Assignment 3 Report

CREDIT CARD FRAUD DETECTION

**INFO 7390**

**Advance Data Science & Architecture**

Professor: Srikanth Krishnamurthy

By: Team 11

Aditya Pawar

Dhairya Jaiswal

RangaChari Vinjaramuri

Contents

1. Information about the dataset.

a. Under sampling

b. SMOTE (Oversampling)

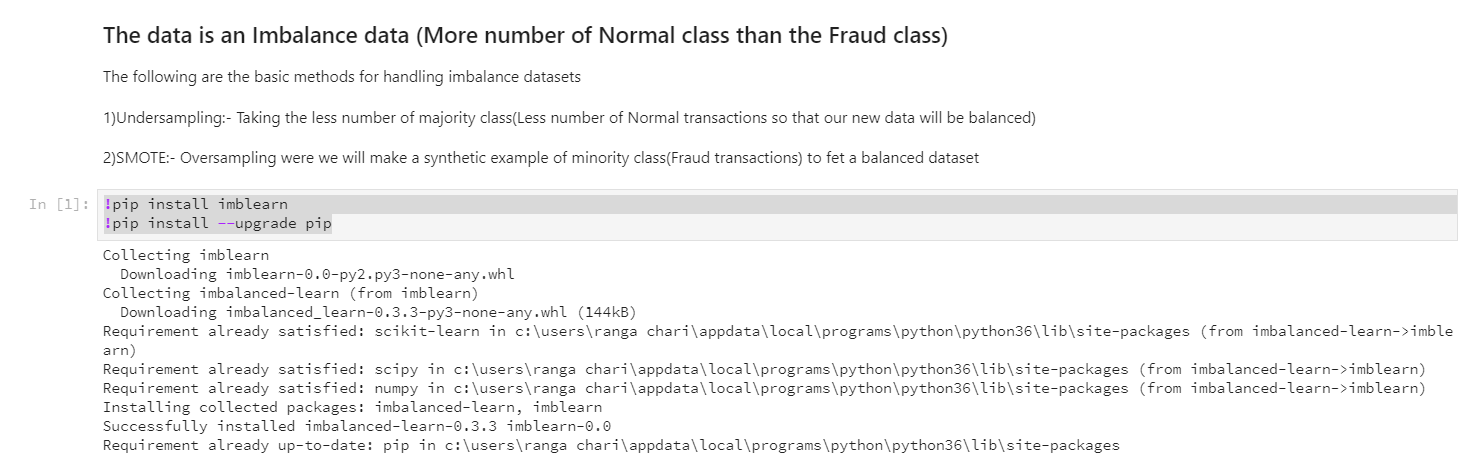
1. Exploratory data analysis Techniques
2. Anomaly detection techniques

Z-score method

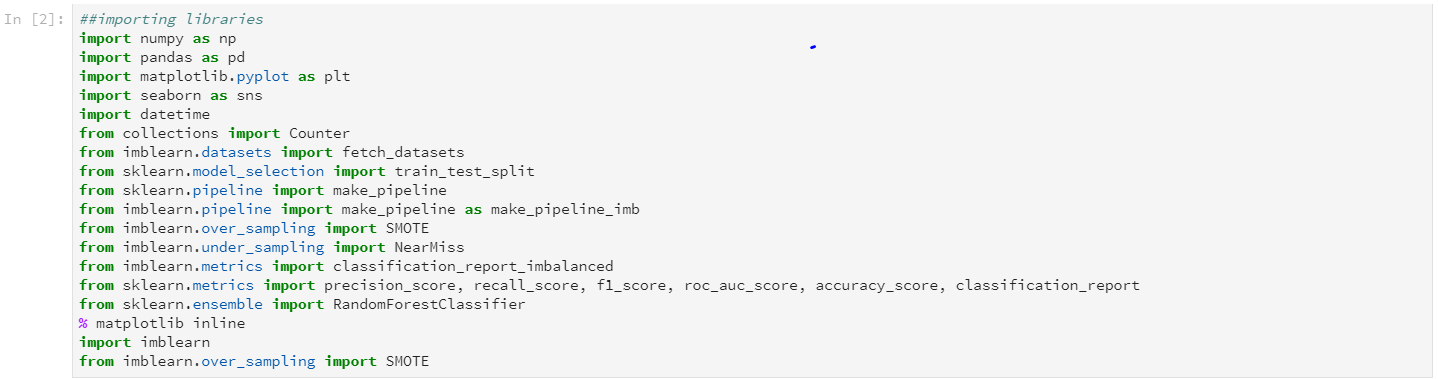
1. Feature engineering and feature reduction
2. SMOTE application
3. Application of Machine Learning Algorithm
4. Logistic regression
5. Random Forest Classification
6. SVM
7. Naïve bayes
8. Accuracy Metric
9. Confusion Matrix
10. Precision, recall and f1 score
11. ROC curve
12. Cross validation
14. Infromation about the dataset

The data consists of Normal Transactions and Fraud transaction. The data is an Imbalance data (More number of Normal class than the Fraud class). So its not possible to split the dataset into train and test dataset. Following are the fuctions carried to handle such type of dataset:

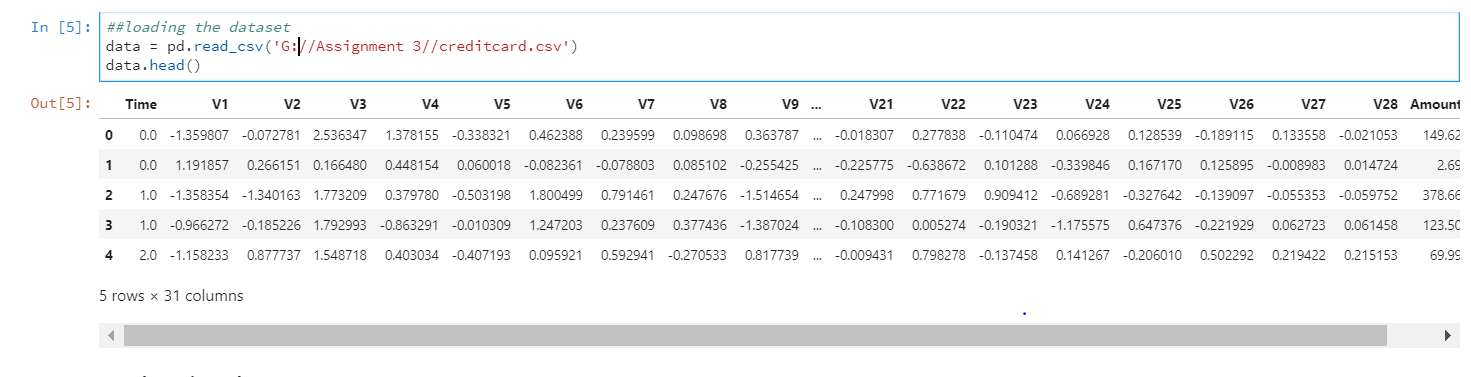
1. Undersampling:- Taking the less number of majority class(Less number of Normal transactions so that our new data will be balanced.
2. SMOTE:- Oversampling were we will make a synthetic example of minority class(Fraud transactions) to fet a balanced dataset



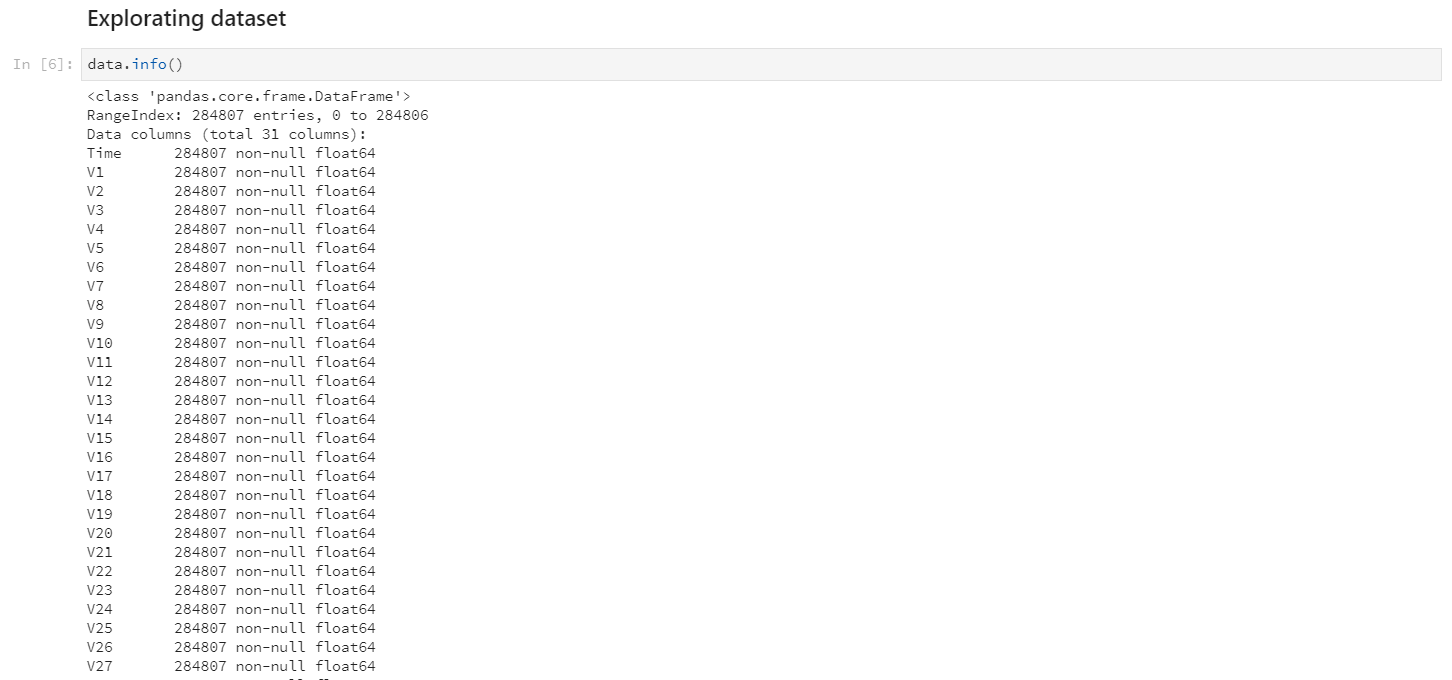
Libraries imported and used:



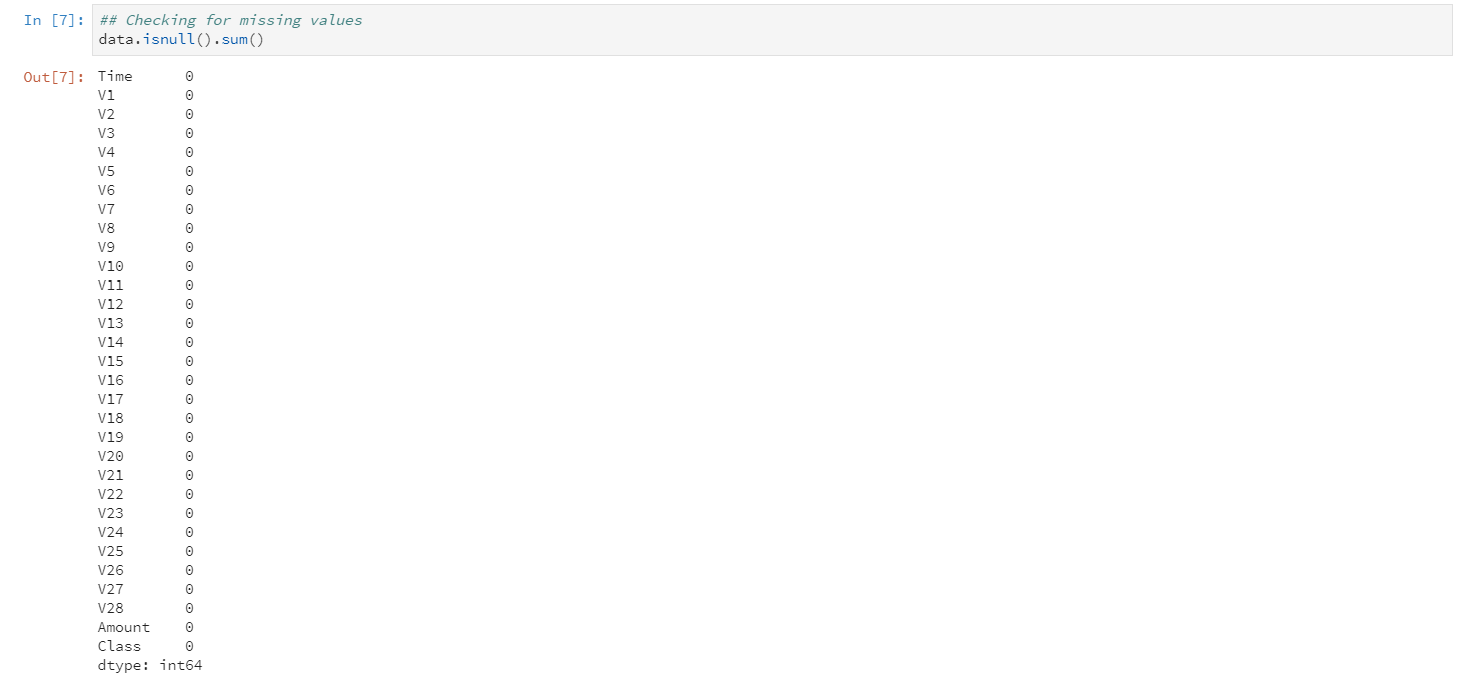
The Data set description:



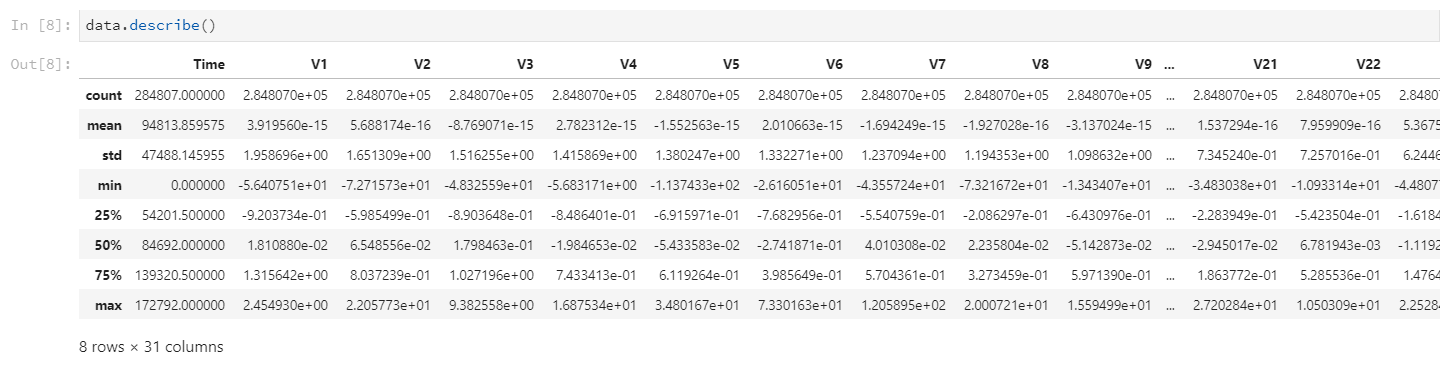
1. Exploratory Analysis:



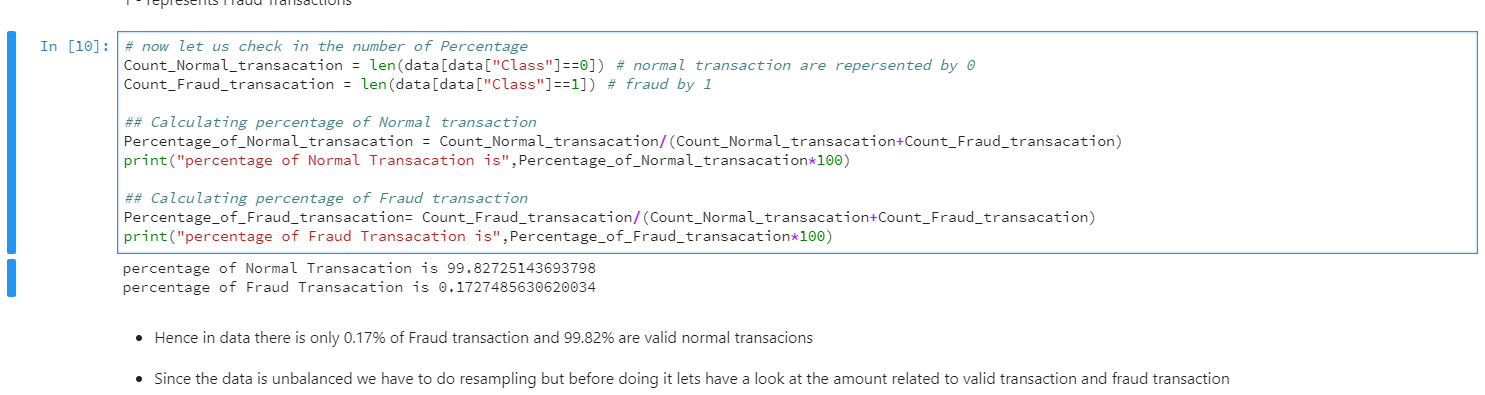
* To check Null values



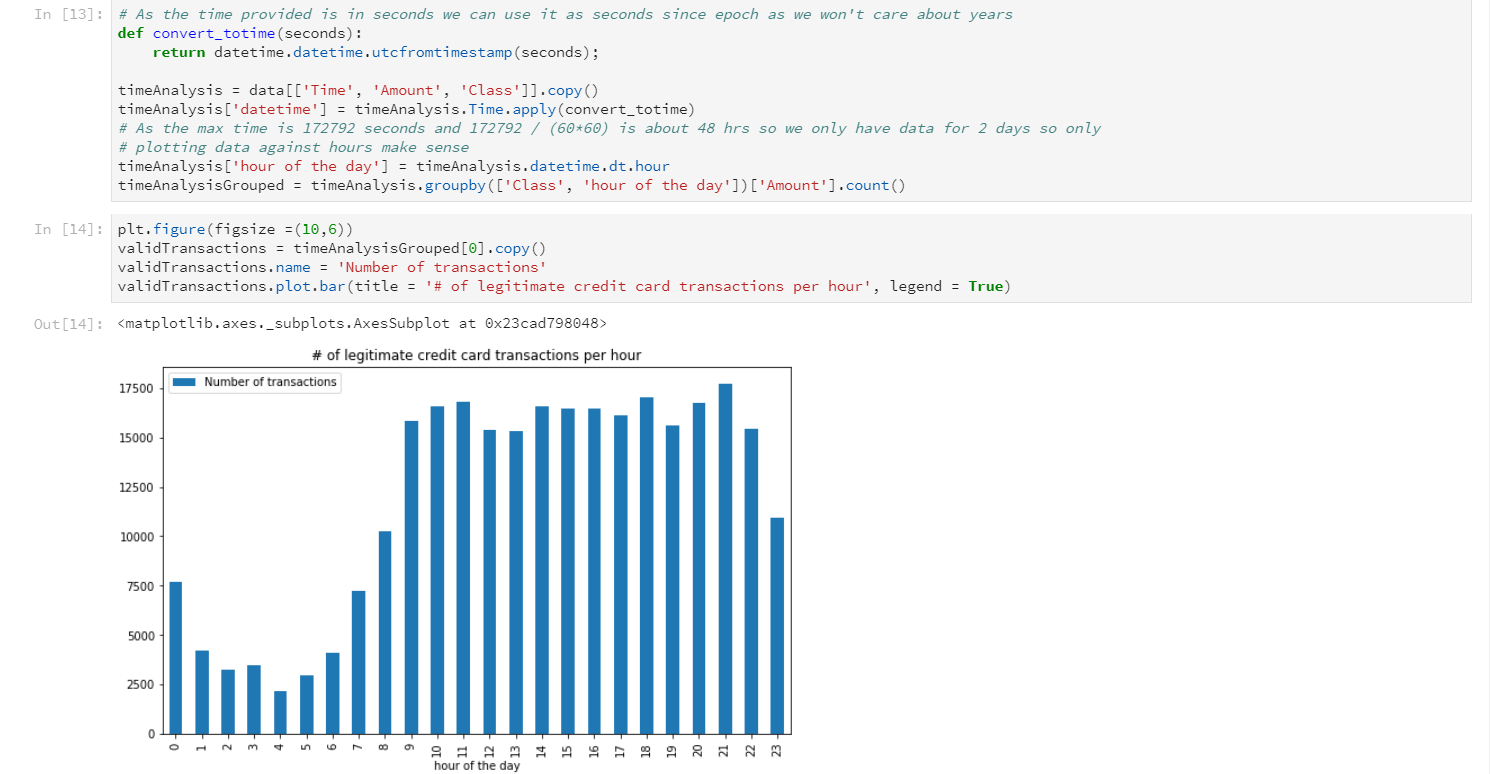
Description of the Data frame:

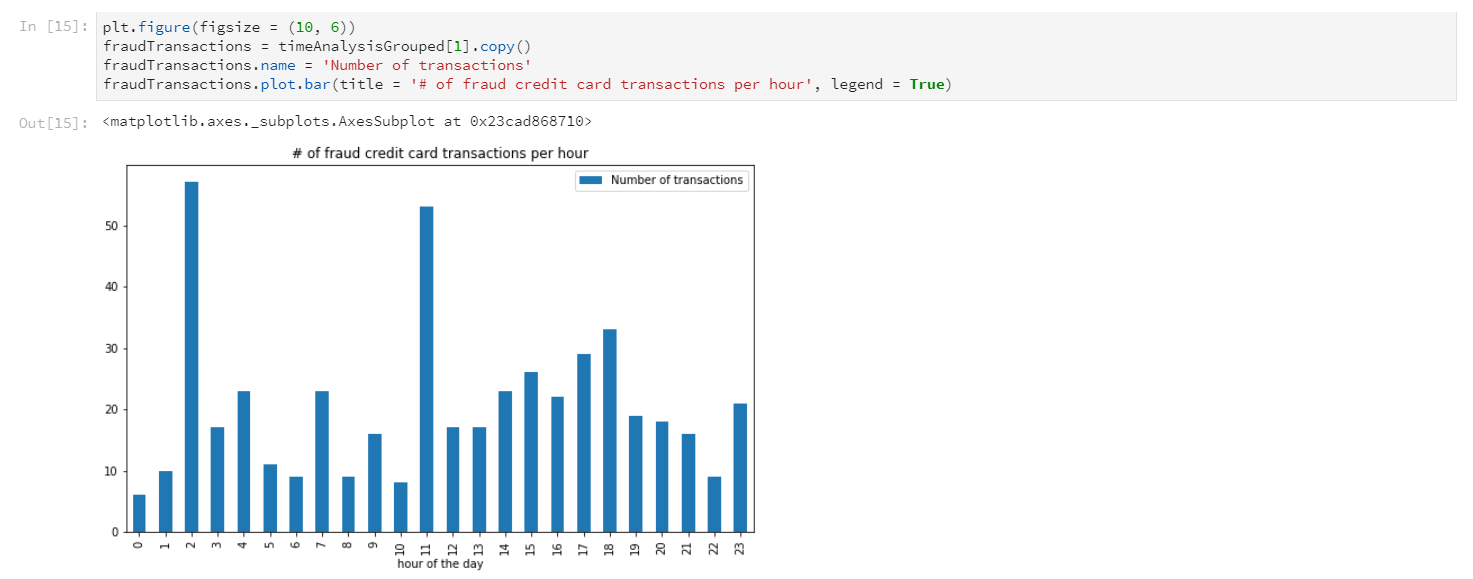


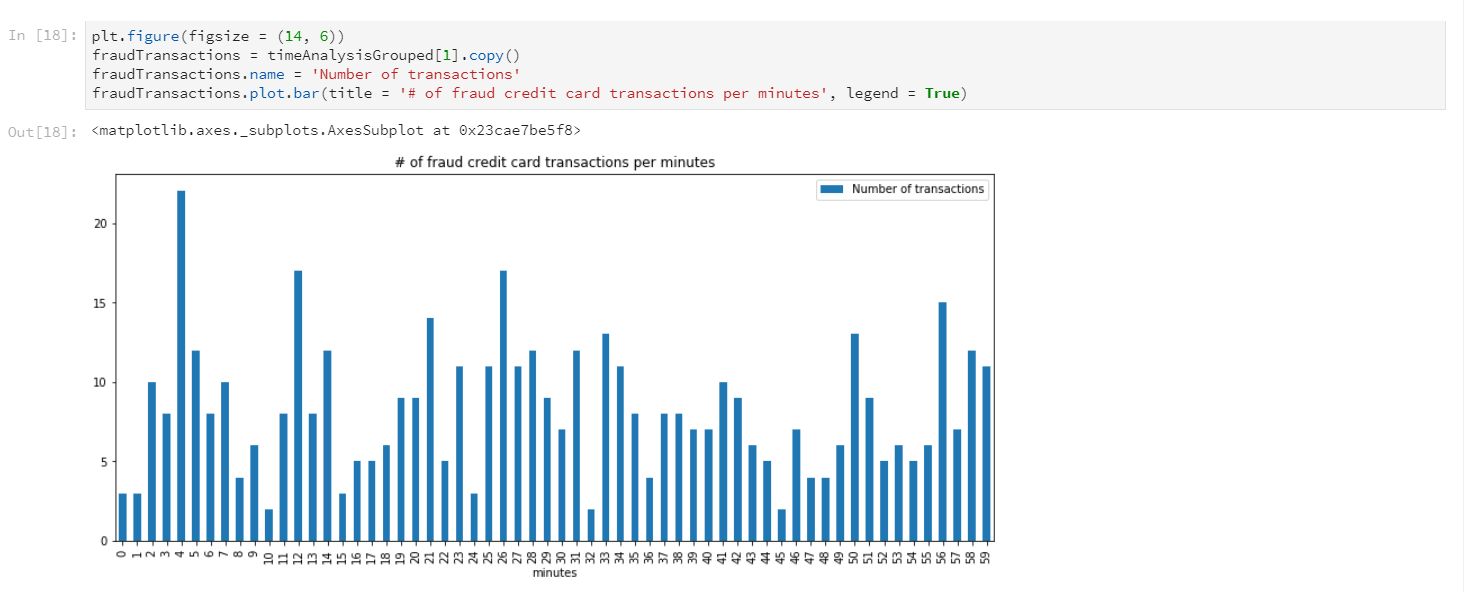


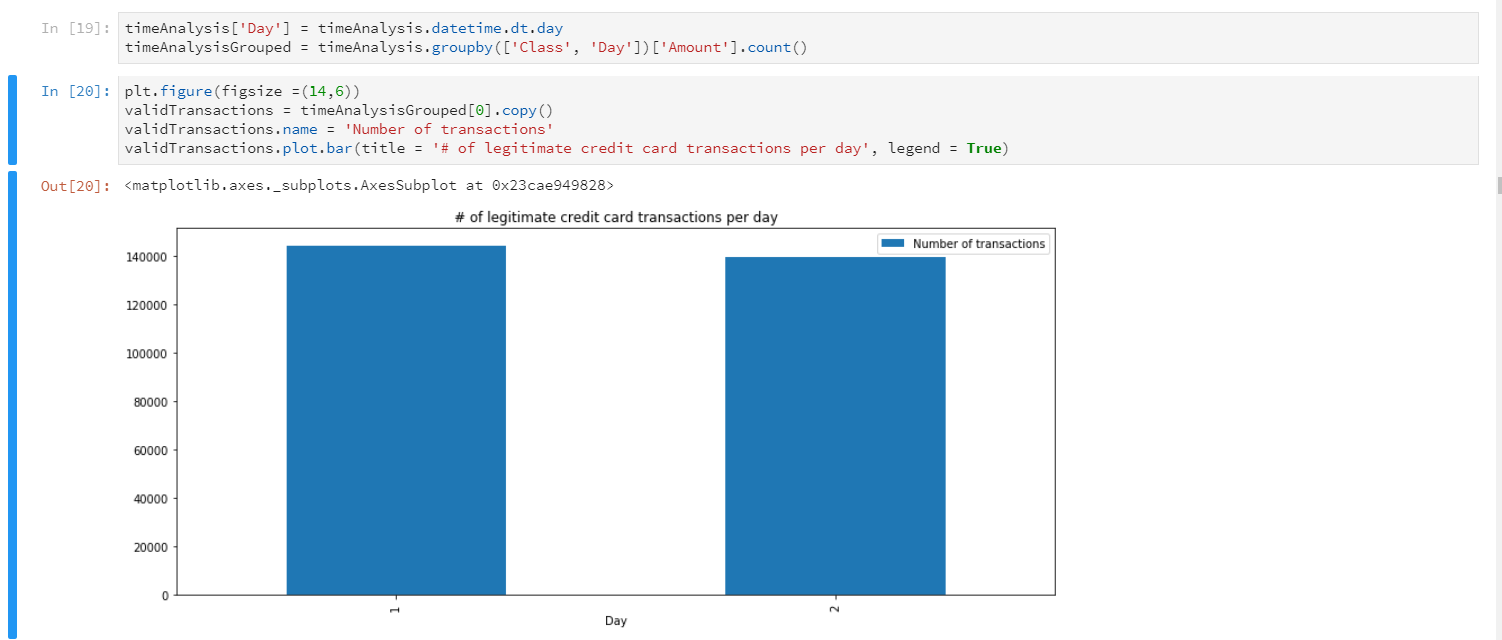


* Hence in data there is only 0.17% of Fraud transaction and 99.82% are valid normal transactions
* Since the data is unbalanced we have to do resampling but before doing it lets have a look at the amount related to valid transaction and fraud transaction
* Here now after exploring data we can say there is no pattern in data, We are not going to perform feature engineering in first instance. The dataset has been already downgraded using PCA in order to contain 30 features (28 anamolised + time + amount).
* Changing the performance metric:
  + Use the confusion matrix to calculate Precision, Recall
  + F1score (weighted average of precision recall)
  + Use Kappa - which is a classification accuracy normalized by the imbalance of the classes in the data
  + ROC curves - calculates sensitivity/specificity ratio.





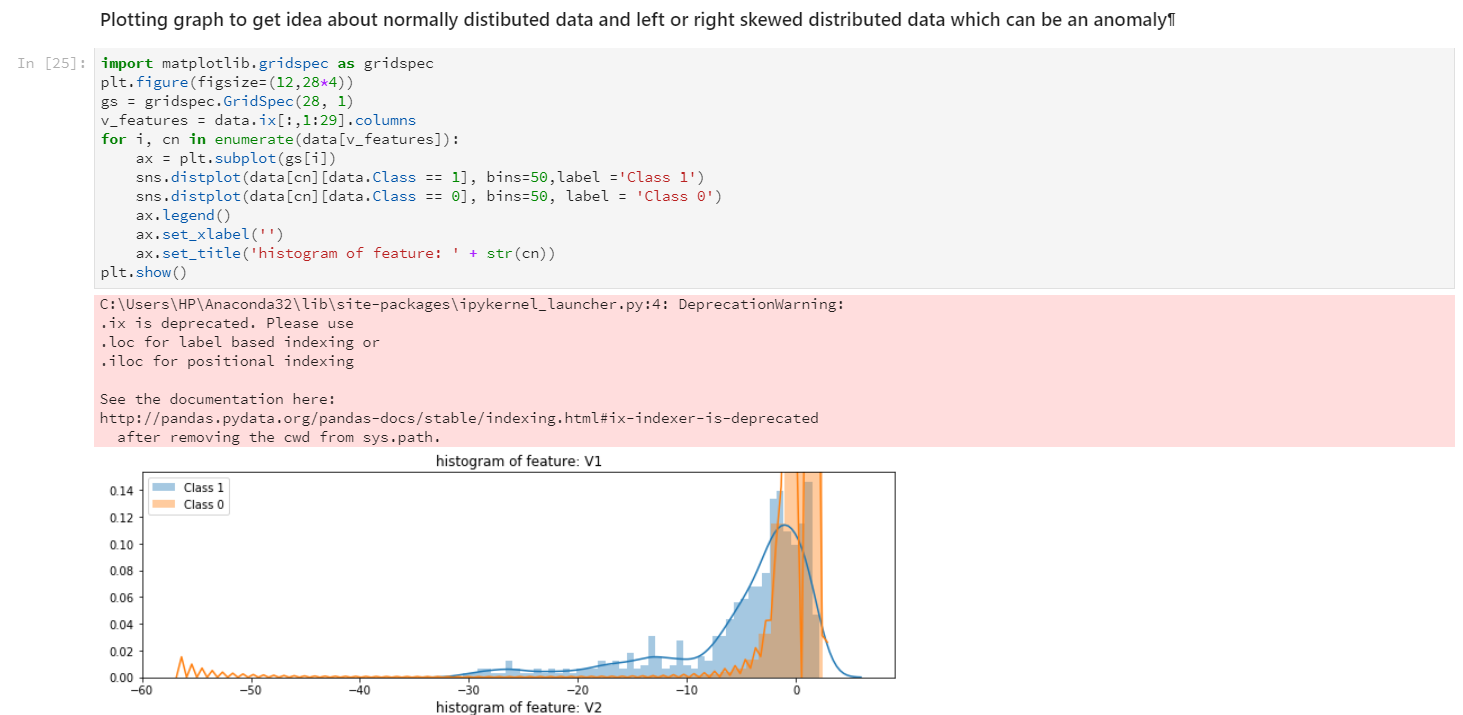


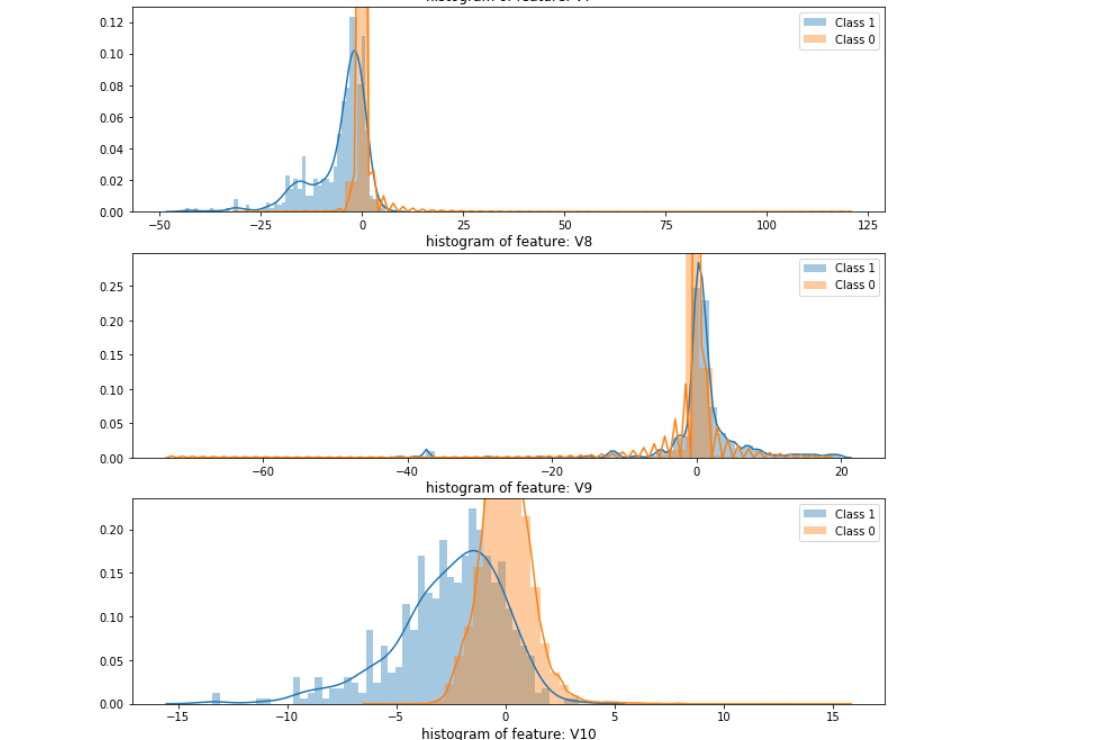
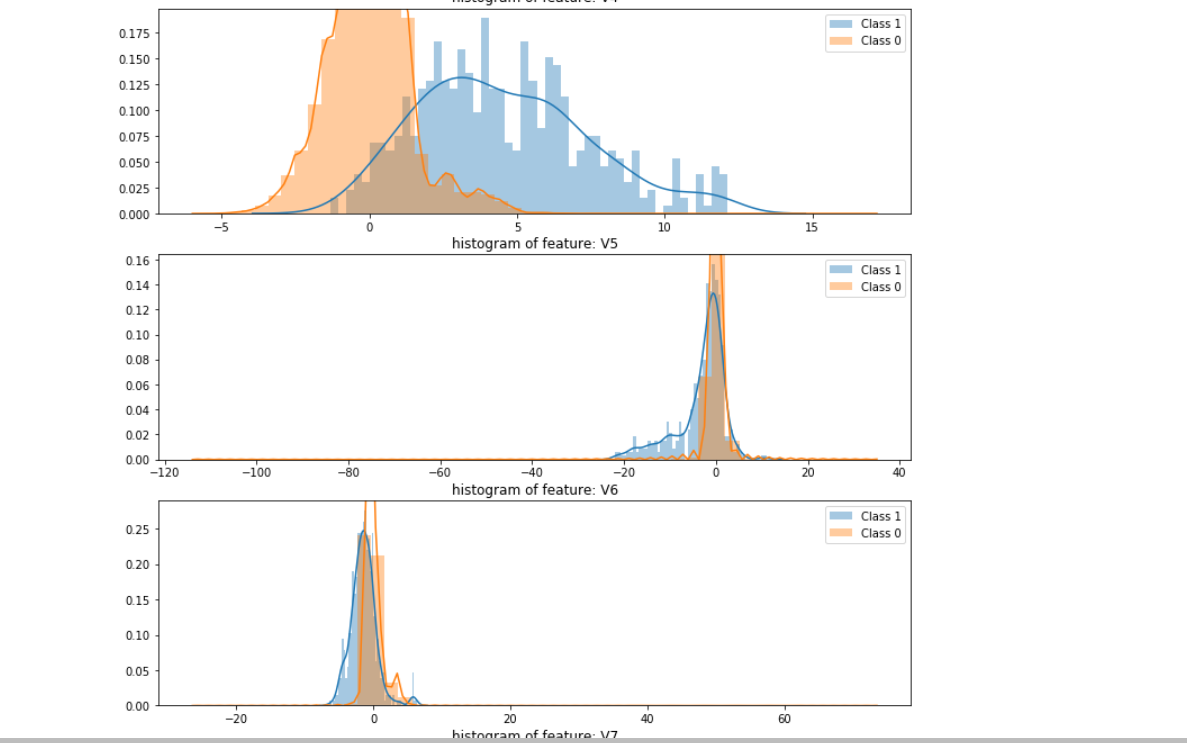
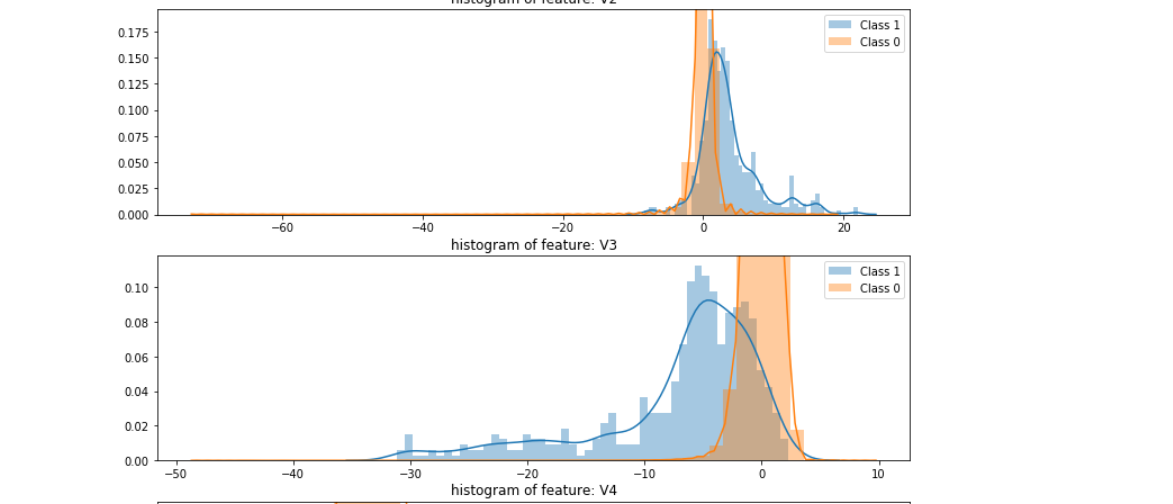


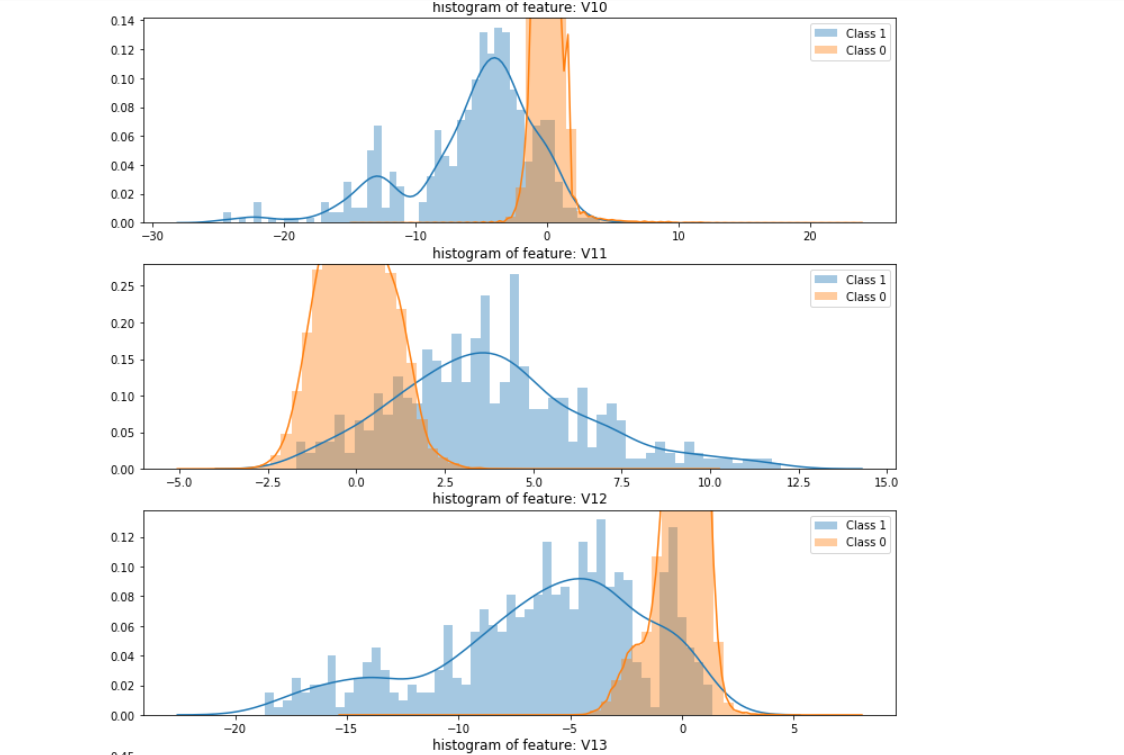


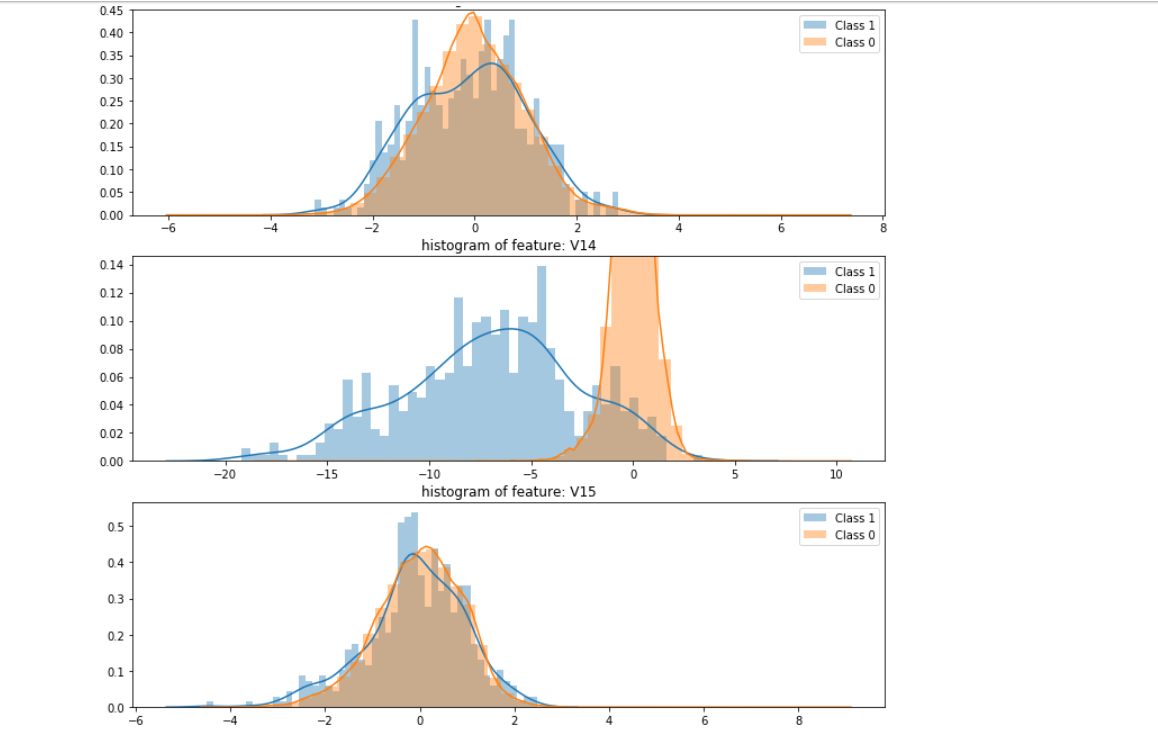
1. Outlier detection Technique

z-score

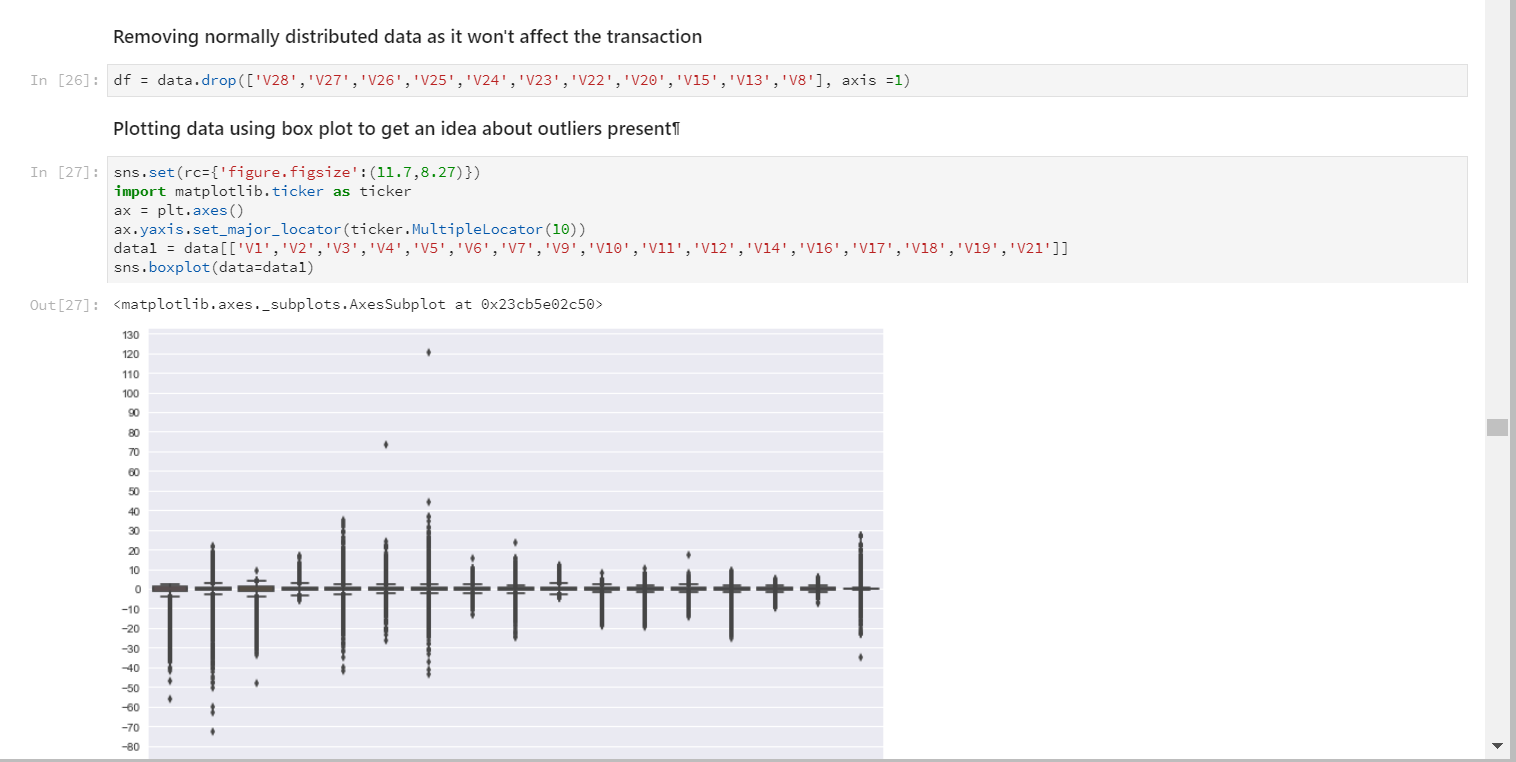




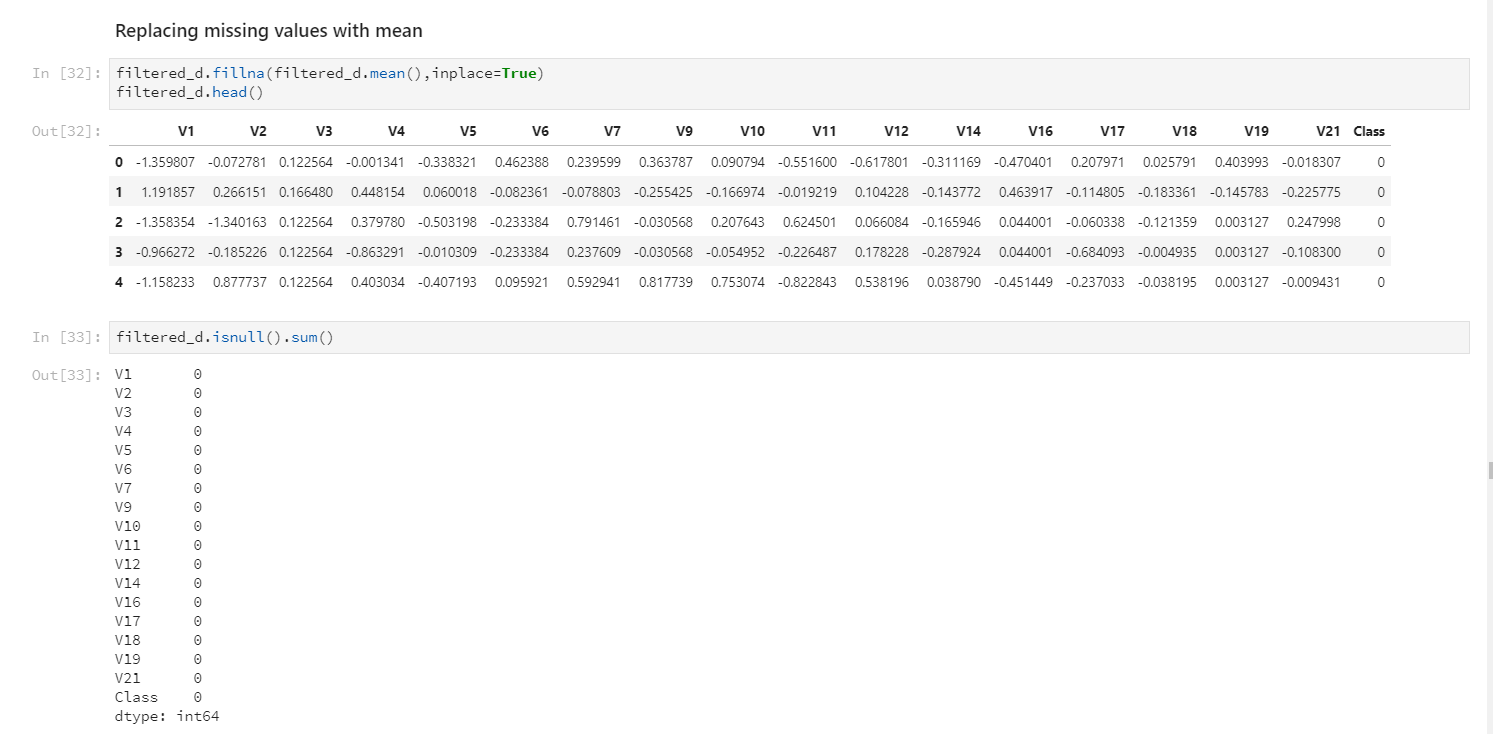


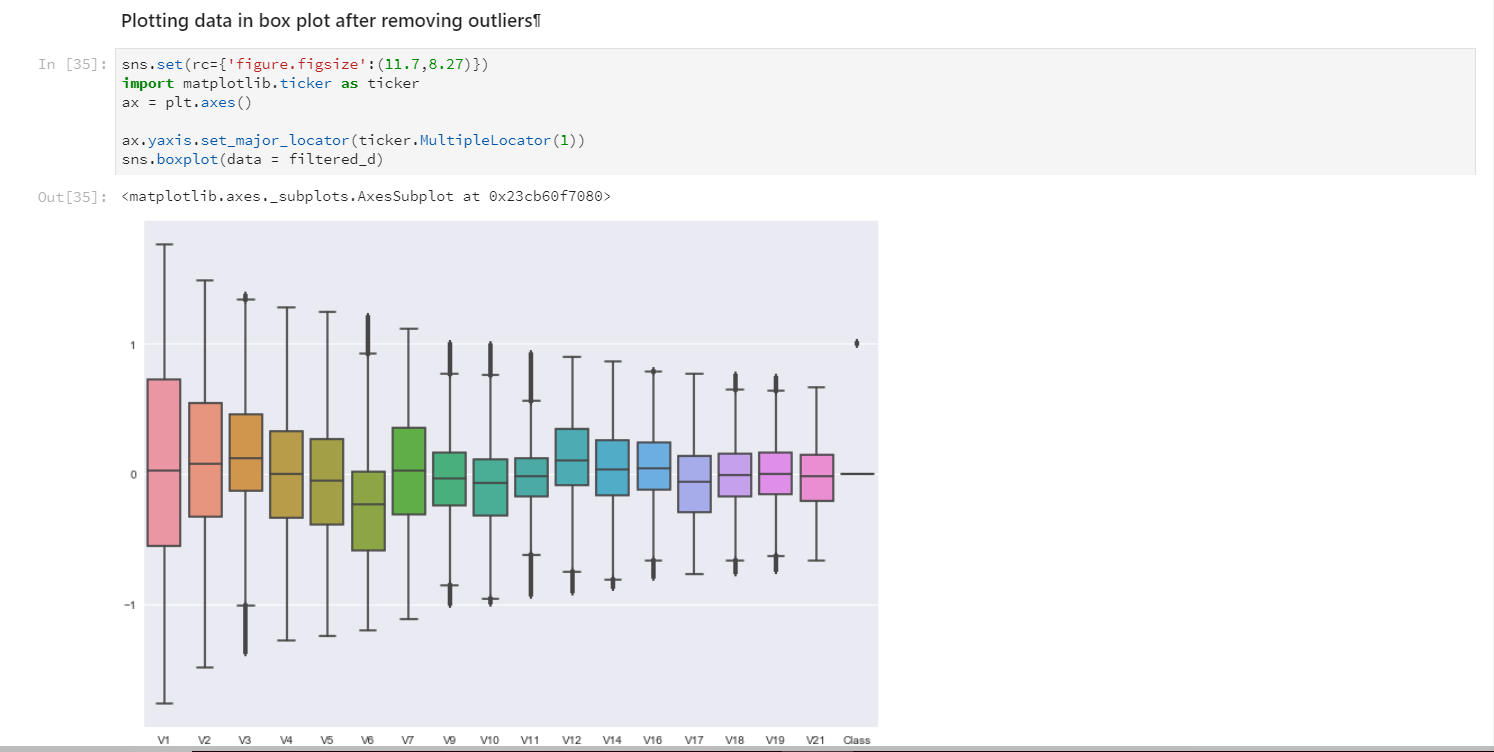


1. Featuring Engineering and Feature reduction



* Replacing missing values in the data after feature reduction with mean





### Application of Algorithms

* We will now compare what happens when using resampling and when not using it. We will test this approach using following models:
  + Logistic Regression
  + SVM
  + Random forest classifier
  + Naïve bayes





### **Accuracy Metrics And Error Metrics**

Before re sampling lets have look at the different accuracy metrices

* Accuracy = TP+TN/Total
* Precison = TP/(TP+FP)
* Recall = TP/(TP+FN)
* TP = True possitive means no of possitve cases which are predicted positive
* TN = True negative means no of negative cases which are predicted negative
* FP = False possitve means no of negative cases which are predicted positive
* FN= False Negative means no of possitive cases which are predicted negative

Now for our case recall will be a better option because in these case no of normal transacations will be very high than the no of fraud cases and sometime a fraud case will be predicted as normal. So, recall will give us a sense of only fraud cases

In this we will resample our data with different size,then we will try to use this resampled data to train our model.Then we will use this model to predict for our original data

