Reg NO:RA1811027010070
Name:Kotha Sai Narasimha
Rao Branch:Cse-BD ‘J1’Sec
Faculty Name:Nithyakani P

Data Science Assignment-1 Credit Card Fraud Detection(Stage-3&4)

Dataset:https://www.kaggle.com/mlg-ulb/creditcardfraud

Stage-3: Model Planning:

The data science team including the stakeholders builds the framework for model building by determining the techniques and methods particularly addressing the business problem. At this stage, the business problem is clearly emphasized and distributed across the departments breaking the data silos. Introducing scientific methods and corresponding key parameter variables are chosen at this stage to solve the business problem. Without the discovery phase and data preparation, a premeditated selection of scientific method will not apply to the business problem.

Model Planning Steps:

1.In this Stage First we are going to Find out The types of attributes in the DataSet:

In our dataset all the features are
Numerical Values

2.Checking Null Values in our Dataset:Null Values are not present In our dataset.

3.By using Visualization techniques we can find the Relation ship between attributes.

Some of the Visualization techniques are:

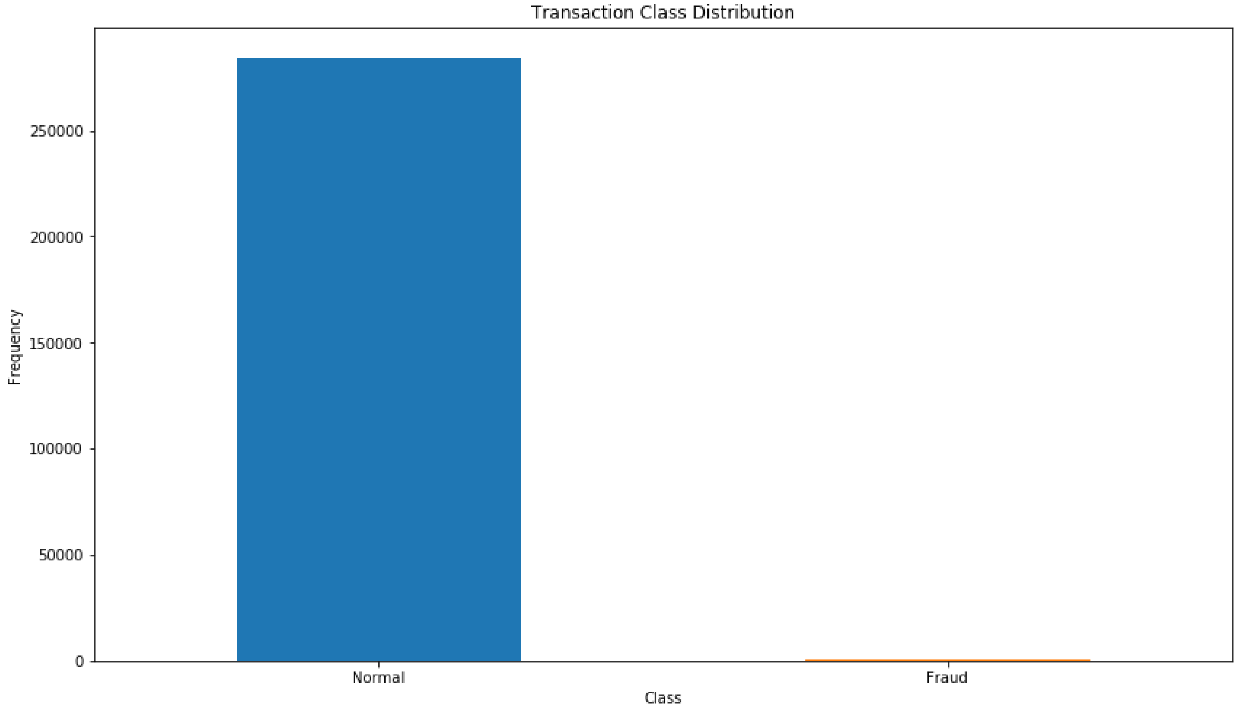
1.Boxplot: **boxplot** is a method for graphically depicting groups of numerical data through their quartiles.

2.Scatterplot:A **scatter plot** (aka **scatter** chart, **scatter graph**) uses dots to represent values for two different numeric variables

3.Heat map:A **heat map** is data analysis software that uses color the way a bar graph uses height and width: as a data visualization tool

These are some of the Visualization techniques.

4.The below graph Describes the Transaction Class distribution(Output variable) In the Graph: ->The normal Transactions are more than 2.5lakh and fraud line transactions are very less.



->The dataset isimbalanced.

->So Resampling should be important to balance the dataset.

The two powerfull Resampling Techiniques are:

1.Over Sampling

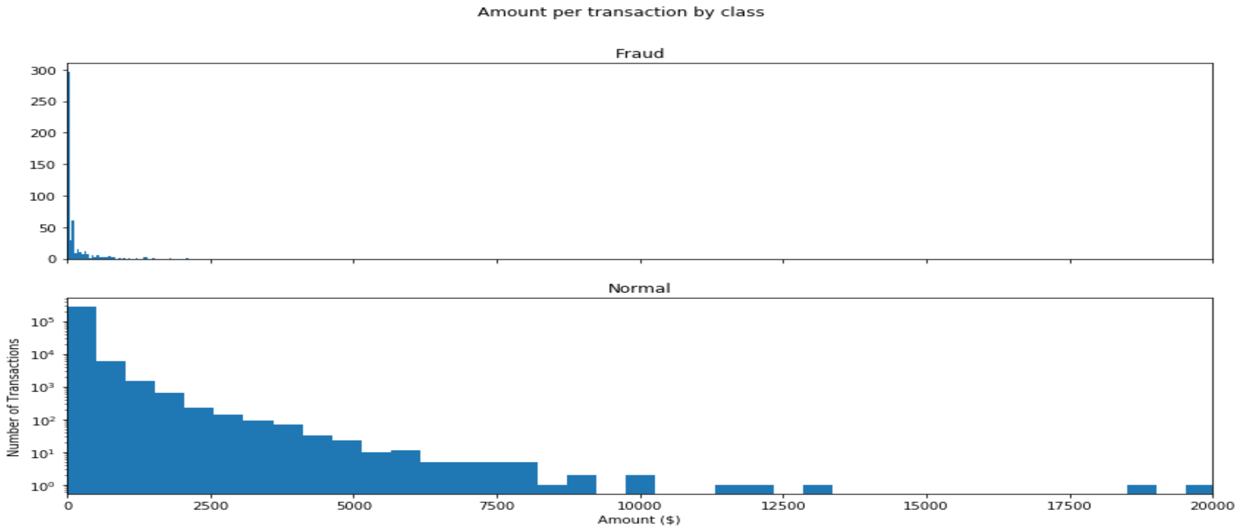
2.Under Sampling

Over Sampling and Resampling: When one class of data is the underrepresented minority class in the data sample, over sampling techniques maybe used to duplicate these results for a more balanced amount of positive results in training. Over sampling is used when the amount of data collected is insufficient. A popular over sampling technique is SMOTE

(Synthetic Minority Over-sampling Technique), which creates synthetic samples by randomly sampling the characteristics from occurrences in the minority class. Conversely, if a class of data is the overrepresented majority class, under sampling may be used to balance it with the minority class. Under sampling is used when the amount of collected data is sufficient. Common methods of under sampling include cluster centroids and Tomek links, both of which target potential overlapping characteristics within the collected data sets to reduce the amount of majority data.

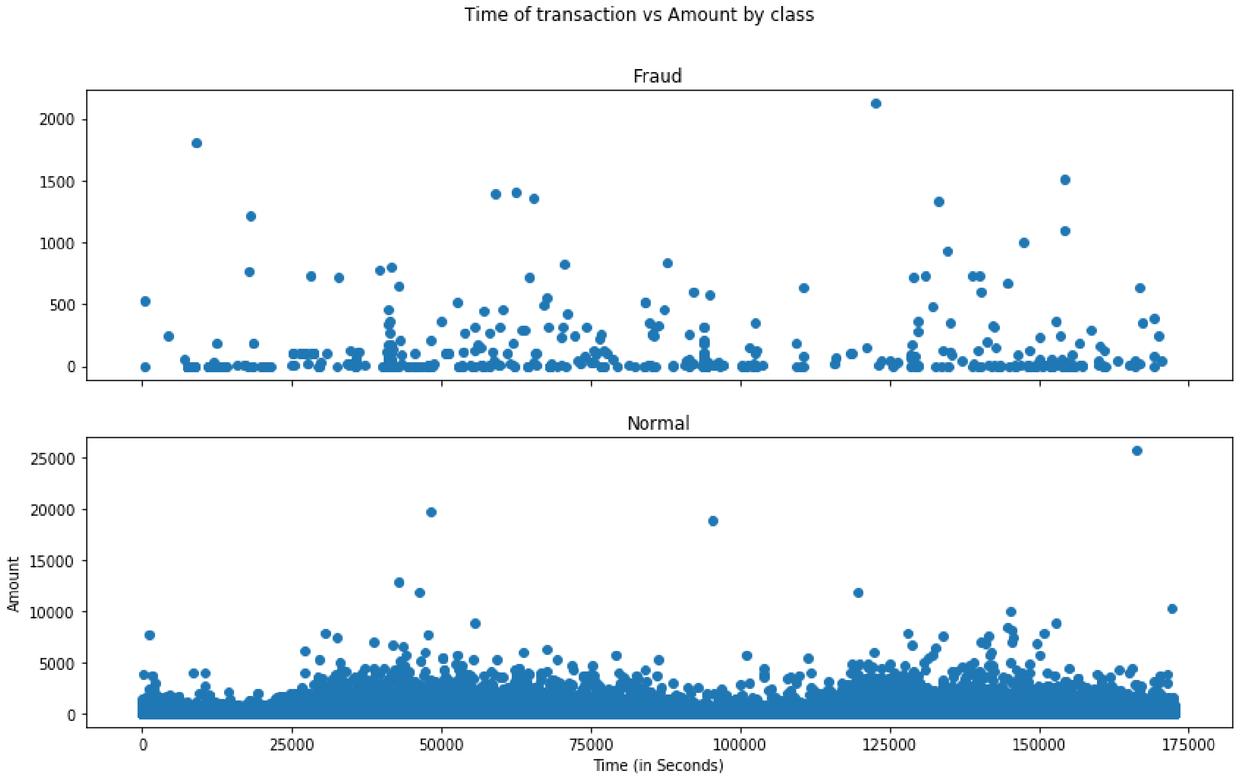
In both over sampling and under sampling, simple data duplication is rarely suggested. Generally, over sampling is preferable as under sampling can result in the loss of important data. Under sampling is suggested when the amount of data collected is larger than ideal and can help data mining tools to stay within the limits of what they can effectively process.

5.The Transcation(Output Variable) wrt to the amount in dollars:



We noticed that from above diagram InFraud Transaction(First Diagram) The transaction w.r.t to the amount is very small transactions. In Normal Transaction the wr.t to amount is very large

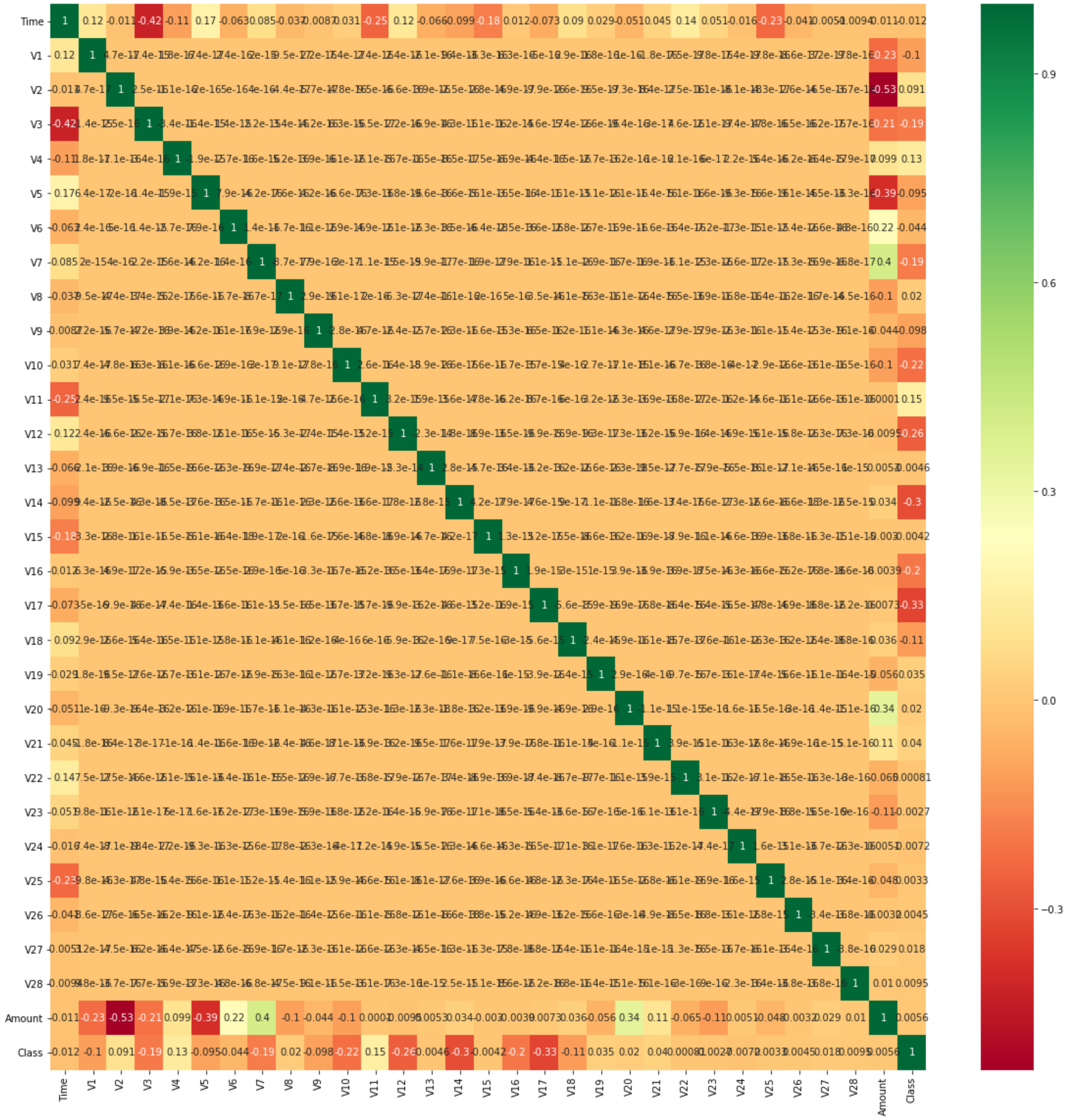
6.The transactions w.r.t to time:



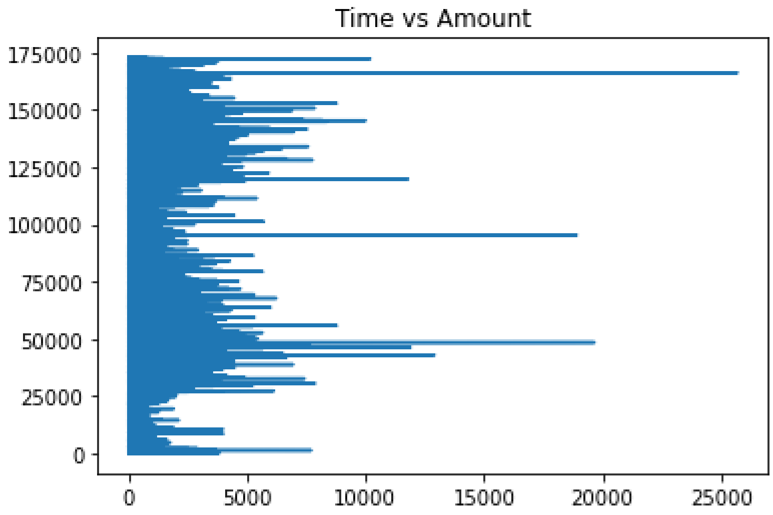
7.The Features w.r.t the outputvariable(corelation matrix):

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

The below corelation matrix describes the relationship between features and output variable.



8.Amount vs Time:



9.The above Graphs describes the relation ship between the Features and Output Variable.

So here we use Classification Technique for building the Model.

Stage-4:- Building the model(Stage-4):

We are building the model by using Classification Technique(Super Vised learning). The different and powerful classification Techniques are:
1.Logistic Regression:

Logistic regression is a statistical **model** that in its basic form uses a **logistic** function to **model** a binary dependent variable, although many more complex extensions exist. In **regression** analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a **logistic model** (a form of binary **regression**).

2.Naive Bayes
Classifiers:

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where
all of them share a common principle, i.e. every pair of features being classified is independent of each other.

3.Support Vector Machine
:

:Support Vector Machine" (SVM) is a supervised machine learning algorithmwhich can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well

· 4.Decision tree

Classifier:

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

5.Random Forest
Classifier:

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn’t depend on one decision tree but multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs..

After Balancing the dataset .We will Train our Model by using the above Classification Techniques.Then we will measure the Accuracy of each
Model.Then we finalize the algorithm from the above algorithms based on accuracy . Accuracy is calculated by using confusion matrix.

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

For the given dataset Logistic Rgeression gives more accuracy compared to other models.

So we finalize the logistic Regression Techinique for building the model.