**Objective :**

To predict the fraud transaction of the payments.

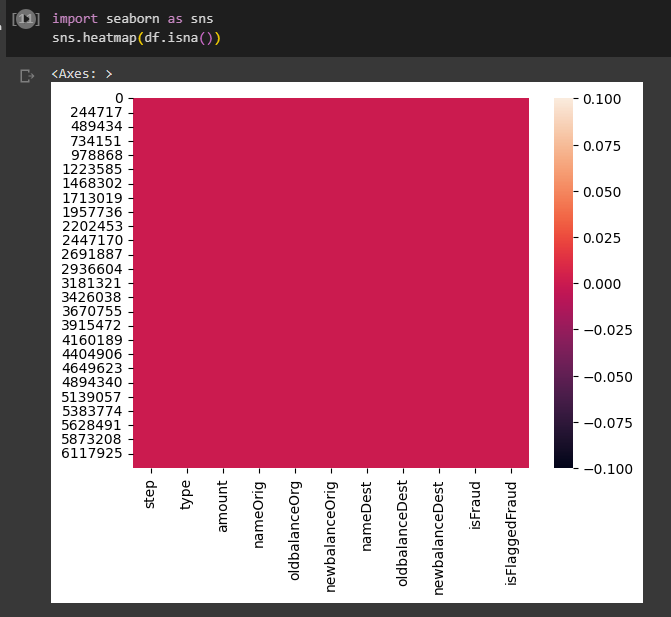
The main objective is to create a predictive model that can accurately classify transactions as either legitimate or fraudulent.

This model will be based on the available data and will be used to make predictions on new, unseen data.

Since, the current system is letting fraud transaction able to pass through a system which is not labeling them as a fraud. The need for a system which can be fast and reliable to mark the transaction which is fraud can help us reduce fraud crime.

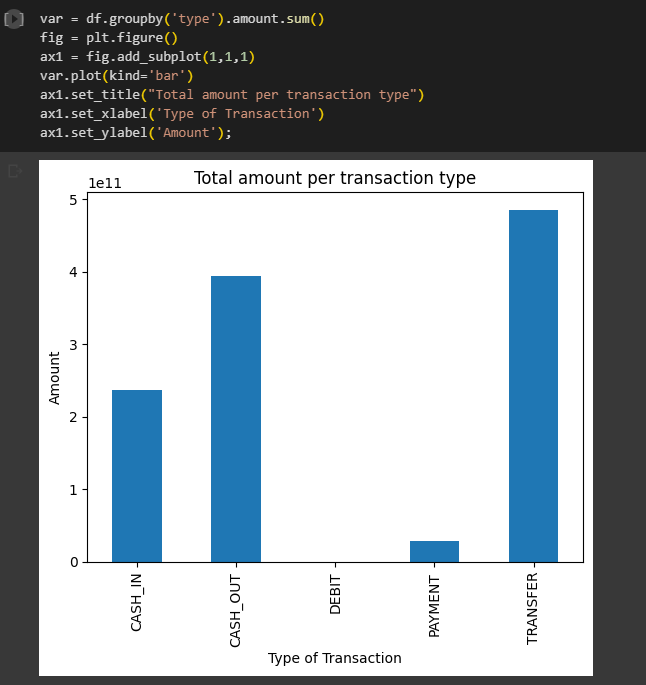
**Steps implemented for predicting whether the data is having fraud transaction or not:**

1. **Data Analysis:**
   1. **Null Values Check :**
      1. The given data set does not contain any null values and we can see the below **heatmap** to identify the same.
      2. the target column is **“isFraud”** which is the actual fraud status of the transactions and isFalggedFraud is the indicator, that is used to flag the transaction using the threshold value.

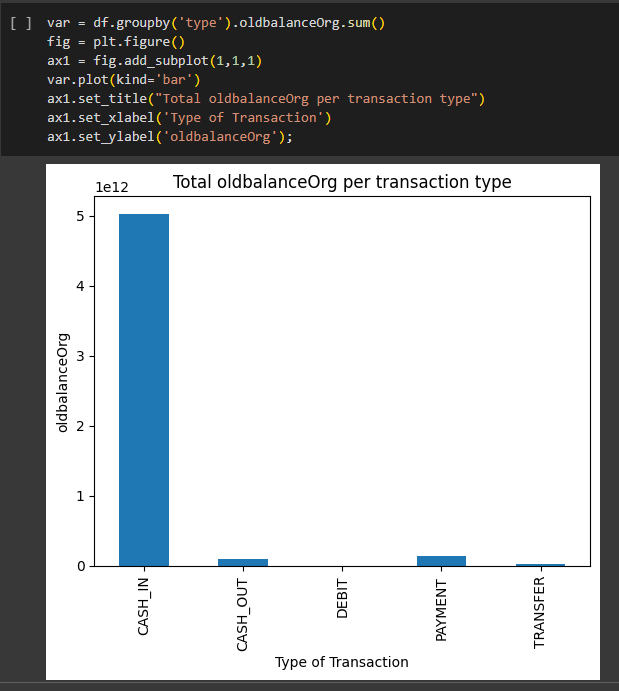


* 1. **Analyzed the columns by plotting bar graph :**

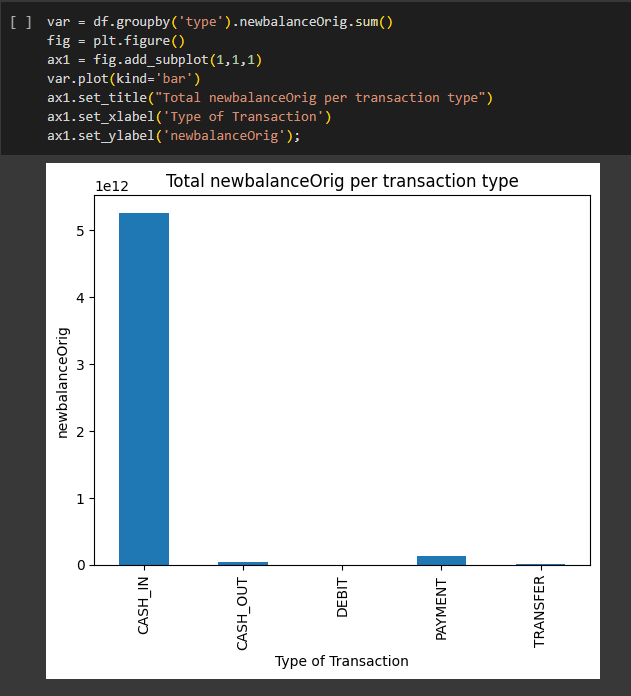
1. Total amount per transaction type.



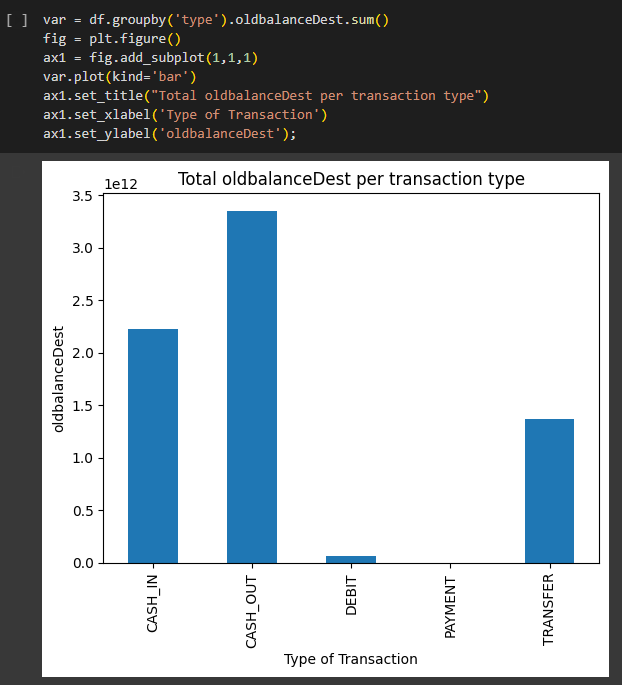
1. Total oldamountDest per transaction type.



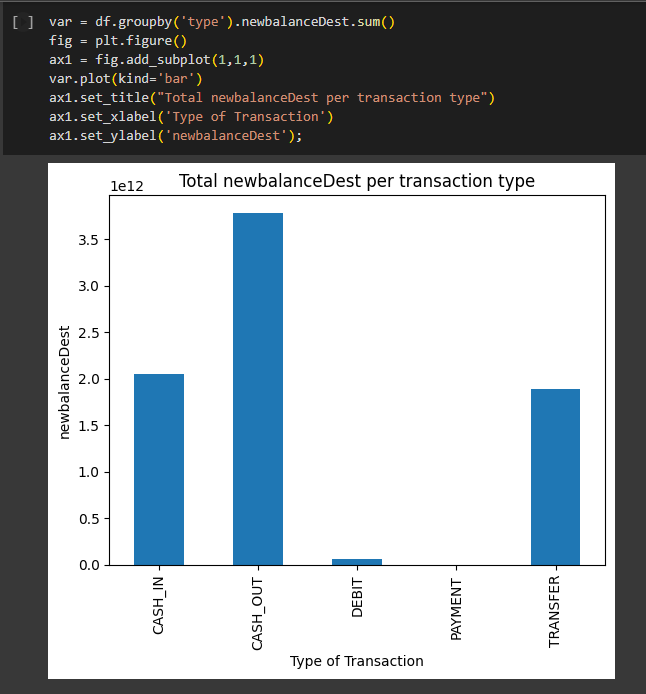
1. Total newamountOrig per transaction type.



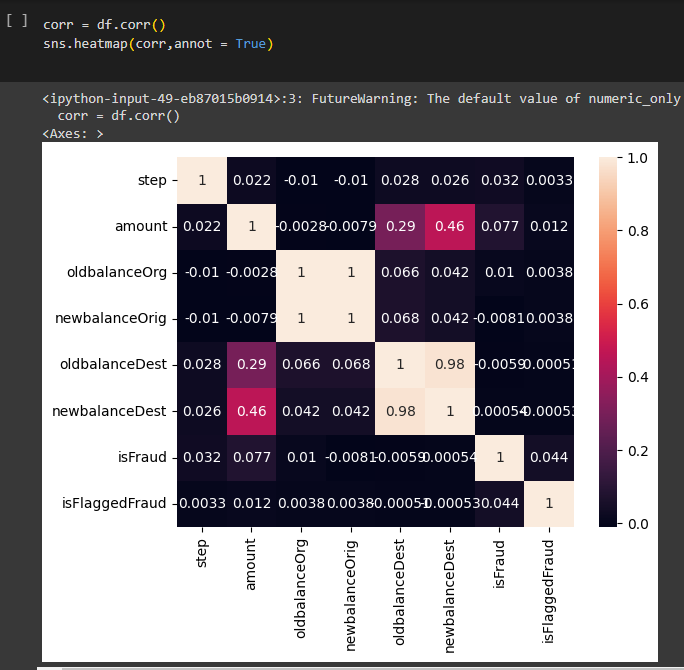
1. Total oldamountDest per transaction type.



1. Total newamountOrg per transaction type.



* 1. **Analysed by the correlation heatmap :**



From the above correlation heatmap:

We can conclude that -

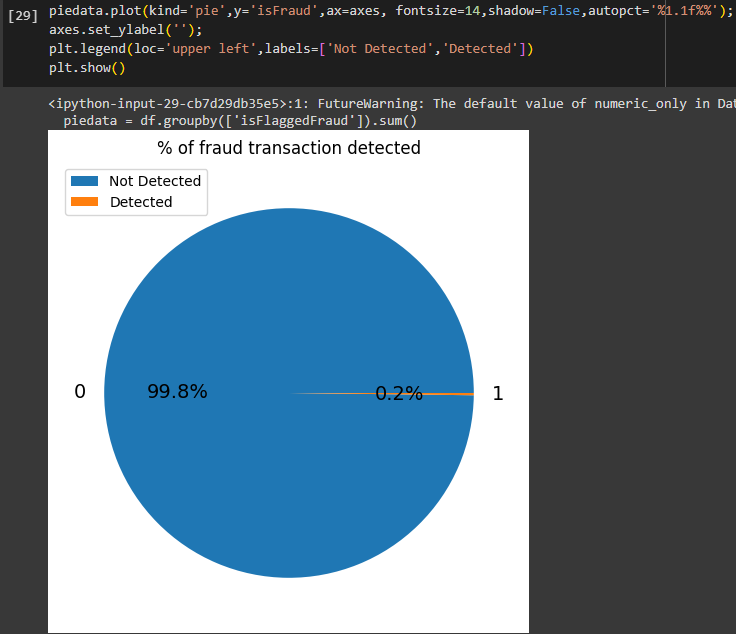
OldbalanceOrg and NewbalanceOrg are highly correlated.

OldbalanceDest and NewbalanceDest are highly correlated.

Amount is correlated with isFraud ----- Target Variable .

There is not much relation between the features, so we need to understand where the relationship between them depends on the type of transaction and amount.

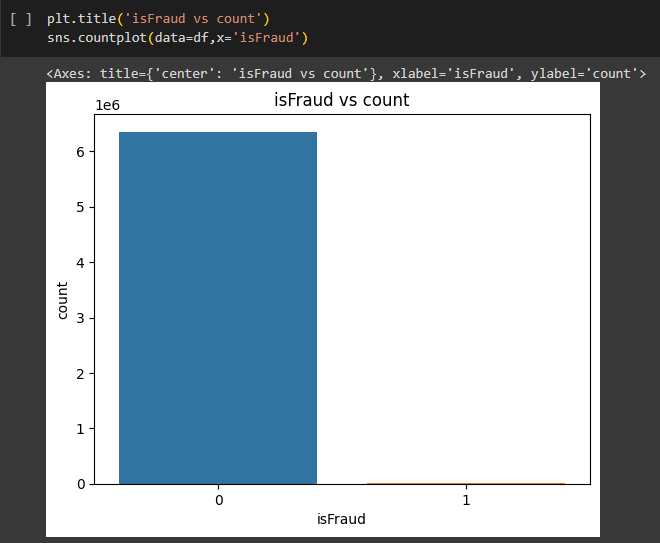
* 1. **Pie chart of ratio Fruad and NotFruad :**



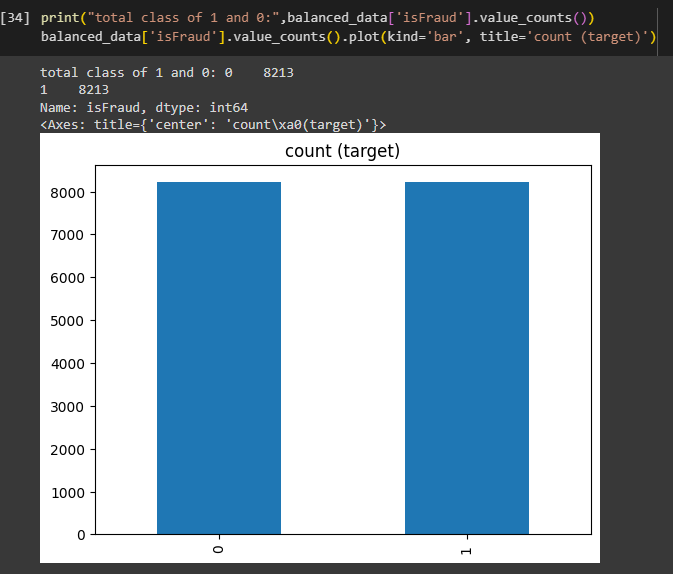
And only 0.2 percent of the data is the fraud , so the data is highly imbalance , hence we need to balance the data.

1. **Data Exploratory :**
   1. **Imbalance data :**

* Data is to much overfitted can get the accuracy as 1.
* Here data is highly imbalance . So we will take the class\_count of the isFraud column means to balance the imbalance data.



* 1. **Balance data:**

****

1. **Splitting the data into training and testing :**

From sklearn used the selection model and by using train\_test\_spilt method splited the data into 80% training data and 20% testing data.

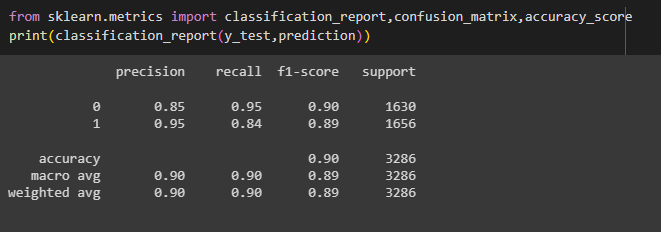
1. **Standard Scaling on the data :**

* Feature scaling on some columns as some of the scaling have the high values.
* From sklearn used preprocessing and by using StandardScaler() and fit\_transform() of x\_train and x\_test data.
* Here we have converted the data between 0 and 1 as the standard input to the models.

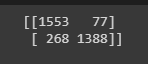
1. **Applying models :**

Applied three types of model on the fraud data :

1. **Logistic Regression :**
   * 1. **Classification report as below :**



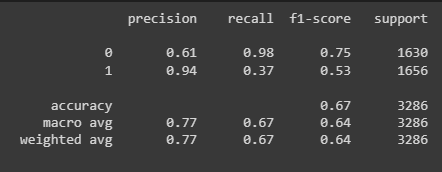
* + 1. **Confusion Matrix :**



* + 1. **Accuracy Score :**

****

* 1. **Decision Tree Classifiers :**
     1. **Classification Report :**

****

* + 1. **Confusion Matrix :**

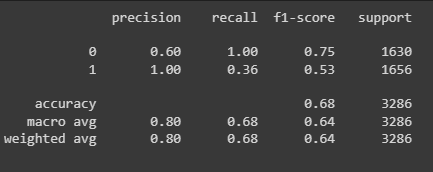
****

* + 1. **Accuracy Score :**

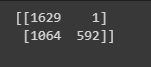
****

* 1. **Random Forest :**

**I. Classification report :**

****

**ii. Confusion Matrix :**

****

**iii. Accuracy Score :**

****

**Conclusion :**

Existing rule-based system is not capable of detection of all the fraud transaction. As the data is having imbalance and if we apply direct model prediction on the imbalance data we would have got the exact 1 accuracy, so the prediction is basically wrong. That is called the overfitting data.

Hence we can get the more accuracy of data by underfitting the imbalance data.

Transfer and cash\_out are two must used mode of transaction and we can see that TRANSFER and CASH\_OUT are also the only way in which fraud happen. Thus, we will focus on this type of transactions.

Logistic model is giving the more accuracy as compare to decision tree and random forest.