

**Predictive Analytics**

***Credit Scoring***

*Siva Yogitha Mokkapati*

*Ramya Muthukrishnan*

**School of Graduate Professional Studies**

Data Analytics

IE-575 – Predictive Analytics

Spring, 2020

**Document Control**

**Work carried out by:**

|  |  |  |
| --- | --- | --- |
| **Name** | **Email Address** | **Task description** |
| Ramya Muthukrishnan | Rzm262@psu.edu | Document created |
| Siva Yogitha Mokkapati | Sqm6303@psu.edu | Document Reviewed |
|  |  |  |

**Revision Sheet**

|  |  |
| --- | --- |
| **Date** | **Revision Description** |
| 02/24/2020 | Initial Version |
| 02/25/2020 | Documented Results |
| 02/25/2020 | Added workflow diagram |
| 02/26/2020 | Final Revision |
|  |  |
|  |  |
|  |  |

**TABLE OF CONTENTS**

[1 Introduction 3](#_Toc33129402)

[2 Problem Statement 3](#_Toc33129403)

[2.1 Challenges 3](#_Toc33129404)

[2.2 Related Works 3](#_Toc33129405)

[2.3 Importance 3](#_Toc33129406)

[3 Data Collection 3](#_Toc33129407)

[4 Data Preprocessing 3](#_Toc33129408)

[5 Methodology 4](#_Toc33129409)

[6 Results 4](#_Toc33129410)

[7 Discussion of Results 4](#_Toc33129411)

[8 References 5](#_Toc33129412)

# Introduction

Credit evaluation is one of the most crucial processes in banks credit management decisions. This process includes collecting, analyzing and classifying different credit elements and variables like socio- economic factors of the individual. Credit scoring is the use of statistical models to determine the likelihood that a prospective borrower will default on a loan. This project is mainly aimed at classifying an individual as good or bad using the modelling techniques learnt in the Predictive Analytics course offered at Penn State University. Credit Score Dataset is available in [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/credit+approval).

# Problem Statement

Classify a customer as either good or bad by analyzing the dataset and demonstrate different analytics techniques, aiming to identify whether a consumer belongs to either a legitimate or a suspicious customer group.

## CHALLENGES

The dataset used in this project has all its attribute names and values changed to meaningless symbols to protect confidentiality of data. Here is the extensive list of steps involved in modelling this dataset

## Data preparation

## Data cleansing, data wrangling

## Ensuring the data format is correct

## Identifying important variables

## Recognizing correlations

## Dealing with imbalanced data

## Understanding how different algorithms work

## Choosing the right algorithm for the right problem

## Deciding the right configurations, the algorithm

## Understanding the output of the algorithm

Although, lot of studies are done using the German (1994), Australian (1992), Japanese (1992) and other available datasets, they do not represent a realistic setup as they have a limited number of variables and without such the realistic data established for benchmarking. Also, finding other public datasets in credit scoring problem is still difficult. This fact indicates that how difficult is to obtain datasets on the credit scoring scenario since there are issues related to maintenance of confidentiality of credit scoring databases

## RELATED WORKS

This dataset is referenced in the following articles:

1. [Credit Scoring Model based on Weighted Voting and Cluster Based Feature Selection](https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050918X00088/1-s2.0-S1877050918307877/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjELH%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJGMEQCIFXpB2dF3u1z%2FWCqiJZBQhC7P5OEUFtsx0nMB3HIw39aAiAYyNss3hsBIrBCUYb0VxMIk%2BAyLB2JZOwRyQzDKe9rGyq0Awh6EAIaDDA1OTAwMzU0Njg2NSIM1%2Fz1PEPfrhehQ89%2FKpEDOF%2Brgi7syglVqoJlI9D%2F%2BvOSB0KEcPIMwtmYvHg2N5TSW6UoVpVzt%2FdI6yTUoCTym6m2%2FIcNAjnIlZLZELH746yWsqY0Vv%2F9ovhMq0m8PFpeqHHO2TkIpzemU5U24RUwih%2FN3Jk7QZ5gbk34yLm9YOHvG8%2BTqibvPPToD6y4901vF3dAQ%2Br9hBQxCRKDDGm7TGqZMJ97fl9C0mousAqHNhKfAH2Ddu6fkXsD2QfTKK0T81tr8PWAHadqD47rbcda2dQkjarlfofyp0566vur02KBBRRaTr2a%2FUcY4ELqQezJXN2FaAeOjuKrPZn5d%2FMrXOBSV4h7oQJODqRz49s24oXk6W1hJEvFVqnPSK8ryYUQhEzb9nnZTagleIYoD6LH4mDcEjvJHuBEP1jUxsXpSLiswqr1%2B11SesReOCJaxNHkwtjGQgLWg4TAZaY4Obwh4o1UEC9WKOIZC3d5A24TMYO%2Bdg0tgIPMDUx%2BgL4f0Go1L8IE5BCYrNj2Ruy8sbZcLc5B3COG2io5RAz1VFPVr5QwhfXP8gU67AHcY9IGgEnp%2Bbss8YDrsNcEAj9KrMMnSnsaO%2FjkpljJZFgSB8%2Bhqk6GMW%2F1kRP5byUbW%2FvZk7hZjhKEFMcFbd5uPO2qhwFZ9UkTMbVy%2FaooowCHoXxGSSFKDp06tPDy1Pvowg6MtsOcILoSzrpNAwZPxykBdfsVbvY58QCkuV1F6%2FejjDiV6%2F9UqVTiUIZFLkR4YXzVYugUbt4p%2Fhbd1TBA4CF40Swm6zcBMNIfLVdWnS37zuuhFS3cbrcj%2BInEEab9jjIHwZOAqV%2BfBO9W%2B1FtpdGFozOoxcXU%2BoGHwdrS6Ha98x3VWp18UeRIfA%3D%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20200224T180343Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTY5NWXRI2A%2F20200224%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=86659c3b5e32a52a34c6f66b86e788d38bfa30ef9ea621550747c817719c05a7&hash=1e9889119a2afa50ac383303671e358630309b80e904e1fd9d1ea5ce95ecbdd0&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1877050918307877&tid=spdf-2fbd71a8-d015-45c8-a1b3-22f4ba949472&sid=94724e4d4468734bdf4bbb96aa9b0015e148gxrqa&type=client)
2. [Credit Scoring Prediction Using Feature Selection and Parallel Random Forest](https://www.researchgate.net/publication/303954641_A_Novel_Credit_Scoring_Prediction_Model_based_on_Feature_Selection_Approach_and_Parallel_Random_Forest)

## IMPORTANCE

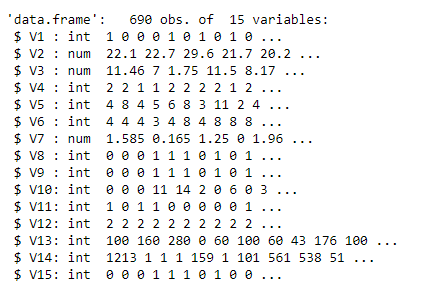
In today’s competitive economy, credit scoring constitutes one of the most significant and successful operational research techniques used in banking and finance. It corresponds to the procedure of estimating the risk related to credit products which is evaluated using applicants credentials (such as annual income, job status, residential status, etc.) and historical data.

The global financial crisis of 2008 (which resulted in the collapse of many venerable names in the industry) caused a ripple effect throughout the economy and demonstrated the potential large losses when a credit applicant defaults on a loan. Therefore, the credit scoring systems are of great interest to banks and financial institutions, not only because they must measure credit risk, but also because any small improvement would produce great profits

# Data Collection

The first step in any analysis is to obtain the [dataset](http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/) and [codebook](http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/crx.names). Both the dataset and the codebook can be downloaded for free from the UCI website. A quick review of the codebook shows that all the values in the dataset have been converted to meaningless symbols to protect the confidentiality of the data. This will still suit our purposes as a demonstration dataset since we are not using the data to develop actual credit screening criteria.

Using str () function in R to understand the type of data in the dataset.

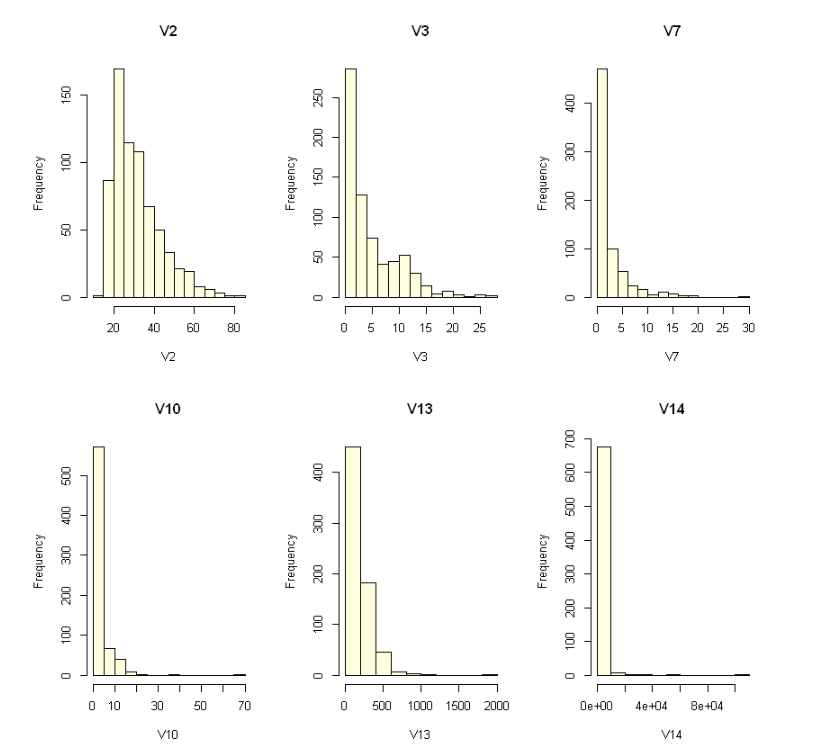


We can see that there are no null values in any attribute. Also, from the summary there are 6 numerical attributes and 9 categorical attributes out of which V15 is the predicted variable.

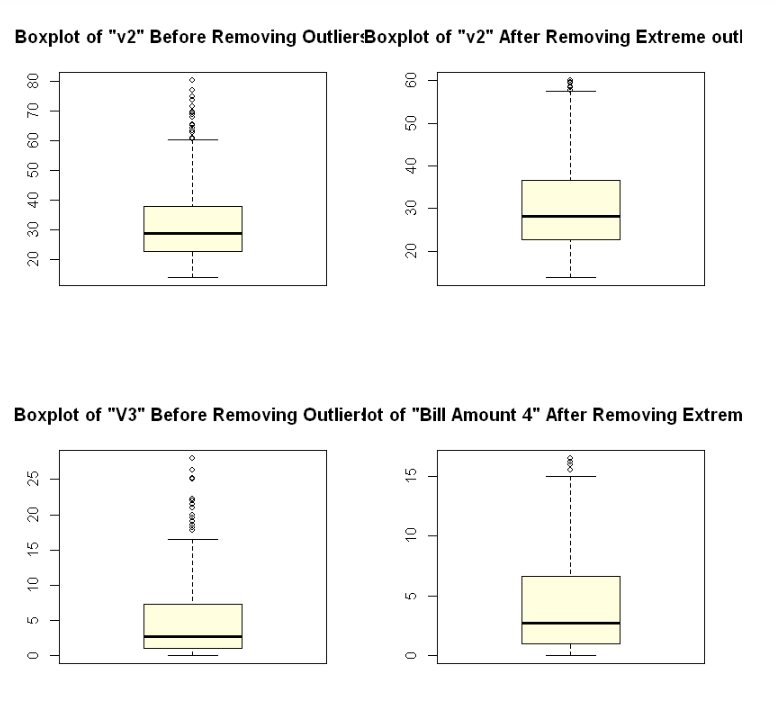
# Data Preprocessing

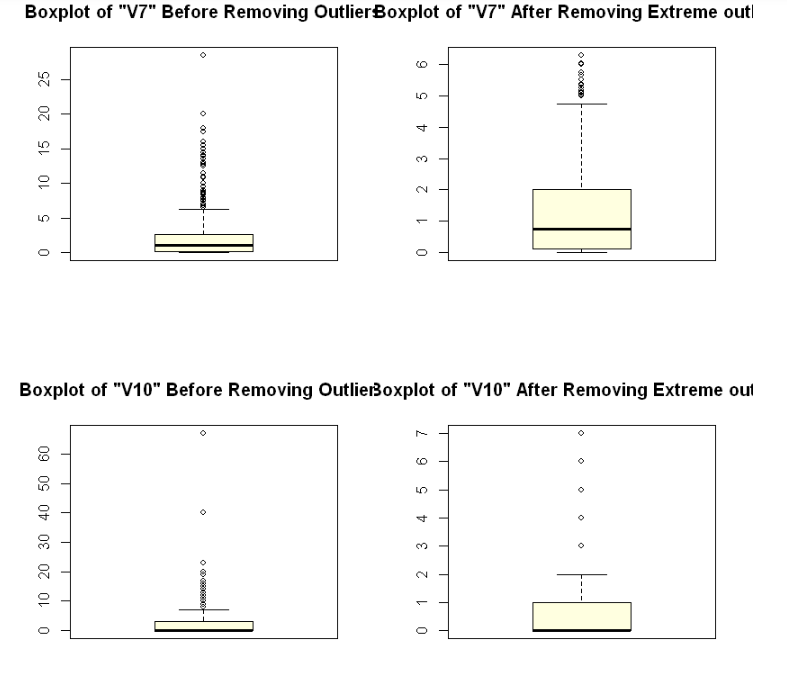
Step1: Visualize the data to see if the data is normalized and to find the outliers. If any, remove them.

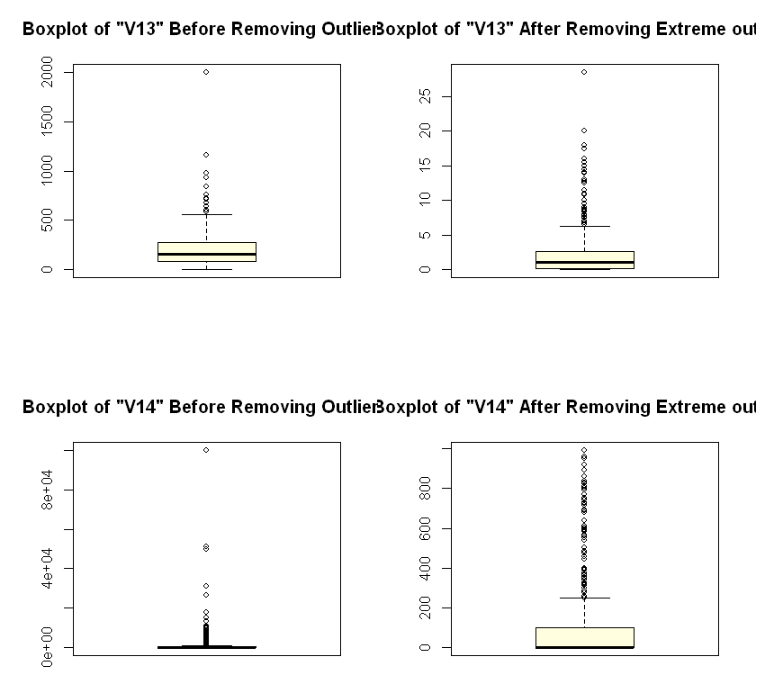
**Histogram of the numerical variables show that the data before pre-processing is skewed towards right for all the attributes:**



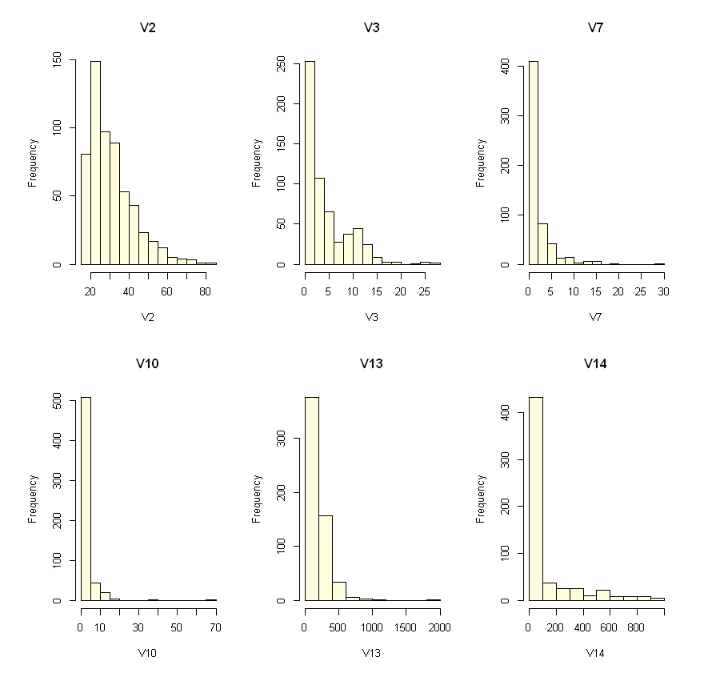
**Outlier detection and removal for numerical attributes:**



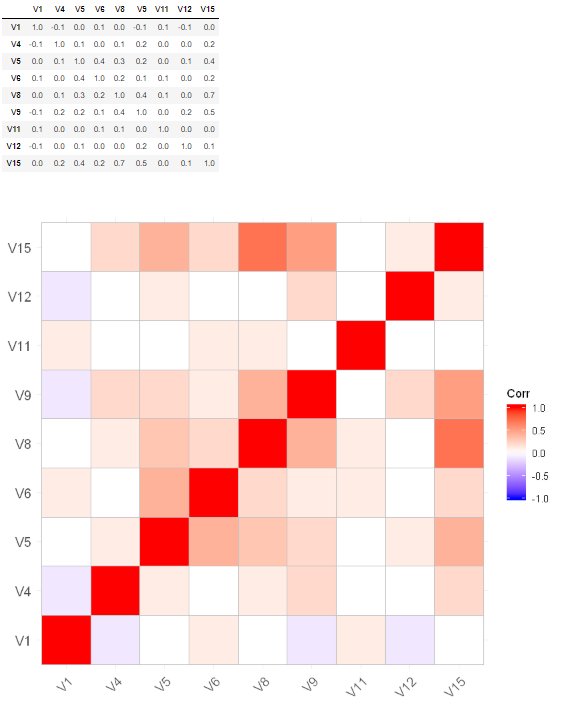




**Histogram after removing outliers is skewed towards right as well. Hence next step is to the normalize the data**



Step 2: Find the correlation between numeric variables and remove them accordingly if correlation exists:



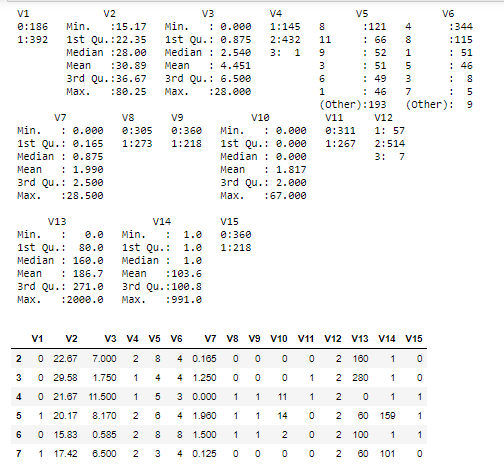
**There is no correlation between the numeric variables**

There is no correlation between numerical attributes. Hence, we don’t have to remove any of them.

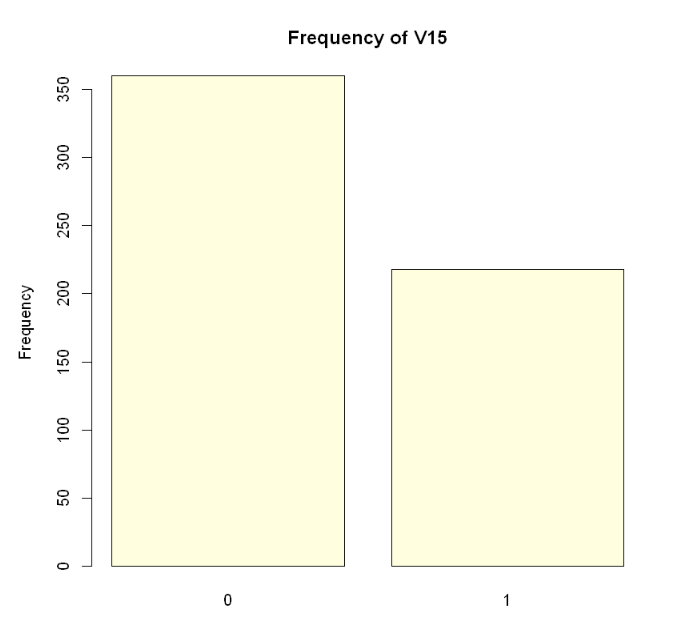
Step3: Convert categorical data to factors:

This is done using the ***as. factor (attribute name).***

Using summary () method after converting the categorical data to numerical factors.



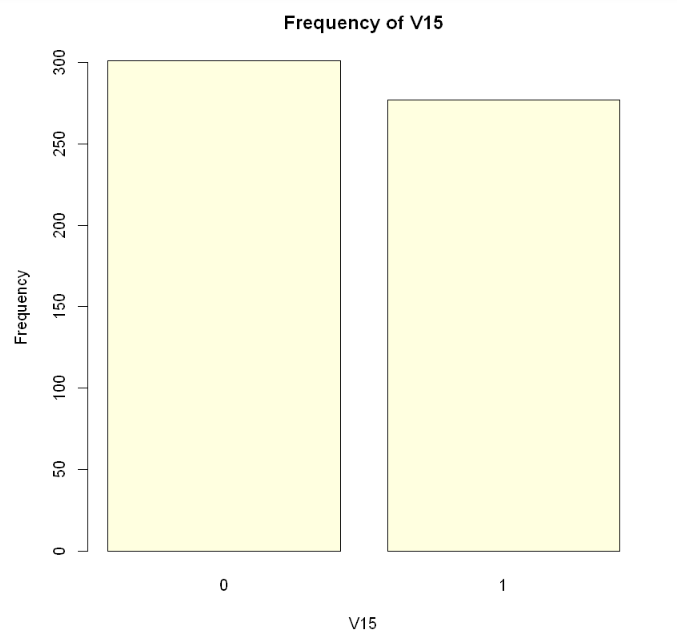
Step 4: Balancing the dataset with respect to predictor variable if its unbalanced:



Before balancing

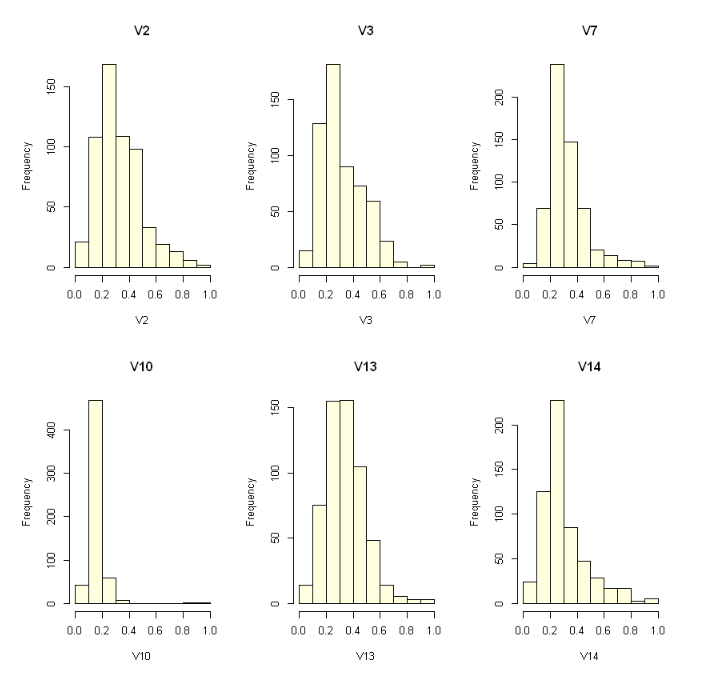
We can see that the dataset is not balanced. Hence, we used ROSE package to balance the dataset.

After balancing using ROSE package



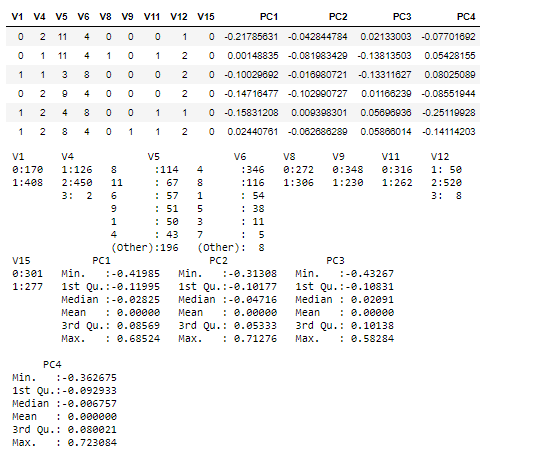
Step 5: Normalizing and Scaling the data:

We used Preprocess () function to normalize the numerical attributes. We can see from the histogram that the data is normalized and scaled after using preprocess function



Step 6: Principle Component Analysis to reduce the dimensionality of the data:

PCA is performed on numerical attributes to reduce the dimensionality of the dataset.



Almost 85% of the data is being explained by only 4 PC components. Hence, we consider only 4 PCA components and remove all the other 2 PC components.

Auto EDA is also used for the purpose of exploring the dataset and a report is generated using that.

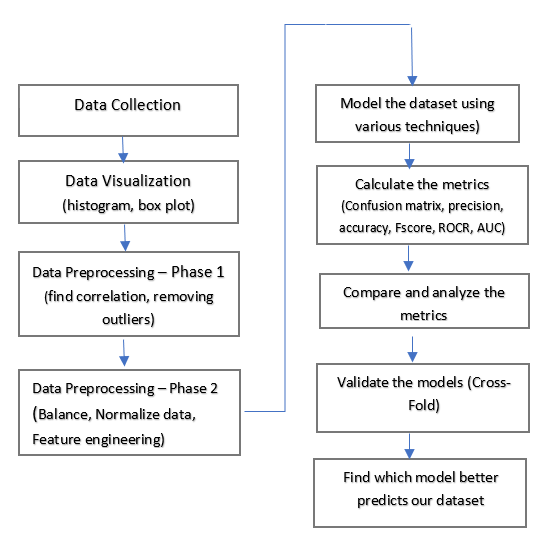
# Methodology

* + We built models to predict if a customer is good or bad using the following datasets:
  + the entire dataset,
  + the dataset that is reduced using PCA
  + the above datasets with feature selection after using variable importance function.
  + Predictive models we used in the project:

1. Logistic Regression
2. Naive Bayes
3. SVM
4. CART
5. C5.0
6. Neural Network
7. Random Forest
8. Bagging
9. Boosting

These models were chosen as they are generally the best possible models for binary classification problem using supervised learning.

* + The workflow of the project:



**Figure 1 - Workflow involved**

The programming language used in the project is R and Jupyter Notebook is used for this purpose along with documenting our comments/findings

# Results

The model building, the performance metrics and cross validation for each model is demonstrated in the Juypter notebook along with necessary explanations.

The summary of the performance metrics on the test data of the models built is presented below:

| **Model** | **Precision** | **Recall** | **Fscore** | **Accuracy** |
| --- | --- | --- | --- | --- |
| **Logistic Regression(balanced)** | 0.9272727 | 0.8500000 | 0.8869565 | 0.8869565 |
| **Logistic Regression(reduced)** | 0.9615385 | 0.8333333 | 0.8928571 | 0.8956522 |
| **Logistic Regression(balanced,imp features)** | 0.9411765 | 0.8000000 | 0.8869565 | 0.8695652 |
| **Logistic Regression(reduced,imp features)** | 0.9600000 | 0.8000000 | 0.8727273 | 0.8782609 |
| **Naive Bayes(balanced)** | 0.8571429 | 0.9000000 | 0.8780488 | 0.8695652 |
| **Naive Bayes(reduced)** | 0.9433962 | 0.8333333 | 0.8849558 | 0.8869565 |
| **Naive Bayes(balanced,important features)** | 0.9795918 | 0.8000000 | 0.8807339 | 0.8869565 |
| **Naive Bayes(reduced,imp features)** | 0.9795918 | 0.8000000 | 0.8807339 | 0.8869565 |
| **SVM(balanced)** | 0.9600000 | 0.8000000 | 0.8727273 | 0.8782609 |
| **SVM(reduced)** | 0.9600000 | 0.8000000 | 0.8727273 | 0.8782609 |
| **SVM(balanced,imp features)** | 0.9600000 | 0.8000000 | 0.8727273 | 0.8782609 |
| **SVM(reduced,imp features)** | 0.9600000 | 0.8000000 | 0.8727273 | 0.8782609 |
| **C5(balanced)** | 0.8666667 | 0.8666667 | 0.8666667 | 0.8608696 |
| **C5(reduced)** | 0.8928571 | 0.8333333 | 0.8620690 | 0.8608696 |
| **Random Forest(balanced)** | 0.9821429 | 0.9166667 | 0.9482759 | 0.9478261 |
| **Random Forest(reduced)** | 0.9814815 | 0.8833333 | 0.9298246 | 0.9304348 |
| **Random Forest(balanced data with tuning)** | 0.9649123 | 0.9166667 | 0.9401709 | 0.9391304 |
| **Random Forest(reduced data with tuning)** | 0.9814815 | 0.8833333 | 0.9298246 | 0.9304348 |
| **Random Forest (balanced,imp features** | 0.9473684 | 0.9000000 | 0.9230769 | 0.9217391 |
| **Random Forest(reduced,imp features)** | 0.9423077 | 0.8166667 | 0.8750000 | 0.8782609 |
| **Bagging CART(balanced)** | 0.9000000 | 0.9642857 | 0.9310345 | 0.9304348 |
| **Bagging(balanced, imp features)** | 0.8666667 | 0.9454545 | 0.9043478 | 0.9043478 |
| **Bagging(reduced)** | 0.8333333 | 0.9615385 | 0.8928571 | 0.8956522 |
| **Bagging(reduced,imp features)** | 0.8666667 | 0.8666667 | 0.8666667 | 0.8608696 |
| **Boosting(balanced)** | 0.9137931 | 0.8833333 | 0.8983051 | 0.8956522 |
| **Boosting(reduced)** | 0.9285714 | 0.8666667 | 0.8965517 | 0.8956522 |
| **Boosting(balanced,imp features)** | 0.9137931 | 0.8833333 | 0.8983051 | 0.8956522 |
| **Boosting(reduced,imp features)** | 0.9230769 | 0.8000000 | 0.8571429 | 0.8608696 |
| **Neural network(balanced)** | 0.9642857 | 0.9000000 | 0.9310345 | 0.9304348 |
| **Neural network(reduced)** | 0.9454545 | 0.8666667 | 0.9043478 | 0.9043478 |
| **Neural network(balanced,imp features)** | 0.8750000 | 0.9333333 | 0.9032258 | 0.8956522 |
| **Neural network(reduced,imp features)** | 0.8644068 | 0.8500000 | 0.8571429 | 0.8521739 |

It is better to consider only few important features that to consider the entire dataset as most of the performance metrics is almost the same considering only few features. There is no need to consider 14 features when we can get almost the same performance metrics with only 4 or 5 features.

Features contributing more towards accuracy of each model are listed below:

Logistic Regression – V4, V5, V8 and V13

Naïve Bayes – V4, V5, V8, V13

SVM – V4, V5, V8 and V13

CART – V5, V8, V10, V12, V13

C5 – V5, V8, V10, V12, V13

Random Forest – V5, V7, V8, V10, V14

Bagging – V5, V8, V10, V12, V13

Boosting – V5, V8, V10, V12, V13, V14

# References

1. PCA Tutorial - <https://www.dezyre.com/data-science-in-python-tutorial/principal-component-analysis-tutorial>
2. [Credit Scoring using Weighted Voting and Clustering](https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050918X00088/1-s2.0-S1877050918307877/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjELH%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJGMEQCIFXpB2dF3u1z%2FWCqiJZBQhC7P5OEUFtsx0nMB3HIw39aAiAYyNss3hsBIrBCUYb0VxMIk%2BAyLB2JZOwRyQzDKe9rGyq0Awh6EAIaDDA1OTAwMzU0Njg2NSIM1%2Fz1PEPfrhehQ89%2FKpEDOF%2Brgi7syglVqoJlI9D%2F%2BvOSB0KEcPIMwtmYvHg2N5TSW6UoVpVzt%2FdI6yTUoCTym6m2%2FIcNAjnIlZLZELH746yWsqY0Vv%2F9ovhMq0m8PFpeqHHO2TkIpzemU5U24RUwih%2FN3Jk7QZ5gbk34yLm9YOHvG8%2BTqibvPPToD6y4901vF3dAQ%2Br9hBQxCRKDDGm7TGqZMJ97fl9C0mousAqHNhKfAH2Ddu6fkXsD2QfTKK0T81tr8PWAHadqD47rbcda2dQkjarlfofyp0566vur02KBBRRaTr2a%2FUcY4ELqQezJXN2FaAeOjuKrPZn5d%2FMrXOBSV4h7oQJODqRz49s24oXk6W1hJEvFVqnPSK8ryYUQhEzb9nnZTagleIYoD6LH4mDcEjvJHuBEP1jUxsXpSLiswqr1%2B11SesReOCJaxNHkwtjGQgLWg4TAZaY4Obwh4o1UEC9WKOIZC3d5A24TMYO%2Bdg0tgIPMDUx%2BgL4f0Go1L8IE5BCYrNj2Ruy8sbZcLc5B3COG2io5RAz1VFPVr5QwhfXP8gU67AHcY9IGgEnp%2Bbss8YDrsNcEAj9KrMMnSnsaO%2FjkpljJZFgSB8%2Bhqk6GMW%2F1kRP5byUbW%2FvZk7hZjhKEFMcFbd5uPO2qhwFZ9UkTMbVy%2FaooowCHoXxGSSFKDp06tPDy1Pvowg6MtsOcILoSzrpNAwZPxykBdfsVbvY58QCkuV1F6%2FejjDiV6%2F9UqVTiUIZFLkR4YXzVYugUbt4p%2Fhbd1TBA4CF40Swm6zcBMNIfLVdWnS37zuuhFS3cbrcj%2BInEEab9jjIHwZOAqV%2BfBO9W%2B1FtpdGFozOoxcXU%2BoGHwdrS6Ha98x3VWp18UeRIfA%3D%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20200224T180343Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTY5NWXRI2A%2F20200224%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=86659c3b5e32a52a34c6f66b86e788d38bfa30ef9ea621550747c817719c05a7&hash=1e9889119a2afa50ac383303671e358630309b80e904e1fd9d1ea5ce95ecbdd0&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1877050918307877&tid=spdf-2fbd71a8-d015-45c8-a1b3-22f4ba949472&sid=94724e4d4468734bdf4bbb96aa9b0015e148gxrqa&type=client)
3. [Novel Credit Scoring Prediction Model](https://www.researchgate.net/publication/303954641_A_Novel_Credit_Scoring_Prediction_Model_based_on_Feature_Selection_Approach_and_Parallel_Random_Forest)