Bank Loan Default Case

AJAY SINGH

Problem Statement

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.

Data

Number of attributes:

Var. #	Variable Name	Description	Variable Type
1.	Age	Age of each customer	Numerical
2.	Education	Education categories	Categorical
3	Employment	Employment status - Corresponds to job status and being converted to numeric format	Numerical
4	Address	Geographic area - Converted to numeric values	Numerical
5	Income	Gross Income of each 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

Methodology

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**. To start this process we will first try and look at all the distributions of the Numeric variables. Most analysis like regression, require the data to be normally distributed.

Univariate Analysis

we have plotted the probability density functions numeric variables present in the data including target variable cnt..

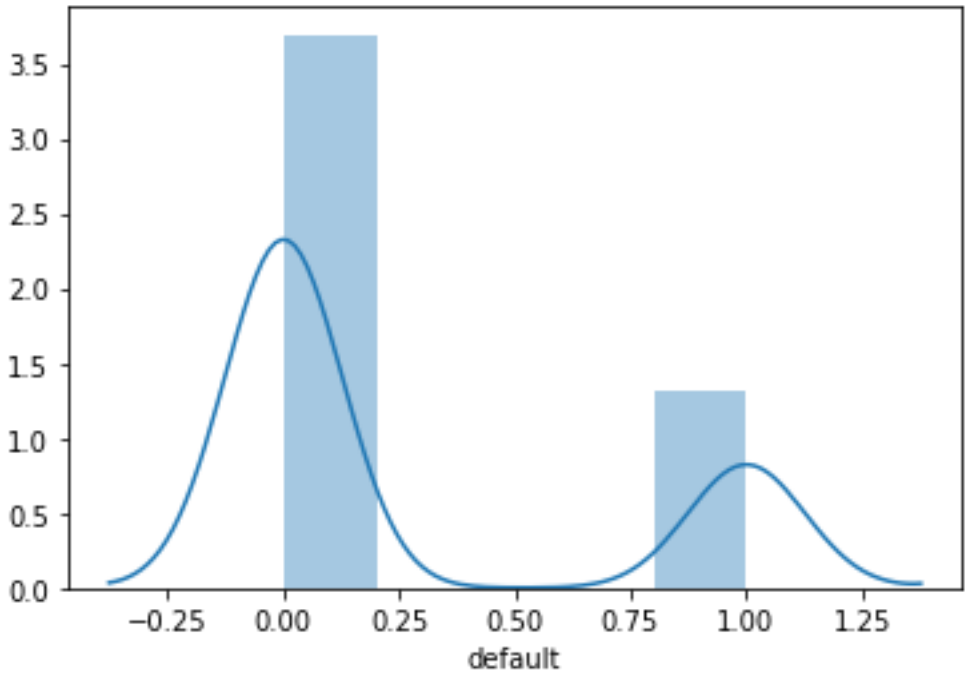
i. Target variable default is normally distributed

ii. Independent variables like 'temp','atemp', and 'regestered' data is distributed normally.

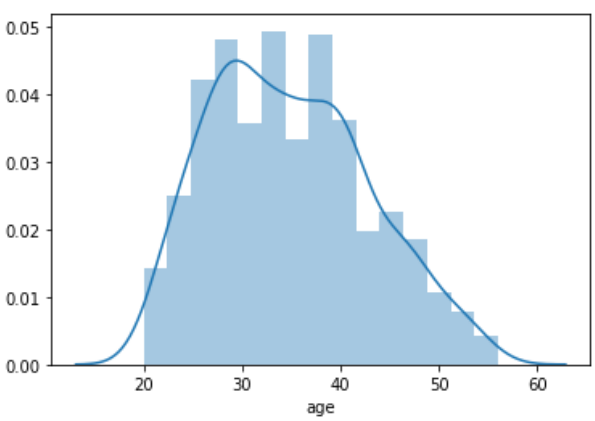
iii. Independent variable 'casual' data is slightly skewed to the right,so, there is chances of getting outliers.

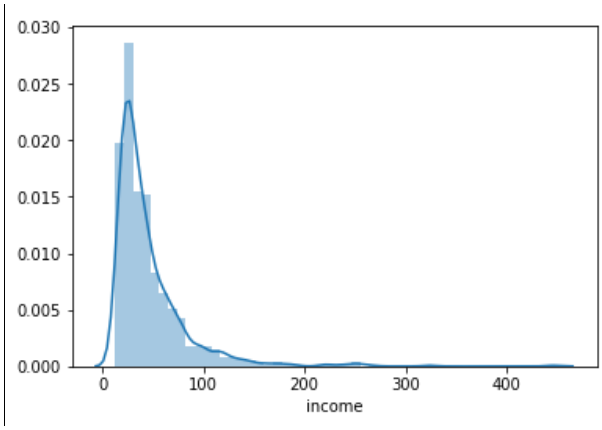
iv. Other Independent variable 'hum' data is slightly skewed to the left , here data is already in normalize form so outliers are discarded.

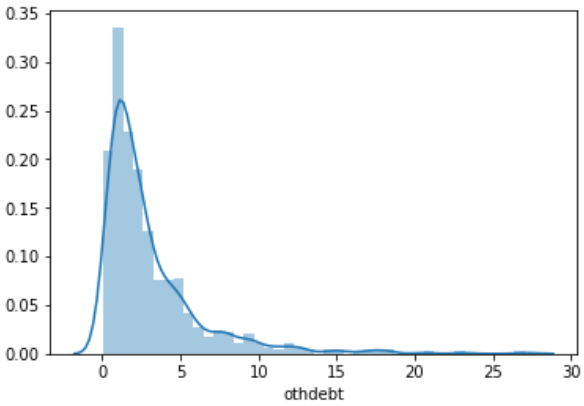
1. Distribution of target variable (default)

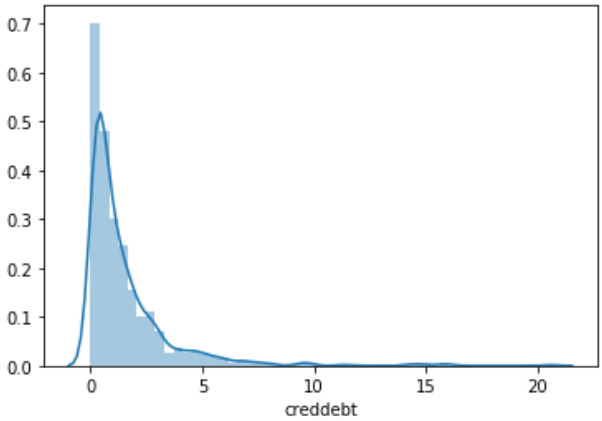


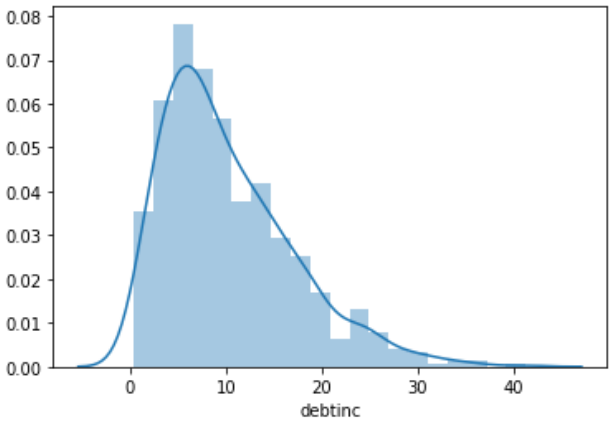
Showing distribution of independent variables:

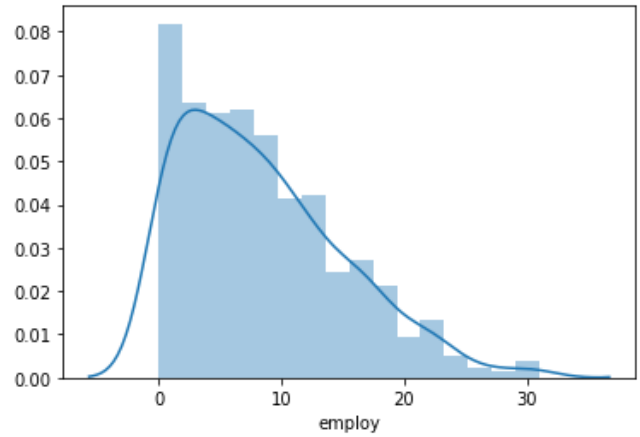

0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
age

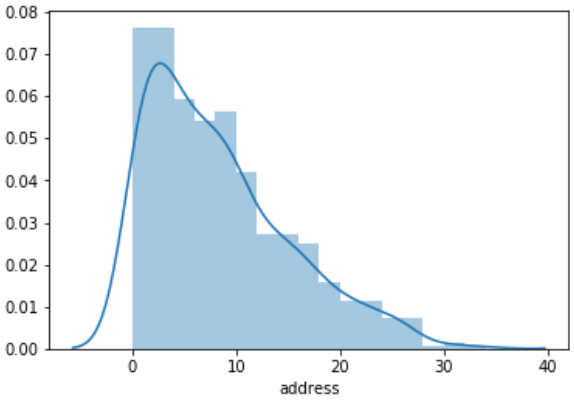

0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
income

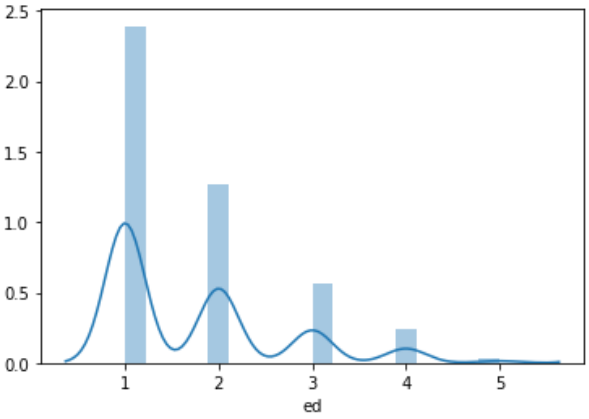

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
25
30
othdebt


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
creddebt


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
debtinc


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
employ


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
address

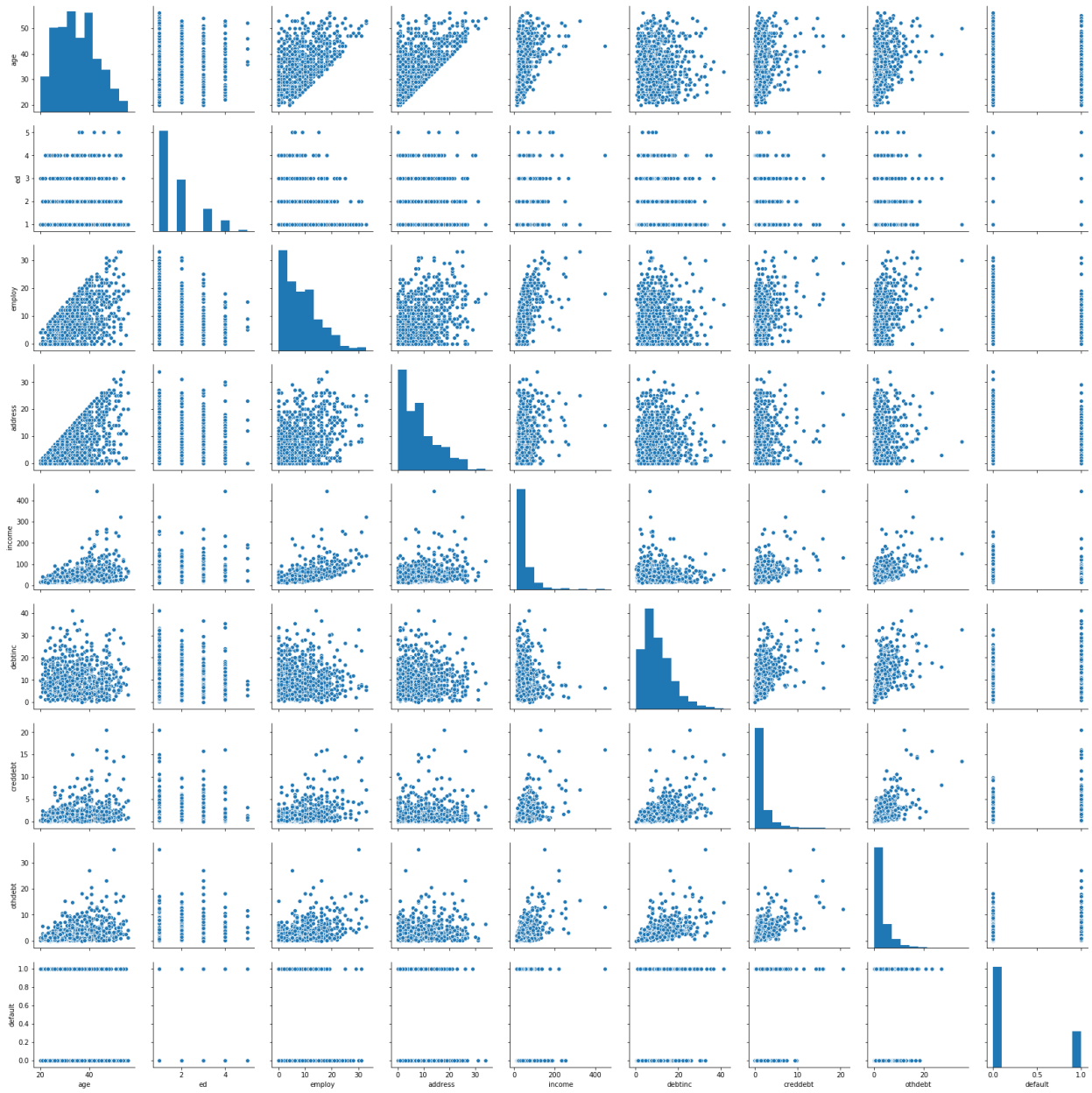

2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
ed

Bivariate Analysis

Ggpair function built upon ggplot2, GGally provides templates for combining plots into a matrix through the ggpairs function. Such a matrix of plots can be useful for quickly exploring therelationships between multiple columns of data in a data frame.The lower and upper arguments to
the ggpairs function specifiesthe type of plot or data in each position of the lower or upper diagonalof the matrix, respectively.For continuous X and Y data, one can specify the smooth option toinclude a regression line.

Below figures shows relationship between independent variables and also with numeric target variable using ggpair

i. Below ggpair graph is showing clearly that relationship between independent variables 'othdebt' and 'debtinc' are strong.
ii. The relationship between 'othdebt' with target variable 'default' is very less.



Missing Value Analysis

Missing values in data is a common phenomenon in real world problems. Knowing how to handle missing values effectively is a required step to reduce bias and to produce powerful models.

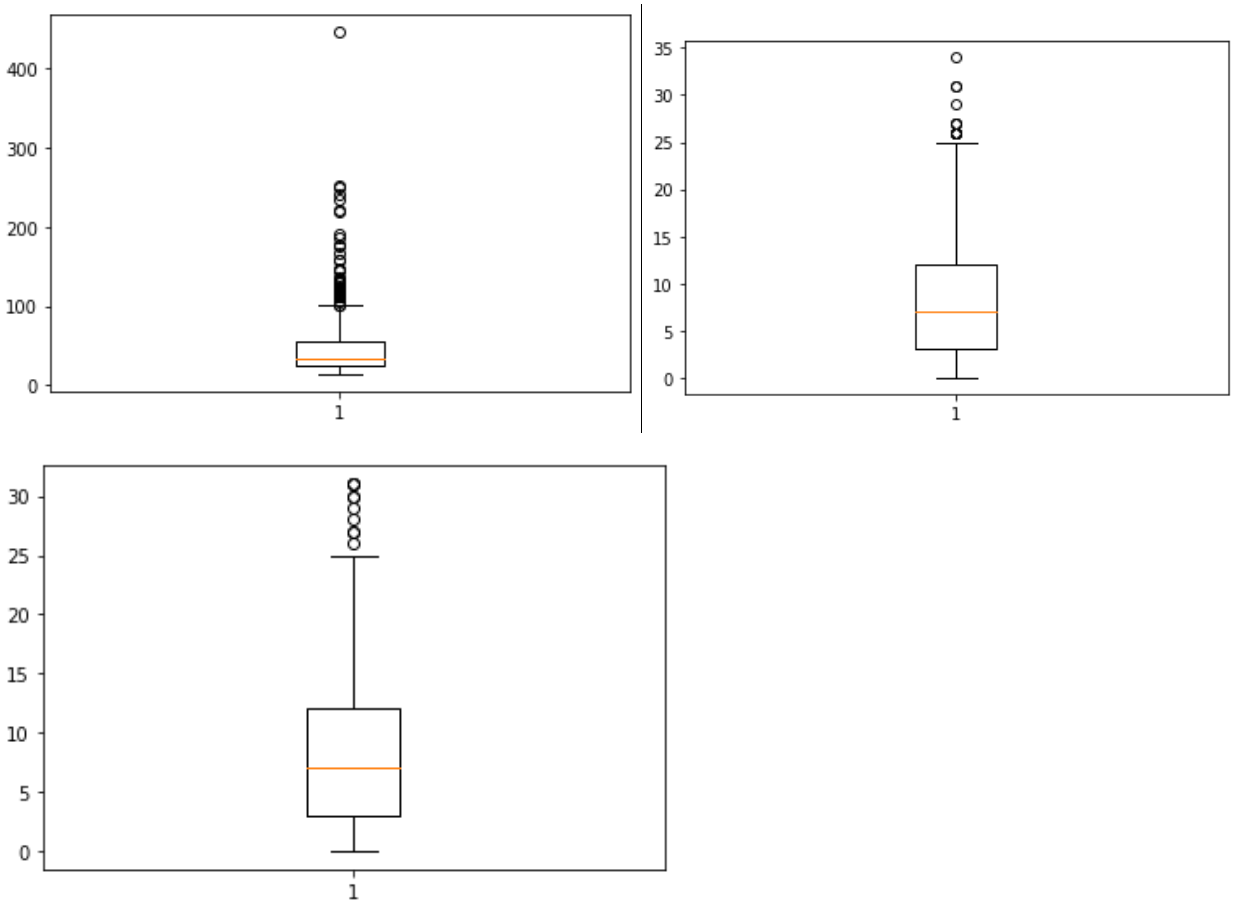
There are 150 missing values in the target value i.e, default.we cannot impute the missing values in the target variable as our model will get bias.so we will separate the data of missing values and make the model on remaining 700 obserbvations.later we can predict the missing values in the target variable after making a model.

Outlier Analysis

The Other steps of Preprocessing Technique is Outliers analysis , an outlier is an observation point that is distant from other observations. Outliers in data can distort predictions and affect the

accuracy, if you don't detect and handle them appropriately especially in regression models..

As we are observed there are outliers in 3 independent variables "income","address","employ":



We removed the ouliers in the 3 variables using the z-score method.All the outliers were treated.

Features Selections

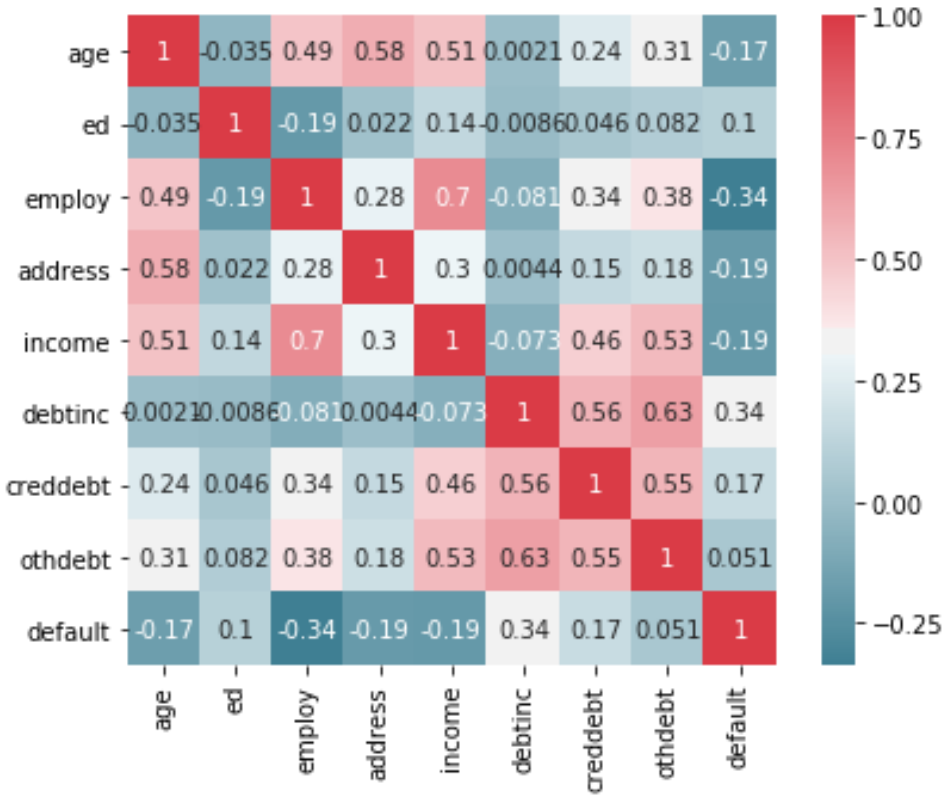
Machine learning works on a simple rule – if you put garbage in, you will only get garbage to come out. By garbage here, I mean noise in data.

This becomes even more important when the number of features are very large. You need not use every feature at your disposal for creating an algorithm. You can assist your algorithm by feeding in only those features that are really important. I have myself witnessed feature subsets giving better results than complete set of feature for the same algorithm or – "Sometimes, less is better!".

We should consider the selection of feature for model based on below criteria

i. The relationship between two independent variable should be less and
ii. The relationship between Independent and Target variables should be high.

Below fig 2.6 illustrates that relationship between all numeric variables using Corrgram plot .



Color dark blue indicates there is strong positive relationship and if darkness is decreasing indicates relation between variables are decreasing.

Color dark Red indicates there is strong negative relationship and if darkness is decreasing indicates relationship between variables are decreasing.

2.2.4 Features Scaling

The word "normalization" is used informally in statistics, and so the term normalized data can have multiple meanings. In most cases, when you normalize data you eliminate the units of measurement for data, enabling you to more easily compare data from different places. Some of the more common ways to normalize data include:

Transforming data using a z-score or t-score. This is usually called standardization. In the vast majority of cases, if a statistics textbook is talking about normalizing data, then this is the definition of "normalization" they are probably using.

Rescaling data to have values between 0 and 1. This is usually called feature scaling. One possible formula to achieve this is.

In our case the was no need for scaling as the variables were not having high values through will our would be affected.

Model Selection

In our case dependent variable is categorical so, the predictive analysis that we can perform

is

classification Analysis

We will start our model building from logistic classification.

Evaluating logistic classification Model:

- Prediction Model – Simple linear regression – Multiple linear regression
- Describe relationship among variables
- The one simple case is where a dependent variable may be related to independent or explanatory variable.

SUMMARY :

Logit Regression Results

Dep. Variable:	default	**No. Observations:**	508
Model:	Logit	**Df Residuals:**	500
Method:	MLE	**Df Model:**	7
Date:	Tue, 03 Dec 2019	**Pseudo R-squ.:**	0.3011
Time:	18:05:59	**Log-Likelihood:**	-197.33
converged:	True	**LL-Null:**	-282.32
Covariance Type:	nonrobust	**LLR p-value:**	2.524e-33

	coef	std err	z	P>|z|	[0.025	0.975]
age	0.0044	0.017	0.253	0.801	-0.030	0.038
ed	0.1020	0.145	0.701	0.483	-0.183	0.387
employ	-0.2481	0.043	-5.763	0.000	-0.333	-0.164
address	-0.1055	0.028	-3.755	0.000	-0.161	-0.050
income	-0.0241	0.018	-1.364	0.173	-0.059	0.011
debtinc	0.0272	0.037	0.731	0.465	-0.046	0.100
creddebt	0.8684	0.169	5.146	0.000	0.538	1.199
othdebt	0.0342	0.130	0.263	0.793	-0.221	0.289

#check accuracy of model

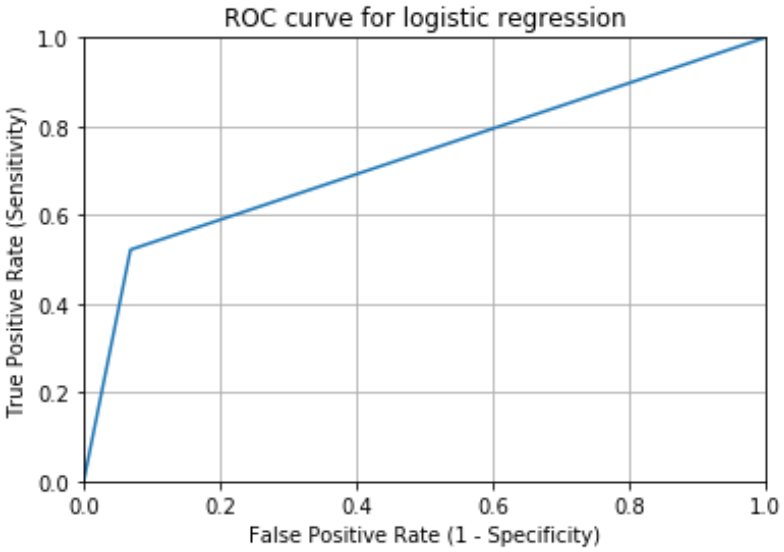
#accuracy_score(y_test, y_pred)*100

((TP+TN)*100)/(TP+TN+FP+FN)

```
ACCURACY = 85.6

# False negative rate ((FN*100)/(FN+TP)) = 47.82608695652174

To prove model stability below is the ROC-CURVE
```



The accuracy is good but the false negative rate is low which is good for the model.

Decision Tree

A tree has many analogies in real life, and turns out that it has influenced a wide area of **machine learning**, covering both **classification and regression**. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.

#check accuracy of model

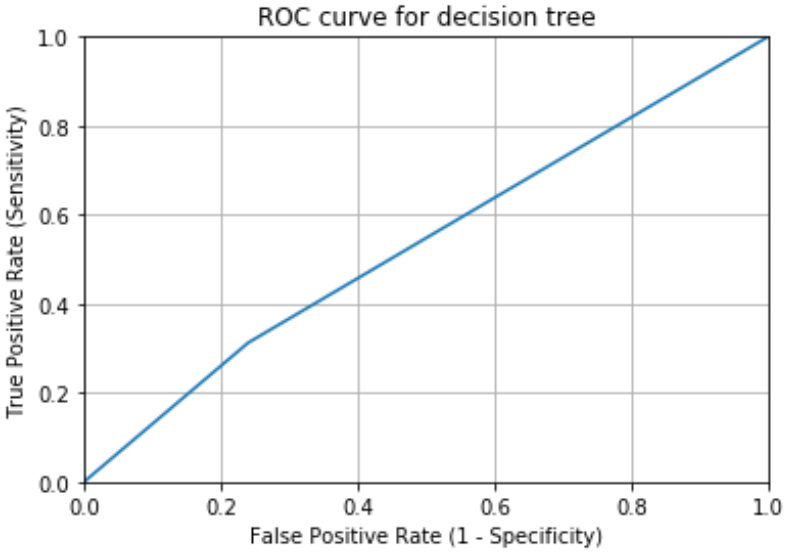
#accuracy_score(y_test, y_pred)*100

((TP+TN)*100)/(TP+TN+FP+FN) = 65.15

#False Negative rate

(FN*100)/(FN+TP) = 68.75

The accuracy of the model is okay but the false negative rate is very high. To validate the model we plot the ROC –CURVE.



Area under the curve is not good. This shows the model is not able to predict the target vriable efficiently.

Random Forest

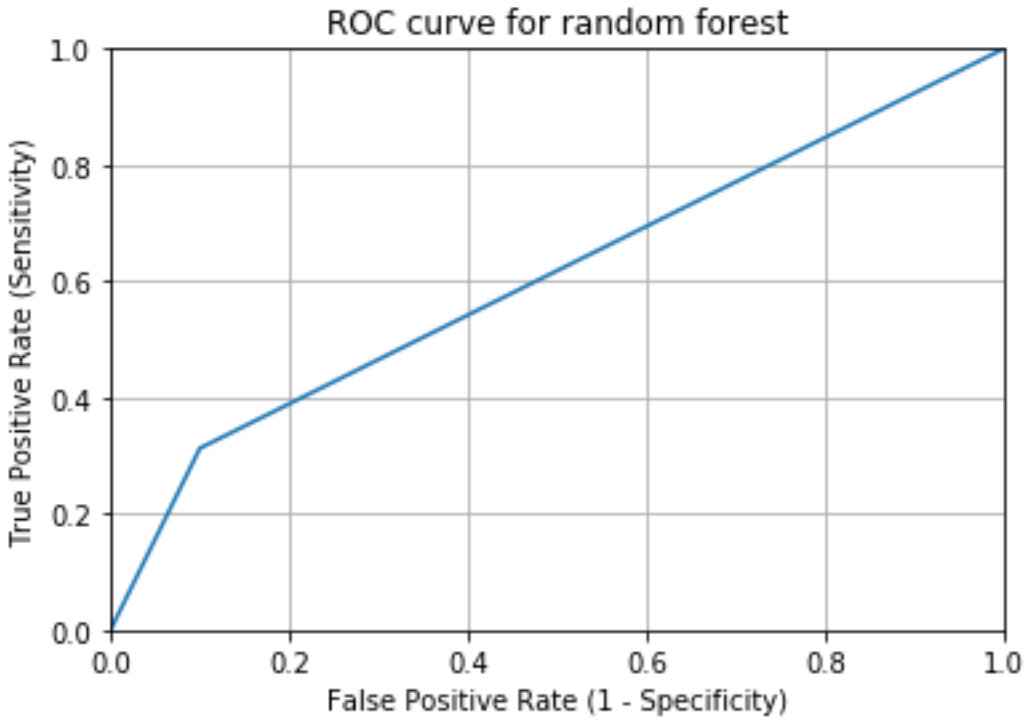
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Random forest functions in below way

i. Draws a bootstrap sample fromtraining data.
ii. For each sample grow a decision tree and at each node of the tree
   a. Ramdomly draws a subset of mtry variable and p total of features that are available
   b. Picks the best variable and best split from the subset of mtry variable
   c. Continues until the tree is fully grown.

#check accuracy of model
#accuracy_score(y_test, y_pred)*100
((TP+TN)*100)/(TP+TN+FP+FN) = 75.75%
#False Negative rate
(FN*100)/(FN+TP) = 68.75%

To check the model validation we plot the ROC –CURVE



The accuracy is good but the false negative rate is also high and the area under the curve is less.

Model Selection

As we predicted defeaut values of bank loan prediction data using four Models Decision Tree, Random Forest,naïve bayes and logistic Regression as accuracy is high and false negative rate is less for the naïve bayes Model so conclusion is

#conclusion : logistic regression is the best suited model for this project as it gives better accuracy,false negative rate is low and roc curve is good.