**CAPSTONE PROJECT SYNOPSIS**

**Proactive detection of fraud**

Table of Contents:

Abstract .....................................................................................................................................2 Acknowledgements................................................................................................................... 3 Certificate of Completion......................................................................................................... 4 CHAPTER 1: INTRODUCTION ........................................................................................... 6 1.1.Title & Objective of the study ........................................................................................... 6 1.2 Need of the Study................................................................................................................ 6 1.3 Data Sources........................................................................................................................ 6 1.4 Attributes Information....................................................................................................... 6 1.5. Tools & Techniques........................................................................................................... 6 CHAPTER 2: DATA PREPARATION.................................................................................. 7 2.1 Phase I – Data Extraction and Cleaning .......................................................................... 7 2.2. Null values Detection .........................................................................................................9 2.3. Exploratory Data Analysis.............................................................................................. 10 2.4. Train and Test Dataset.................................................................................................... 12 CHAPTER : 3.Model Building.............................................................................................. 13 3.1.Linear Regression ............................................................................................................ 14 3.2 OPTIMIZERS.................................................................................................................. 16 3.3.VIF ....................................................................................................................... 17

3.4. KEY FACTORS……………………………………………………………………. 18

CONCLUSION....................................................................................................................... 19

CHAPTER 1: INTRODUCTION

* 1. Title & Objective of the study:

“Proactive fraud detection” is the project title and objective is to find the patterns of transactions performed and help algorithms learn those patterns in identifying the fraudulent transactions and flag them.

1.2 Need of the Study

Ever since the advent of internet the digital revolution has rising and has creeped into all aspects to our lives. One of the most important digital revolution happened in financial system and especially transacting money to someone from any part of the world digitally. Digital transactions have become a part of daily life like purchasing a product online, sending money to friends, depositing cash in bank account, investment purposes etc., They had a lot of benefits so does paved way for fraudulent activities. People started using digital money transactions medium to launder money and make the money look like it comes from a legal source.

1.3 Data Sources

The Dataset under study consists of the information about the company such as step,type, amount,nameOrig,oldbalanceOrig,newbalanceOrig,nameDest,oldbalanceDest,newbalanceDest, isFraud and isFlaggedFraud.

1.4 Attributes Information:

The Dataset contains 6362620 rows and 11 columns.

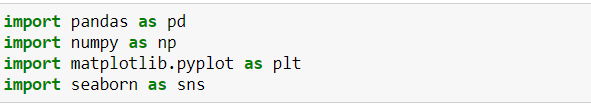
1.5. Tools & Techniques: Tools: Jupiter Notebook.

Techniques: Linear Regression, Optimizers

CHAPTER 2: DATA PREPARATION

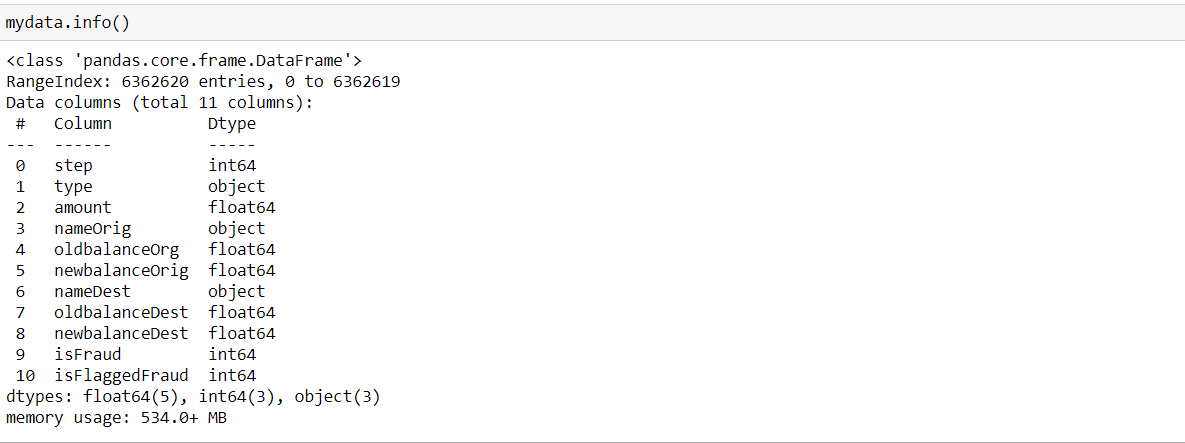
2.1 Phase I – Data Extraction and Cleaning:

At first let us import the necessary Libraries that we need for computation and processing

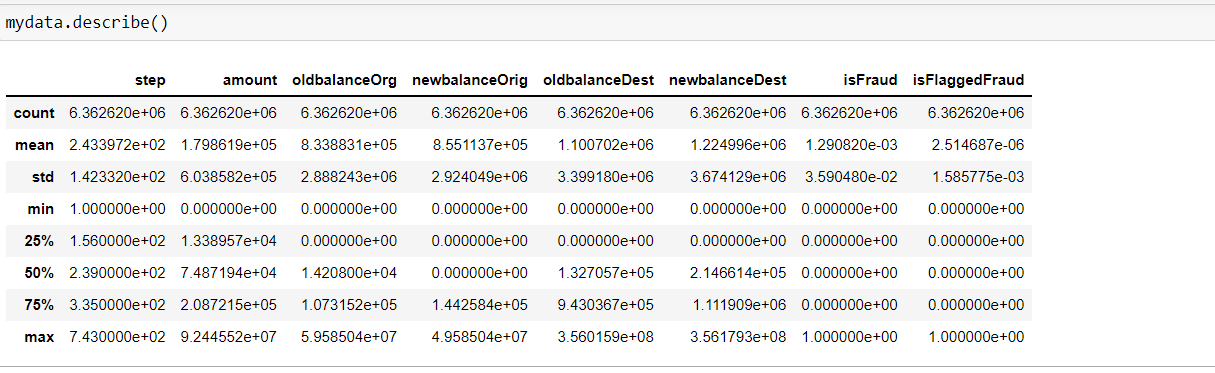


The Dataset contains 6362620 records and 11 Columns. And at first import the Dataset into python by using pandas function. Since, it‟s a CSV file we have to do as follows, 

And the data types are

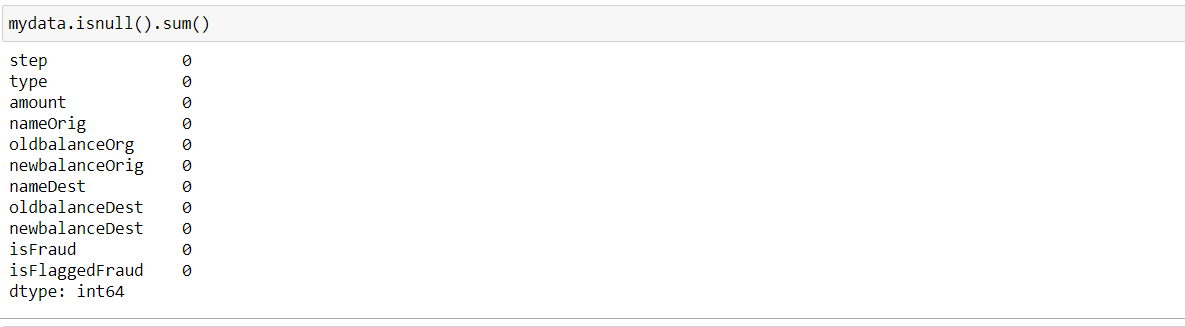


Basic statistic information of the Dataset:



2.2. Null values Detection :

Now we have check whether the Dataset consists of any null values that might be affecting the model. In order to identify null values,



We don’t have null values in any of the columns. The data is clean without null values hence So we can proceed with the computation.

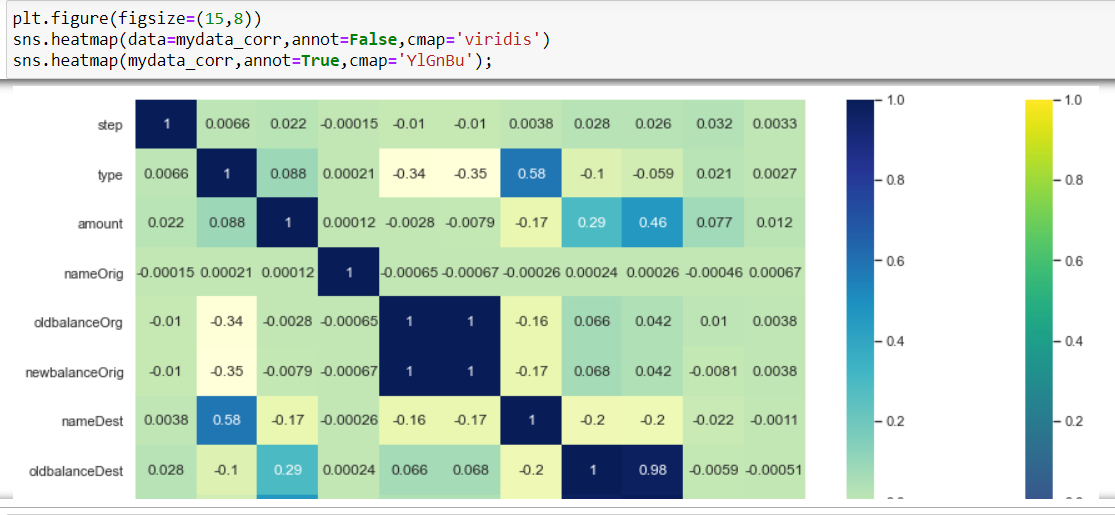
2.4. Exploratory Data Analysis:

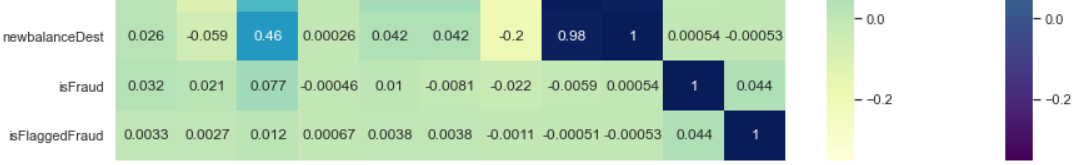
Through EDA process, we can visualize the dataset and could easily observe the insights of the data. At first let us check whether the variables among the data have multicollinearity.

#Heat map

A correlation heatmap is a heatmap that shows **a 2D correlation matrix between two discrete dimensions**, using colored cells to represent data from usually a monochromatic scale. The values of the first dimension appear as the rows of the table

while of the second dimension as a column

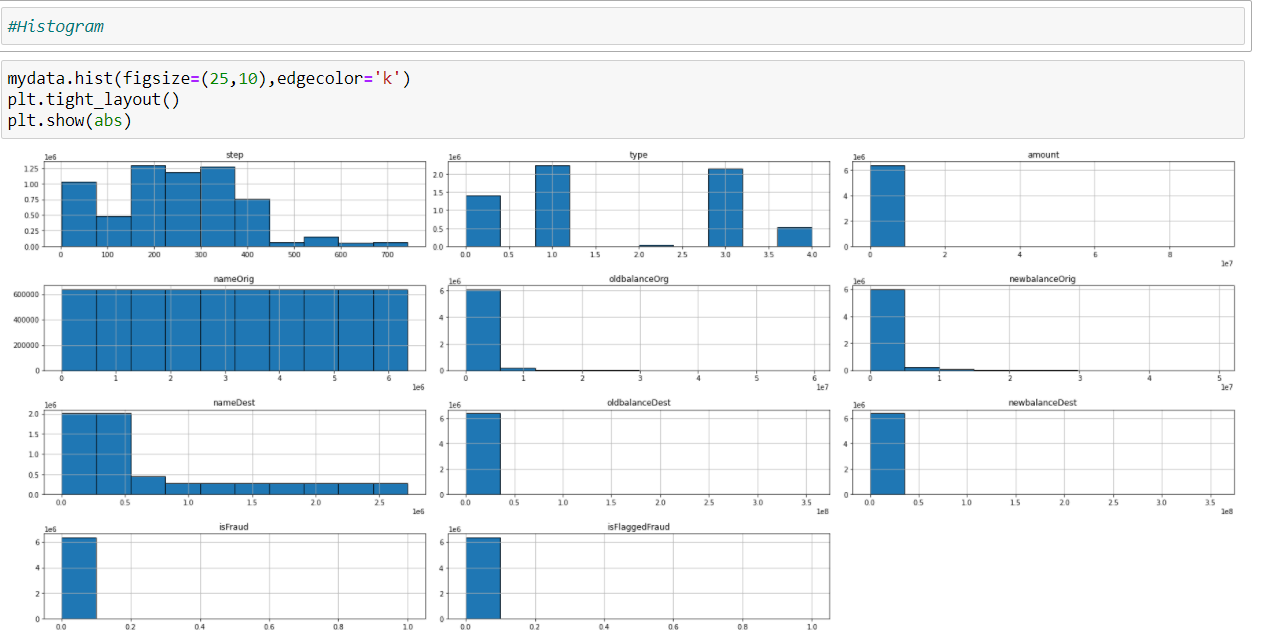




#Histogram

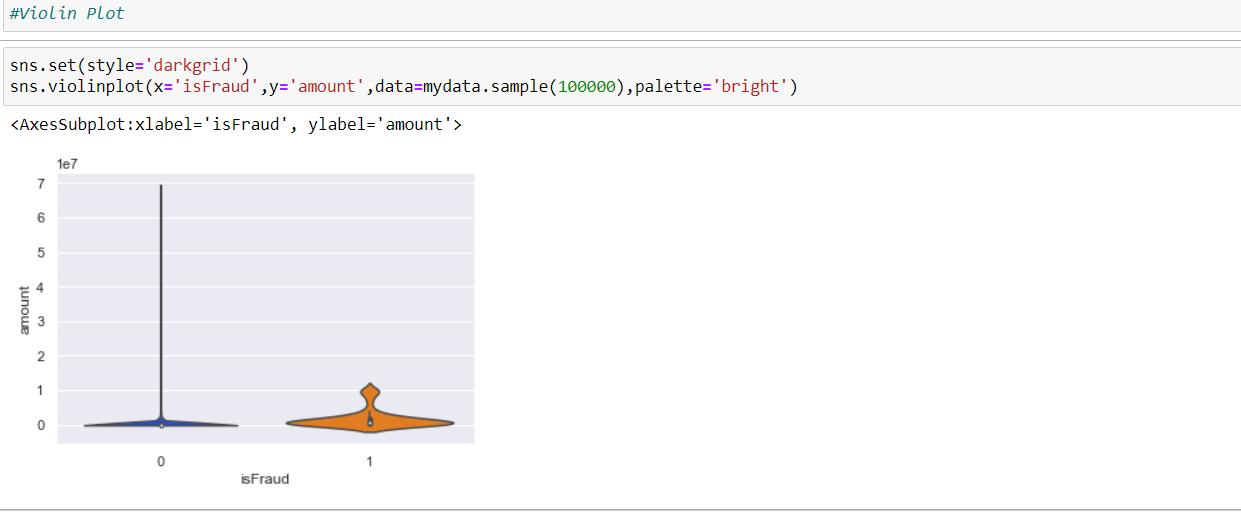
It is a type of bar plot where X-axis represents the bin ranges while Y-axis gives information about frequency.

Below histogram shows the graph of all the variables of the data set.



# Violin Plot

Violin plots are similar to box plots, except that they also show the probability density of the data at different values. It shows the relation how the data points are spread.



#Joint Plot

This function provides a convenient interface to the 'JointGrid' class, with several canned plot kinds ,below showing the spread of data point between variables isFlaggedFraud and amount.



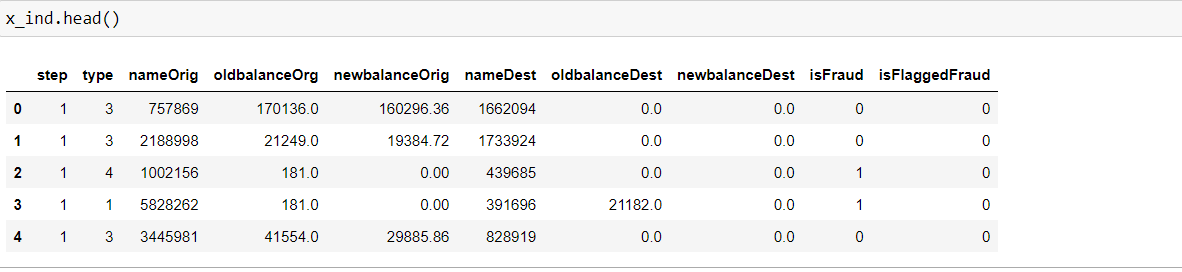
2.5. Train and Test Dataset:

In order to proceed with model building, first we have split our Dataset into train and test parts. From our Dataset, the dependent variable is amount and all other variables are independent variables.

Let’s first split the variables into Independent and Dependent datasets.



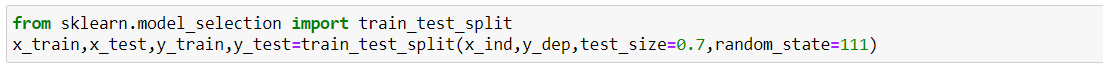
All the independent variables will be assigned under x\_ind,



The dependent variable will be assigned under y\_dep which is amount,



We have x\_ind , and y\_dep and now we have to split the dataset into train and test . For this, We can import „train\_test\_split‟ function from scikit learn.

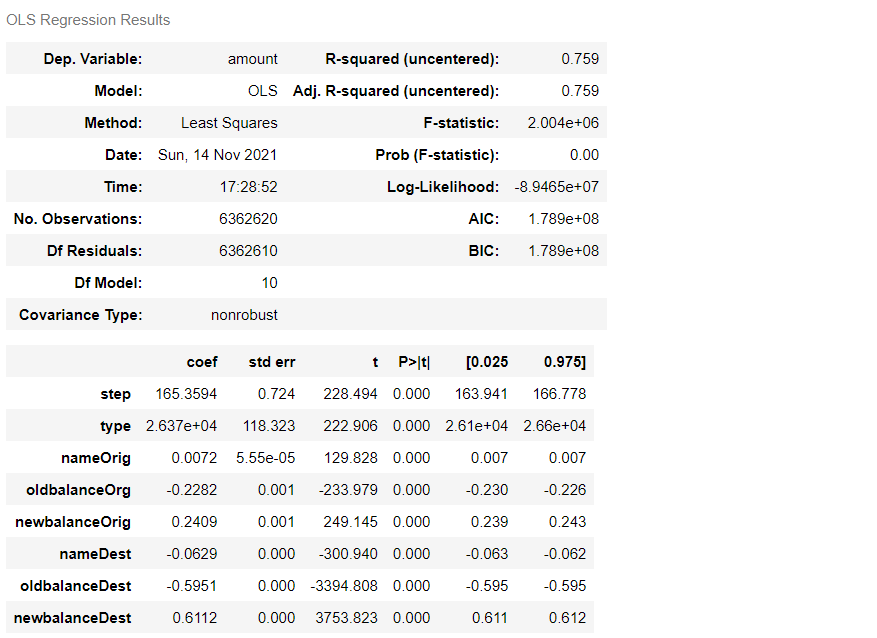


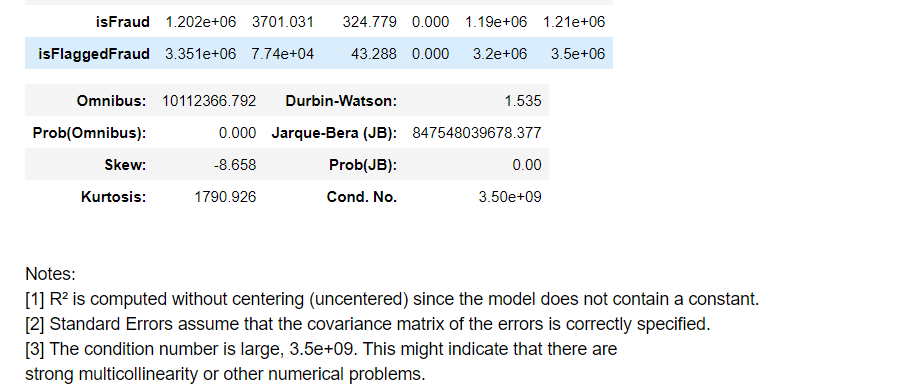
Now , we got x\_train , x\_test ,y\_train ,y\_test in order for the model building purpose.

CHAPTER: 3.Model Building

3.1. Linear Regression:

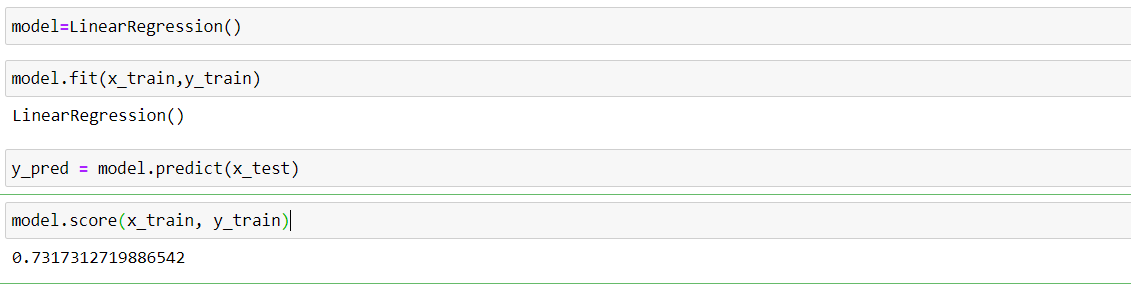






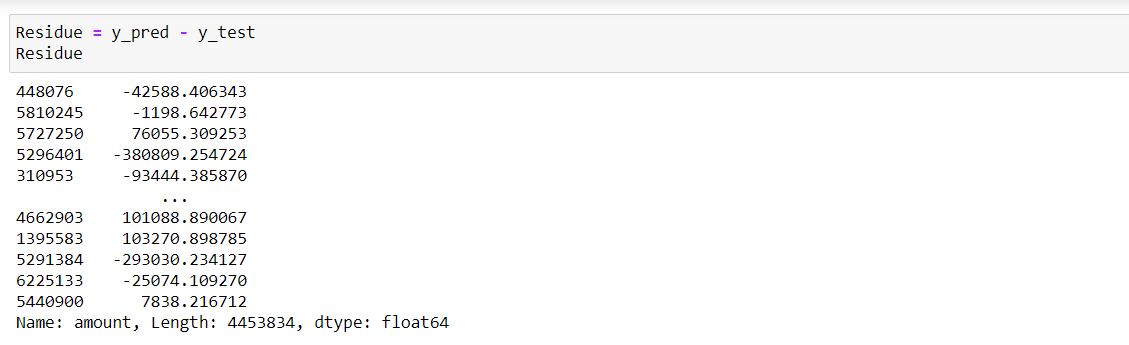
From the summary the R-square value is 0.759 and not even a single variable has P-value greater than 0.05.

Imported Linear Regression from Sklearn and we fit the x\_train and y\_train into the model.

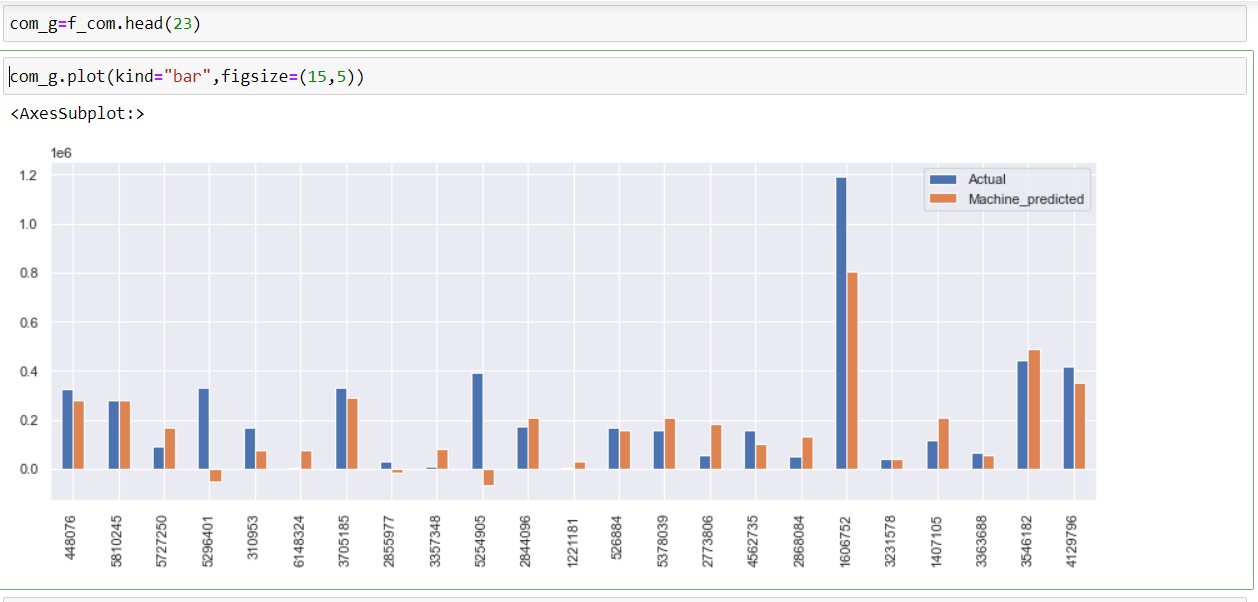


Comparison of actual and predicted weights

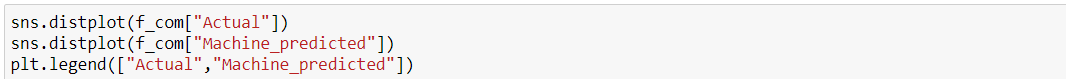
RESIDUAL:

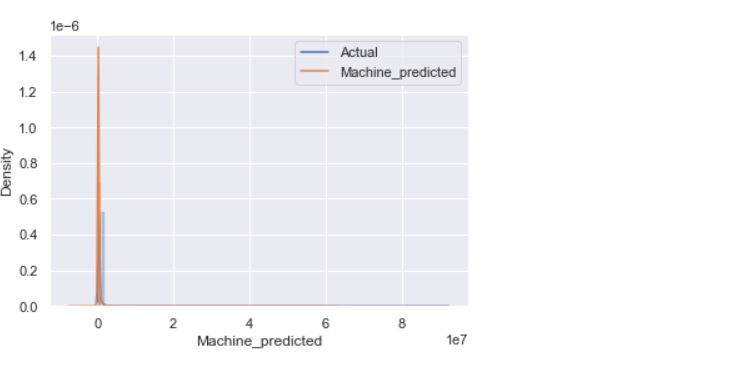


Graph showing the prediction between actual and machine predicted values,



Distplot showing the graph for machine predicated and actual values,



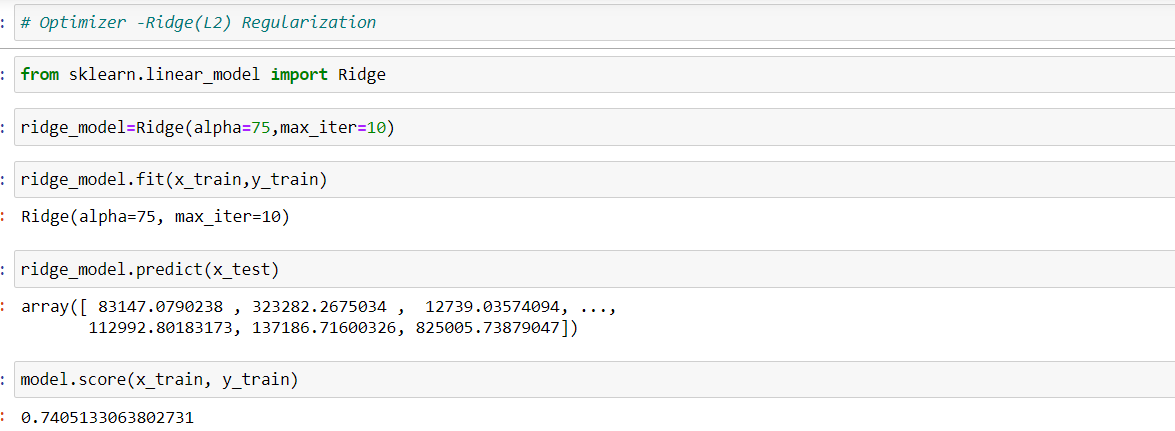


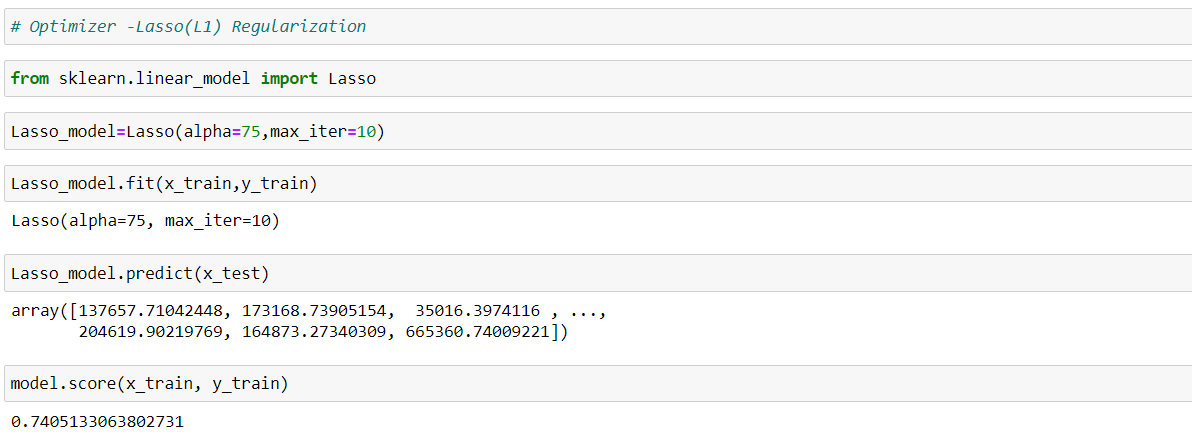
**3.2 OPTIMIZERS**

**Ridge and Lasso Regression are types of Regularization techniques**

**Regularization techniques are used to deal with overfitting and when the dataset is large**

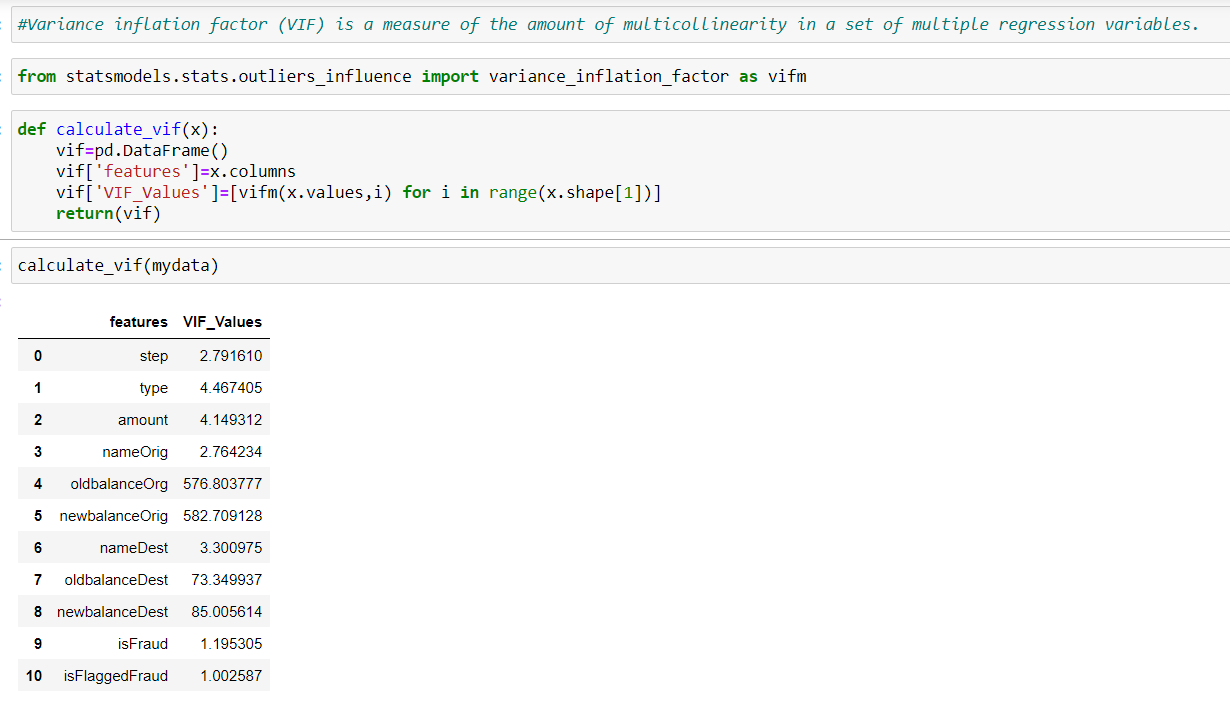
**Ridge and Lasso Regression involve adding penalties to the regression function**





Accuracy score is 74% as per the Linear regression ,optimizers are used for regularization as this is huge dataset.

**3.3 V.I.F**



The higher the value, the greater the correlation of the variable with other variables.

**3.4 KEY FACTORS**

# Key factors that predict fraudulent customer and Prevention measures to be adopted.

1. **Difference in balance:** It is an universal truth that the amount debited from senders account gets credited into the receivers account without any deviation in cents. But what if there is a deviation incase of the amount debited and credited. Some could be due to the charges levied by the service providers, yet we need to flag such unusual instances.
2. **Surge indicator:** Also we have to trigger flag when large amount are involved in the transaction. From the distribution of amount we understood that we have a lot of outliers with high amount in transactions. Hence we consider the 75th percentile(450k) as our threshold and amount which is greater than 450k will be triggered as a flag.
3. **Frequency indicator:** Here we flag the user and not the transaction. When there is a receiver who receives money from a lot of people, it could be a trigger as it can be for some illegal games of chance or luck. Hence it is flagged when there is a receiver who receives money for more than 20 times.
4. **Merchant indicator:** The customer ids in receiver starts with 'M' which means that they are merchants and they obviously will have a lot of receiving transactions. So we also flag whenever there is a merchant receiver