PROJECT REPORT

BANK LOAN DEFAULT CASE

RISHABH TEWARI

2 JAN 2020

Contents

Chapter 1
Introduction

1.1 Problem Statement

The Bank share it's Dataset for Loan Defaulters .We have to develop a model using their data set.

New applicants for loan application can also be evaluated on the basis of the model developed using the existing dataset and classified as a default or non-default .

1.2 Data

There are 9 variables in our data in which 8 are independent variables and 1 is dependent variable. Since our target variable is categorical in nature, this is a classification problem. out of 8 independent variable 7 are continuous variable and 1(edu) is categorical variable.

Variables Information:

Var. #	Variable Name	Description	Variable Type
1.	Age	Age of each customer	Numerical
2.	Education	Education categories	Categorical
3	Employment	Employment status	Numerical
4	Address	Geographic area	Numerical
5	Income	Gross Income of Customer	Numerical
6	debtinc	Individual's debt payment to his or her gross income	Numerical
7	creddebt	debt-to-credit ratio is a measurement of how much you owe your creditors as a percentage of your available credit (credit limits)	Numerical
8	othdebt	Any other debts	Numerical
9	**DEFAULT**	**TARGET VARIABLE**	**Categorical**

1.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics. In the given data set there are 9 variables and data types is numeric. There are 850 observations and 9 columns in our data set. Missing value is also present in our data.

List of columns and their number of unique values -

VARIABLE	NO OF UNIQUE VALUES
Age	37
Edu	5
Employ	33
Address	32
Income	129
debtinc	245
creddebt	842
othdebt	848
default	2

From EDA we have concluded that there are 7 continuous variable and 2 categorical variable in nature.

Chapter 2

DATA PREPROCESSING

Before feeding the data to the model we need to clean the data and convert it to a proper format. It is the most crucial part of data science project we spend almost 80% of time in it.

2.1 Missing Value Analysis

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a column has more than 30% of data as missing value either we ignore the entire column or we ignore those observations

In our Data set there are missing value in our target variable and we cannot apply Missing value analysis on our target variable since the predicted value are not exact and it will affect the performance of our model.

So we will drop those observation for which there are missing values in our target variable

COLUMN	VARIABLE	Missing Percentage
9	Default	17.64706
1	Age	0.00000
2	Ed	0.00000
3	Employ	0.00000
4	Address	0.00000
5	Income	0.00000
6	Debtinc	0.00000

COLUMN	VARIABLE	Missing Percentage
7	Creddebt	0.00000
8	Othdebt	0.00000

There are 150 observation for which values are missing in our Target Variable . So we will drop those observation

Now we are left with 700 observations.

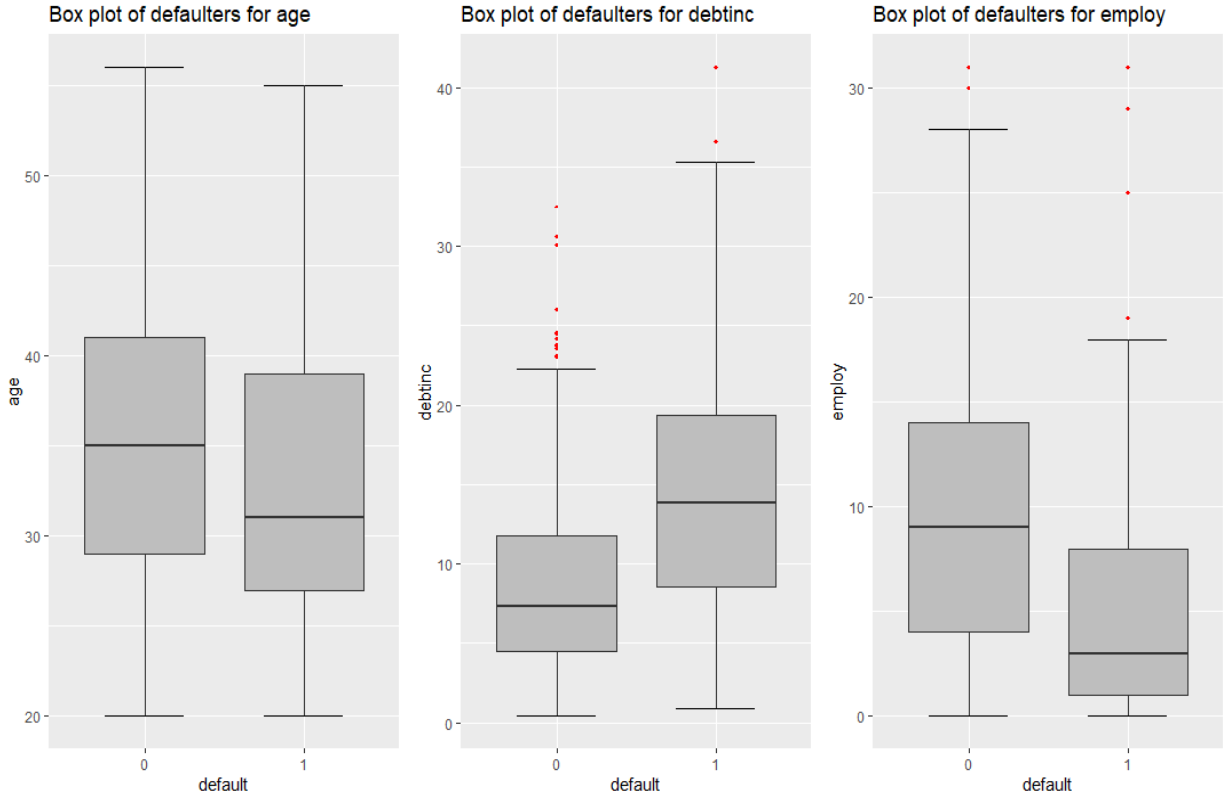
2.2 Outlier Analysis

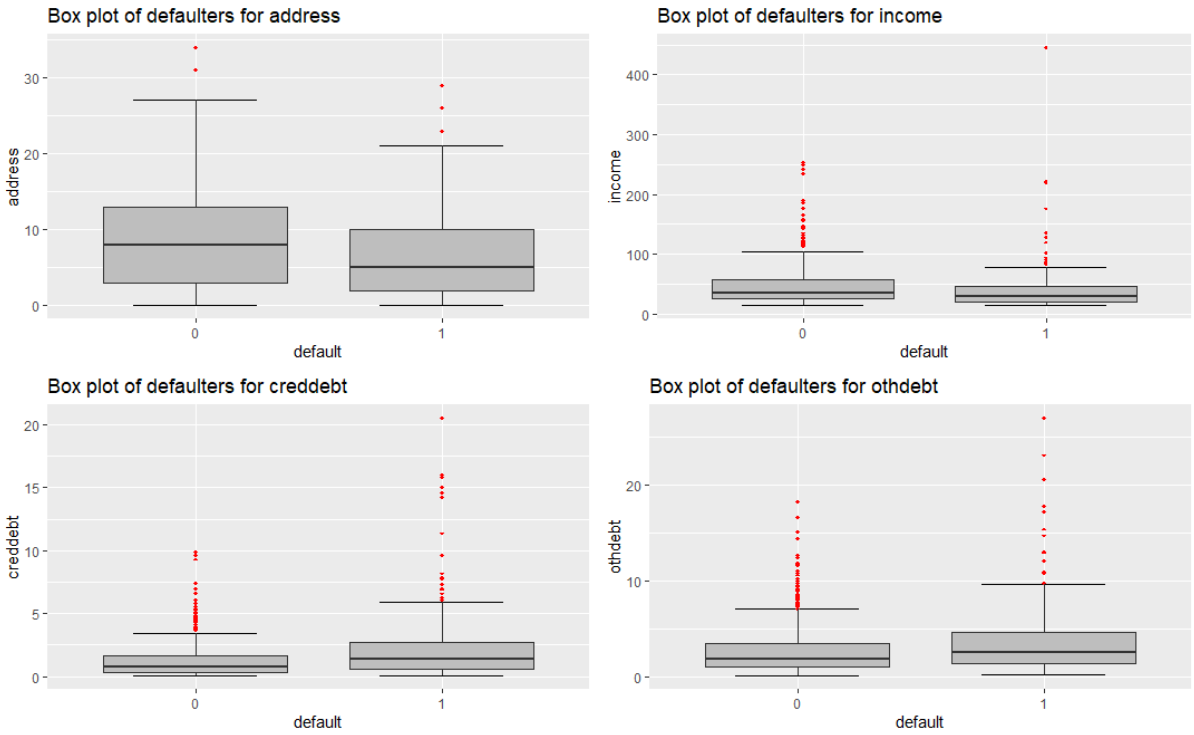
Outlier Analysis can only be applied on Numeric continuous data.
So first we have to separate our Categorical and Continuous Data.
Here I have convert edu and Default from Numeric data type to Factor data type since they are categorical Variable.

An **outlier** is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The **analysis** of **outlier** data is referred to as **outlier analysis**

We can use different methods to detect the outliers. Here we use Box plot to detect the outliers available in our DataSet

In figure we have plotted the boxplots of the 7 predictor variables with respect to **Default** (Our Target Variable). A lot of useful inferences can be made from these plots. First as you can see, we have a lot of outliers and extreme values in each of the data set.

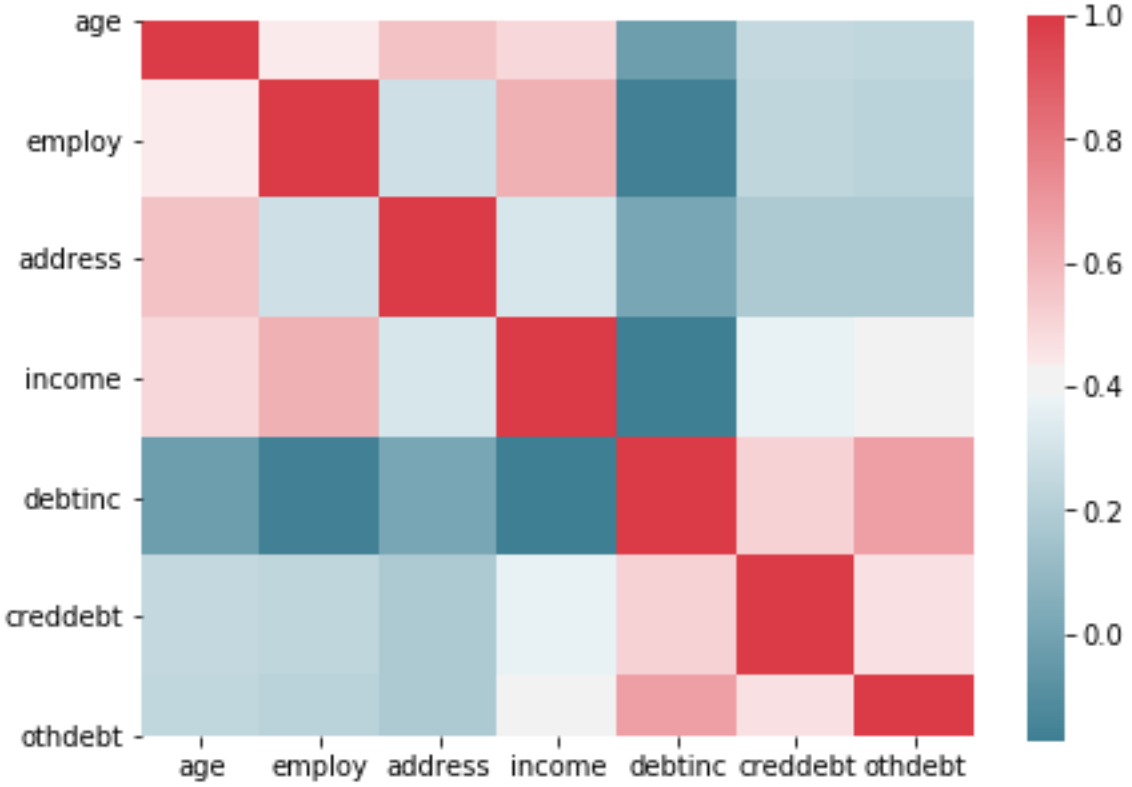




We have drop those Observations for which outliers are present. After dropping those Observations we are finally left with 546 observations and 9 variable.

2.3Feature Selection

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction.Selecting subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. This in turn helps us to avoid the problem of multi collinearity. In this project we have selected **Correlation Analysis** for numerical variable and **CHI SQUARE TEST** for categorical variable.



It can be clearly seen that no independent variable is highly correlated with any other independent variable.

Also for our Categorical Variable **ed** the value is 0.0497 which is also less than 0.05 So we have to reject our Null Hypothesis and consider our categorical variable also.

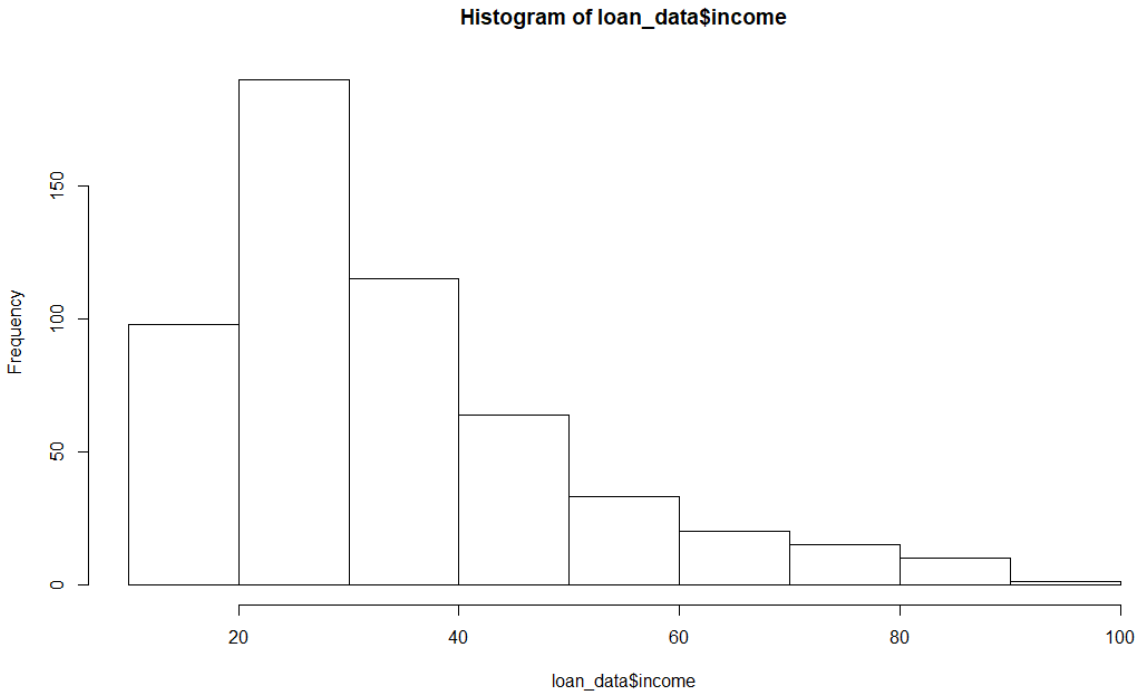
So after applying feature selection process we have 9 variable (8 indpendent and 1 dependent)

2.4Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

We have to check whether our data is uniformly distributed or not. if our data is uniformly distributed then we apply STANDARDIZATION otherwise we will go with NORMALIZATION.

HISTOGRAM FOR INCOME VARIABLE TO CHECK WHETHER OUR DATA IS UNIFORMLY DISTRIBUTED OR NOT



Since our data is not uniformly distributed we will use **Normalization** as Feature Scaling Method.

Chapter 3
ERROR METRICS

Error Metrics is used to evaluate the performance of our model .It is basically of two types

CLASSIFICATION METRICS-This is used to evaluate our model when target variable is categorical or for classification purpose.

REGRESSION MATRICS- This is used when our target variable is continuous.

for our data set we will use classification metrics because our target dependent variable is categorical in nature

We will use Confusion Metrics which is a type of classification metrics and used to evaluate the performance of certain model.

CONFUSION MATRICS 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.

		PREDICTED VALUES	
		CLASS NO	**CLASS YES**
ACTUAL VALUES	**CLASS NO**	**TN**	**FP**
	CLASS YES	**FN**	**TP**

The above table is the confusion Metrics on the basis of which we evaluate our model.

- **true positives (TP):** These are cases in which we predicted yes and they are actually yes
- **true negatives (TN):** We predicted no, and they are actually no.
- **false positives (FP):** We predicted yes, but they are actually no

- **false negatives (FN):** We predicted no, but they are actually yes.

Accuracy: It tells Overall, how often the classifier is correct or accurate.

It can be calculated as-

$$= (TP+TN)/(TP+TN+FP+FN)$$

False Negative Rate A **false negative Rate**, is a test result that indicates that a condition does not hold, while in fact it does. It can be calculated as

$$=FN/(FN+TP)$$

RECALL: It is actually positive cases that are correctly classified. It can be calculated as

$$=TP/(TP+FN)$$

PRECISON : Precision talks about how precise/accurate your model is out of those predicted positive, how many of them are actual positive

It can be calculated as:

$$=TP/(TP+FP)$$

Chapter 4
MODELING

After a thorough preprocessing we will be using some regression models on our processed data to predict the target variable. Following are the models which we have built –

4.1 Regression Analysis

In statistical modeling, **regression analysis** is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables .

Mainly Regression Analysis is of Two types:

Linear Regression- used for Regression When our target Variable is Continuous.

Logistic Regression –Used for Classification when our Target Variable is Categorical.

In our Data set we **use Logistic Regression** analysis since our Target variable is Categorical.

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 .

The Accuracy value and False Negative Rate(FNR) value for our project in R and Python are –

Logistic Regression	R	PYTHON
Accuracy(%)	73.14	77.22
FNR(%)	61.53	69.23

4.2 Naïve Bayes

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.
Naïve Bayes works on both Continuous and Categorical data but best work when our data is Categorical.

P(C/X)=(P(X/C) * P(C))/P(X)

X=independent variable.
C=target Class.

The Accuracy value and False Negative Rate(FNR) value for our project in R and Python are-

Naïve Bayes	R	PYTHON
Accuracy(%)	66.67	78.89
FNR(%)	53.84	48

4.3 KNN

KNN (K — Nearest Neighbors) is one of many supervised learning algorithms used in data mining and machine learning, it's a classifier algorithm where the learning is based "how similar" is a data from other .

KNN is an algorithm that is useful for matching a point with its closest k neighbors in a multi-dimensional space. It can be used for **data** that are **continuous**, discrete, ordinal and **categorical** which makes it particularly useful for dealing with all kind of missing **data**.

KNN	R K=1	R K=3	R K=5	PYTHON K=1	PYTHON K=3	PYTHON K=5
Accuracy(%)	70.37	71.3	70.37	67.88	75.22	76.14
FNR(%)	60.71	60.86	65	60	60	63

4.4 Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each branch connects nodes with "and" and multiple branches are connected by "or". It can be used for classification and regression. It is a supervised machine learning

algorithm. Accept continuous and categorical variables as independent variables. Extremely easy to understand by the business users. The Accuracy value and FNR value for our project in R and Python are –

Decision Tree	R	PYTHON
Accuracy(%)	68.52	64.22
FNR(%)	73	60

4.5 Random Forest

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree.

Random Forest	R N=100	R N=500	R N=700	PYTHON N=100	PYTHON N=500	PYTHON N=700
Accuracy(%)	61.53	74.07	73.15	71.55	76.14	75.22
FNR(%)	73.15	57.69	57.69	72	68	68

Chapter 5
CONCLUSION

5.1 Model Evaluation

In this chapter we are going to evaluate our models, select the best model for our dataset.

In the previous chapter we have seen the Accuracy and **False Negative Rate(FNR)** Value of different models.

The **accuracy** can be defined as the percentage of correctly classified instances

$$(TP + TN)/(TP + TN + FP + FN).$$

where TP, FN, FP and TN represent the number of true positives, false negatives, false positives and true negatives, respectively.
False Negative Rate can be defined as Number of items wrongly identified as negative out of total true positives-

$$FN/(FN+TP)$$

where FN and TP are False Negative and True positive.

Lower values of **FNR** and higher value of **ACCURACY** indicate better fit.

5.2 Model Selection

5.2.1 Model Selection in R

From the observation of all **Accuracy Value** and **FNR** Value we have concluded that **Random Forest Model** has minimum value of FNR(57%) and its Accuracy Value is also maximum (i.e.74.07).
So our Final Model is random Forest.

Although the FNR in Naïve Bayes is 53% which is 4% less than in the case of Random forest but there is large difference in the accuracy(around 8%) for both models, that's why we consider Random forest as our final model.

5.2.2 Model Selection in Python

From the observation of all **Accuracy Value** and **FNR** Value we have concluded that **Naïve Bayes Model** has minimum value of FNR(48%) and its Accuracy Value is also maximum (i.e.78.89%).

So our Final Model is Naïve Bayes.