Data Analysis

Loans.csv



* **Name: Rohit Pravin Sapkal**
* **Course: DataScience For Business**
* **Subject: Data Analytics**

Table of Contents

* **Introduction**
* **Crisp-DM Methodology**
* **Understanding of Business**
* **Data Understanding**
* **Data Preparation**
* **Modeling**
* **Evaluation**
* **Deployment**
* **Understanding of Business**
* **Determination of Business Objective**
* **Business Success Criteria**
* **Situational Assessment**
* **Data Summary (Exploratory Data Analysis)**
* **Data Preparation**
* **Modeling**
* **Logistic Regression**
* **Decision tree**
* **Neural network**
* **Evaluation and Deployment**

**Introduction**

Data evaluation is the manner of collecting information, manipulating information, cleansing up information, after which making use of a few algorithms with the precise precision of our carried out algorithms, this means that our version is examined on each check information and educated information. Totally depending on the analyst, however, the maximum of the time we take is 25% for train out and 75% for the test. In this we see that our version is operating properly or not, otherwise, we need to follow new algorithms with a few increase algorithms in our information units in order that we are able to are expecting or provide our purchaser the proper decision.

**Problem statement:**

The Stir bank is in difficult situation they give a loan to the people who promised to pay it back but they did not so Stir bank want to identified those people who are not able to pay the loan amount using their track record or the data In simple word they want to know whether this loan aspirant person is eligible for the loan or not. For this, Stir bank give a data of 2000 customer to predict or build model which help bank to identified the person who is eligible to for the loan or not so that, they cannot meet the fraud people who promise to pay the loan amount but they are not. Stir Bank wants the model which able to find out which customer is likely to repay their loan?

To solve this Stir Bank problem I am going to use three algorithms namely

1. Logistic Regression
2. Decision tree
3. Neural Network

The Stir Bank want to classified the customer so Logistic regression is the classification algorithm and also the part of supervised learning.

Decision tree is the algorithm which is suits for predictive and classification modeling problems.

Neural Network is almost similar to the operation of human brain to identify the relationships also used in fraud detection and risk management.

To solve this problem or to build the model which recognize whether person is able to pay the loan or not. When I use Python, it is important to have a basic understanding of Python. I have used Pandas, NumPy, Matplotlib, ScikitLearn, Seaborn libraries,tensorflow,keras. Your system must have Python API, Jupyter Notebook, or Google Collab. Together with the data provided.

**Crisp-DM Methodology**

Crisp –DM Methodology is the procedure of series facts from any database like Oracle, MySQL, and

Many greater after that we want to recognize enterprise and the query that customer need us to solve.

In this client is Stir Bank and they wanted to know the eligibility of person towards Loan repayment. This method has a six phases that describe the data analytics project life cycle.

1. **Understanding the business:**

At this stage, we need to understand what is needed for business growth.

1. **Understanding the data** :

At this stage, we need to understand the data, which means whether we have clean data, what data do we have.

1. **Data Preparation** :

We need to organize our data or make our data ready for modeling.

1. **Modeling** :

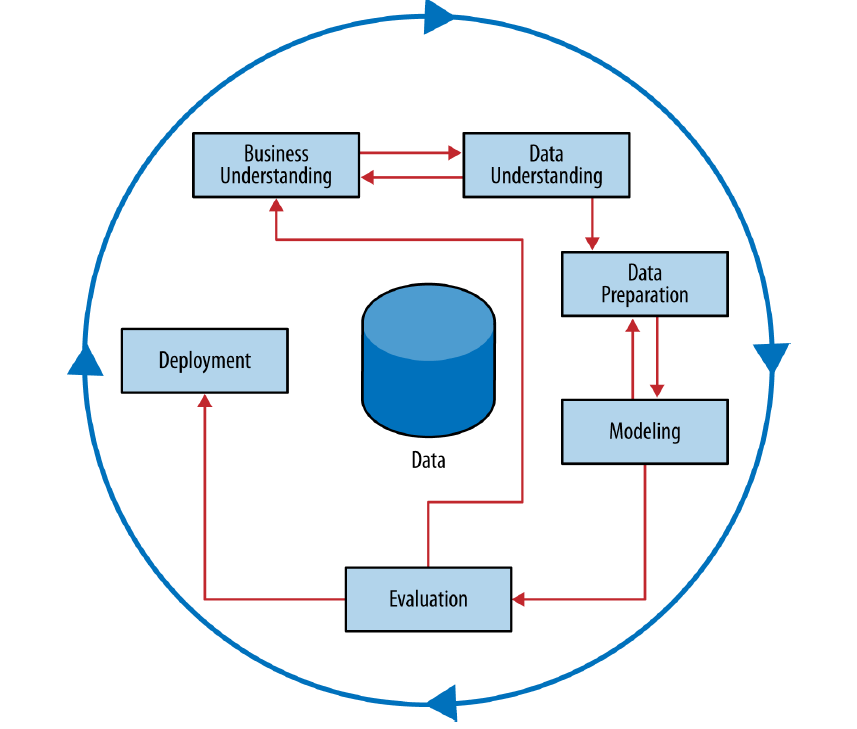
At this stage, we need to determine which modeling techniques are suitable for our business growth.

1. **Evaluation** :

Then we must evaluate which model is suitable for our business

1. **Deployment** :

This is the final step in this we finally provide result to our stakeholder.



**Business Understanding**

It is one of the fundamental stages in the records evaluation assignment lifestyles cycle. Once we recognize our commercial enterprise records and their goal which assists us to develop their commercial enterprise in a worthwhile direction. Business expertise manner we want to recognize fundamental hassle of commercial enterprise or expertise the consumer commercial enterprise. In the given dataset fundamental goal is to decide which human beings are capable of pay off their mortgage through the use of a few music records. This segment additionally has 4 steps to recognize commercial enterprise efficiently that's as follows

1. Decide business Objective
2. Evaluate Situation
3. Decide data mining goals
4. Product plan of the Project

Variables: The variable is the quantity that adjustments are primarily based totally at the context and the motion taken on the time. They are divided into structured variables and unbiased variables, unbiased variables are unbiased of different variables or quantities, and structured variables are laid low with adjustments in different variables. The overall performance right here is the structured variable the others are unbiased.

Logistic regression fashions use logistic functions (additionally referred to as sigmoid functions) to generate output from zero to 1. When drawing, it creates an S-formed curve. It is given as 1 / (1 + ex). The end result received after the use of a characteristic with unbiased variables, e is Euler's number.

A Decision tree is a tree version of selection-making and its viable consequences, along with the effects of random events, aid costs, and benefits. In our case, the customer can pay or not pay his mortgage, in the near future; the loan will be paid yes or no.

The neural network consists of several layers, where the output of the past layer is inserted into the next layer as input. You can use a variety of trigger functions to determine the output. The first layer where our input arrives is called the input layer. The layer where we get the result is called the output layer. The middle layer is called the hidden layer.

**Determination of Business Objectives:**

This client want to stop giving loan to those people who are not able to pay their loan amount according to the data that they have or get before giving approval to the customer in that many things like customer age, Years at the address, Income and many more. Using this dataset I need to build the model which is effective towards identifying the people who is eligible for the loan and which one is not eligible so that, Stir Bank may remain safe from the frauds.

**Business Success Criteria:**

There are plenty of factors to discover the fulfillment standards of enterprise for this it's far obligatory to study the records and diagnosed the enterprise troubles and remedy that during a right manner. For this we want to comply with the following

1. Access to the project throughout.
2. Monitor cycle times and project cycle times.
3. Process and workflow optimization.
4. Project status across the portfolio.
5. Monitor risk and roadblocks.
6. Dashboard and report with real time data.
7. Visualize planned vs. actuals ongoing.
8. Share repots with stakeholders.

**Situational Assessment:**

A situation evaluation is executed at the start of the project and its miles a part of the inquiry segment with inside the project. At the cease of the mission overview of the development and achievements is vital to gather moreover, with this information replace of the project is depended with this we get to recognize the following

1. Project Vision
2. Understand the current situation of the project health
3. Understand the problem of the project health
4. Understand the main problem of the project

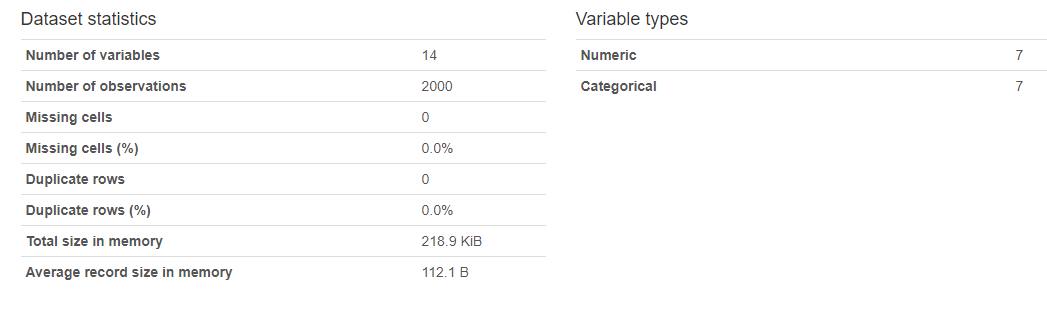
**Data summary:**

Stir Bank are in trouble they are not able to distinguish which customer is able to pay the loan so that they want a machine learning model which able to identified the fraud customer using some of their data and that data Stir bank gives to me to study and give them a solution using some machine learning algorithms.

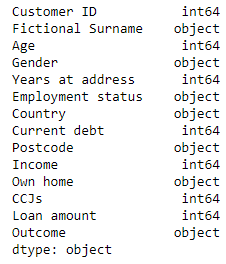
The dataset contain 2000 customer information that include both defaulter as well as innocent customer who pay their loan on time. The dataset contain the following information.

EDA (Exploratory Data Analysis) in this we gone find the all information about our datasets like mean, mode, median, how many missing values and many more. Even we gather the full knowledge of each and every column as well.

Basic Overview:



Data Type:

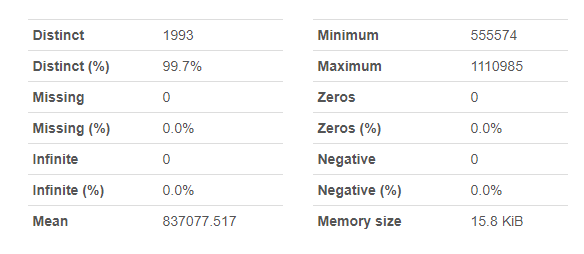


**Variables:**

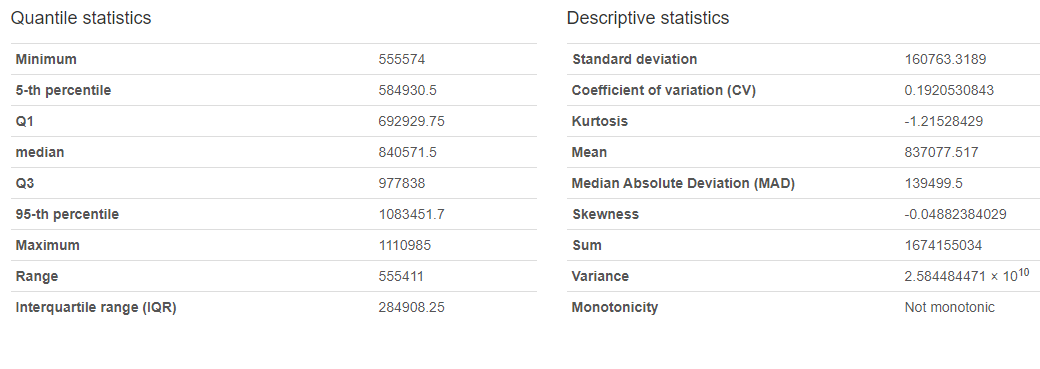
1. **Customer ID**:
   * + Type: Identifier variable (Numeric).
     + Description:

The account number used to uniquely identify a customer. We don’t need this table for our analysis because it is not used to identify the customer this column has no impact on the model building.

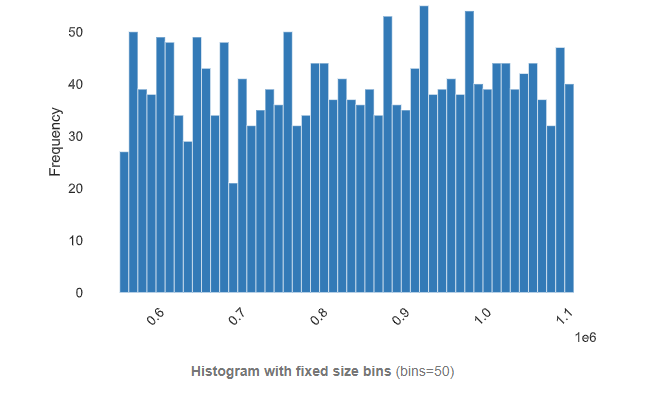
* + - EDA:



* Statistics:



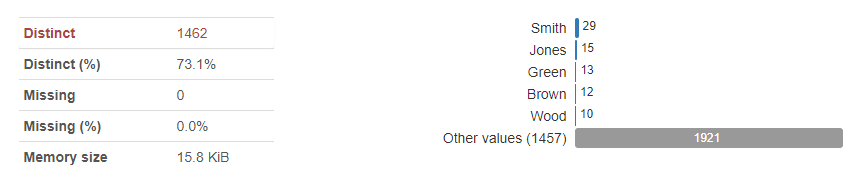
* Histogram:

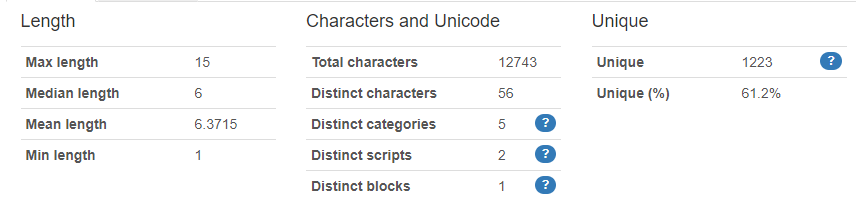


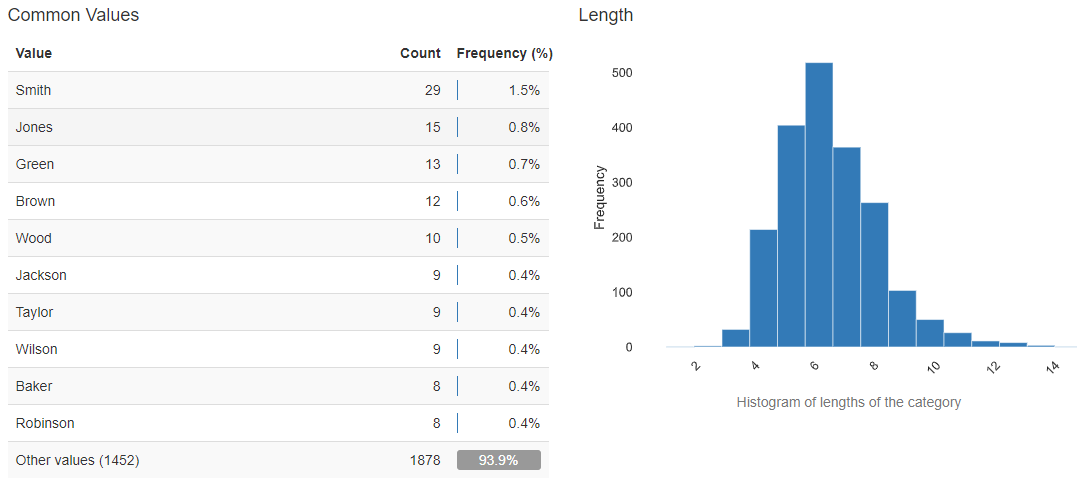
1. **Fictional Surname**:
   * + Type: Categorical Variable
     + Description:

Customer's family name. we have 73.1% Distinct values. Which means we have same names in this Bank dataset. It is no needed as it has very less impact and isn’t contributed much in the model building.

* + - EDA:





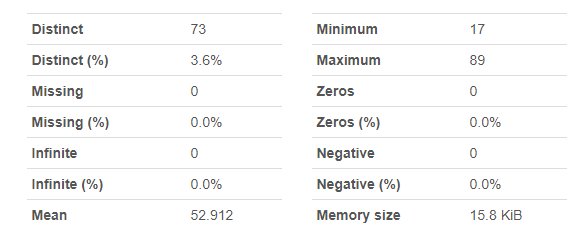


1. **Age** :
   * + Type: Discrete Numeric Variable
     + Description:

Customer's age in years. In this Dataset minimum age is 17 and maximum age is 89. This

variable we need for building models so that I keep this to build our models. This variable has high correlation with ‘years at address’ column.

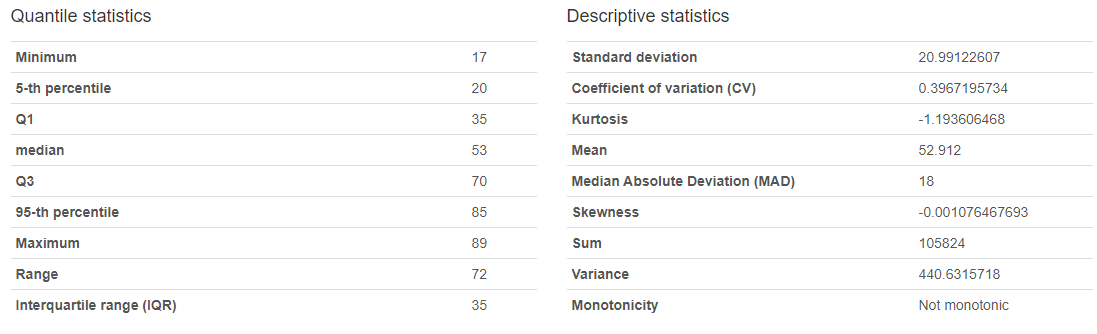
* + - EDA:



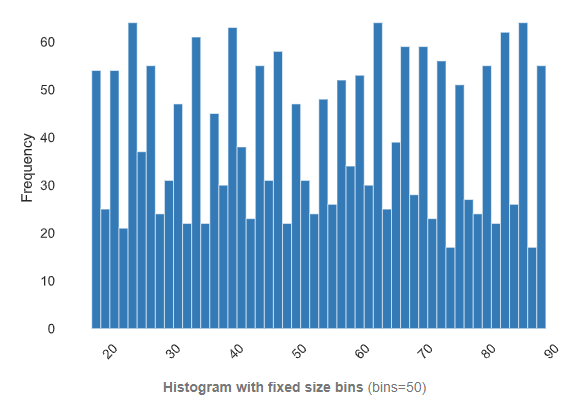
* Common values:



* Statistics:



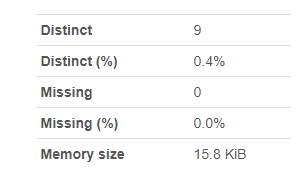
* Histogram:



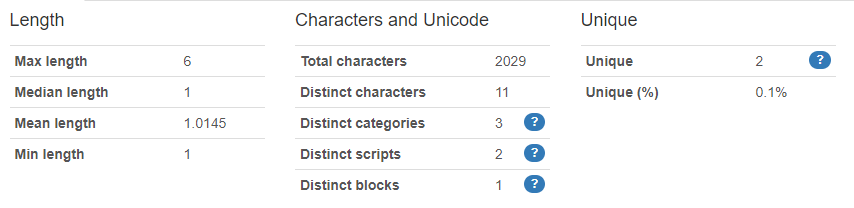
1. **Gender**:
   * + Type: Categorical Variable
     + Description:

Customer's gender. For the model building we don’t need this variable because this has almost no impact on model so that we don’t need this table for analysis.

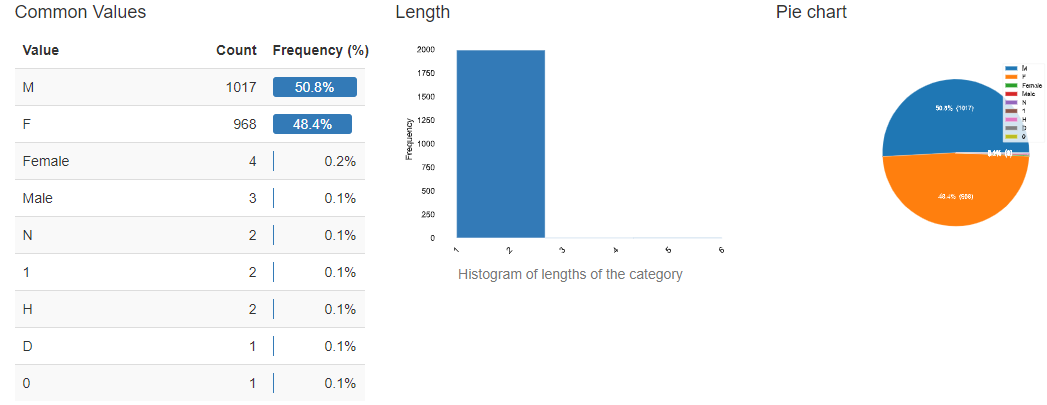
* + - EDA:



* Overview:



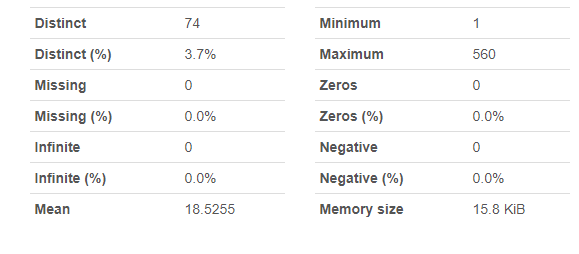
* Common values:



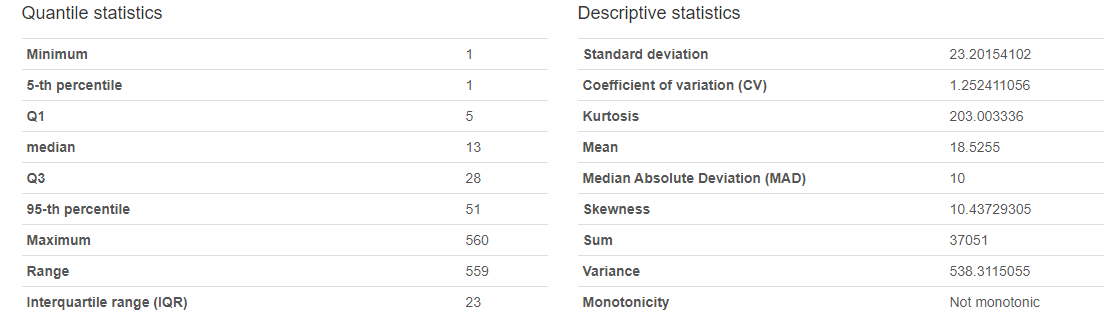
1. **Years at address**:
   * + Type: Discrete Numeric Variable
     + Description:

Number of years the customer has lived in their present home. It is important table for our model building because it has direct impact on loan aspirants. So that I will keep this table it also gives us the information about customers’ location. This table shows high correlation with ‘Age’ column.

* + - EDA:



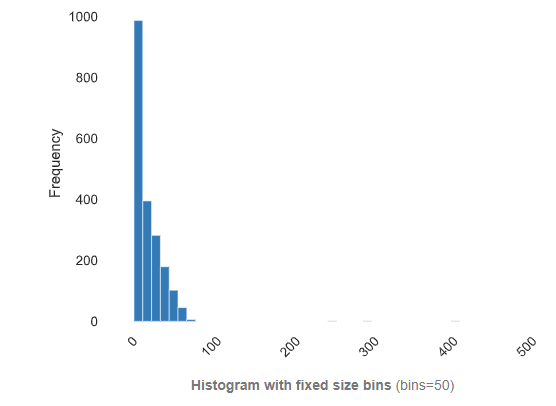
* Statistics:



* Common values:



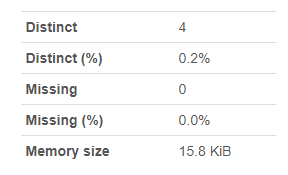
* Histogram:



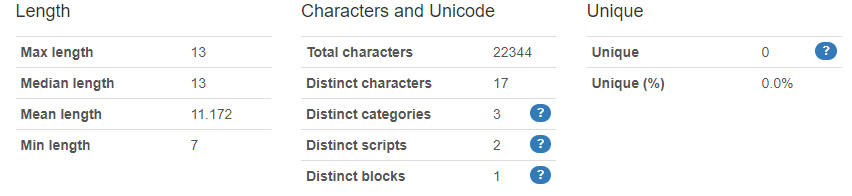
1. **Employment status**:
   * + Type: Categorical Variable
     + Description:

Is the customer employed by a company, self-employed, or out of work? This table is also very important we have four distinct categories to differentiate customers and get to know that is that person is able to pay the loan or not.

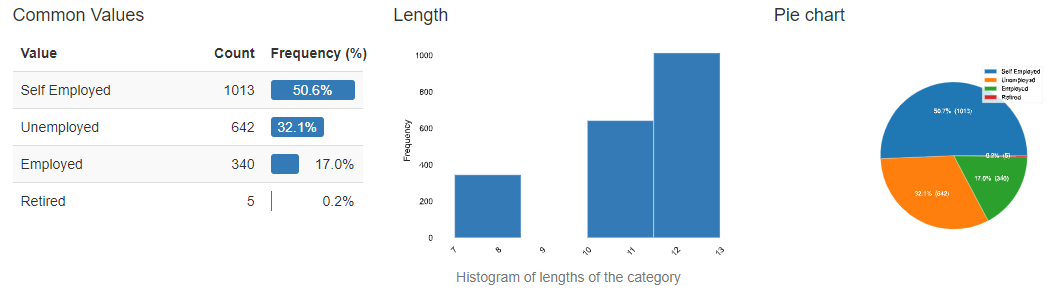
* + - EDA:



* Overview:



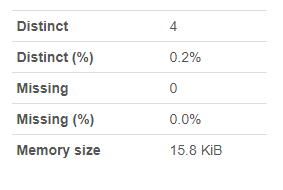
* Common Values:



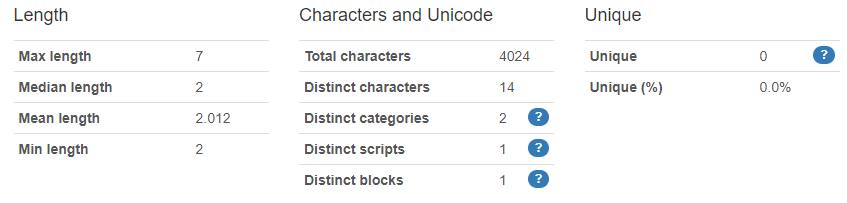
1. **Country**:
   * + Type: Categorical Nominal variables
     + Description:

The country the customer lives in. It contains 4 distinct values and give us information about the location of customer although a majority of the customer live in the UK (1994 Customers) and we will remove this variable in the next part also we have three more countries France, Spain, Germany.

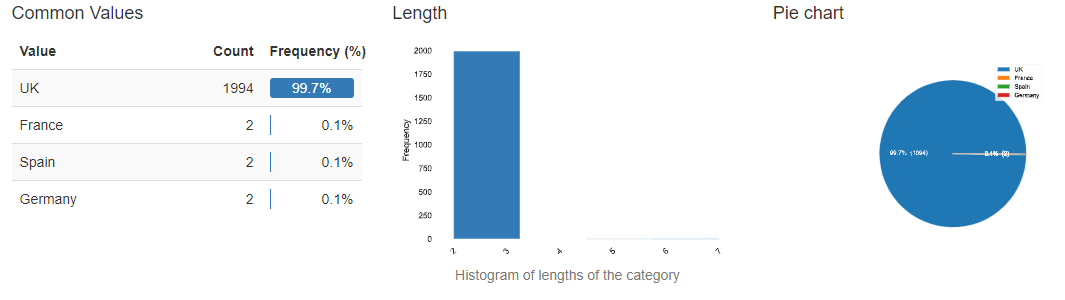
* + - EDA:



* Overview:



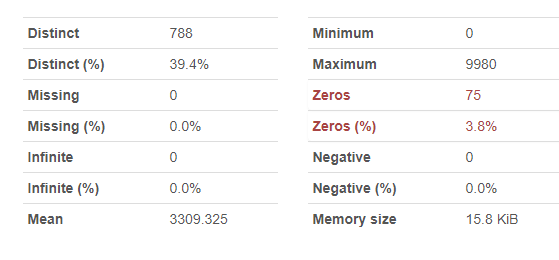
* Common Values:



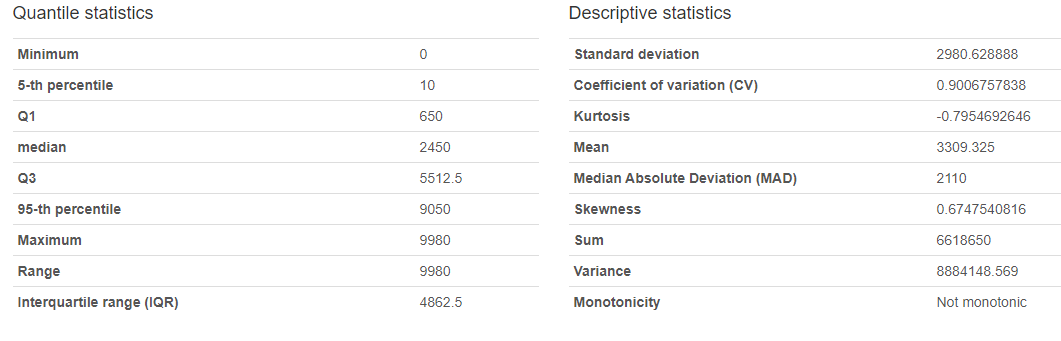
1. **Current debt**:
   * + Type: Discrete Variable
     + Description:

Total value in pounds of any existing debts owed by the customer. We need this table because it plays an important role in model building the minimum current debt is 0 and maximum is 9980.

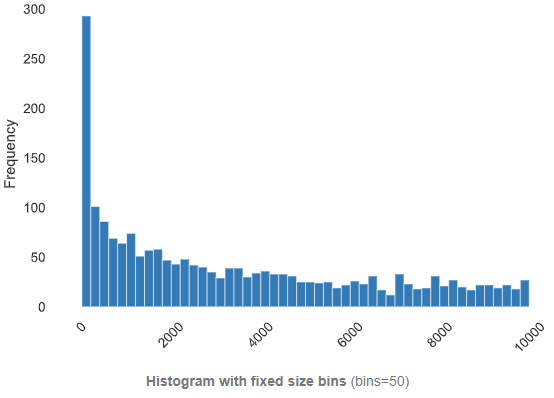
* + - EDA:



* Statistics:



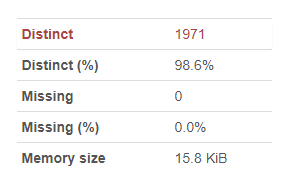
* Histogram:



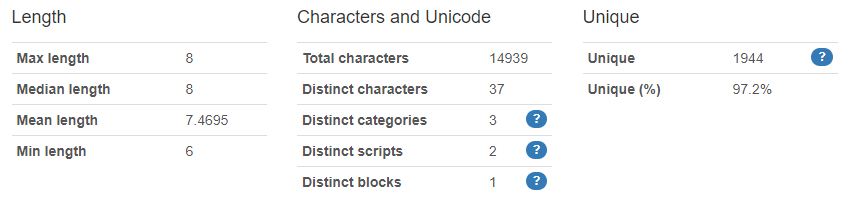
1. **Postcode**:
   * + Type: Categorical Variable
     + Description:

The postcode of the customer's home address. It has impact on model building and we have 1971 distinct values. Is also gives the areal information about the customer. Highly Cardinality and Uniform in nature.

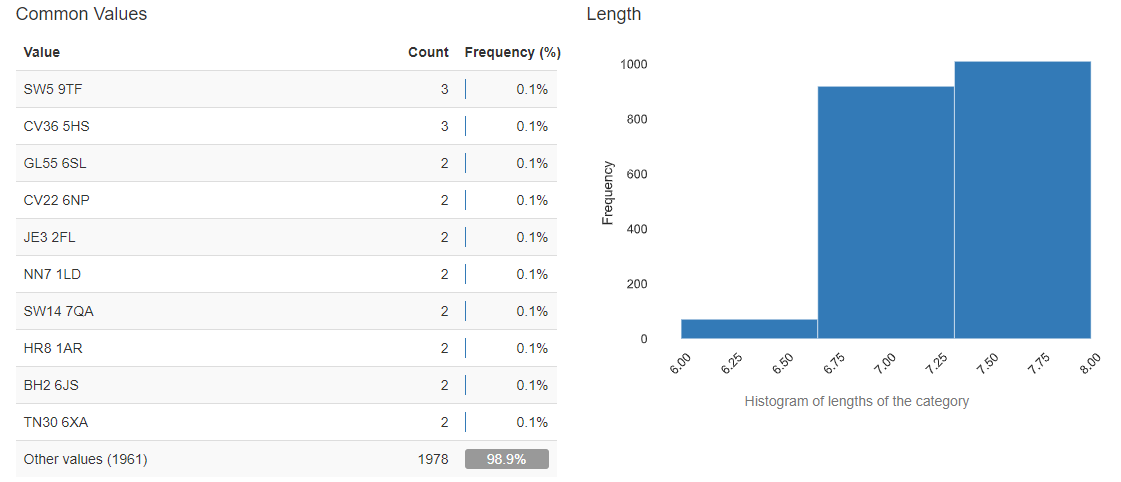
* + - EDA:



* Overview:



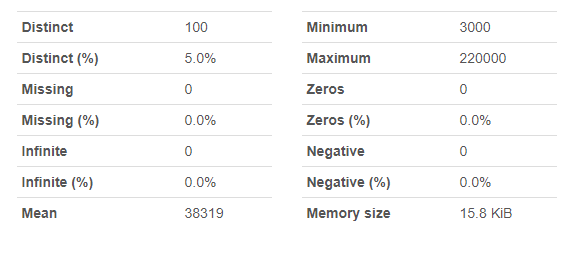
* Common values:



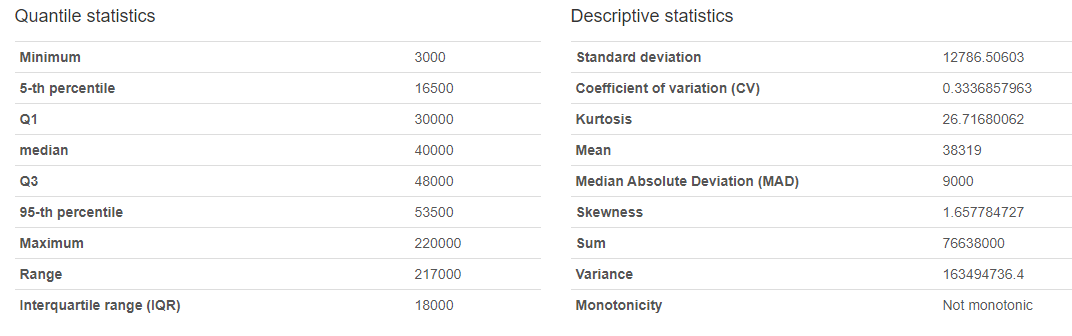
1. **Income**:
   * + Type: Discrete Numeric Variable
     + Description:

The income of customers is in pounds. This table is useful for the model building because it gives us the information about customer wealth. We have minimum Income 3000 and maximum 220000. Highly correlation with ‘loan amount’ column.

* + - EDA:



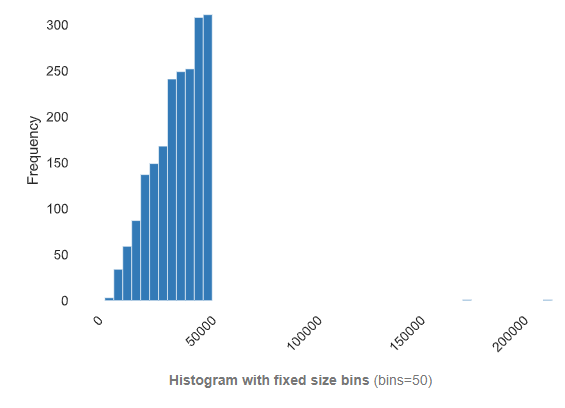
* Statistics:



* Common values:



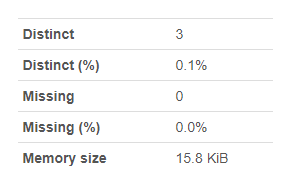
* Histogram:



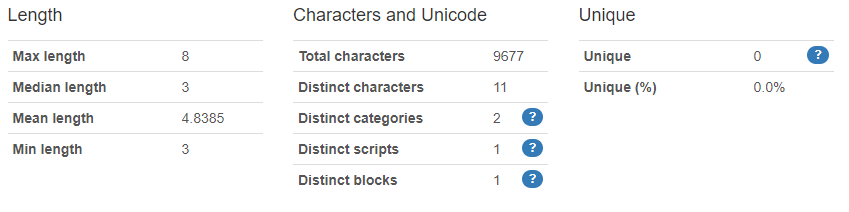
1. **Own home**:
   * + Type: Categorical Variable.
     + Description:

We get the information about customers where they are live in a rented, own house. We have three distinct values in this table which is Own, Mortgage, and Rent. It gives us the information about customer living status. We will need this table.

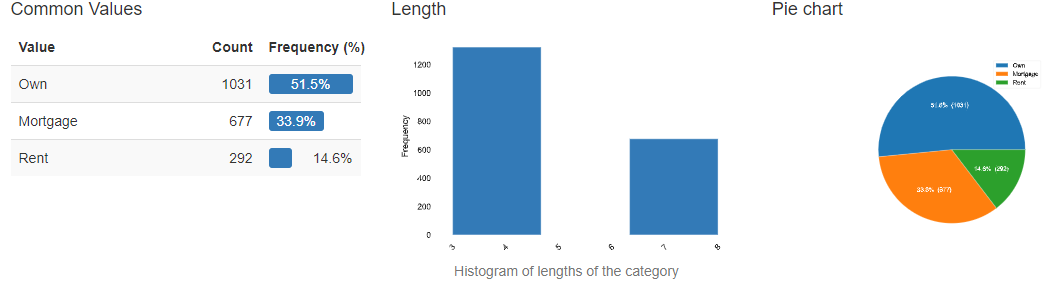
* + - EDA:



* Overview:



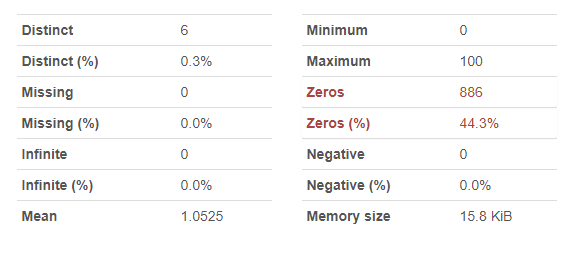
* Common values:



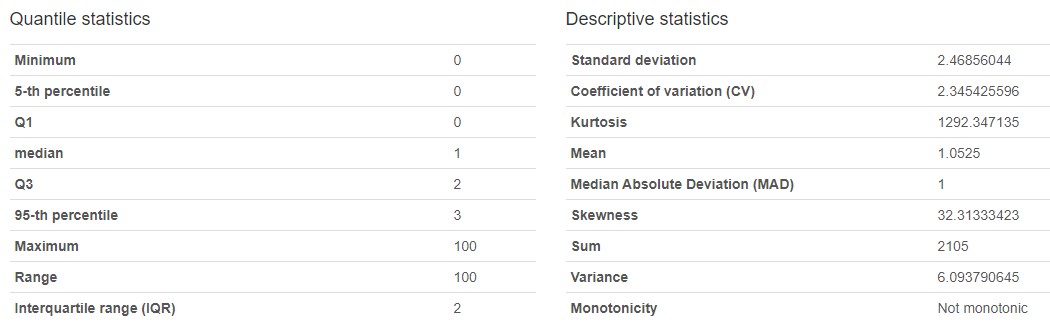
1. **CCJs**:
   * + Type: Categorical nominal Variable
     + Description:

The number of County Court Judgments which means the customers has had against them for previous unpaid debts.it gives the information about customer that how many Court Cases against unpaid debts. We have six distinct numbers of cases which is 0, 1, 2, 3, 10, and 100.

* + - EDA:



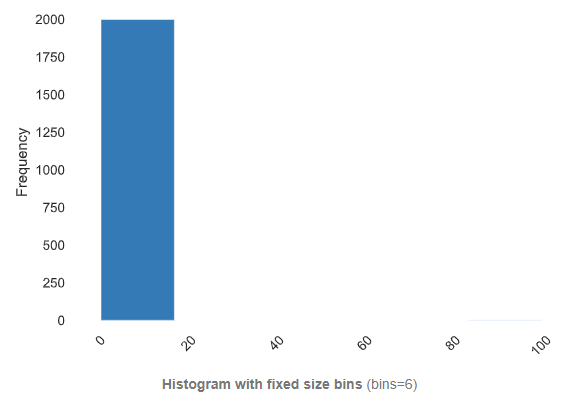
* Statistics:



* Common values:



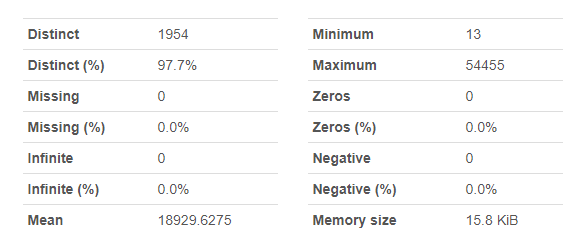
* Histogram:



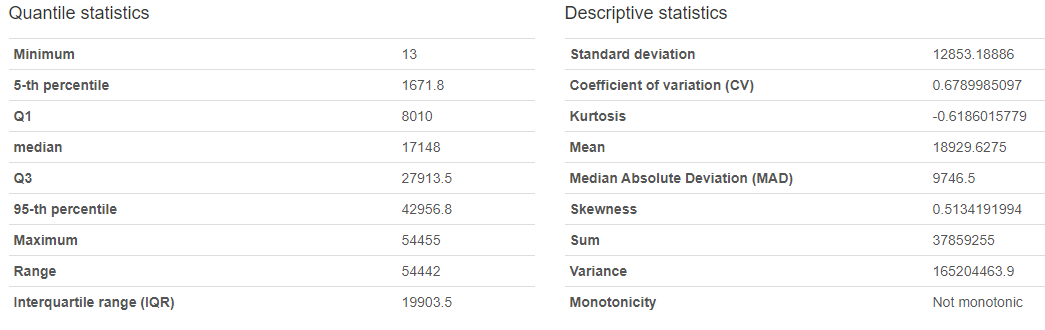
1. **Loan amount**:
   * + Type: Discrete Numeric Variable
     + Description:

The total amount requested in the loan, in pounds. It gives the customer acceptation towards loan amount in which minimum loan amount is 13 and maximum is 54455. We have 1954 distinct values. Highly correlation with ‘Income’ column.

* + - EDA:



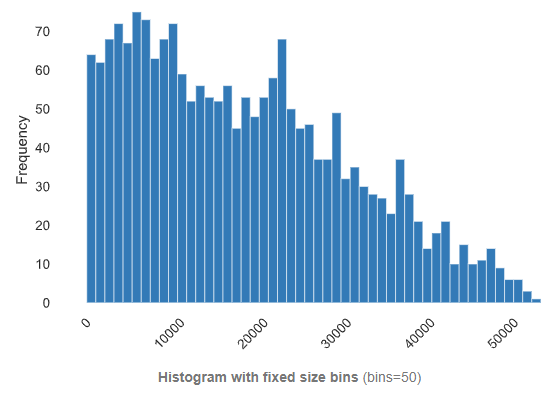
* Statistics:



* Common values:



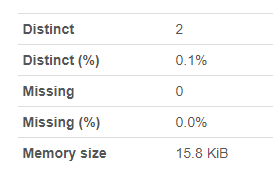
* Histogram:



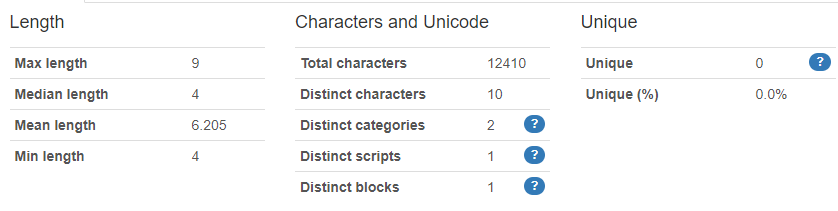
1. **Outcome**:
   * + Type: Categorical Nominal Variable
     + Description:

Was the loan repaid (paid) or not repaid (defaulted). It gives the information about customer behavior toward loan payments. We have two distinct values in this table although we get the information about how many customers is able to pay their loan and how many are not able to pay their loan. 1118 number of customers is able to pay their loan and 882 numbers of people are not.

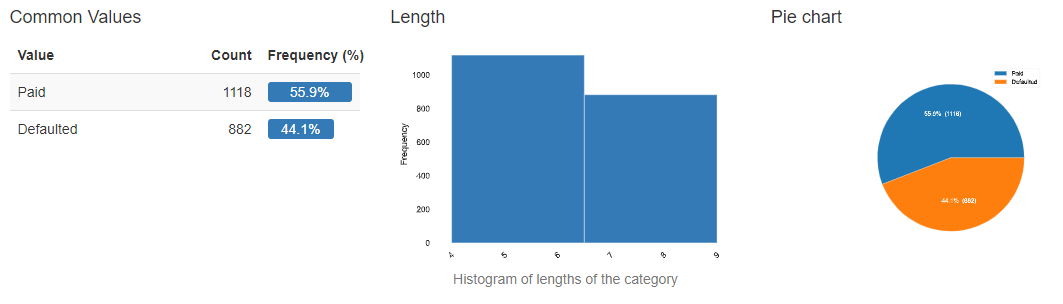
* + - EDA:



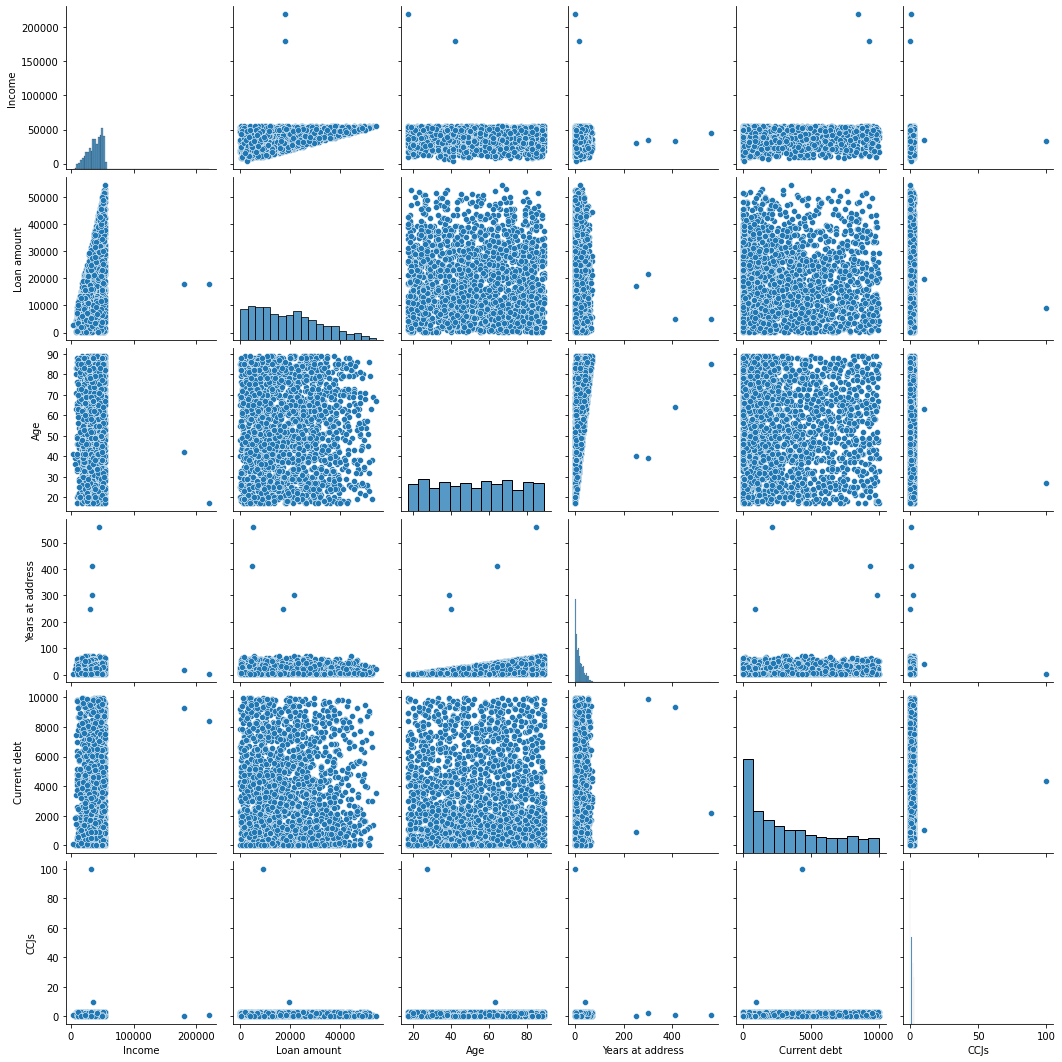
* Overview:



* Common values:



All Data

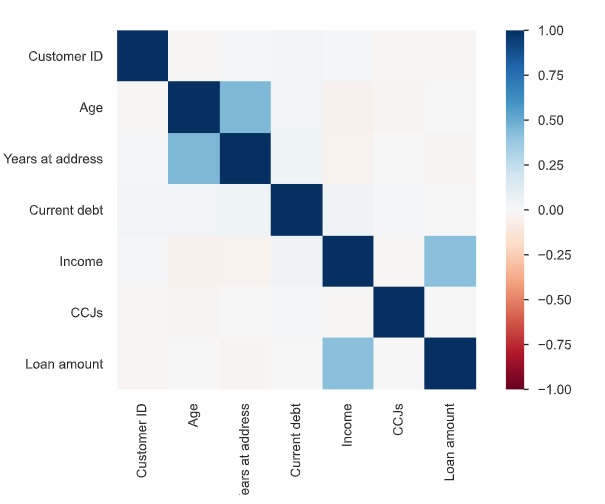


To gather the information about the variable have any relationship between them or are correlated with each other, I run the Pearson’s r, Phik (φk) correlation test.

Pearson’s r:

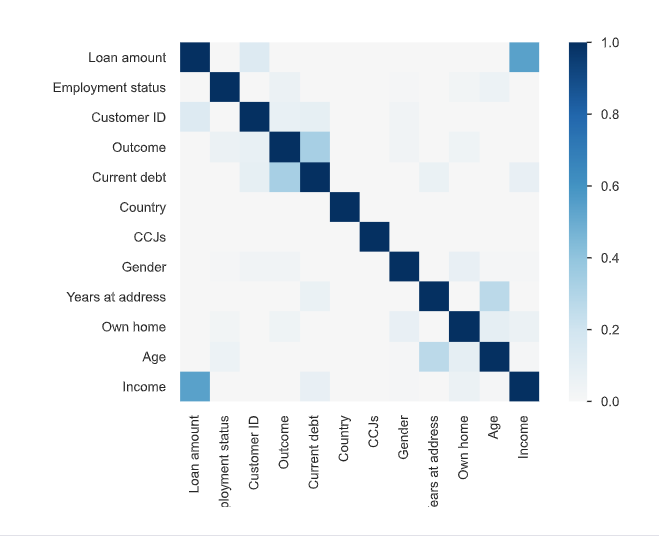
The Pearson correlation (r) may be alive of linear correlation amongst variables. It' values lie amongst -1 and +1, -1 indicating fashionable negative linear correlation, zero indicating no linear correlation, and one indicating fashionable incredible linear correlation. Furthermore, r is invariant below separate adjustments in place and scale of the 2 variables, implying that for a linear carry out the mindset to the coordinate axis could now no longer have an effect on r.

To calculate r for 2 variables X and Y, one divides the variance of X and Y with the useful resource of the usage of the goods of their trendy deviations.



Phik (φk):

Phik (φk) can be a new and realistic parameter statistic. It works systematically between the classification, order, and periodic variables of the programming language. It records the non-linear dependencies and simply in the case of normal distribution. Use Pearson correlation coefficient. Enter the quantity.



The values continually lie among -1 and 1. A correlation among a variable and itself 1. The issues correlation some of the variables display us that the variables are closely enthusiastic about each other. It offers upward push to more than one correlations some of the information given and offer inessential significance to a few variables while constructing a version. This can construct the version over-fit, assign weights incorrectly and manufacture an incorrect result.

* 0 means no relationship at all.
* 1 represents there is a strong positive correlation between the variables.
* -1 represents there is a strong negative correlation between the variables.

**Data Preparation:**

While doing data preparation we need to spilt data in to test and train data set I make it 25% test and 75% train. Before applying models we need to clean our data or may need to change variables like replacing Paid with 1 and Unpaid with 0 but the both are in string format.

* **Train dataset:** we need this data for learning process or in the other word we need this data for training so that we use these successful algorithms on test as well as live data to make predictions. Our model learns from this dataset. In this dataset we have all the independent variable and the target variable.
* **Test dataset:** This dataset is not dependent on the training dataset but follow the same algorithms as the training dataset and if the model fits into the training dataset then it also fit to the test dataset. We need this dataset because this is the best example for the live data so that it makes prediction process simple. In this dataset we have all the independent variable but not the targeted variable. We use this to predict the targeted variable.
* **Sample Submission:** This file we have the format in which we are going to submit our predictions.

Following result we have after splitting dataset:

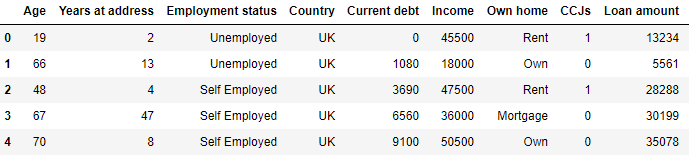
X\_train: (1600, 9),

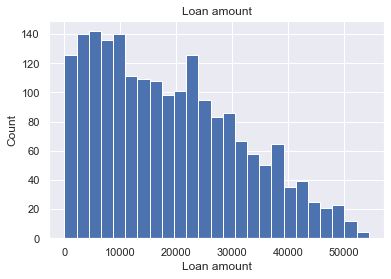
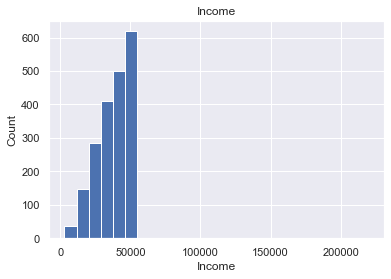
X\_test : (400, 9),

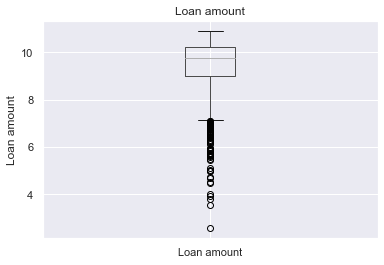
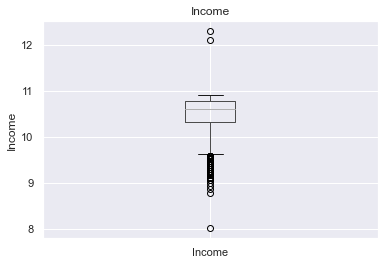
y\_train: (1600, 1),

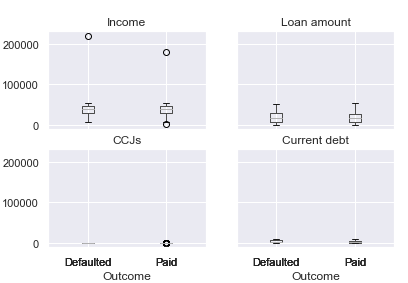
y\_test: (400, 1)

After splitting dataset we work on train dataset we drop some of the columns that does not make any effects on our model Moreover we make some changes on columns as well to make our table more compatible towards our model building process. Before moving forward we need to make some valuable changes in dataset which is useful for the model building. Dataset look like this after cleaning.​





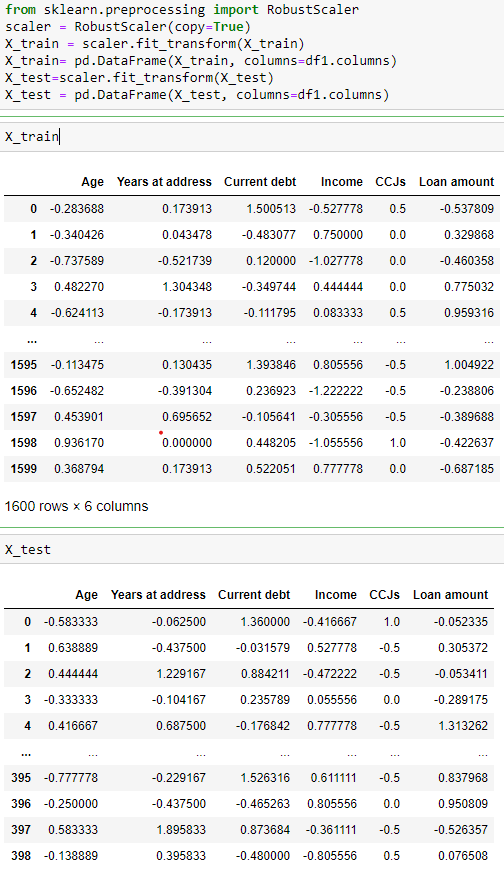


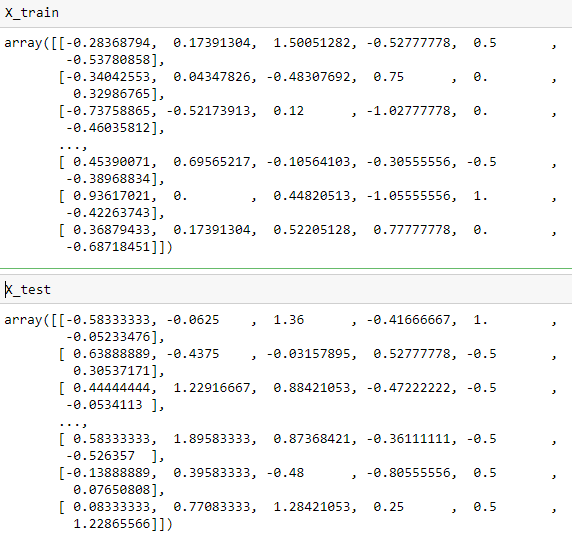
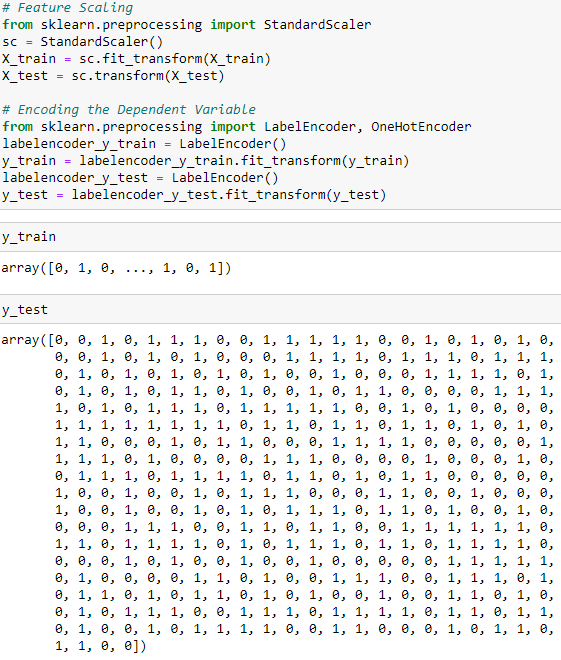


We see from the above graphs how data is dispersed but we don’t afford to lose the data so now we do scaling.

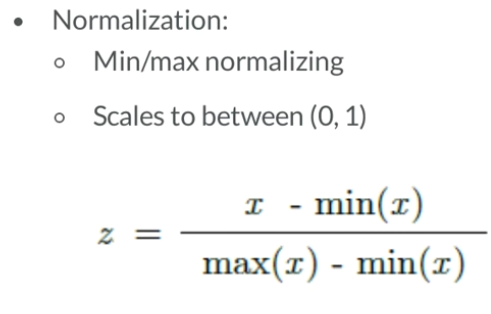
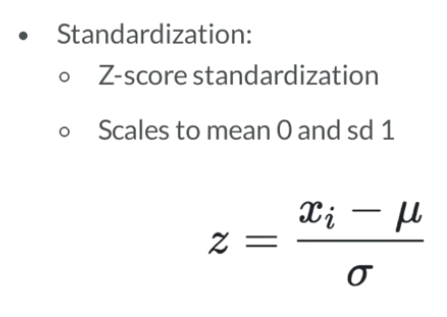
Scaling:

To avoid data leakage we scale X\_train and X\_test separately. In this use case we have small dataset(2000 entries) with outliers. We cannot lose the data so I am going to use Robust Scaling which scales our dataset in the inter quartile range (IQR) and ignore the effect of outliers. We need to remove this outlier for designing machine learning model it has a purpose to detect anonymous behavior of data. Before going further I also use Standard Scaling with Encoding the Dependent Variable which helps us to avoid data leakage as well.





While scaling we need to do balancing to make our data compatible to model building for this I use the following standardization and normalization technique which is already done by scaling functions.



**Modeling:**

For this stir bank problem we are going to use three modeling technique which is

Logistic Regression

Design Tree

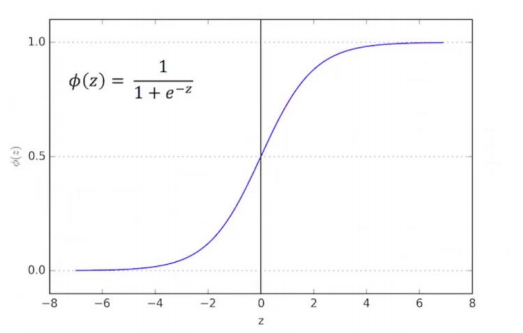
Neural Network

**Logistic regression:**

This is the one of the machine learning algorithms that comes under supervised learning. We use this algorithm to classify the problem which means in our problem statement we need to classify the customer who is able to pay the loan amount or who is not. This set of rules is relevant for specific structured variables by the use of a few given unbiased variables. The final results or end result have to be a specific or discrete price this means that output may be both Yes or No, zero or 1, True or False, etc. however these algorithms will now no longer offer the precise price as zero or 1, it offers a few probabilistic values that lie among zero and 1. This set of rules use logistic characteristic/sigmoid characteristic that is an S-Shaped curve.

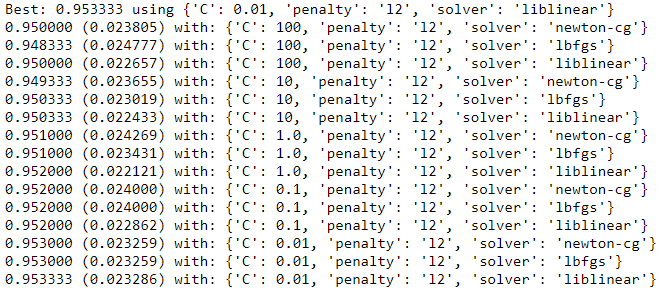
**Sigmoid function**:

G(z)= 1/(1+e^-z)

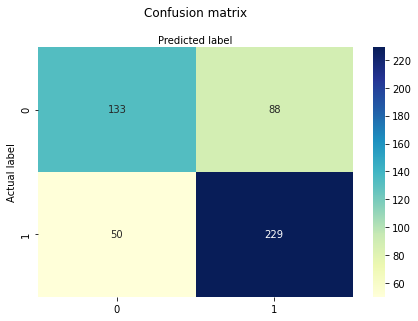


**Hyper parameter:**

For logistic regression I use GridSearchCv Strategies for hyperparameter tunning . The model is evaluates for the a range of hyperparameter values because it able to search the best set of hyperparameters from a grid of hyperparameters values.



**Confusion matrix:**



This means 133 and 229 are actual prediction although, 88 and 50 wrong prediction.

The following result I get is on test dataset

When applying this algorithm we get 72% classification rate which is considered as good accuracy

**Precision** means how accurate is our model is and we got 72% Precision.

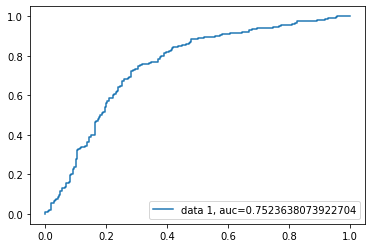
**Recall** if there are customers who are not going to pay the loan in the test set and our Logistic

Regression model can identify it 82% of the time.

Almost similar result that I get in train dataset also.

**Receiver Operating Characteristic(ROC):**

This is the plot shows the tradeoff between sensitivity and specificity with pos\_label=’1’.



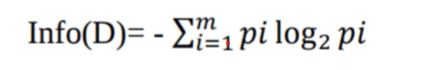
AUC score for our case is 0.75. AUC score 1 represent our perfect classifier although almost greater than 0.8 represent worthless classifier.

**Design Tree:**

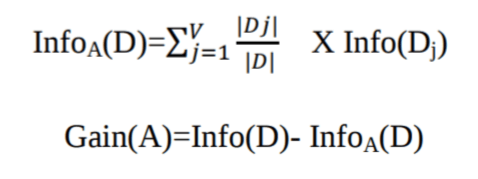
It is also the part of supervised algorithms in machine learning this algorithm first choose the fine characteristic the use of Attribute choice measures(ASM) to break up the information then make that characteristic a choice node and breaks the dataset into smaller subsets then begin doing this above matters repetitively for every toddler till one of the following circumstances will match.

* All tuples belongs to the same attribute value
* No remaining attribute available
* No more instance available

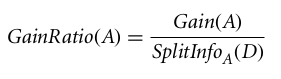
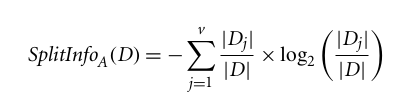
This algorithm use information gain.



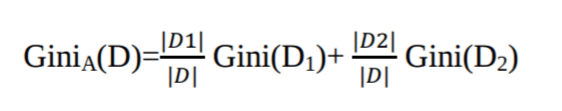
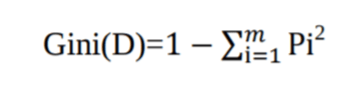
Pi = probability of arbitrary tuple in Ci class



**Gain ratio:**

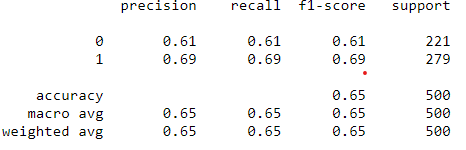
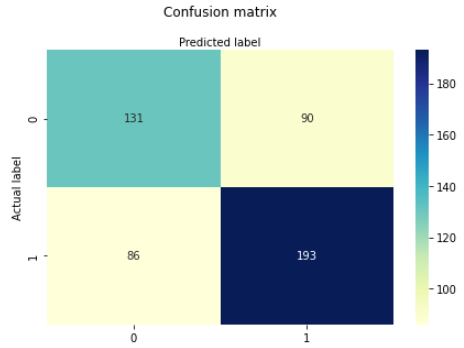


**Gini Index:**



C:\Users\ROHIT\Desktop\DA\%RS\Main DA bank\DT\gini.png

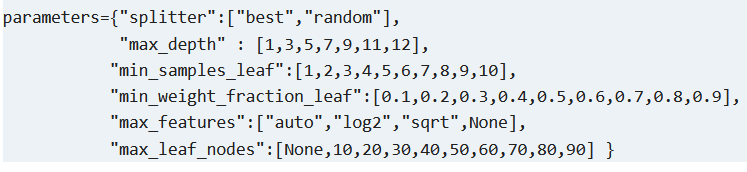
Confusion Matrix:

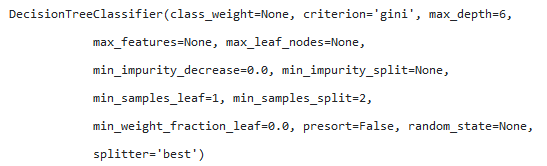


F1 score takes both false positive and false negative this is useful when you have uneven class distribution.

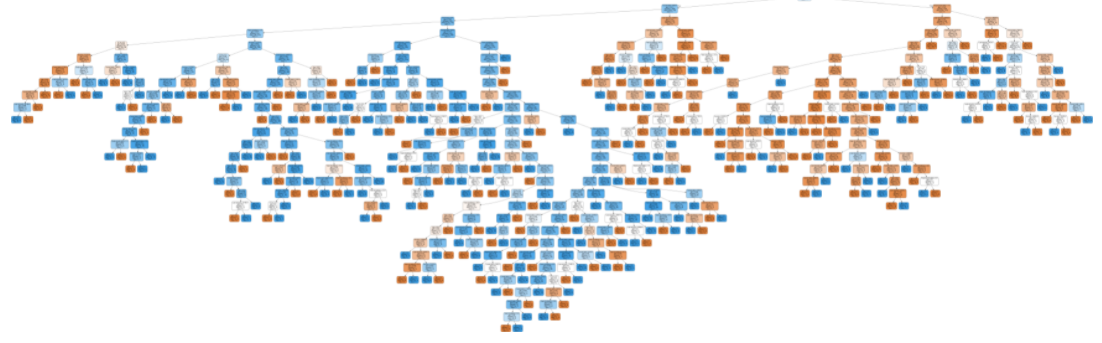
F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

Using all this we got 65% accuracy and it consider being a good accuracy. Hyper parameter that I use to boost the accuracy.





Whole tree diagram



Now we are going to optimize the Decision tree Performance

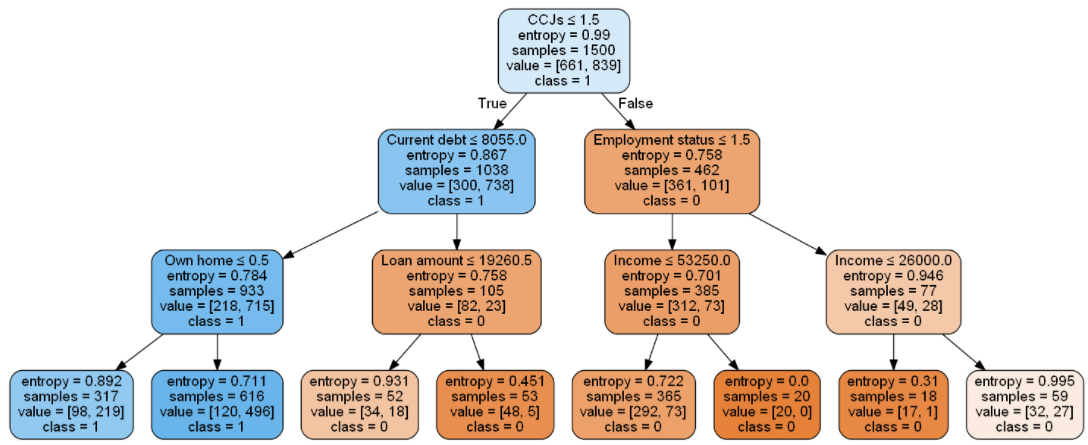
Choosing attribute selection measurement this is supported criteria are “gini” for the gini index and “Entropy” for the information gain.

Splitter or split Strategy this will help for best split and random to choose the best random split.

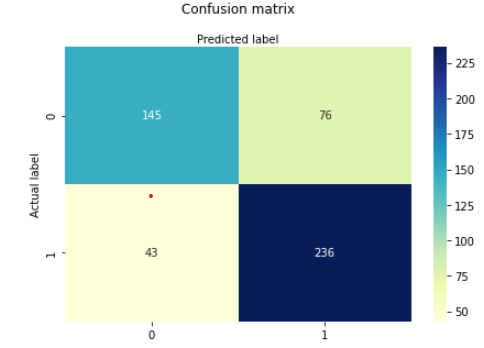
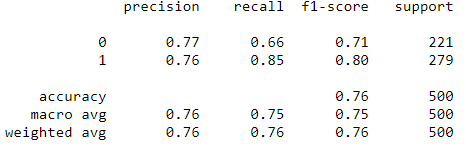
Maximum depth of a tree if it is none, then nodes get expanded until all the leaves and the higher value of the maxi-depth causes the over fitting although, lower value causes under fitting.

I am going to use max\_depth = 3 for same data.

Because of this my classification rate increases to 76.2%, which is better accuracy than our previous model.



After optimistic we get the confusion matrix and new Accuracy with great increment.



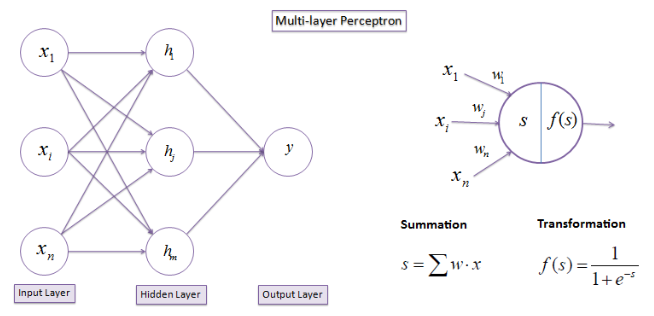
**Neural Network:**

A neural network (NN) is a community form comprising more than one related unit. It consists of 3 layers of units: enter layers, hidden layers, output layers. When the wide variety of hidden layers is extended or extra than or identical to 2 then the community is known as deep neural community. In our case I am going to use the deep neural network to increase the accuracy of our model although, I am going to use the normalization technique to normalize moreover; it will increase the performance of my model on test data set. For the development of this model I am going to use both tensor flow and keras.

For the first two layer we use Relu(Rectified linear unit) for this we need to use following functions:

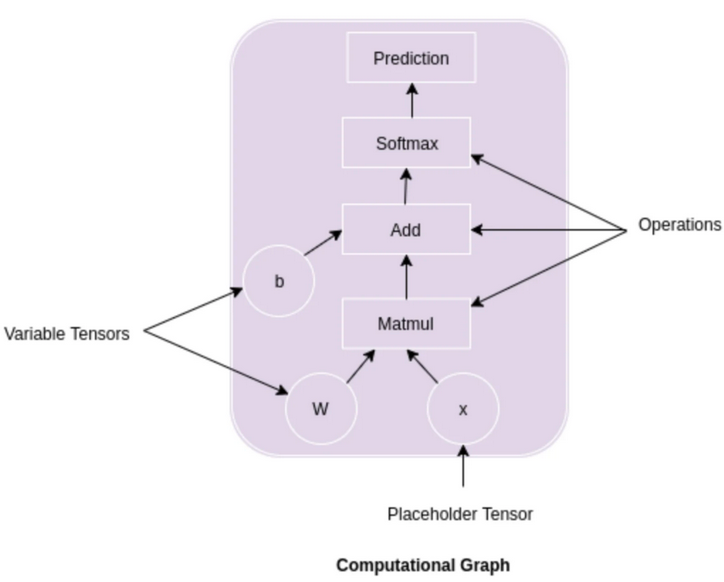
f(x) = 0 if x <=0

f(x) = 1 if x > 0

****

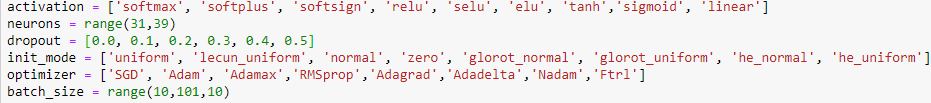
For making prediction in tensor flow

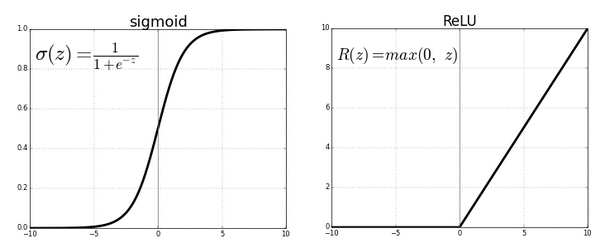
Prediction = tf.nn.softmax(tf.matmul(W,x) + b)



Hyper parameter that I use in this model:

Activation = 'relu',’elu’, last\_Activation = 'sigmoid’, dense= 12, 8,1, Alpha = 0.0001, batch\_size = 'auto', beta\_1 = 0.9, beta\_2 = 0.999, losses = [logcosh, binary\_crossentropy], early\_stopping = False, epsilon = 1e-08, hidden\_layer\_sizes = (100,), learning\_rate o = 'constant', learning\_rate\_init = 0.001, max\_fun = 15000, max\_iter = 5000, momentum = 0.9, n\_iter\_no\_change = 10, nesterovs\_momentum = True, power\_t = 0.5, random\_state = None, shuffle = True, solver = 'adam', tol = 0.0001, validation\_fraction = 0.1, verbose = False, warm\_start = False, input\_shape, loss, optimizer, epoch, metrics, fit()

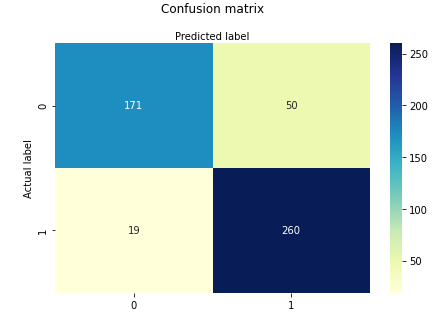




I get 84% accuracy on this model.

I use K-Fold cross-validation to select the best parameter values.

Confusion matrix:



Precision\_score: 83.87%

Recall\_score: 93.18%

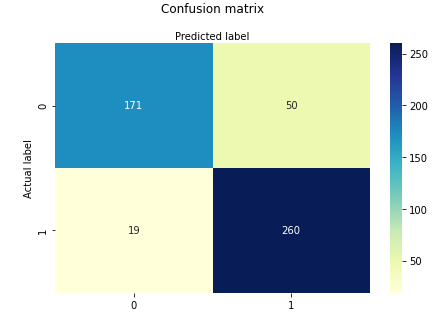
F1\_score: 88.28%

**Evaluation:**

**Test results:**

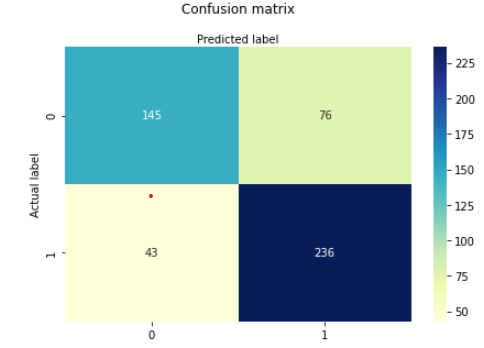
1. **Neural network.**

* **Confusion matrix:**



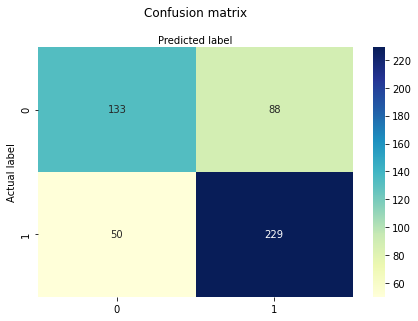
1. **Decision tree.**

* **Confusion matrix:**



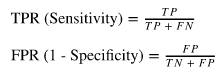
1. **Logistic regression.**

* **Confusion matrix:**



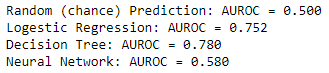
In this we are now working on ROC curve as well as their score it may varies from the original accuracy Moreover, Original Accuracy is based on specific cut point, while ROC tries all of the cut points and plots the sensitivity and specificity.

Receiver Operating Characteristic curve predict the performance of a classification model at all classification thresholds. In the Other word, ROC curve plot the False Positive Rate (FPR) on X-axis and True Positive Rate (TPR) on Y-axis.

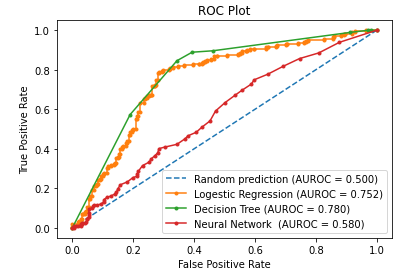


To performing this I am going to use some steps which as follows.

1. I will fit all the models in X-train and y-train.
2. Predict Probability and Probability for positive Outcome
3. Calculating AUROC and ROC curve values in that I get AUROC score like



1. Plot ROC curve



According to the ROC curve Decision tree performace Is the best on our stirbank dataset after this we see logistic regreesion model performance is good after that the lowest performance of our model is Neural network algorithms Moreover the area under the curve in Decision tree is larger as compaire to the other models.

After all this analysis the last step is to deploy our model.

This is the whole EDA using some python code

