**Project 2:**

***Bank***

***Loan***

***Default***

***Case***

**Submitted by- Tanya Shrivastava**

**Submitted on -09/11/2019**

**Contents**

**1 Introduction**

1.1 Problem Statement

1.2 Data

1.3 Software and hardware Requirement

**2 Methodology**

2.1 Pre Processing

2.1.1 Exploratory Data Analysis

2.1.2 Missing Value Analysis

2.1.3 Outlier Analysis

2.1.4 Feature Selection

2.2 Model Developement

2.2.1 Model Selection

**3. Conclusion**

**1.** **INTRODUCTION**

**1.1** **Problem statement**

Predictive Analytics is the stream of the advanced analytics which utilizes diverse techniques like data mining, predictive modelling, statistics, machine learning and artificial intelligence to analyse current data and predict future.

Loans default will cause huge loss for the banks, so they pay much attention on this issue and apply various method to detect and predict default behaviours of their customers.

The loan default dataset has 8 variables and 850 records, each record being loan default status for each customer. Each Applicant was rated as “Defaulted” or “Not-Defaulted”. New applicants for loan application can also be evaluated on these 8 predictor variables and classified as a default or non-default based on predictor variables.

**What is Classification?**

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.

**1.2 Data**

Our task is to build classification model which will predict that the new applicant for loan application can be classified as default or non-default depending on yhe 8 predictor variables.

**In a loan risk prediction situation** of a loan financing company, the company would be interested in metrics such as how long it takes customers with certain attributes to pay back their loans and also, what is the possible risk of a default.

Generally, the company stands a higher risk of default from customers who have a bad credit rating or who have certain bad spending habits. In this situation, the company is very keen to find out if a customer will default or not. So, the past data observations gathered by the company are used to group customers into categories such as “Defaulter” or “Non-defaulter”.

Given below is a sample of the data set that we are using :

Table 1.1: Sample Data (Columns: 1-8)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **age** | **ed** | **employ** | **address** | **income** | **debtinc** | **creddebt** | **othdebt** | **default** |
| **0** | 41 | 3 | 17 | 12 | 176 | 9.3 | 11.359392 | 5.008608 | 1.0 |
| **1** | 27 | 1 | 10 | 6 | 31 | 17.3 | 1.362202 | 4.000798 | 0.0 |
| **2** | 40 | 1 | 15 | 14 | 55 | 5.5 | 0.856075 | 2.168925 | 0.0 |
| **3** | 41 | 1 | 15 | 14 | 120 | 2.9 | 2.658720 | 0.821280 | 0.0 |
| **4** | 24 | 2 | 2 | 0 | 28 | 17.3 | 1.787436 | 3.056564 | 1.0 |

Variables present in given dataset are 'age', 'ed', 'employ', 'address', 'income',

'debtinc', 'creddebt', 'othdebt' and 'default'.The details of variable present in the dataset are as follows :

1. Age

Age of each customer Numerical

2. Education

Education categories Categorical

3. Employment

Employment status - Numerical

Corresponds to job

status and being

converted to numeric

format

4. Address

Geographic area - Numerical

Converted to numeric

values

5 Income

Gross Income of each Numerical

customer

6. debtinc

Individual’s debt Numerical

payment to his or her

gross income

7. creddebt

debt-to-credit ratio is a Numerical

measurement of how

much you owe your

creditors as a

percentage of your

available credit (credit

limits)

8. othdebt

Any other debts Numerical

# **1.3 Software and Hardware Requirement**

a) R 3.6.1 for 64 bit

b) Anaconda 3 for 64 bit

c) Rstudio

d) 64 bit OS

e) Python 3

f) Jupyter Notebook 6.0

g) Windows 10

h) 4 GB of RAM

**2. Methodology**

**2.1 Pre Processing**

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms *looking at data* refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis.

**2.1.1 Exploratory Data Analysis**

**Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.**

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand,before getting them dirty with it.

**2.1.2 Missing Value Analysis**

Missing value analysis is done to check is there any missing value present in given dataset. Missing values can be easily treated using various methods like mean, median method, knn method to impute missing value.

In R function(x){sum(is.na(x))} is the function used to check the sum of missing values.

In python df.isnull().sum() is used to detect any missing value

There are 150 missing values here

The percentage of missing values is 17.64%

age 0

ed 0

emp 0

address 0

income 0

debtinc 0

creddebt 0

othdebt 0

default 150

So , we remove missing values using df=df.dropna() in python .

Only 700 observations were left.

**2.1.3 Outlier Analysis**

Outlier analysis is done to handle all inconsistent observations present in given dataset. As outlier analysis can only be done on continuous variable.

Figure 2.1 and 2.2 are visualization of numeric variable present in our dataset to detect outliers using boxplot and distribution plots. Outliers will be detected with black color.According to above visualizations there all features columns shows outliers. All independent variables are right skewed/positively skewed.

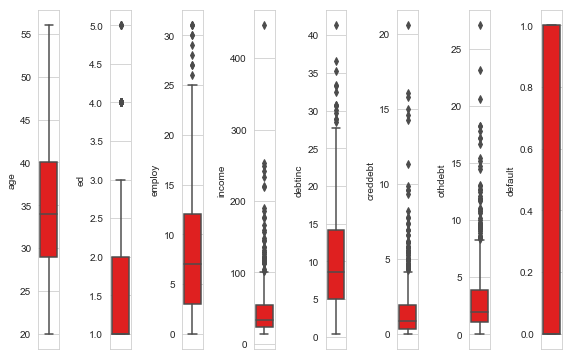


Figure 2.1 Boxplot graph of variables

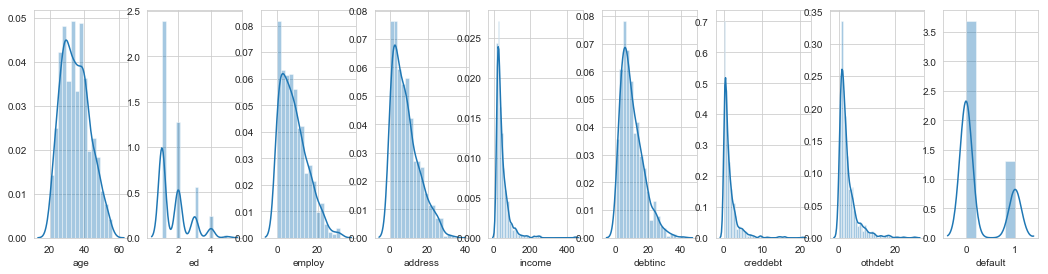


Figure 2.2 Distribution plots -To check distribution-Skewness

**2.1.4 Feature Selection**

**Feature Selection** is the process of selecting the attributes that can make the predicted variable more accurate or eliminating those attributes that are irrelevant and can decrease the model accuracy and quality.

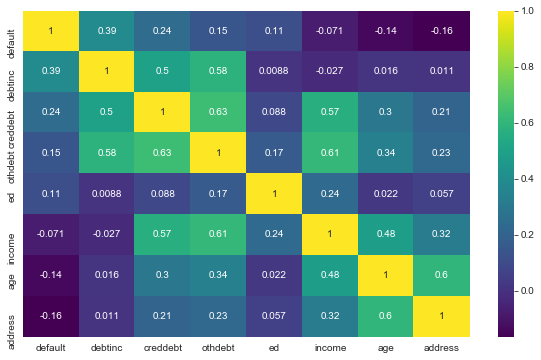
Data and feature correlation is considered one important step in the feature selection phase of the data pre-processing especially if the data type for the features is continuous.

**Positive Correlation:** means that if feature **A** increases then feature **B** also increases or if feature **A** decreases then feature **B** also decreases. Both features move in tandem and they have a linear relationship.

**Negative Correlation:** means that if feature **A** increases then feature **B** decreases and vice versa.

**No Correlation:** No relationship between those two attributes.

Each of those correlation types can exist in a spectrum represented by values from 0 to 1 where slightly or highly positive correlation features can be something like 0.5 or 0.7. If there is a strong and perfect positive correlation, then the result is represented by a correlation score value of 0.9 or 1.If there is a strong negative correlation, it will be represented by a value of -1.

 Figure 2.4 correlation plot

**2.1.4 Feature Scaling**

Feature scaling includes two functions normalization and standardization. It is done reduce unwanted variation either within or between variables and to bring all of the variables into proportion with one another.

In given dataset all numeric values are already present in normalized form.

**Confusion Matrix**

A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known. 

Here,

• Class 1 : Positive

• Class 2 : Negative

**Definition of the Terms:**

• Positive (P) : Observation is positive (for example: is an apple).

• Negative (N) : Observation is not positive (for example: is not an apple).

• True Positive (TP) : Observation is positive, and is predicted to be positive.

• False Negative (FN) : Observation is positive, but is predicted negative.

• True Negative (TN) : Observation is negative, and is predicted to be negative.

• False Positive (FP) : Observation is negative, but is predicted positive.

**Classification Rate/Accuracy:**

Classification Rate or Accuracy is given by the relation:



However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

**ROC CURVES and AUC**

* **Receiver operating characteristic (ROC) curve:** plots the true positive rate (TPR) versus the false positive rate (FPR) as a function of the model’s threshold for classifying a positive
* **Area under the curve (AUC):** metric to calculate the overall performance of a classification model based on area under the ROC curve

**2.2 Model Development**

**2.2.1 Model Selection**

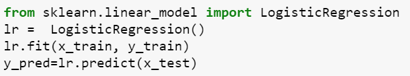
In this case we have to Each Applicant was rated as “Defaulted” or “Not-Defaulted”. New applicants for loan application can also be evaluated on the 8 predictor variables and classified as a default or non-default based on predictor variables.Model having less error rate and more accuracy will be our final model.In these we have divided the dataset into train and test part using random sampling. For this model we have divided the dataset into train and test part using random sampling Where train contains 80% data of data set and test contains 20% data and contains 8 variable where 8th variable is the target variable.

Classification Models used are-

**2.2.2 Logistic Regression**

Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.

Advantages: Logistic regression is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outcome variable.Disadvantages: Works only when the predicted variable is binary, assumes all predictors are independent of each other, and assumes data is free of missing values.



Accuracy 84.28571428571429

FNR 50.0

Defaulted 22

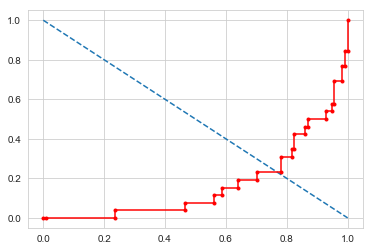
Non-defaulted 118

Confusion matrix

105 9

13 13

ROC -CURVE



**2.2.3 Decision Trees**

Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

**Advantages:** Decision Tree is simple to understand and visualise, requires little data preparation, and can handle both numerical and categorical data.

**Disadvantages:** Decision tree can create complex trees that do not generalise well, and decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

Accuracy 74.28571428571429

FNR 65.38461538461539

Defaulted 28

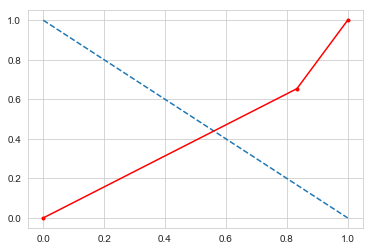
Non-defaulted 112

Confusion Matrix

95 19

17 9

Roc-Curve



**2.2.4 Random Forest Classifier**

Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

Advantages: Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.

Disadvantages: Slow real time prediction, difficult to implement, and complex algorithm.

Accuracy 83.57142857142857

FNR 65.38461538461539

Defaulted 15

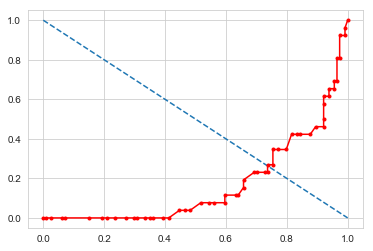
Non-defaulted 125

Confusion Matrix

108 6

17 9

Roc-Curve



**2.2.5 Naive Bayes**

Naive Bayes algorithm based on Bayes’ theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering.

**Advantages:** This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.

**Disadvantages:** Naive Bayes is is known to be a bad estimator.

Accuracy 81.42857142857143

FNR 76.92307692307692

Defaulted 12

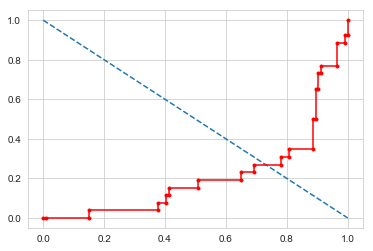
Non-defaulted 128

Confusion Matrix

108 6

20 6

Roc-Curve



***CONCLUSION***

|  |  |  |
| --- | --- | --- |
| **Classification Algorithms** | **Accuracy** |  |
| **Logistic Regression** | **84.20%** |  |
| **Naïve Bayes** | **81.42%** |  |
| **Decision Tree** | **74.23%** |  |
| **Random Forest** | **83.57%** |  |
| **XGB** | **81.42%** |  |
|  |  |  |
|  |  |  |