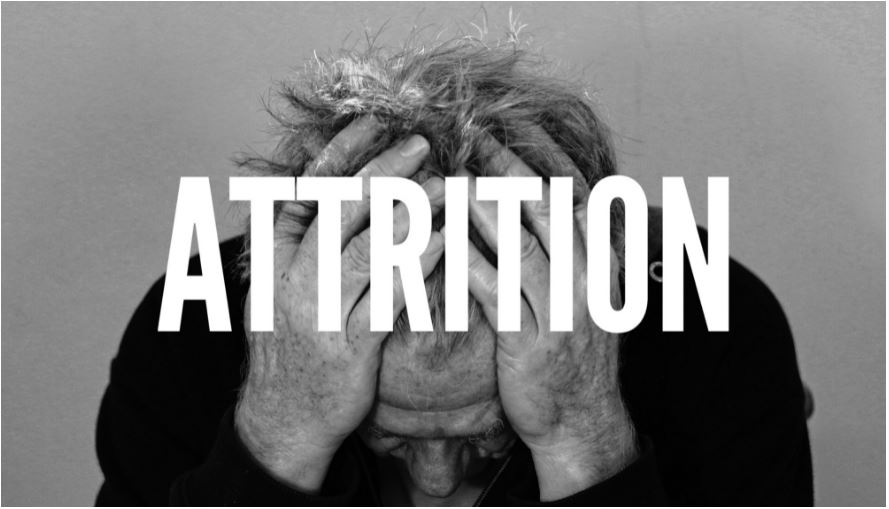
**Blog Submission**

**Project-1**

**HR Analytics Project**



*This is my first machine learning project. Recently I have completed this project. I am here to describe how i solved the case study in a very detailed manner.*

**Submitted by: Sachin Gupta**

**Acknowledgement**

*Many thanks to Data Trained for providing me this project to understand about the Real Time Field work present in Data Science Industry.*

*As this project of Churn Prediction was very sufficient and preferable for me to acknowledge about the live customer churned out due to various use of Social Networking Platforms in ultimate users which predict the telecom industries.*

**Contents**

1. [**Introduction**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#c099)
2. **Work Steps**
3. [**EDA**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#030a)

1. **[Building Machine Learning Models](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d" \l "8341)**
2. **[Conclusion](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d" \l "13e2)**

**1.Introduction**

*Employee attrition is defined as the natural process by which employees leave the workforce – for example, through resignation for personal reasons or retirement – and are not immediately replaced.Some forms of attrition are unavoidable, like if an employee is retiring or is moving to another city. But after a certain threshold, attrition can make a big dent in your company’s bottom line as well as its culture. In this primer, we tell you all you need to know about employee attrition, and how to measure it accurately. Attrition is an inevitable part of any business. There will come a time when an employee wants to leave your company – for either personal or professional reasons. But when attrition crosses a particular threshold, it becomes a cause for concern. For example, attrition among minority employee groups could be hurting diversity at your organization. Or, attrition among senior leaders can lead to a significant gap in organizational leadership.Employee attrition occurs when the size of your workforce diminishes over time due to unavoidable factors such as employee resignation for personal or professional reasons. Employees are leaving the workforce faster than they are hired, and it is often outside the employer’s control. For example, let’s say that you have opened a new office designated as the Sales Hub for your company. Every salesperson must work out of this office – but a few employees cannot relocate and choose to leave the company. This is a typical reason for employee attrition. But there are other reasons for attrition as well, including the lack of professional growth, a hostile work environment, or declining confidence in the company’s market value. Weak leadership is another factor that often drives attrition among employees.*

*In this article, I would like to present how to predict employee attrition with machine learning. For analysis I will use a data set created by IBM data scientists, which is available here. However, I will do a split into train and test samples to better explain you how machine learning methods can be applied to this problem. The splitted data is available at my github. The train set represents historical data about employees. In this data each sample (row) describes the employee with parameters like: age, department, distance from home, marital status, income, years at company. You can check all used descriptors here. For each employee in the train set the attrition is known (it is historical value). In test data we have employees descriptors available, however the attrition is unknown and we want to predict (compute) it with our machine learning model. (To be honest, the attrition values in test data are available, but for better explanation let’s assume that it is missing).*

**2.Work Steps**

**2. Work Steps What Methods I wiil going to solve ?**

**Attrition Prediction Methodology:**

***Loading The Dataset***

***Feature Selection Method Using Info Gain Of Features.***

***Train Model on Training Dataset***

***Test the model on Testing Dataset***

***Result Analysis***

**MACHINE LEARNING METHODS:**

**Six machine learning classification models have been used for prediction of Attrition Prediction:**

***Decision Trees***

***KNN – K neighbour’s Classifier***

***AdaBoostClassifier***[***¶***](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/HR%20Analytics%20Project.ipynb#AdaBoostClassifier)

***RandomForestClassifier***

***Logistic Regression***

***XGBClassifier.***

**3. Details about the dataset**

**Employee is one of the most important resource in company, where a high attrition rate indicates that the company is unable to maintain their employees. In a short term, with high attrition rate, company must pay a great money to cover the cost of turnover. While in a long term, this will affect the company's performance as employees come and go the company's performance will decline.**

**Description about the data**

**Education 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'**

**EnvironmentSatisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'**

**JobInvolvement 1 'Low' 2 'Medium' 3 'High' 4 'Very High'**

**JobSatisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'**

**PerformanceRating 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'**

**RelationshipSatisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'**

**WorkLifeBalance 1 'Bad' 2 'Good' 3 'Better' 4 'Best'**

**Additinal Description Which will help to understand the features**

**Hourly Rate: An hourly employee is paid a wage for each hour worked, unlike a salaried employee.**

**EmployeeNumber: An Employee Number is a unique number that has been assigned to each current and former State employee and elected official in the Position and Personnel DataBase (PPDB).**

**JobLevel: Job levels, also known as job grades and classifications, set the responsibility level and expectations of roles at your organization. They may be further defined by impact, seniority, knowledge, skills, or job title, and are often associated with a pay band. The way you structure your job levels should be dictated by the needs of your unique organization and teams.**

**StockOptionLevel: Employee stock options, also known as ESOs, are stock options in the company’s stock granted by an employer to certain employees. Typically they are granted to those in management or officer-level positions. Stock options give the employee the right to buy a certain amount of stock at a specific price, during a specific period of time. Options typically have expiration dates as well, by which the options must have been exercised, otherwise they will become worthless.**[**¶**](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/HR%20Analytics%20Project.ipynb#StockOptionLevel:-Employee-stock-options,-also-known-as-ESOs,-are-stock-options-in-the-company%E2%80%99s-stock-granted-by-an-employer-to-certain-employees.-Typically-they-are-granted-to-those-in-management-or-officer-level-positions.-Stock-options-give-the-employee-the-right-to-buy-a-certain-amount-of-stock-at-a-specific-price,-during-a-specific-period-of-time.-Options-typically-have-expiration-dates-as-well,-by-which-the-options-must-have-been-exercised,-otherwise-they-will-become-worthless.)

**4.Importing All Necessary libraries**

#### **By Importing above All Necessary Libraries I got the idea of implementing all the Libraries in one go for better model Prediction in this dataset.**

**5.Reading the Csv file**



**Here I am loading the data set into a variable i.e. “df” and processing the first 5 rows.¶**

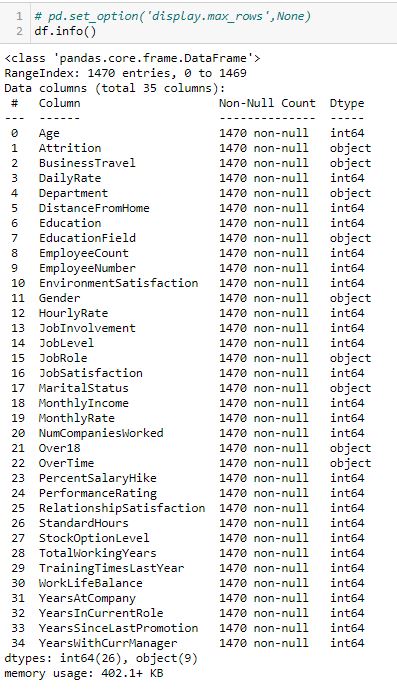
**After Reading the csv file I find out that Attrition are label columns while remaining are the Features columns.**

#### F:\DATA TRAINED\RESUBMIT FILES\Blog\1\4.JPG **By using shape I got 614 rows and 13 columns.**

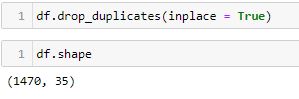
#### 6.EDA - Exploratory Data Analysis **In statistics, exploratory data analysis is an approach of analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.**

**EDA which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed.**

# 



**Also, most of the column are int64 and few columns are object data type**



***From the above code I can analysed there is no duplicates in this dataset.***

##### F:\DATA TRAINED\RESUBMIT FILES\Blog\1\7.JPG ***By finding (df. Unique()) it distinguish the difference between continuous & categorical columns in this dataset.***[***¶***](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/HR%20Analytics%20Project.ipynb#By-finding-(df.unique())-it-distinguish-the-difference-between-continious-&-categorical-columns-in-this-dataset.)

##### F:\DATA TRAINED\RESUBMIT FILES\Blog\1\8.JPG ***By the above code it gives detail information about all the columns included in this dataset.***

#### F:\DATA TRAINED\RESUBMIT FILES\Blog\1\9.JPG **Above statistics data show that their multiple outliers.There is also difference between mean and 50% value in some of the columns which used to get fix for better prediction. Also, number of rows in each column are same, means there are no null values in the data set Also, the mean and 50%value of most of the column are same and the STD and mean are having the huge difference. Most of the column statistics data are near to 0 values that need to be replaced by mean and mode. By checking the difference between the 75% and max value there are outliers in some of the column, I will check it soon.**

**Removing Useless Features**

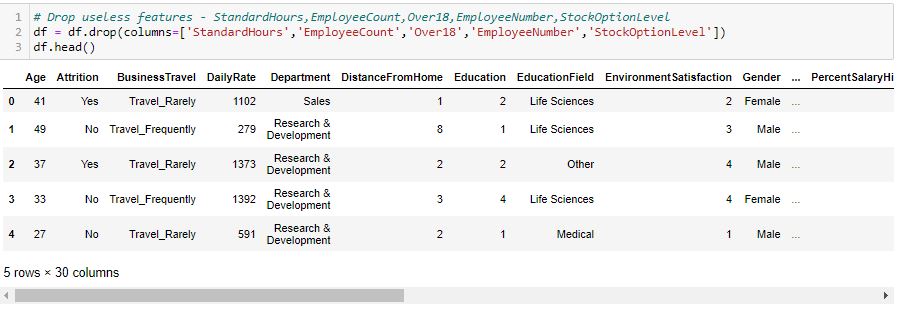
***1.StandardHours***

***2.EmployeeCount***

***3.Over18***

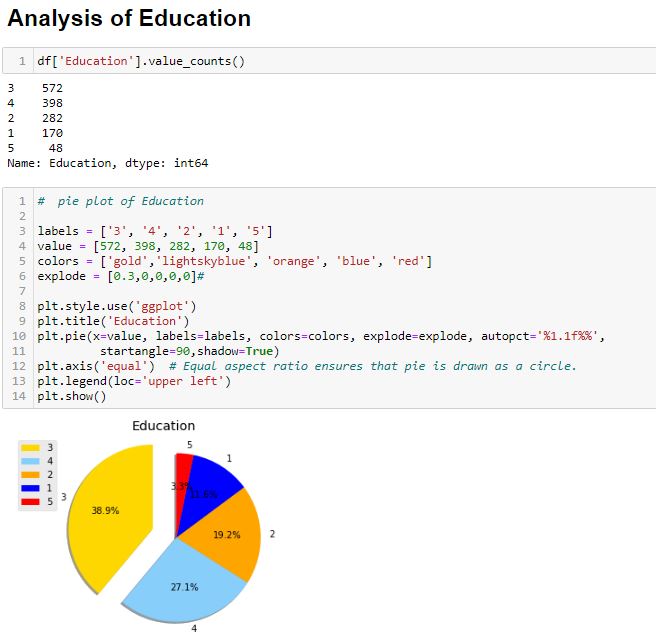
***4.EmployeeNumber[¶](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/HR%20Analytics%20Project.ipynb" \l "4.EmployeeNumber)***

***5.StockOptionLevel***

****

# 7.Data visualization

##### **F:\DATA TRAINED\RESUBMIT FILES\Blog\1\11.JPG F:\DATA TRAINED\RESUBMIT FILES\Blog\1\12.JPG *From the above plots Male employees are more likely to leave the company than female employees.***

****

### **F:\DATA TRAINED\RESUBMIT FILES\Blog\1\19.JPG From the above plots I can analyze five Education in the company.**

***Employees who are at 3 education level have the most attrition number which makes up the 38.9% of attrition rate in the company.***

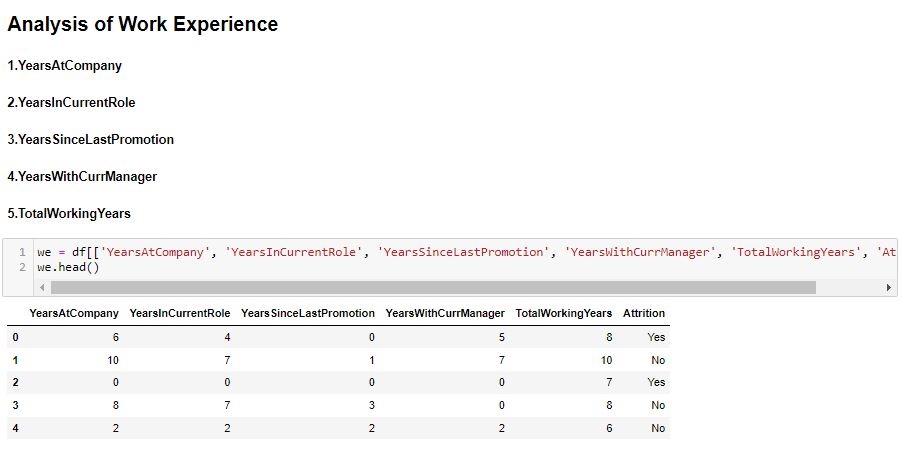
***Employees who have the master, college, and below college degrees are follower of employees who have bachelor degrees in terms of the attrition number in the company respectively.***

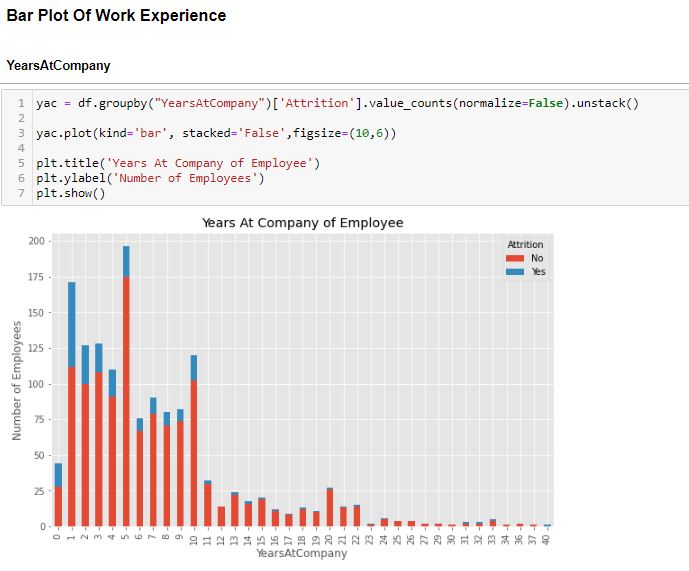
****

**From the above Box & Pie plots I can analyze five EducationFeild in the company.¶**

***Employees who have Life Science education level have the most attrition number which makes up the 41.22% of attrition rate in the company.***

***Handling Continuous Column***

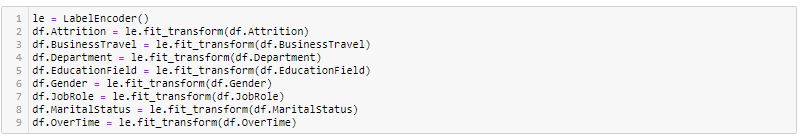
****

****

***It is observed that fresher’s leave the company very likely so it's important that company creates a new policy to handle fresher’s so they don't leave the company from the start***

**Label Encoder**

**Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.**

****

#### From the Above statistics data show that their multiple outliers mostly in many columns.

#### There is also difference between mean and 50% value in some of the columns which used to get fix for better prediction.

#### a) Also, number of rows in each column are not same, means there is no null values present in the data set.

#### b) Also, the mean and 50%value of most of the column are same and the STD and mean have a major difference to each other.

#### c) Most of the column statistics data are near to 0 values.

#### d) By checking the difference between the 75% and max value there are outliers in some of the column, I will check it soon.

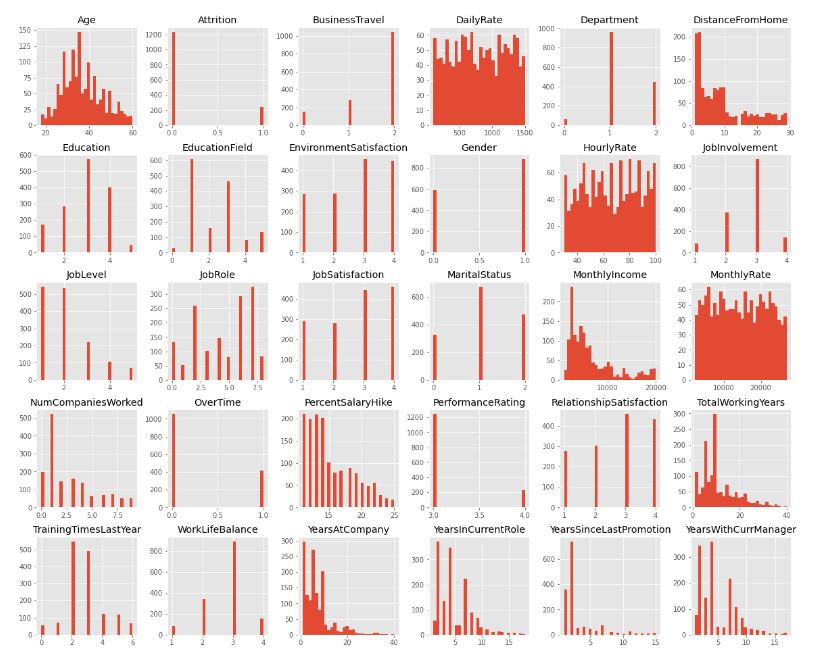
****

***After understanding the Describe I got that in this dataset there are zero values which I have replace with mean because all the zero value column are continuous in nature.***

**Histogram Plot**

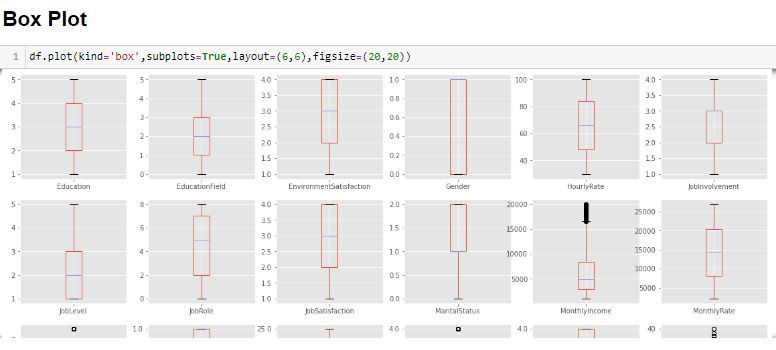
**Histograms group the data in bins and is the fastest way to get idea about the distribution of each attribute in dataset.The following are some of the characteristics of histograms:**

**a)provides us a count of the number of observations in each bin created for visualization. ¶b)From the shape of the bin, I can easily observe the distribution i.e. Attrition it is Gaussian, skewed or exponential. ¶Histograms also help us to see possible outliers.**

****

**By making Histogram Plot I can see there are some skewness in this dataset.**

**From plotting this histogram, I used the bin size as 30, we can take any bin size (suited as per as data).**

****

#### From above image I can clear see that there are number of black dots in most of the column which are referring to the outliers, so it means most of the data are outside the distribution.

#### So now I detect the outliers now the second step is to remove the outliers, there are different way to remove the outliers that are zscore values.

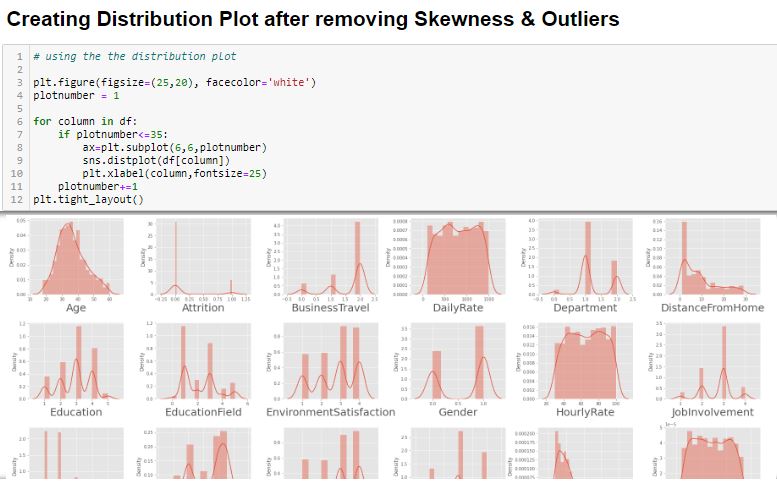
# Removing Outliers



### After using zscore value then I again check if there are some of the outliers then I will remove it by replacing the outliers with the mean value of that column.

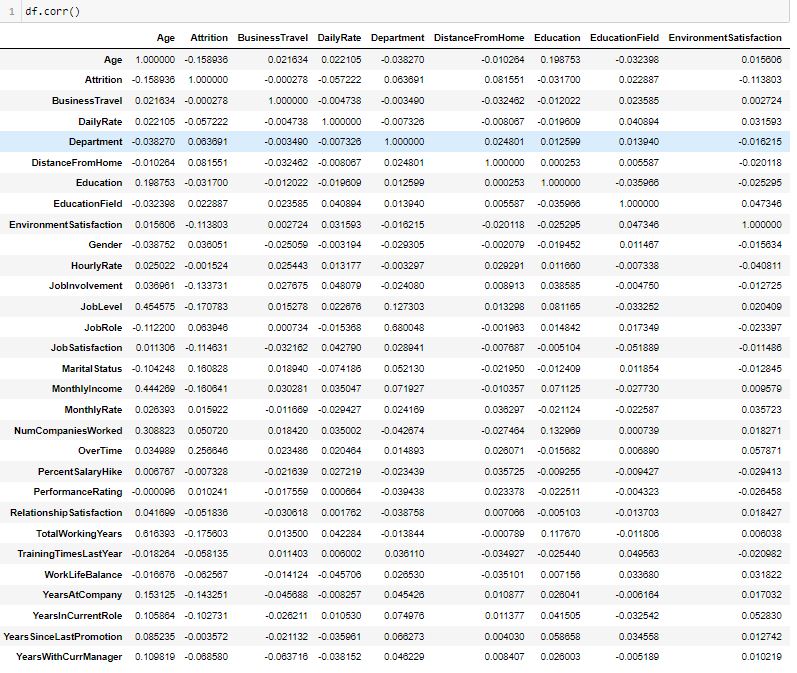
### So, I first find the zscore value and then I decide to make one threshold value as 3 which is standard of industry recommend value and then I remove all the outliers which zscore value is greater than 3.

### After, removing the outlier’s final there are 1383 and 30 column presents in the data set.



#### Hence After Removing Outliers I make distribution plot which shows few outliers are removed.

#### F:\DATA TRAINED\RESUBMIT FILES\Blog\1\77.JPG **I have checked the relationship between label and different features.Most of the features are in minus value only 12 columns are positively related to target value.**

****

***Correlation Matrix is basically a covariance matrix. A summary measure called the correlation describes the strength of the linear association. Correlation summarizes the strength and direction of the linear (straight-line) association between two quantitative variables.¶***

Denoted by r, it takes values between -1 and +1.

# Heatmap

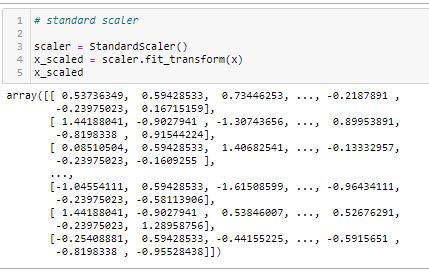
# F:\DATA TRAINED\RESUBMIT FILES\Blog\1\79.JPG

# 8.Machine Learning

# F:\DATA TRAINED\RESUBMIT FILES\Blog\1\80.JPG

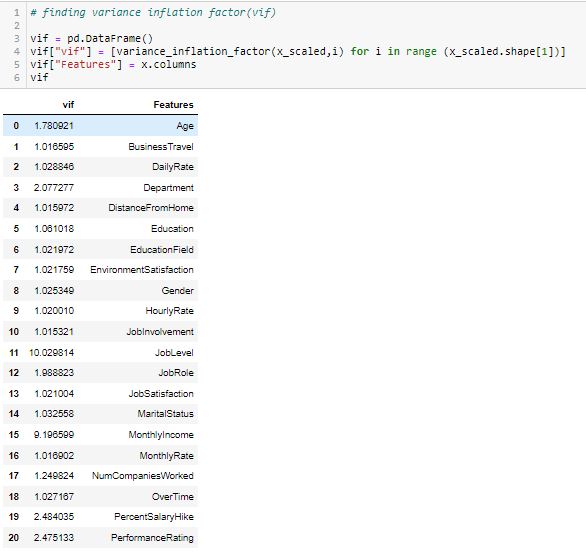
#### Separate response variable and features in x & y.

# StandardScaler

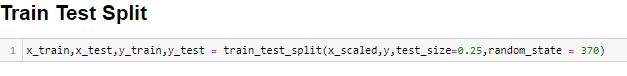


#### Applying the standard scaler to scaled the dataset in one range & on x variable.[¶](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/HR%20Analytics%20Project.ipynb#Applying-the-standard-scaler-to-scaled-the-dataset-in-one-range-&-on-x-variable.)

# VIF - variance inflation factor



**According to my concern about vif in this dataset all the values are less than 10 and overall model variance there is no multicollinearity exists.**



**Splitting the dataset into the Training set and Test set.**

**The data should be divided into train and test data.**

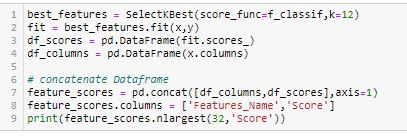
**I will use train\_test\_split feature to divide the data and I will be using testing set of 25% and a training set of 75% of the data.**

**I can set the random state of the split to ensure consistent results.**

**KNN**

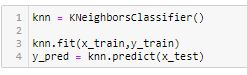
**The abbreviation KNN stands for “K-Nearest Neighbour”. It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.**

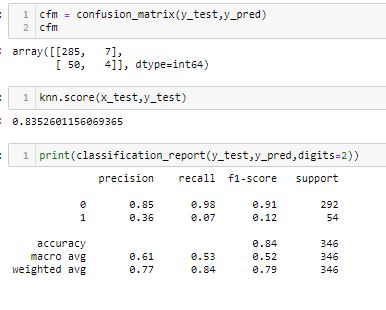
**Using SelectKBest Method for best features selection**

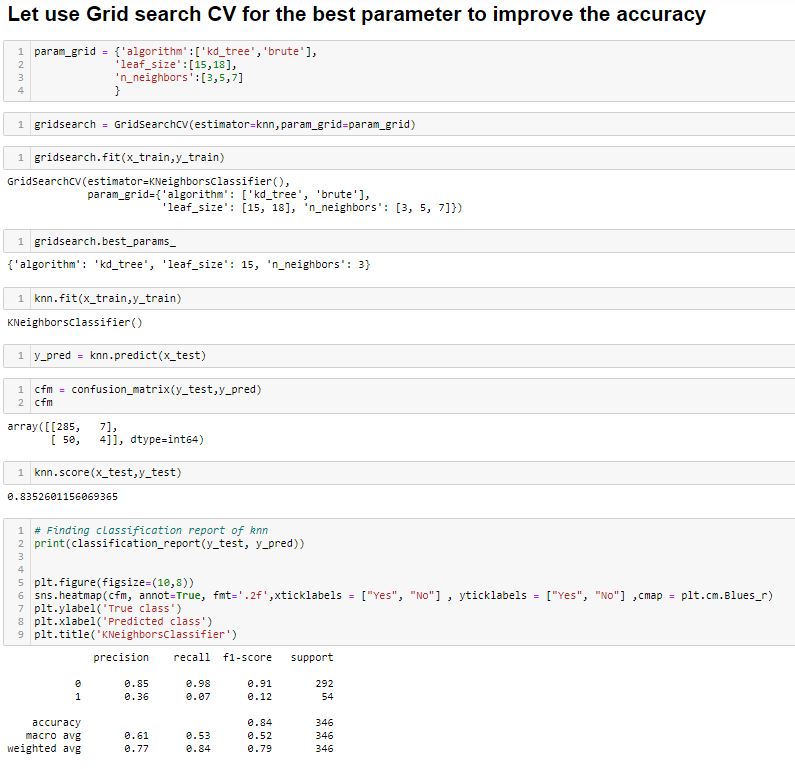


# KNN model Building

#### k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically







#### After modelling KNeighborsClassifier then my next step is to find Roc Score which will help me to make Roc\_Auc Curve.

#### Below is the Roc Score:

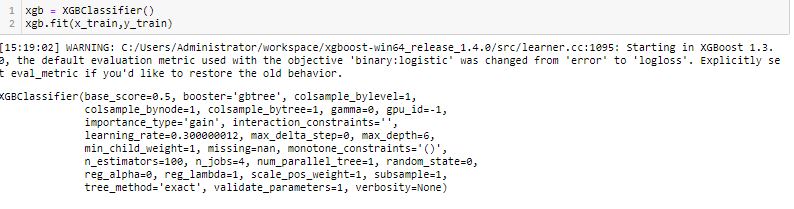
 ***By using Grid Search CV I improve the Accuracy Score : 52.50%***

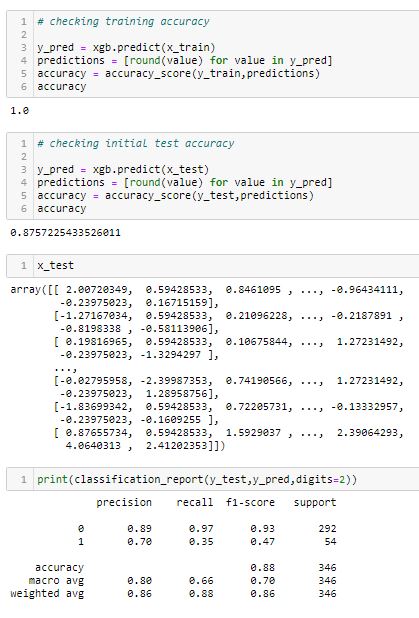
**In prediction KNeighborsClassifier with Attrition: I had done this prediction by taking Attrition as an output variable which is Classification in nature.While calculating the best random state the 0 is best state which providing the highest accuracy value for this model.After calculating confusion matrix and classification report of KNeighborsClassifier - 84%.Therefore i use hyper tuning method by using Grid Search Cv to improve my Accuracy Score and this gives a no increasing value of KNeighborsClassifier - 84%**

**Then I found roc\_auc\_score of KNeighborsClassifier which is 52.50%**

**XGBoost Model**

***XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. XGBoost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.***



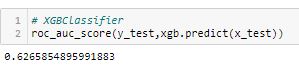


# Let use Grid search CV for the best parameter to improve the accuracy

# F:\DATA TRAINED\RESUBMIT FILES\Blog\1\91.JPG

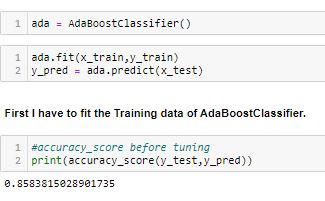
# F:\DATA TRAINED\RESUBMIT FILES\Blog\1\92.JPG

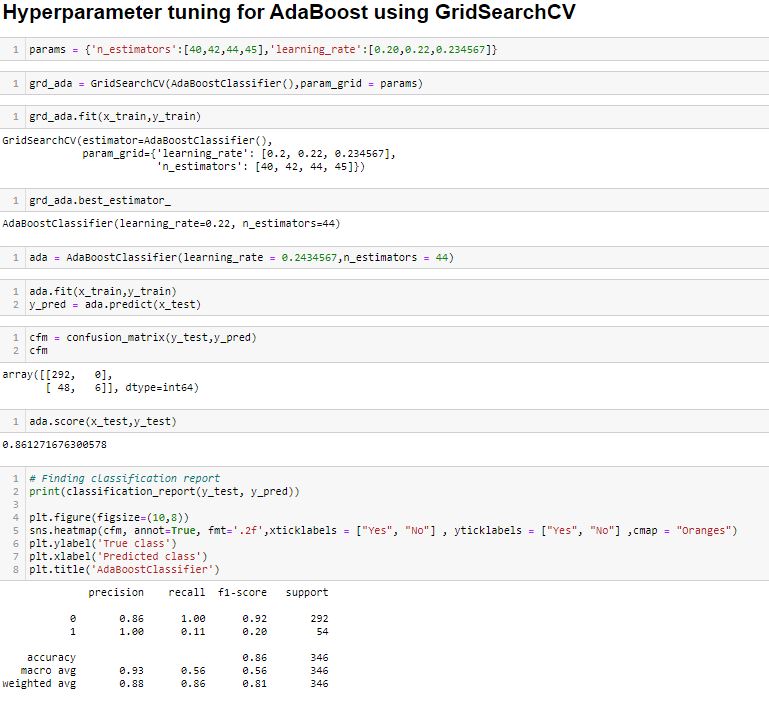
**By using Grid Search CV I improve the Accuracy Score : 84%After modelling XGBClassifier then my next step is to find Roc Score which will help me to make Roc\_Auc Curve .**Below is the Roc Score:

**In prediction XGBClassifier with Attrition: I had done this prediction by taking Attrition as an output variable which is Classification in nature.While calculating the best random state the 370 is best state which providing the highest accuracy value for this model.After calculating confusion matrix and classification report of XGBClassifier - 88%.Therefore i use hyper tuning method by using Grid Search Cv to improve my Accuracy Score and this gives a decreasing value of XGBClassifier - 84%¶Then I found roc\_auc\_score of XGBClassifier which is 62.65%.**[**¶**](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/HR%20Analytics%20Project.ipynb#Then-I-found-roc_auc_score-of-XGBClassifier-which-is-62.65%.)

**AdaBoostClassifier**

***AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem. The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level.***[***¶***](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/HR%20Analytics%20Project.ipynb#AdaBoost-can-be-used-to-boost-the-performance-of-any-machine-learning-algorithm.-It-is-best-used-with-weak-learners.-These-are-models-that-achieve-accuracy-just-above-random-chance-on-a-classification-problem.-The-most-suited-and-therefore-most-common-algorithm-used-with-AdaBoost-are-decision-trees-with-one-level.)

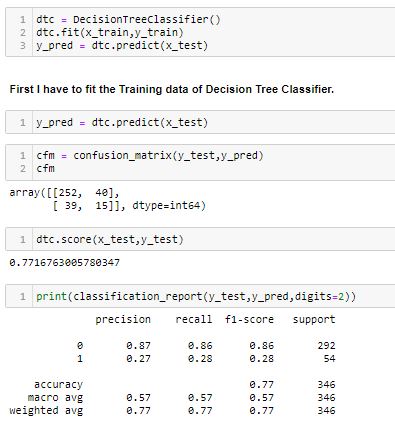


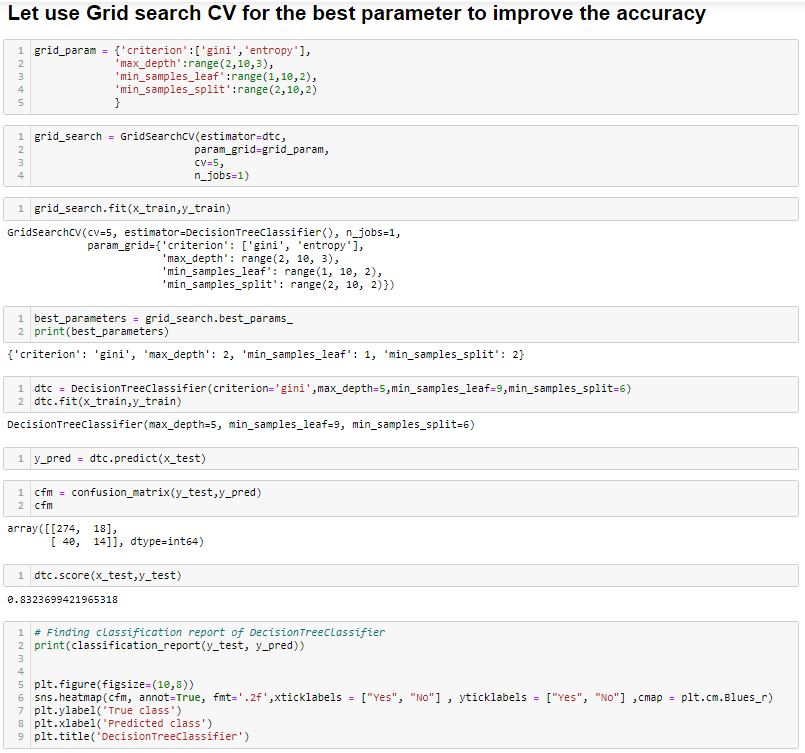


### F:\DATA TRAINED\RESUBMIT FILES\Blog\1\96.JPG In prediction AdaBoostClassifier with Attrition:*I had done this prediction by taking Attrition as an output variable which is Classification in nature.While calculating the best random state the 3700 is best state which providing the highest accuracy value for this model.After calculating confusion matrix and classification report of AdaBoostClassifier - 85.83%.*[*¶*](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/HR%20Analytics%20Project.ipynb#After-calculating-confusion-matrix-and-classification-report-of-AdaBoostClassifier---85.83%.)*Therefore i use hyper tuning method by using Grid Search Cv to improve my Accuracy Score and this gives a increasing value of AdaBoostClassifier - 86%Then I found roc\_auc\_score of AdaBoostClassifier which is 55.55%.*

**Decision Tree Classifier**

**Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.The tree's branches contain the logic for a decision rule, meaning your data is continually split given the input features. The decision tree classifier is commonly used for image classification, decision analysis, strategy analysis, in medicine for diagnosis, in psychology for behavioural thinking analysis, and more.**

**¶** ****





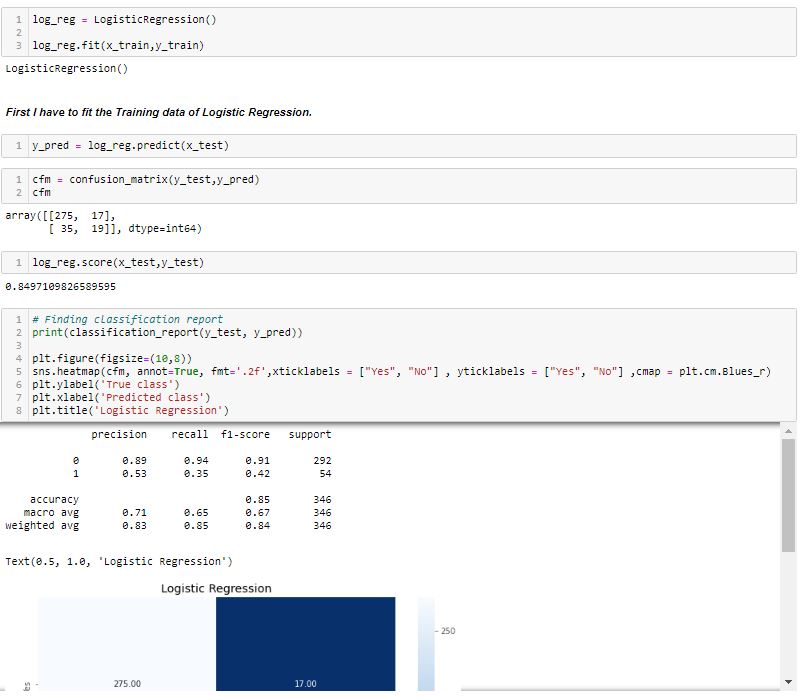
**In prediction Decision Tree Classifier with Attrition:*I had done this prediction by taking Attrition as an output variable which is Classification in nature.While calculating the best random state the 0 is best state which providing the highest accuracy value for this model.***

***After calculating confusion matrix and classification report of Decision Tree Classifier - 77%.Therefore i use hyper tuning method by using Grid Search Cv to improve my Accuracy Score and this gives a increasing value of Decision Tree Classifier - 83%***

***Then I found roc\_auc\_score of RandomForestClassifier which is 59.88%.***

**Logistic Regression**

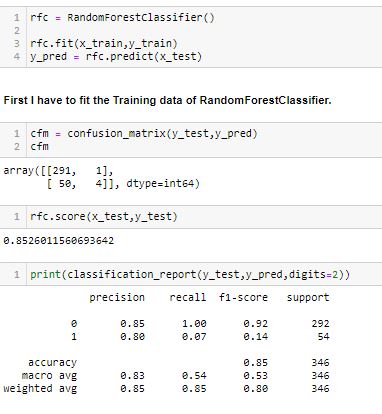
**Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regressionis estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variablesthe independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable.The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.**

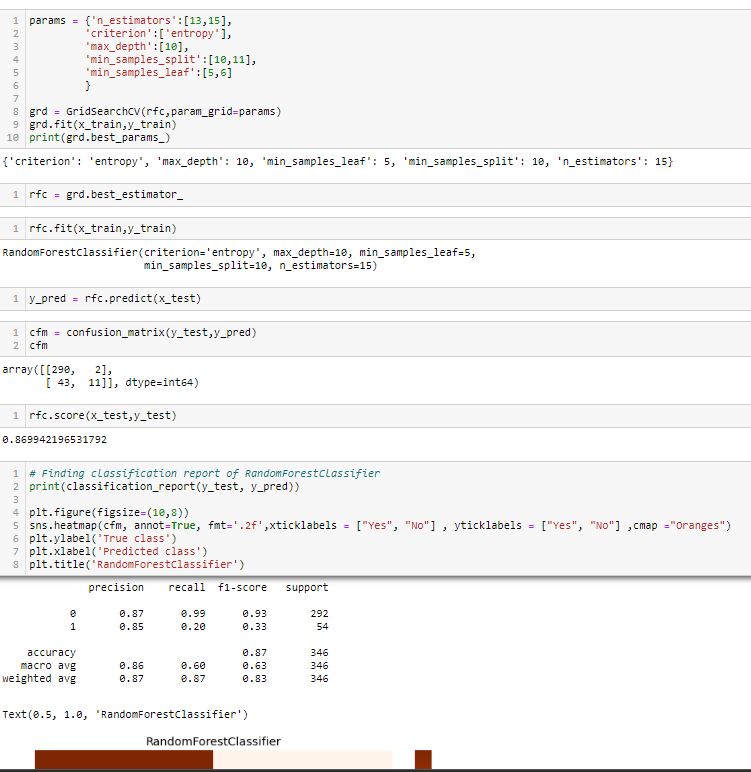


#### F:\DATA TRAINED\RESUBMIT FILES\Blog\1\101.JPG In prediction Logistic Regression with Attrition had done this prediction by taking Attritions as an output variable which is Classification in nature.While calculating the best random state the 0 is best state which providing the highest accuracy value for this model.After calculating confusion matrix and classification report I used Heat Map of Logistic Regression - 85%Then I found roc\_auc\_score of Logistic Regression which is 64.68%

# RandomForestClassifier

#### A random forest classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.It works in four steps:Select random samples from a given dataset.Construct a decision tree for each sample and get a prediction result from each decision tree.Perform a vote for each predicted result. Select the prediction result with the most votes as the final prediction.





#### F:\DATA TRAINED\RESUBMIT FILES\Blog\1\104.JPG In prediction RandomForestClassifier with Attrition:I had done this prediction by taking Attrition as an output variable which is Classification in nature.While calculating the best random state the 0 is best state which providing the highest accuracy value for this model.After calculating confusion matrix and classification report of RandomForestClassifier - 81%.Therefore i use hyper tuning method by using Grid Search Cv to improve my Accuracy Score and this gives a increasing value of RandomForestClassifier - 87%Then I found roc\_auc\_score of RandomForestClassifier which is 59.84%.

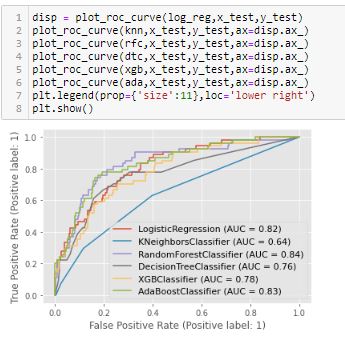
# Roc Curve

### In Machine Learning, performance measurement is an essential task.

### So when it comes to a classification problem, I can count on an AUC - ROC Curve.

### When I need to check or visualize the performance of the multi-class classification problem, I use the AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve.

### It is one of the most important evaluation metrics for checking any classification model’s performance. It is also written as AUROC (Area Under the Receiver Operating Characteristics).



# Saving the Best Model

#### F:\DATA TRAINED\RESUBMIT FILES\Blog\1\106.JPG **From the Above Coding I Get best model Random Forest Model Accuracy Score. So I use Random Forest Model to Save the Best Model.**[**¶**](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/HR%20Analytics%20Project.ipynb#So-I-use-Random-Forest-Model-to-Save-the-Best-Model.)

# Conclusion

**The conclusion from the above-mentioned analysis would be the following:**

**The person which is most likely to leave is under 30, single and works overtime. Stock option level also determines the likelihood of employee attrition, as the employee with no option to purchase company stocks has probably lower interest in its overall success than those who can. Money really plays a big role here as it can be seen, that the income of an average employee is almost double that of an average leaver. The most common job roles of leavers — Laboratory Technician, Sales Executive and Sales Representative — are ones with the lowest job level. It can be clearly seen the feature engineering effects on each individual nine model in 'Comparison of ML Classifiers with Effects of Future Engineering' Roc Curve chart above. I’ll briefly discuss how I approached this problem of predicting Attrition prediction.**

**a) I have Attrition Prediction dataset from which I had to extract information.**

**b ) I had used pandas library to read the Dataset which provide me to explore & visualize the Data properly based on Rows & Columns.**

**c) After from all datasets, I pre-processed the data using imputer function to fill the missing values and replacing all zeros.**

**d) I did exploratory data analysis on main data frame and tried to remove all the possible outliers.**

**e) Based on visualization knowledge, I use various EDA TECHNIQUES to plot the graphs and Z - SCORE.**

**f) I use VIF and Standard Scalar to scaled the data and by VIF I make a relationship between Features & Features.¶**

**g) After from all these i split the Features & Labels into 2 parts.**

**h) On this data, I have applied our machine learning models such as logistic regression, Random forest, Decision Tree,ADABoost Classifier, XGBoost Classifier and KNN.In the modeling Building section, we have seen the hyper parameter tuned Random Forest has the highest Best Score (0.87).**

**It can be clearly seen the feature engineering effects on each individual nine model in 'Comparison of ML Classifiers with Effects of Future Engineering' Roc Curve chart above.**[**¶**](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/HR%20Analytics%20Project.ipynb#It-can-be-clearly-seen-the-feature-engineering-effects-on-each-individual-nine-model-in-'Comparison-of-ML-Classifiers-with-Effects-of-Future-Engineering'-Roc-Curve-chart-above.)

***GitHub Link:***

<https://github.com/sachingupta2015s/NEW-EVALUATION-PROJECT/blob/main/HR%20Analytics%20Project.ipynb>

**Thank you for reading!**

*If you enjoyed this article, check out my other articles on Data Science, Math and Programming. Follow me on GITHUB for the latest updates.*

*Thank you again for reading!*

--------------------------------------------------------xxxxxxxxxx---------------------------------------------------------------

**Blog Submission**

**Project - 2**

# Loan Application Status Prediction

*This is my Second machine learning project. Recently I have completed this project. I am here to describe how i solved the case study in a very detailed manner.*

**Submitted by: Sachin Gupta**

**Acknowledgement**

*Many thanks to Data Trained for Providing me this project to understand about the Real Time Field work present in Data Science Industry.*

*As this project of Loan prediction was very sufficient and preferable for me to acknowledge about the live customers who are eligible for getting Loan or not.*

**Problem Statement**

*This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.*

**Independent Variables:**

**- Loan\_ID**

**- Gender**

**- Married**

**- Dependents**

**- Education**

**- Self\_Employed**

**- ApplicantIncome**

**- CoapplicantIncome**

**- Loan\_Amount**

**- Loan\_Amount\_Term**

**- Credit History**

**- Property\_Area**

**Dependent Variable (Target Variable):**

**- Loan\_Status**

*I have to build a model that can predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset.*

**TYPE OF PROBLEM**

*The above problem is a clear classification problem as we need to classify whether the Loan\_Status is yes or no. So this can be solved by any of the classification techniques like:*

1. ***Logistic Regression***
2. ***Decision Tree Algorithm***
3. ***Random Forest Technique***
4. ***Decision Tree***

*I have mentioned only few. I will be dealing with each of techniques later in this blog.*

**Introduction**



*A loan is when you receive money from a friend, bank or financial institution in exchange for future repayment of the principal, plus interest. The principal is the amount you borrowed, and the interest is the amount charged for receiving the loan. Since lenders are taking a risk that you may not repay the loan, they have to offset that risk by charging a fee known as interest.*

*Financial companies have been using predictive analytics for quite a long time.*

*Recently, due to the availability of computational resources and tremendous research in machine learning made it possible to better data analysis hence better prediction. In the series of articles, I explain how to create a predictive loan model that identifies a bad applicant who is more likely to be charged off. In a step by step processes, I show how to process raw data, clean unnecessary part of it, select relevant features, perform exploratory data analysis, and finally build a model****.***

**Label:**

*In this project my goal is to Predict the loan Status and has to estimate the risk involved before granting a loan to a customer.*

**Description about the Data Columns**

*There are altogether 13 columns in our data set. Of them Loan Status is the response variable and rest all are the variables /factors that decide the approval of the loan or not. Now let us look in to the each variable and can make some assumptions.*

*Loan ID -> As the name suggests each person should have a unique loan ID.*

*Gender -> In general it is male or female. No offence for not including the third gender.*

*Married -> Applicant who is married is represented by Y and not married is represented as N.*

*The information regarding whether the applicant who is married is divorced or not has not been provided. So we don’t need to worry regarding all these.*

*Dependents -> the number of people dependent on the applicant who has taken loan has been provided.*

*Education -> It is either non -graduate or graduate. The assumption I can make is “ The probability of clearing the loan amount would be higher if the applicant is a graduate”.*

*Self\_Employed -> As the name suggests Self Employed means , he/she is employed for himself/herself only. An applicant who is self employed is represented by Y and the one who is not is represented by N.*

*Applicant Income -> Applicant Income suggests the income by Applicant. So the general assumption that i can say that “The one who earns more have a high probability of paying loan amount and would be highly eligible for loan ”*

*Co Applicant income -> This represents the income of co-applicant. can also assume that “ If co applicant income is higher , the probability of being eligible would be higher “*

*Loan Amount -> This amount represents the loan amount in thousands. One assumption I can make is that “ If Loan amount is higher , the probability of repaying would be lesser and vice versa”.*

*Loan\_Amount\_Term -> This represents the number of months required to repay the loan.*

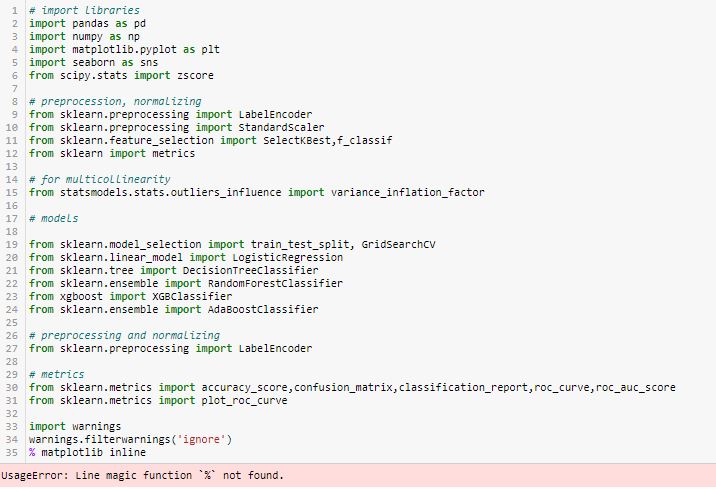
*Credit\_History -> When I googled it , I got this information. A credit history is a record of a borrower’s responsible repayment of debts. It suggests → 1 denotes that the credit history is good and 0 otherwise.*

*Property\_Area -> The area where they belong to is my general assumption as nothing more is told. Here it can be three types. Urban or Semi Urban or Rural*

*Loan\_Status -> If the applicant is eligible for loan it’s yes represented by Y else it’s no represented by N.*

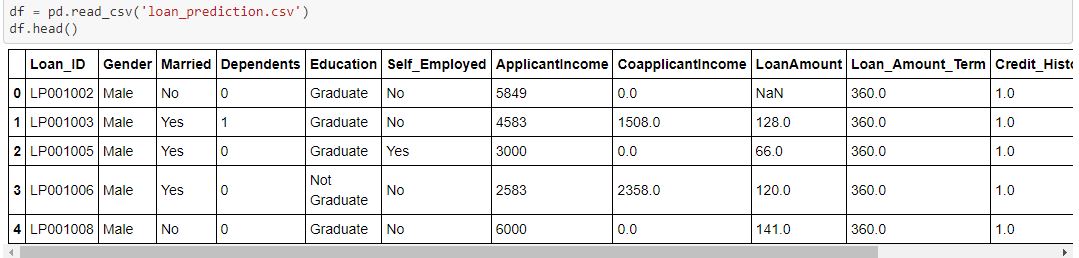
**Importing Libraries**

*I am importing all the library which I required for EDA, visualization, prediction and finding all matrices. The reason of doing this is that it become easier to use all the import statement at one go and we do not require to import the statement again at each point.*



**Loading Data Set into variable**

*Here I am loading the data set into a variable i.e. “df” and processing the first 5 rows.*



**3.Exploratory Data Analysis**

*Exploratory Data Analysis (EDA) ensures the readiness of the data for Machine Learning. In fact, EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them.*

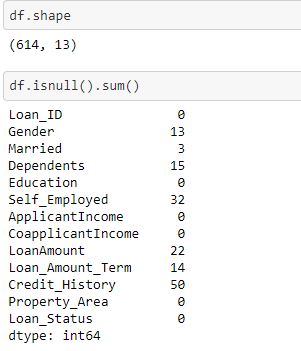
*By looking at the columns description in the above paragraph, i can make many assumptions like : 1. The one whose salary is more can have a greater chance of loan approval.*

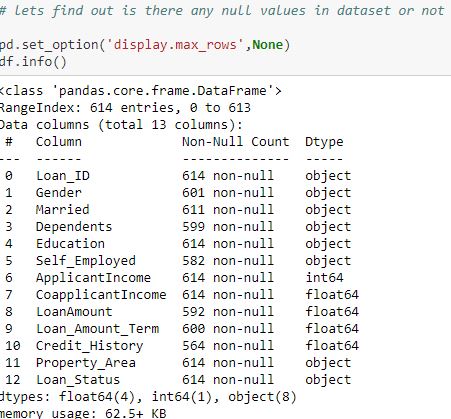
*2. The one who is graduate has a better chance of loan approval.*

*3. Married people would have a upper hand than unmarried people for loan approval .*

*4. The applicant who has less number of dependents have a high probability for loan approval.*

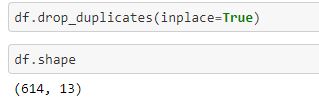
*5. The lesser the loan amount the higher the chance for getting loan.*





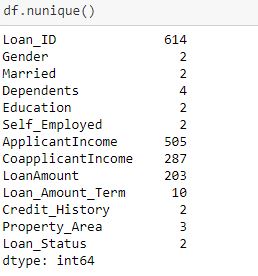
***There are 614 rows and 13 columns in the dataset. There are null values in few columns which I will fixed it soon.***

**Checking duplicates**



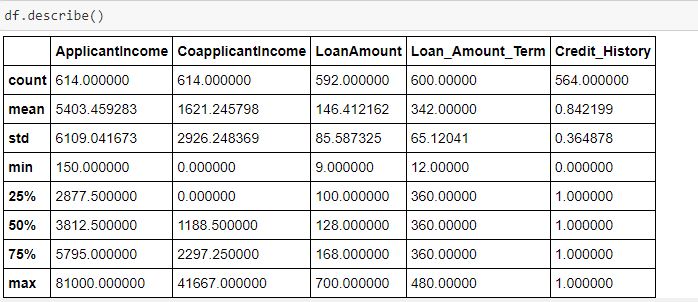
***There are no duplicates values in the dataset.***

**Identifying the columns**



***It is a mixed dataset as it contains few columns which are categorical in nature and few are continuous in nature.***

**Describing the columns**



*Above statistics data show that there multiple outliers mostly in Applicant Income. There is also difference between mean and 50% value in some of the columns which used to get fix for better prediction.*

*Also, number of rows in each column are not same, means there are null values in the data set.*

*Also, the mean and 50%value of most of the column are not same and the STD and mean have a major difference to each other.*

*Most of the columns of min value are near to 0 values.*

*By checking the difference between the 75% and max value there are outliers in some of the column, I will check it soon.*

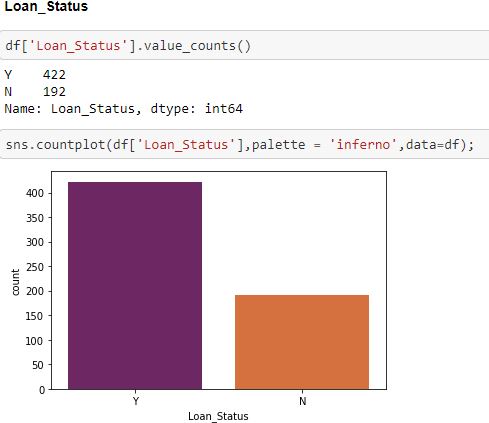
***Count Plot & Pie Plot Definition***

***Count Plot***

*Count Plot function is used for visualizing data of a machine learning and data science project. The Count Plot is majorly used for showing the observational count in different category based bins with the help of bars.*

**Making Plots of Categorical Columns**

**Loan\_status**



***According to the Count Plot of "Loan\_Status" column i can see "Y" shows the Approval and "N" give Disapproval of Loan in the Dataset. There are 422 approved loan and 192 loan not approved. It is balanced dataset.***

**Gender**



***There are 489 males applied for the loan and 112 female applied for the loan. I can say that both male and female loan rejection status are in same ratio.***

**Married**



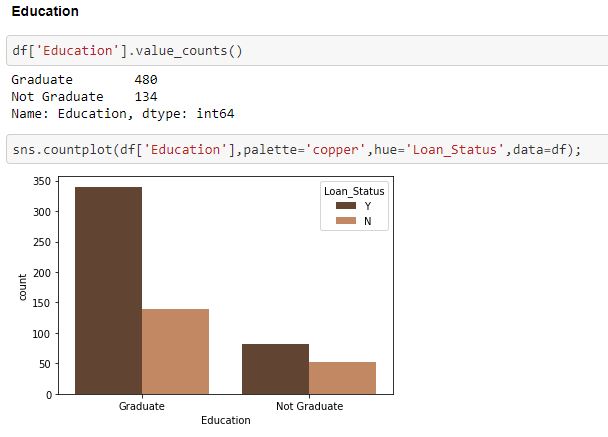
***There are 398 people applied for the loan are married and 213 are unmarried. From the graph I can say that those who are unmarried are more in numbers in terms of loan rejection than those who are married.*** *In short I can say that Married Couples have a higher opportunity of Getting Loan rather than Unmarried ones****.***

**Dependents**



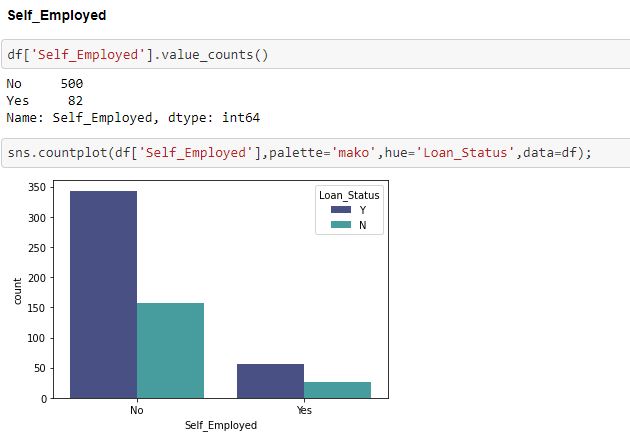
***I can say that those people who have no dependent are getting there loan approved easily as compared to those who have more than 1 family members as a dependent. I can say the loan rejection chances is high those people who have 3+ members as dependent.***

**Education**



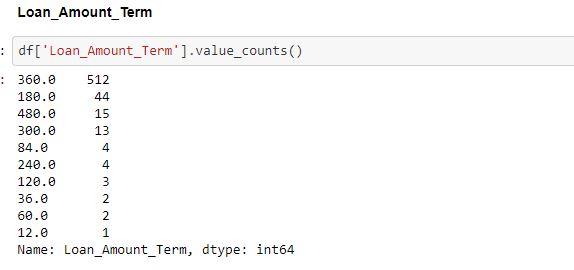
***Education graph Shows that there are more number of graduates than non-graduates in this dataset. There are 480 graduates and 134 under graduates applied for the loan. Moreover people who are not graduates are less in numbers as compared to those who are graduates in terms of approved loan.***

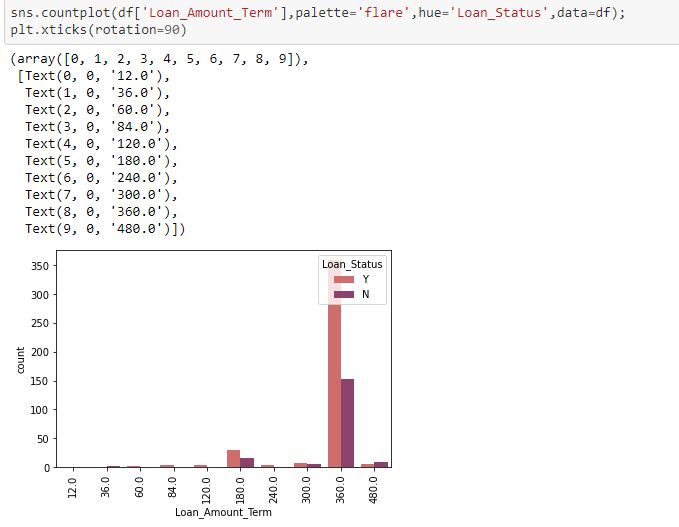
**Self\_Employed**



***There are only 82 people in the dataset who are self employed and 500 people are not self-employed. So, from this graph I can say that those who are salaried people they have more loan requirements as compared to self-employed.***

**Loan\_Amount\_Term**





***I can say that maximum people have applied for loan for the 360-month terms. Which means they are going to replay the loan on EMI basis for 3 years.***

**Credit History**



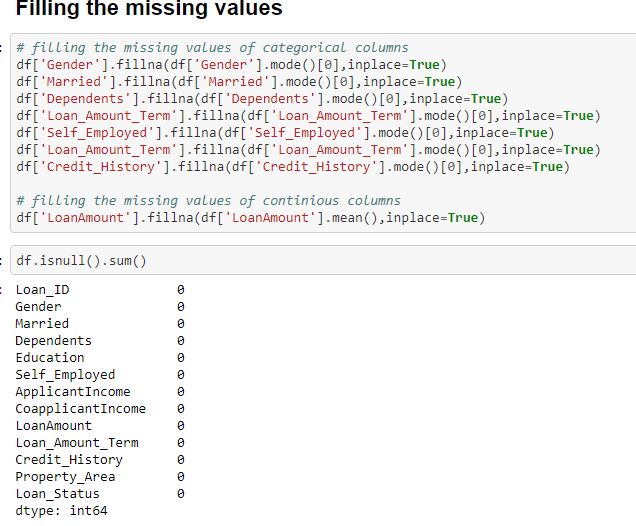
*In the above plot i can say that Larger Percentage of people have a good credit history. 0 represent poor credit history which means their loan rejection chances are very high.*

**Property\_Area**



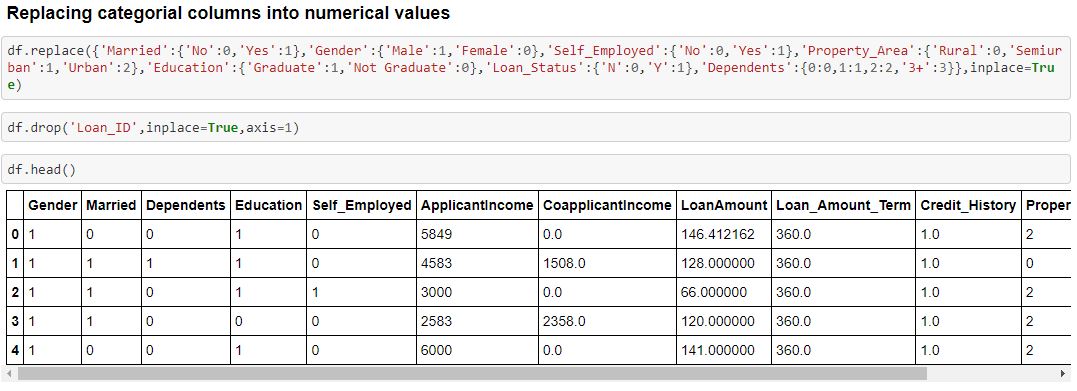
***From the above count plot i can see that Semi-Urban peoples required Loan more than the remaining Urban & Rural Property Areas people. I can say that the loan rejection chances is more in rural area than urban and semi urban. This might be situation because rural people have less source of income so loan rejection chance is high.***

**Filling The Missing Values**



*There were few null values in the dataset which I have removed with the fillna command. Continuous column is filled with mean & Categorical Column is Filled with mode.*

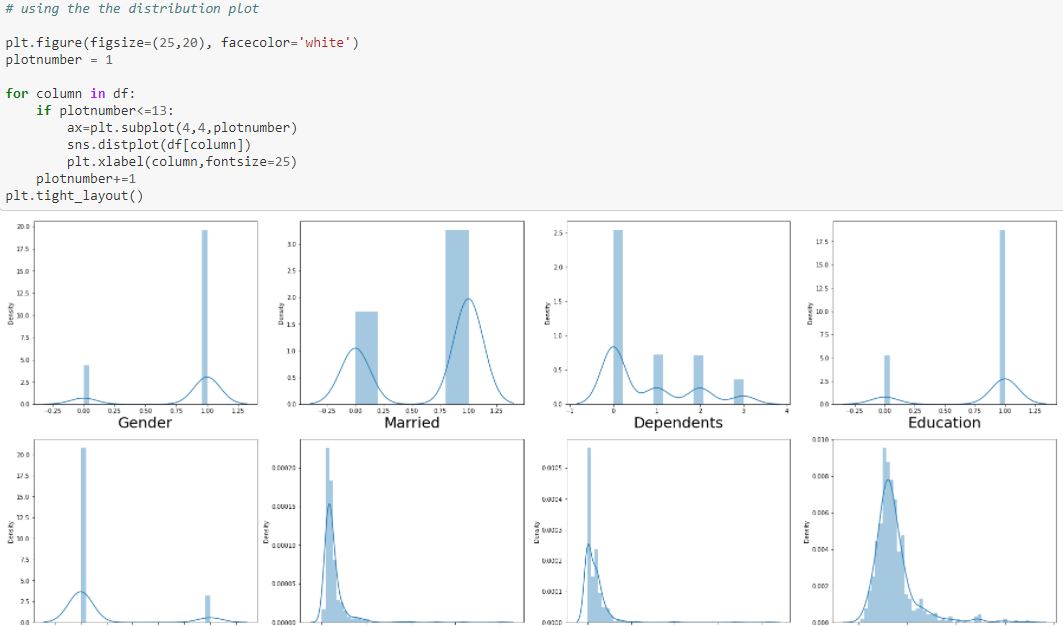
**Replacing Categorical columns into numerical values**



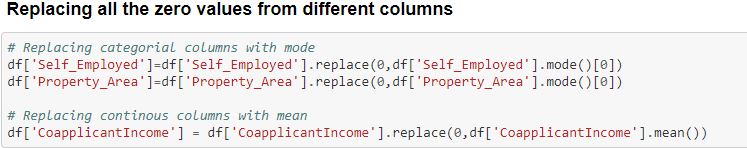
*I have converted all the string columns to numerical values for best prediction and visualization.*

**Distribution Plot**

*It is used basically for univariate set of observations and visualizes it through a histogram that is only one observation and hence we choose one particular column of the dataset.*



*I have plot distribution plot to check the skewness of the dataset. There is skewness in few columns which I will get it fixed soon.*

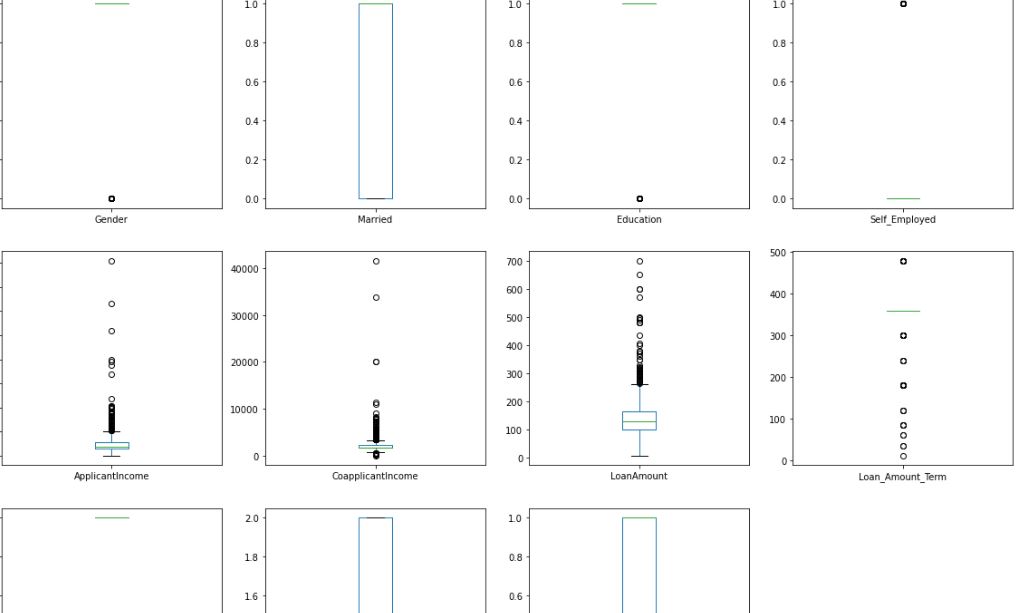


***According to the Dataset these 3 columns (Self\_Employed, Property\_Area, CoapplicantIncome) have zero values which i have replace it with the help of mean and mode.***

**Box Plot**

#### **Box** **Plot** is the visual representation of the depicting groups of numerical data through their quartiles. **Boxplot** is also used for detect the outlier in data set.

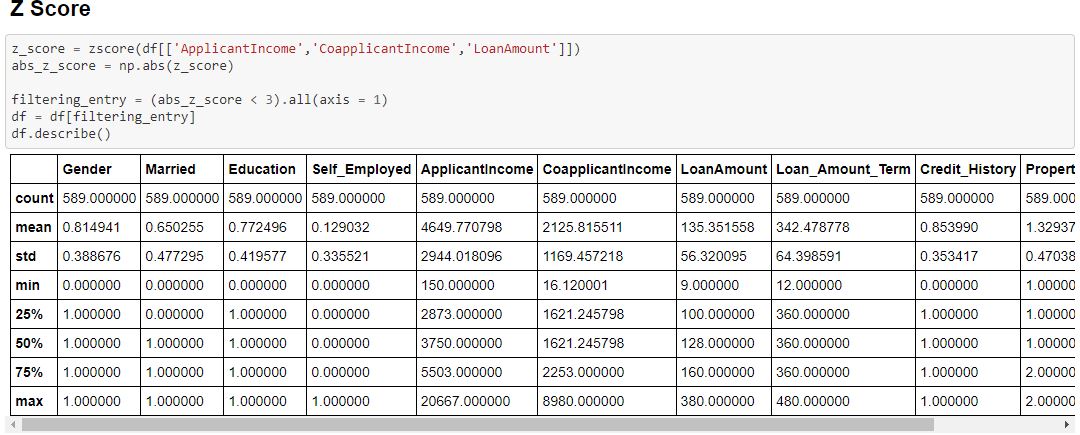
*I used box plot in this dataset because It captures the summary of the data efficiently with a simple box and whiskers and allows me to compare easily across groups.*



***From above image we can clear see that there are number of black dots in most of the column which are referring to the outliers, so it means most of the columns have outliers.***

**ZScore**

*Z score is also called standard score. This score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, Z score tells how many standard deviations away a data point is from the mean.*



***So now we detect the outliers now the second step is to remove the outliers, there are different way to remove the outliers i will be using z score above to remove outliers.***



***After Removing Outliers & Skewness I make Distribution Plot to understand how much my data is ready for Model Prediction.***

[**4.Building Machine Learning Models**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#8341)

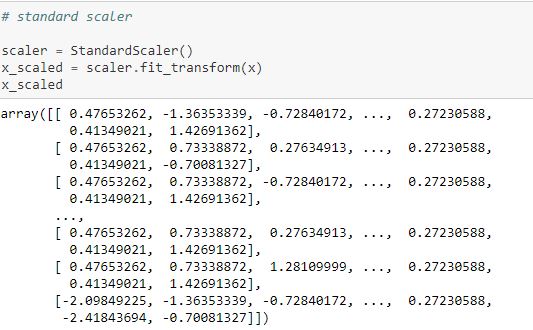
*Here's the role of the label that comes into existence. A label is given to such features to distinguish them from other features. Thus, I have obtain labels as output when provided with features as input*.



*Dividing the features and label into x and y.*

**STANDARD SCALER**

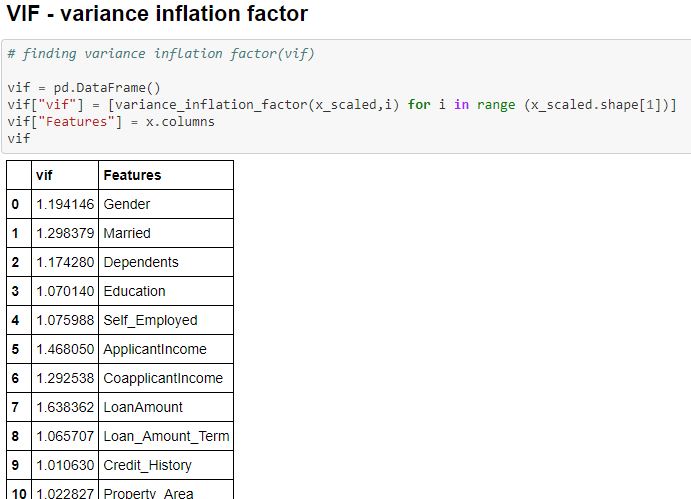
*In Machine Learning, Standard Scaler is used to resize the distribution of values ​​so that the mean of the observed values ​​is 0 and the standard deviation is 1.*



#### By using standard scaler, I have scaled the data in one range.

**VIF – VARIANCE INFLATION FACTOR**

*VIF is preferred as it can ‘****show the correlation of a variable with a group of other variables’****. Dropping one of the correlated features will help in bringing down the multicollinearity between correlated features: The image contains the original VIF value for the dataset.*



# *VIF score for all the columns is below 5 which means Multicollinearity doesn’t exist.*

# Train Test Split

# *The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.*

# *I used train test split as it is a fast and easy procedure to perform, the results of which allow me to compare the performance of machine learning algorithms for my predictive modelling problem.*

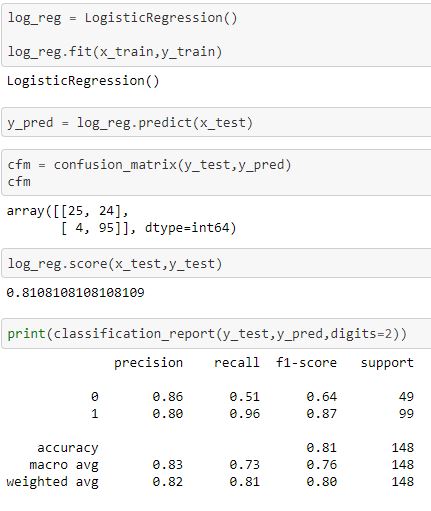
# 

#### Therefore, I have split the dataset into training and testing. I have used 25% of data for testing and 75% of the dataset for training. I used random state for consistent result.

**LOGISTIC REGRESSION**

*Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is discontinuous ,which means there would be only two possible classes.*

*Generally, logistic regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it.*



#### So, I get Auc Score from Logistic Regression is 81.08%.

**AUC SCORE**

*The AUC for the ROC can be calculated using the roc\_auc\_score() function , the AUC function takes both the true outcomes (0,1) from the test set and the predicted probabilities for the 1 class. It returns the AUC score between 0.0 and 1.0 for no skill and perfect skill respectively.*

 So, I get Auc Score from Logistic Regression is 73.49%.

**KNN**

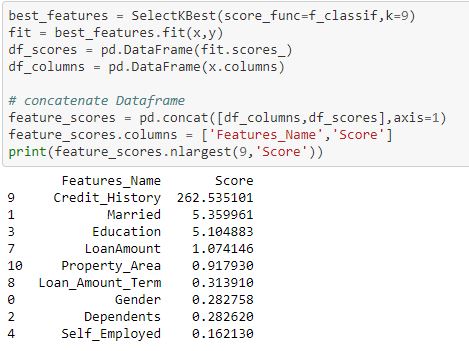
**Features Selection Method**

*Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve.*

*Irrelevant or partially relevant features can negatively impact model performance.*

*Feature selection and Data cleaning should be the first and most important step of your model designing. Therefore, I used Select K Best Method for best features selection.*

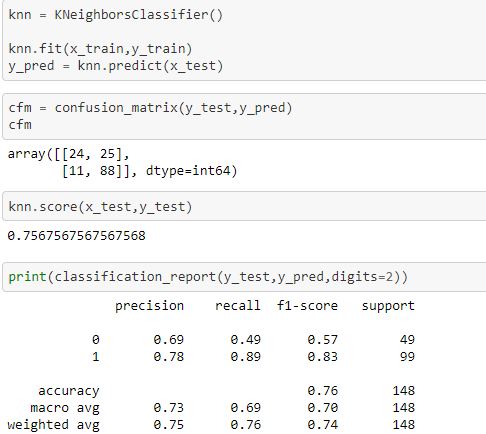
### **Using SelectKBest Method for best features selection**



*According to the SelectKBest Method Credit\_History plays an important role in Approving loans to the ultimate individuals.*

**KNN - K-Nearest Neighbour**

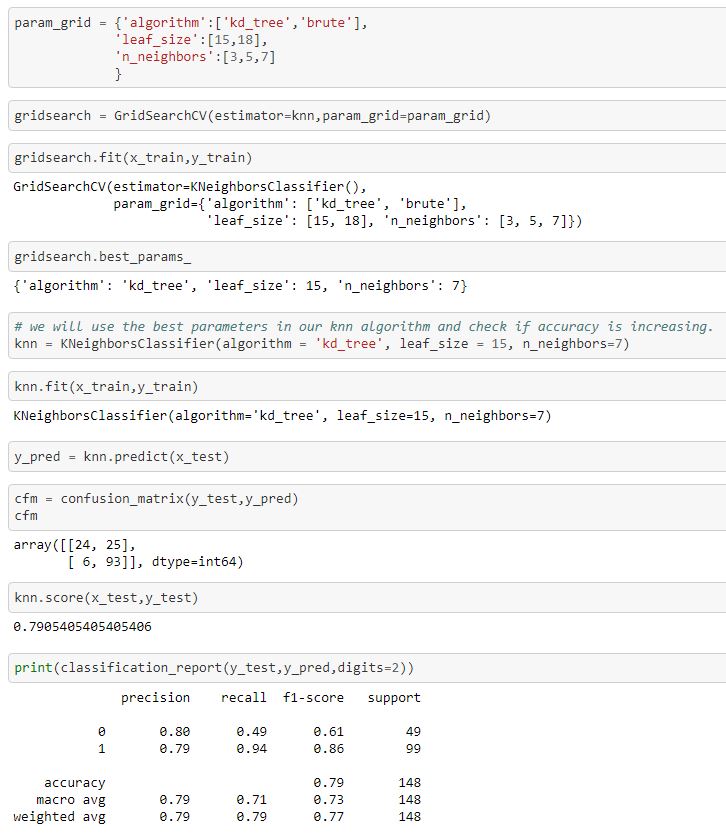
*It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, you can use the KNN algorithm for applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure.*



*By using KNN I Get Accuracy Score : 76%.*

# Let use Grid search CV for the best parameter to improve the accuracy

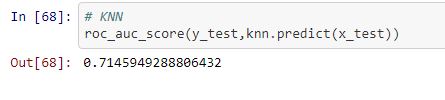
# *GridSearchCV is a library function that is a member of sklearn's model\_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyperparameters.So I used Grid search CV to choose the best parameters in this knn algorithm*

****

**By using Grid search CV in KNN I improve Accuracy Score : 79%**

**Auc Score of KNN**

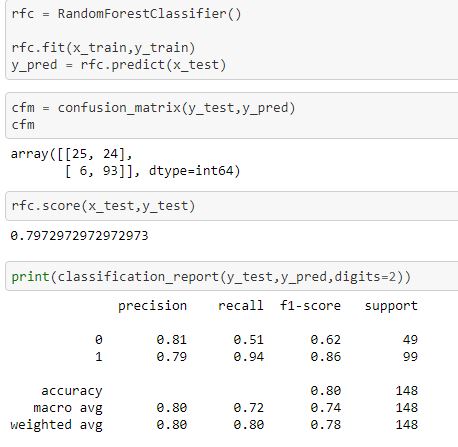
*The AUC for the ROC can be calculated using the roc\_auc\_score() function , the AUC function takes both the true outcomes (0,1) from the test set and the predicted probabilities for the 1 class. It returns the AUC score between 0.0 and 1.0 for no skill and perfect skill respectively.*

****

*So, I get Auc Score from KNN is 71.45%.*

**Random Forest Classifier**

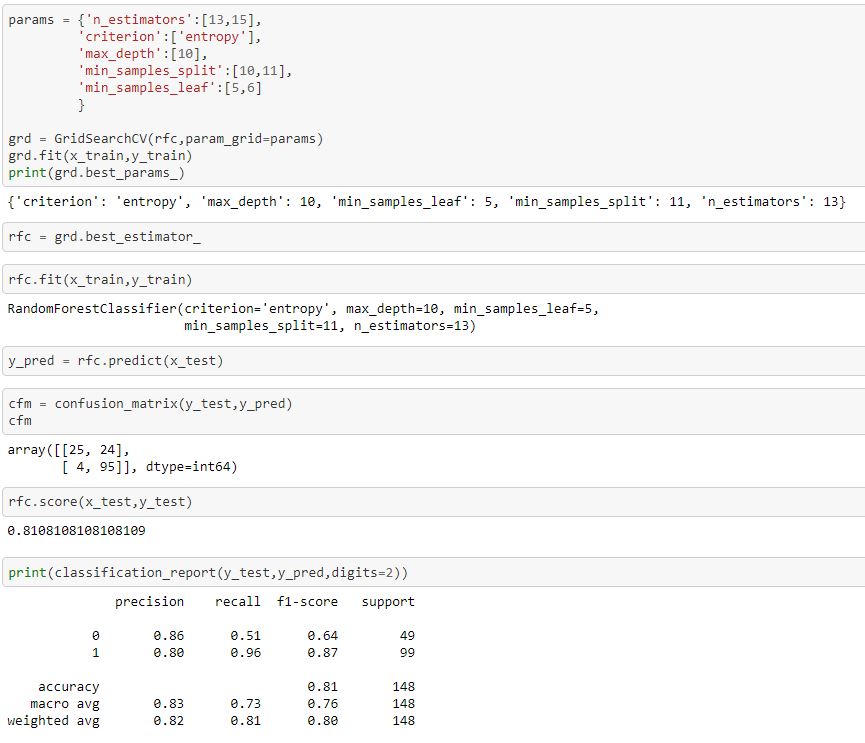
*A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. So I have used 3rd algorithm Random Forest Classifier to predict the model.*

**

*By using Random Forest Classifier I Get Accuracy Score : 80%.*

**Let use Grid search CV for the best parameter to improve the accuracy**

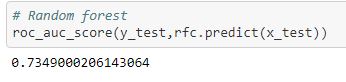
*GridSearchCV is a library function that is a member of sklearn's model\_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So**I used Grid search CV to choose the best parameters in this random forest algorithm.*



**By using Grid search CV in random forest I improve Accuracy Score to 81%**

**Auc Score of Random Forest classifier**

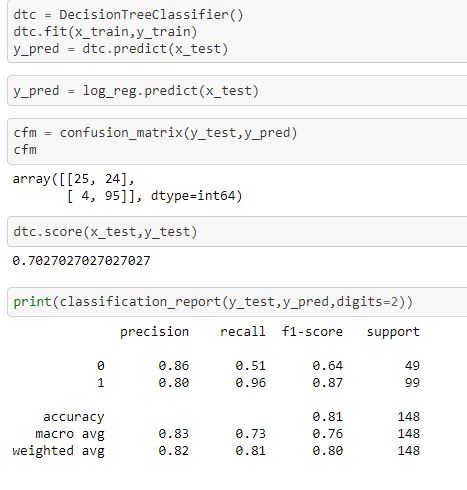
*The AUC for the ROC can be calculated using the roc\_auc\_score() function , the AUC function takes both the true outcomes (0,1) from the test set and the predicted probabilities for the 1 class. It returns the AUC score between 0.0 and 1.0 for no skill and perfect skill respectively.*



*So, I get Auc Score from Random Forest classifier is 73.49%.*

**Decision Tree Classifier**

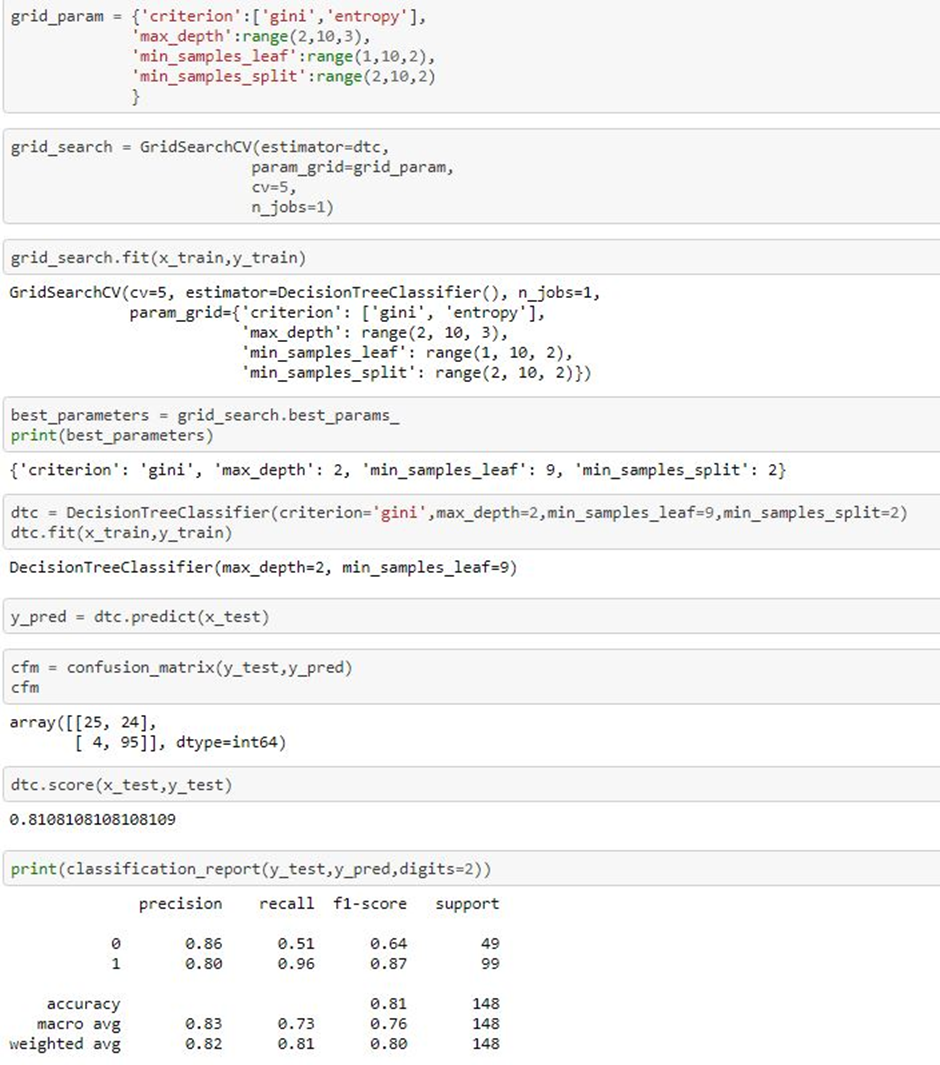
*Decision tree is one of the predictive modelling approaches used in Machine Learning. It is one of the most widely used and practical methods for supervised learning. The decision tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for regression problem.* *Therefore Decision tree is the 4th Algorithm used by me.*



*By using Decision Tree Classifier I Get Accuracy Score : 81%*. *Now I have to improve the accuracy score by using best hyperparameters.*

**Let use Grid search CV for the best parameter to improve the accuracy.**

*GridSearchCV is a library function that is a member of sklearn's model\_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So I used Grid search CV to choose the best parameters in this Decision tree algorithm.*

**

**By using Grid search CV in Decision tree has no improve in Accuracy Score to 81%**

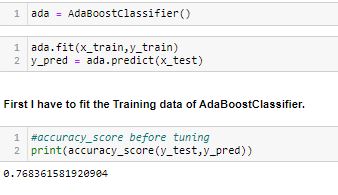
**Auc Score in Decision tree Classifier** *The AUC for the ROC can be calculated using the roc\_auc\_score() function , the AUC function takes both the true outcomes (0,1) from the test set and the predicted probabilities for the 1 class. It returns the AUC score between 0.0 and 1.0 for no skill and perfect skill respectively.*

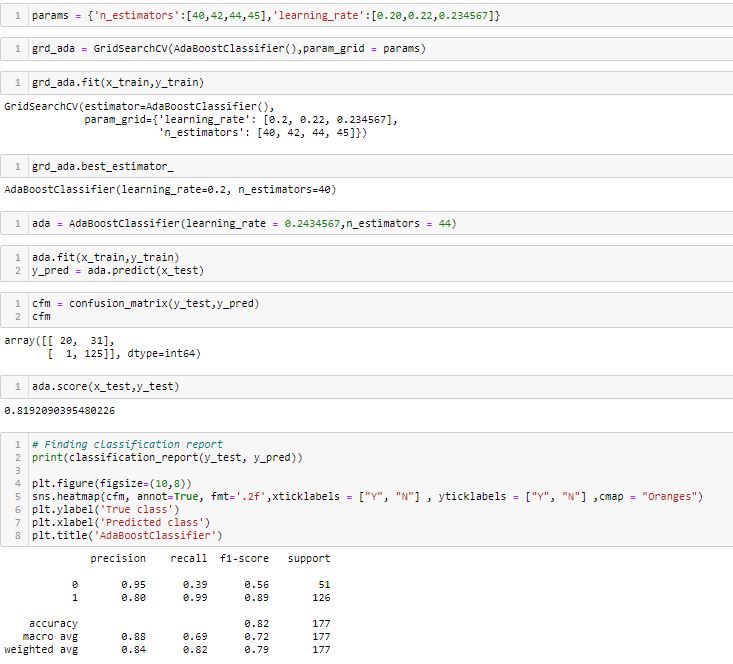


*So, I get Auc Score from Decision Tree classifier is 73.49%.*

**AdaBoostClassifier**

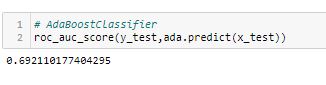
***AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem. The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level.***

**

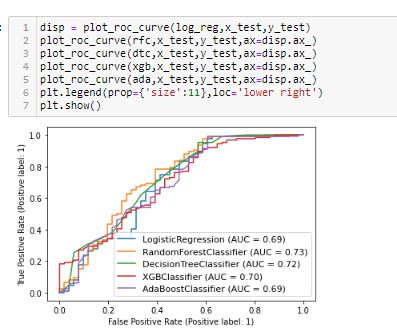
**

**In prediction AdaBoostClassifier with Loan Status:**

**I had done this prediction by taking Loan Status as an output variable which is Classification in nature.While calculating the best random state the 0 is best state which providing the highest accuracy value for this model.After calculating confusion matrix and classification report of AdaBoostClassifier - 76.83%.Therefore i use hyper tuning method by using Grid Search Cv to improve my Accuracy Score and this gives a increasing value of AdaBoostClassifier - 82%**[**¶**](http://localhost:8888/notebooks/RE-SUBMIT%20EVALUATION%20PRJECTS/Loan%20Application%20Status%20Prediction.ipynb#Therefore-i-use-hypertuning-method-by-using-Grid-Search-Cv-to-improve-my-Accuracy-Score-and-this-gives-a-increasing-value-of-AdaBoostClassifier----82%)**Then I found roc\_auc\_score of AdaBoostClassifier which is 69.21%.**

**

**Roc Curve**

*An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate & False Positive Rate. ROC is a probability curve, and AUC represents the degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy Higher the AUC, better the model is at distinguishing between individuals and Loan amount whether to approve the loan or n.*

*According to my Roc\_Auc Curve I can say that Random Forest Model with 81% is the best predictive model in this dataset.*

**Saving the Best Model**

*In machine learning I need to save the trained models in a file and restore them in order to reuse it to compare the model with other models, to test the model on a new data. I have use pickle method to save the best model in this dataset.*



*Saving Random forest Model as it accuracy score is 82% and Roc\_Auc Curve value is 69.21%.*

**Observation**

*I’ll briefly discuss how I approached this problem of predicting Churn.*

1. *I have Loan\_Prediction dataset from which I had to extract information.*

*b ) I had used pandas library to read the dataset which provide me to explore & visualize the data properly based on Rows & Columns.*

*c) I have pre-processed the data using replacing all zeros.*

*d) I did exploratory data analysis on main data frame and tried to see all visualizations.*

*e) Based on visualization knowledge, I used various EDA TECHNIQUES to plot the graphs and Box Plot.*

*f) I used VIF and Standard Scalar to scaled the data and by VIF I make a relationship between Features & Features.*

*g) I split the Features & Labels into 2 parts.*

*h) On this data, I have applied our machine learning models such as logistic regression,* *Random forest, KNN, and Decision Tree Classifier .*

*i) After which I found Random forest has the highest accuracy score(81.08%) and best among l the models.*

*j) Based on the model scores, to predict Loan\_Status Random forest seems to be the best model for this dataset.*

**Conclusion**

**In the modeling Building section, we have seen the hyper parameter tuned AdaBoostClassifier with 82%**

**It can be clearly seen the feature engineering effects on each individual nine model in 'Comparison of ML Classifiers with Effects of Future Engineering' Roc Curve chart above.**

**I’ll briefly discuss how I approached this problem of predicting loan prediction.**

**a) I have Loan Prediction dataset from which I had to extract information.**

**b ) I had used pandas library to read the Dataset which provide me to explore & viualize the Data properly based on Rows & Columns.**

**c) After from all datasets, I preprocessed the data using imputer function to fill the missing values and replacing all zeros.**

**d) I did exploratory data analysis on main data frame and tried to remove all the possible outliers.**

**e) Based on visulization knowledge, I use various EDA TECHNIQUES to plot the graphs and Z - SCORE.**

**f) I use VIF and Standard Scalar to scaled the data and by VIF I make a relationship between Features & Features.**

**g) After from all these i split the Features & Labels into 2 parts.**

**h) On this data, I have applied our machine learning models such as logistic regression, Random forest,Decision Tree and KNN.**



**Links References**

I have build a model that has predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset.

**Dataset Link**

<https://github.com/dsrscientist/DSData/blob/master/loan_prediction.csv>

**Github Link**

<https://github.com/sachingupta2015s/NEW-EVALUATION-PROJECT/blob/main/Loan%20Application%20Status%20Prediction.ipynb>

**Thank you for reading!**

*If you enjoyed this article, Follow me on GITHUB for the latest updates.*

*Thank you again for reading!*

-----------------------------------------------------xxxxxxxx----------------------------------------------------------