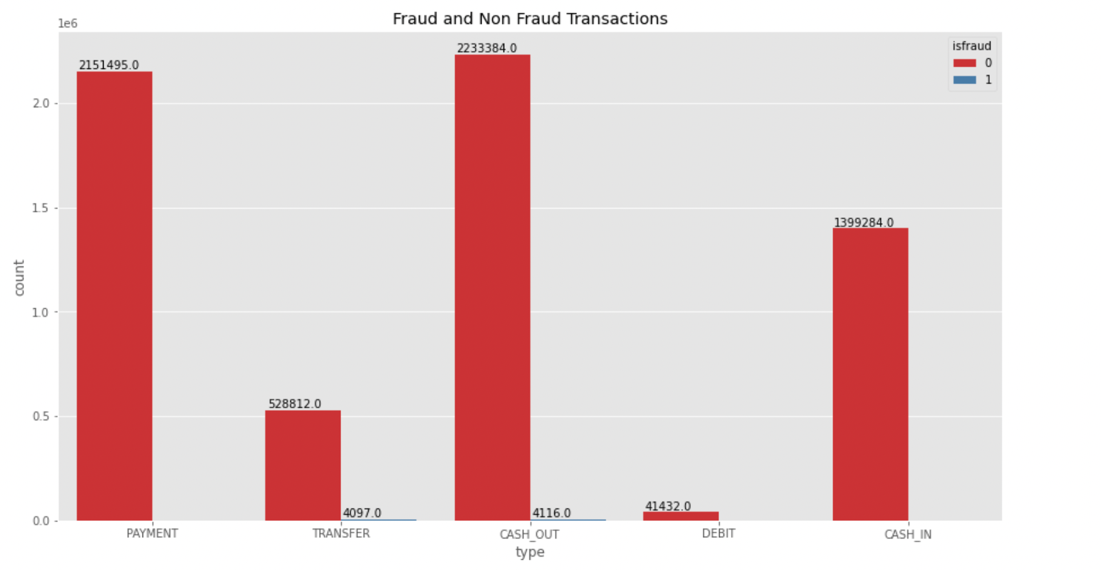
1. **Exploratory results: include summary statistics, charts, relationships:**

Before proceeding with the analysis, we present the summary statistics of the variables. In case of numeric variables, we evaluate the mean, standard deviation and the range of values at different percentiles. In case of categorical variables, we evaluate only the number of unique categories, the most frequent category and its frequency.



1. **Graphical representation of data.**

The following plot shows the count of fraud and non-fraud transactions for different transaction types:

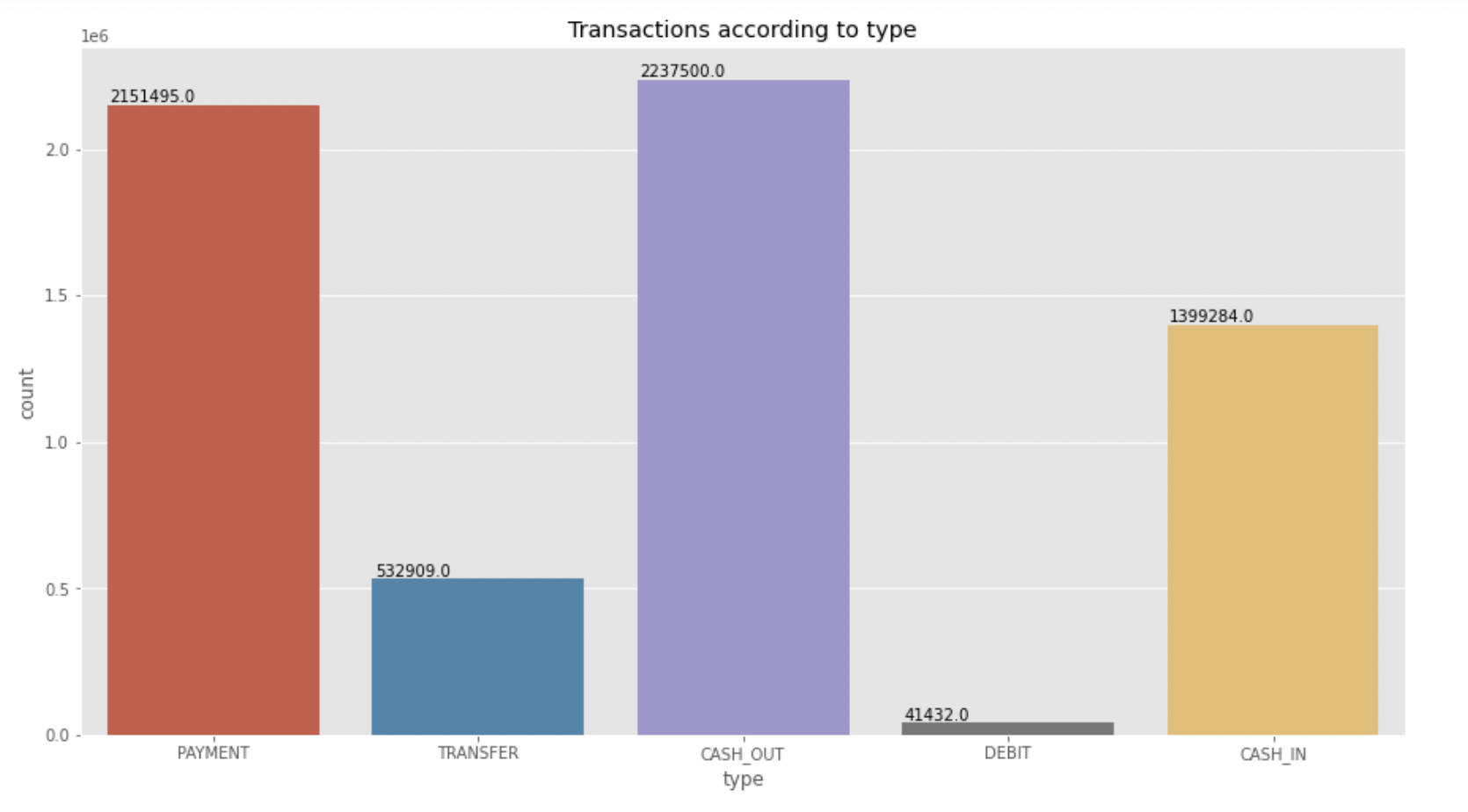


The following plot shows the count of fraud and valid transactions.

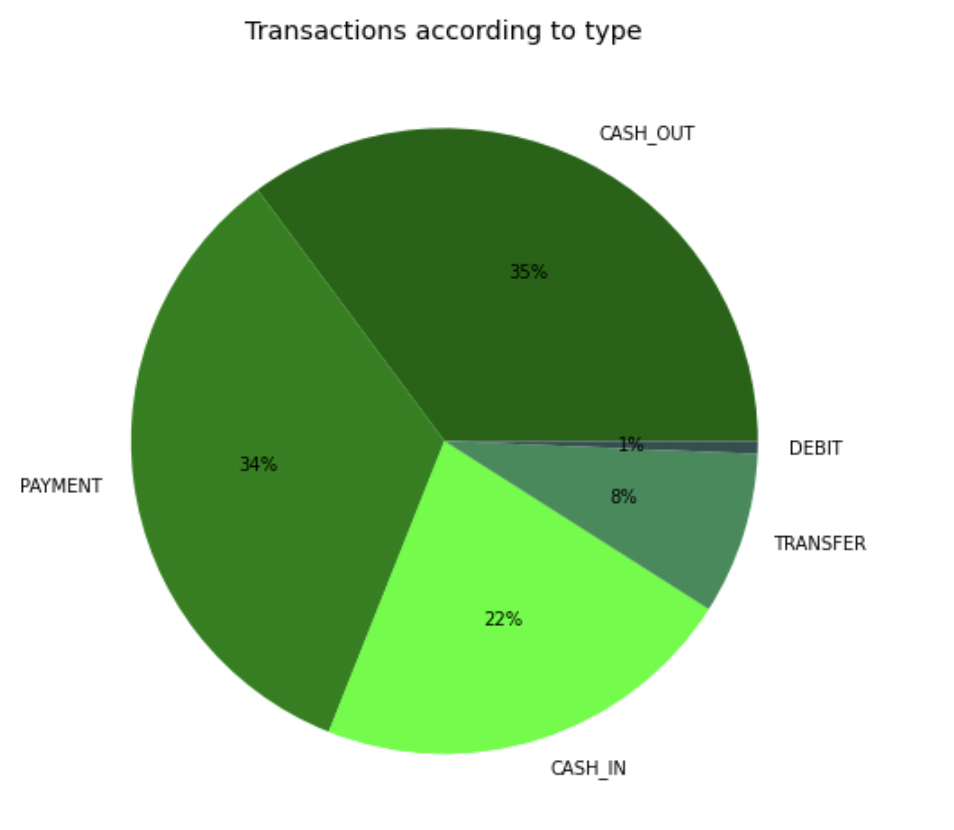
Chart, bar chart

Description automatically generated

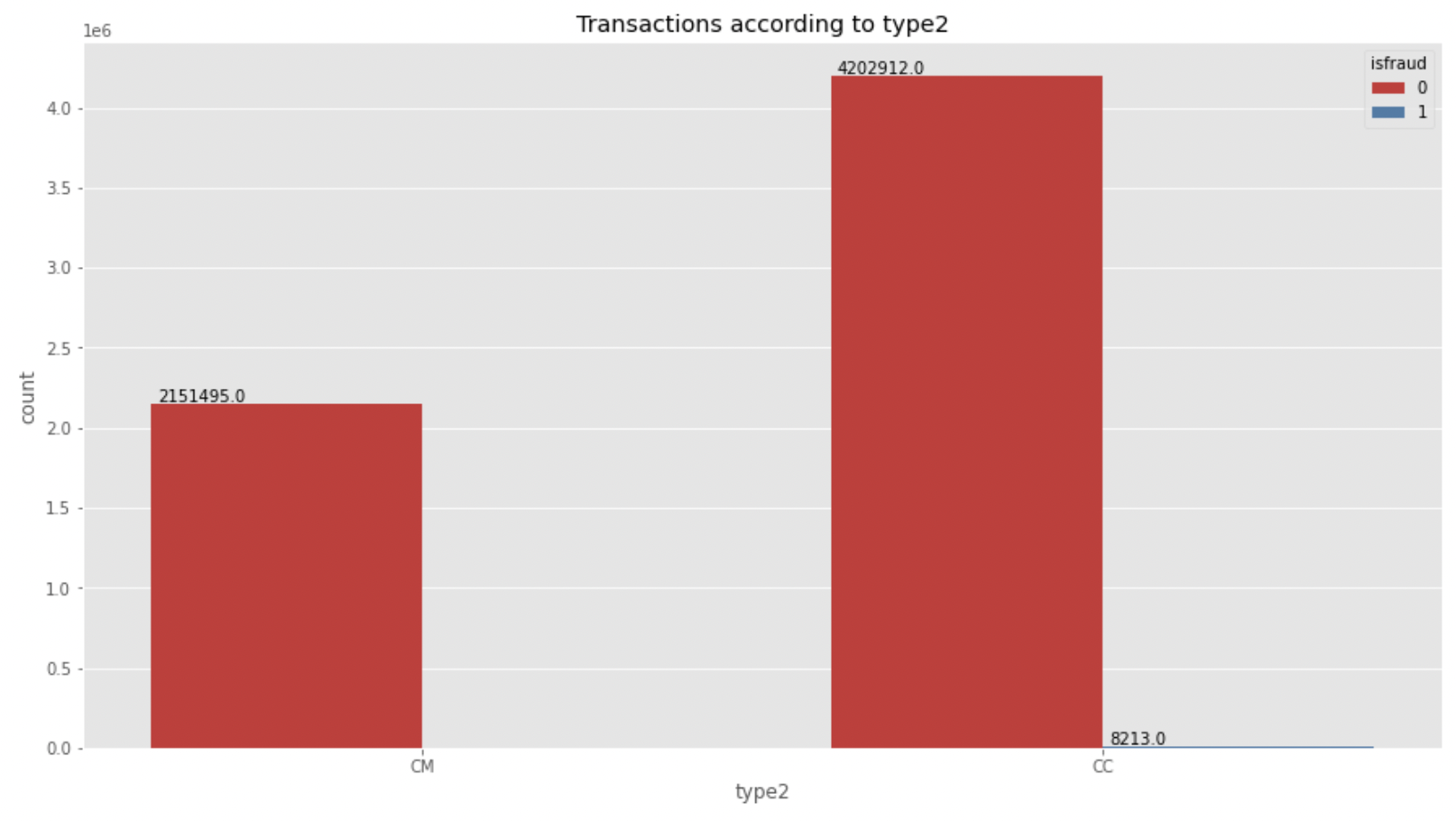
The following plot shows the count of transactions for each transaction type.



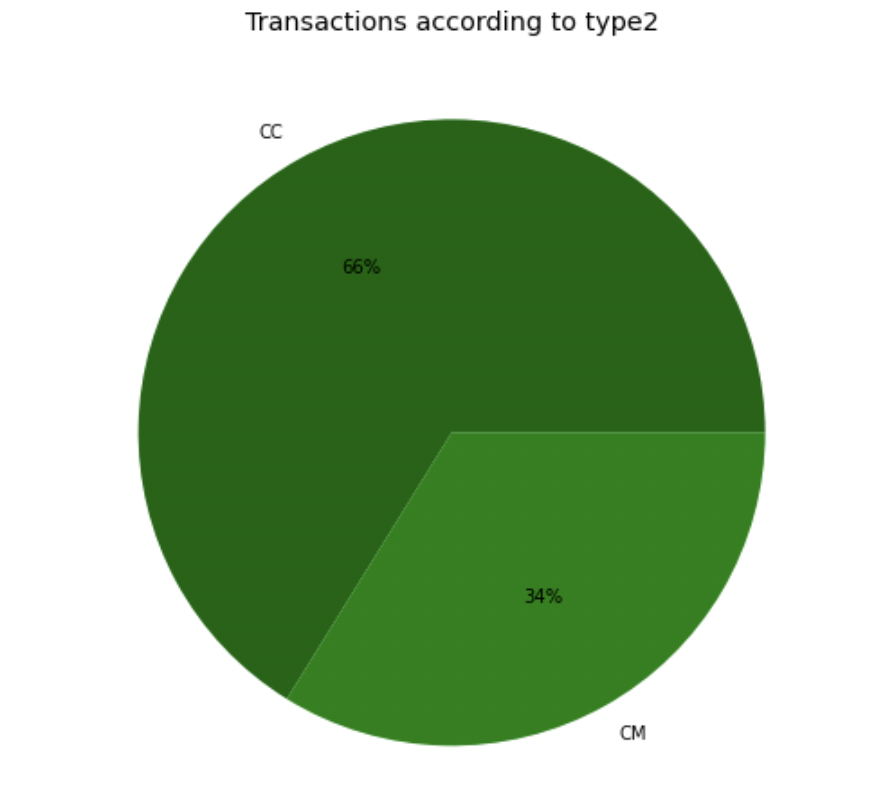
The following pie chart shows the percentage of transactions for each transaction type.



The following plot shows the count of transactions based on the type2.



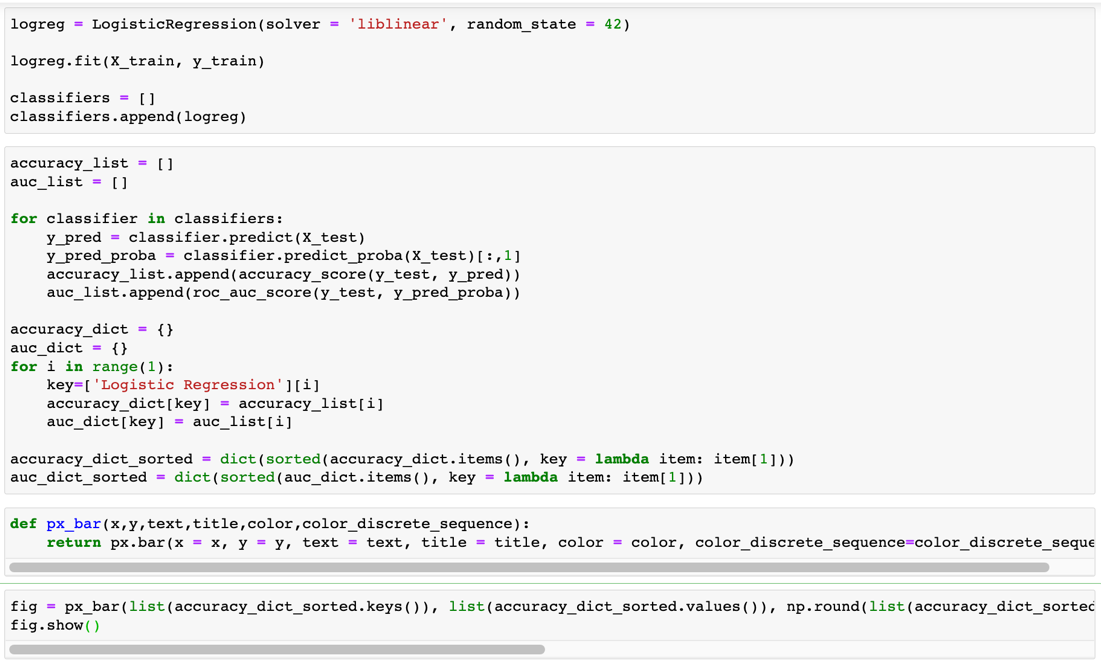
The following plot shows the percentage of transactions based on the type2.



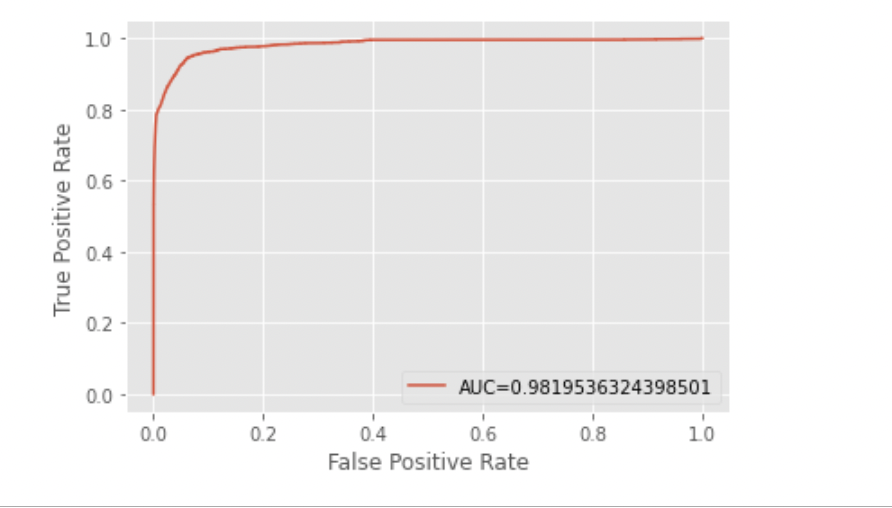
1. **Modeling:**

The following code snippet is used to define and build the logistic regression model.

we train the logistic regression model. This parameter will serve as a benchmark for further experiments.







After building the Logistic regression model and finding out the accuracy and AUC curve the below analysis can be made.

The total number of fraud transactions was **8213** out of **6362620** transactions.

These fraud transactions were either **CASH\_OUT** or **DEBIT** and were made from a **customer-to-Customer** account.

We trained in Logistic Regression algorithm. It gave the **score** of **0.9**.

## **4. Discussion and Conclusion**

1. **Interpret the results in detail and make recommendations:**

Through this experiment, we showed that despite the high-class imbalance, it is possible to identify fraudulent transactions in financial transactions data with extremely high accuracy.

To select the proper features, dispersion and scatter plots are crucial for visualizing the difference between fraudulent and legitimate transactions.

1. **Limitations of the paper**

In this study, we evaluated the effectiveness of using specific supervised machine learning techniques to solve the problem of fraud detection in financial transactions. The limitations of the methods applied in this study are as follows:

* We used a pre-labeled dataset to train the algorithms. However, usually, it is difficult to find labeled data and thus applying supervised machine learning techniques may not be feasible. In such cases, we should evaluate unsupervised techniques which were beyond the scope of this study.
* This study considers digital transactions data that includes amount transacted, the balance of recipient and originator, and time of transaction. These variables that helped in detecting fraud may not apply to other types of financial transactions, such as credit card fraud.
* We evaluated machine learning algorithms – Logistic Regression. Although the result of the study using these algorithms is good, it is necessary to evaluate other techniques to determine which algorithm works best for this application.
* Due to the large size of the data, we were limited by computation capacity to explore different techniques such as grid search for parameter tuning, SMOTE sampling technique. These techniques may help in further improving the results of this study.

1. **Future work and conclusions:**

In conclusion, we successfully developed a framework for detecting fraudulent transactions in financial data. This framework will help understand the nuances of fraud detection such as the creation of derived variables that may help separate the classes, addressing class imbalance and choosing the right machine learning algorithm.

We were successful in creating a system for spotting fraudulent transactions in financial data, to sum up. By resolving class imbalance, selecting the appropriate logistic regression method, and creating derived variables that may help separate the classes, this framework will aid in understanding the complexities of fraud detection.

We can use different machine learning models like Random Forest which may give us a better result. As new machine learning models arrive and evolve, we can use them with this dataset to get better outputs in future.

## **5. References:**

E. Ngai et.al., The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature, Decision Support Systems. 50, 2011, 559–569

Jain R., Gour B., Dubey S., A hybrid approach for credit card fraud detection using rough set and decision tree technique, International Journal of Computer Applications 139(10) (2016).

Carneiro E.M., Dias L.A.V., Da Cunha A.M., Mialaret L.F.S., Cluster analysis and artificial neural networks: A case study in credit card fraud detection, 12th International Conference on Information Technology-New Generations (2015), 122-126.

Varre Perantalu K., Bhargav Kiran, Credit card Fraud Detection using Predictive Modeling (2014).

Phua et.al., Minority Report in Fraud Detection: Classification of Skewed Data. ACM SIGKDD Explorations Newsletter 2004; 6: 50-59.

Wedge et.al., Solving the False Positives Problem in Fraud Prediction Using Automated Feature Engineering, Machine Learning and Knowledge Discovery in Databases, pp 372-388, 2018

Dharwa et.al., A Data Mining with Hybrid Approach Based Transaction Risk Score Generation Model (TRSGM) for Fraud Detection of Online Financial Transaction, International Journal of Computer Applications 2011; 16: 18-25.

Chan P.K., Stolfo S.J., Toward Scalable Learning with NonUniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection, In KDD (1998), 164-168.

Sorournejad et.al., A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective, 2016