Attrition data set

1-Introduction: -

1.1- Background and the purpose of the analysis: -

An issue that every company deals with is attrition. The IBM data scientist created this data to get a better idea of the reason why employees leave the company. The purpose of this analysis is to create a model which can predict the attrition class i.e. YES, NO, based upon the the independent variables. Some variables which might look quite prevalent otherwise, have very low weightage on the attrition class. This research analysis tries to find a good machine learning algorithm which can predict the Attrition rate.

1.2 - Structure of the Data sets.

The dataset has one CSV files named Employee Attrition. The data set has 1470 observation with 35 variables including Attrition. The main feature of the dataset are:-

1- Age
2-. Attrition
3- Distance from Home
4- Employee Count
5- Employee Number
6- Environment Satisfaction
7- Hourly Rate
8- Job Involvement
9- Job Satisfaction
10-MonthlyIncome
11-TotalWorkingYears
12-TrainingTimesLastYear
13-WorkLifeBalance
14-YearsInCurrentRole
15-YearsSinceLastPromotion

Apart from these continuous variables there are some categorical variables which show the various employees segments the data set is divided into.

1- MaritalStatus- "Single", "Married"
2- JobRole- "Sales Executive", "Research Scientist" etc
3- OverTime- "Yes" ,"No"
4- EducationField- "Life Sciences" , "Other" etc
5- Department- "Sales" "Research & Development" "Research & Development"
6- BusinessTravel "Travel_Rarely" , "Travel_Frequently" etc
7- Education- “Below College”, “College”, “Bachelor” etc

The variable importance and variable relation are covered in the later part of the report. The attrition is the target variable which will be predicted in this research analysis.

2- Examination of the Data and Initial Data reconfiguration: -

The dataset given is very clean and and does not require any significant data reconfiguration. There are no NA values and thus does not require replacement.

The target variable is in the second column which should be either in first column or the last for better analysis. This is done by direct column shuffle in R. All the the categorical variables are in character which should be converted I factor before analysis for better model fitting. *Lapply* function in the R studio is used for doing that. It should be noted that direct conversion of character into factor using *as.factor()* will create garbage value in the data set and thus should be used judiciously.

3-Data pre-processing and examination: -

3.1- Variable Importance: -

The variable importance of the dataset can be found using the *varImp()* function of library *randomForest*.

The table of the variable importance is given below.

Feature	Importance
Age	22.958664
Business Travel	6.809578
Daily Rate	20.283931
Department	3.909125
Distance From Home	17.318321
Education	7.254066
Education Field	12.42122
Employee Count	0
Employee Number	18.098085
Environment Satisfaction	11.280757
Gender	2.430959
Hourly Rate	18.228197
Job Involvement	9.684742
Job Level	8.003529
Job Role	17.648654
Job Satisfaction	10.462939
Marital Status	8.506924

Monthly Income	28.89256
Monthly Rate	18.391056
Num Companies Worked	12.40862
Over18	0
Over Time	20.520649
Percent Salary Hike	13.053421
Performance Rating	1.980224
Relationship Satisfaction	8.930243
Standard Hours	0
Stock Option Level	10.920601
Total Working Years	19.336666
Training Times Last Year	9.81672
Work Life Balance	9.809122
Years At Company	15.633714
Years In Current Role	9.944709
Years Since Last Promotion	9.561381
Years With Curr Manager	11.74171

Table 1 Variable importance for predicting Attrition

The Pictorial representation of the the variable importance is also given below. It can be clearly visible that Monthly Income plays the most important role in deciding the attrition of the employee. On the other hand features like standard hours and over 18 have no weightage whatsoever. Age is also an important feature while predicting the Attrition. This might suggest that retirement is also counted in attrition. Other important features are over rate daily rate, total working year, Monthly rate etc. As discussed earlier some features which may look like play an important role in deciding the attrition like percent salary hike has comparatively lower variable importance. Thus we will base our analysis solely on the algorithm.

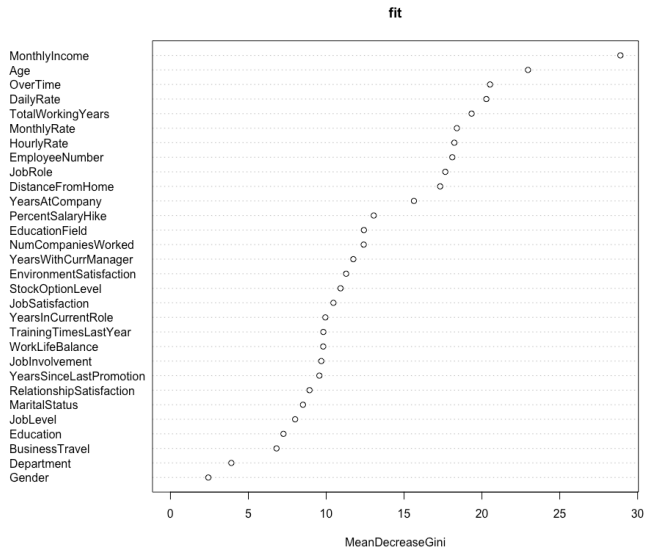

Fig 1- Variable importance of the features

Next correlation plot is studied to find any redundant variables. It is very much visible that most of the variables have very low correlation with each other which is a good sign and will make our model quite good. Although some variables like total working year, years with current manager and years at company have relatively high correlation and thus could be dropped from the analysing.

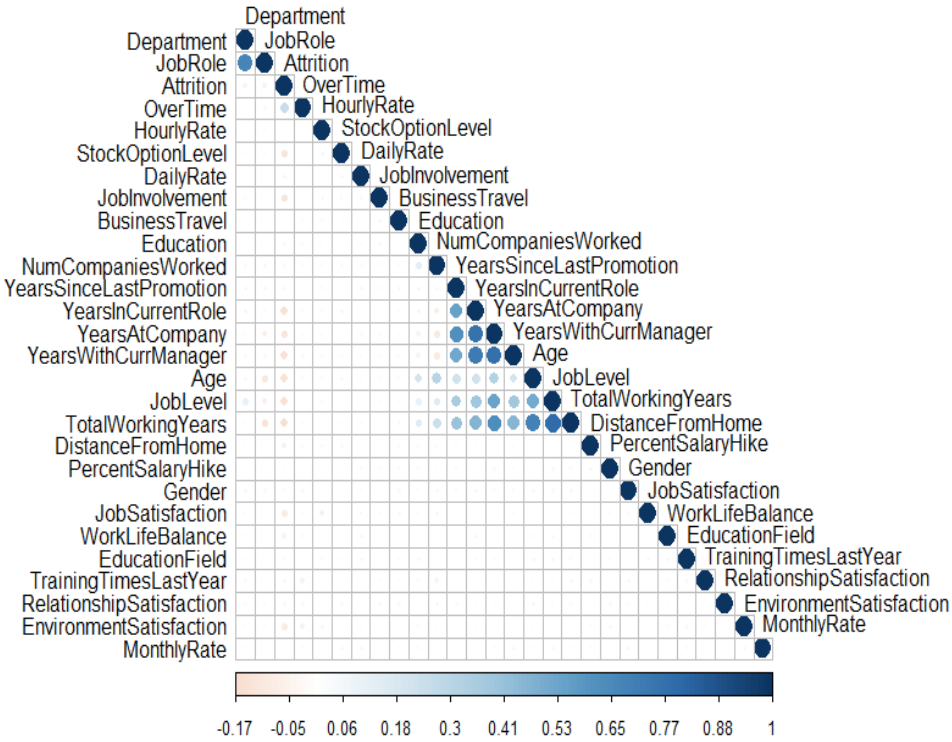

Fig2- Correlation plot of the dataset

Further analysis of some more variables shows the following: -

1. Job Role: There are different job roles present in the employee dataset such as sales representative, laboratory technicians, human resource, sales executives, research scientists, manufacturing directors, healthcare representatives, managers and research directors among which sales representative are having highest tendency to leave the organization followed by the laboratory technicians while the Research Directors are the least tend to move.

2. Over Time: This Variable is one of the categorical variables which contains values of ‘yes’ for the people who does overtime and ‘no’ for the people who does not. People who do overtime are more likely to leave the organization as compare to the people who do not do overtime.

3. Training Time: This variable contains the level from one to six which indicates the training time period. It has been analysed that the time level of 2 are having most tendency to leave the organisation which is followed by level 3 while both the people who spent least and most time in training are in very less percentage of leaving the organisation and tends to stay within the organisation.

3.2- Plots for feature variation: -

3.2.1- Employee satisfaction count

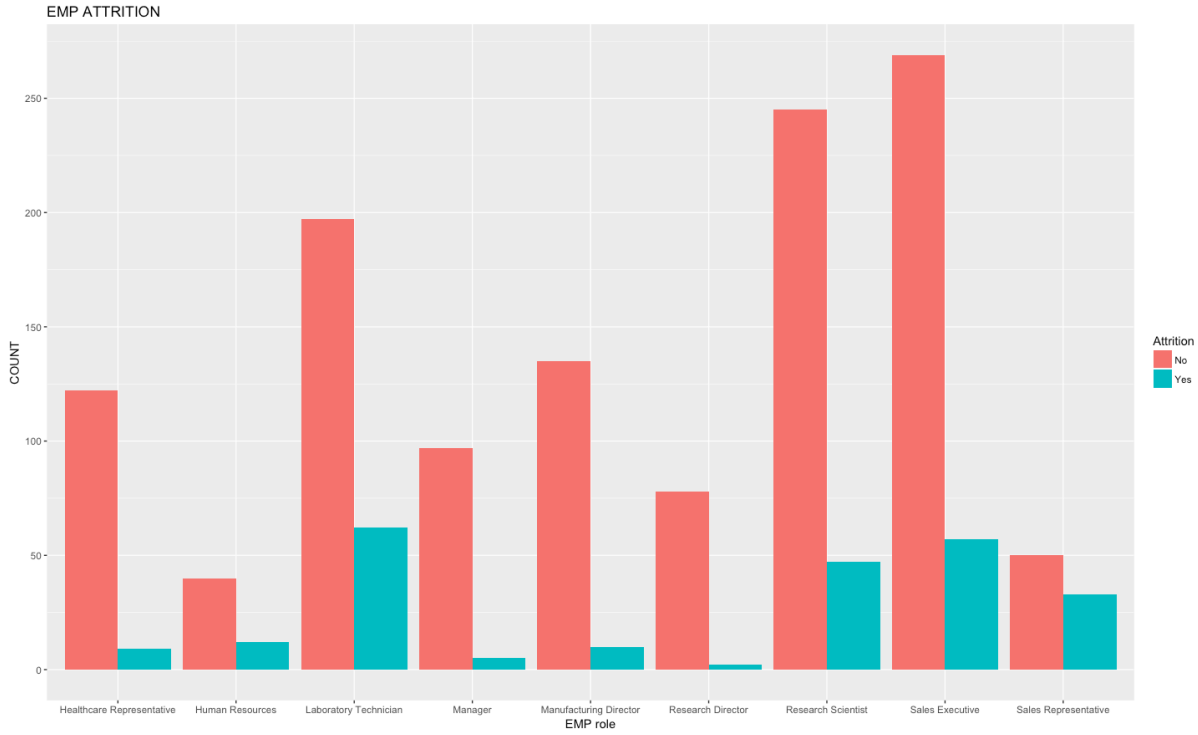

Fig 3- Histogram of Employee role with Attrition

It can be observed that the maximum percentage of attrition is with lab technician job role, followed by sales executive and research scientist.
The least percentage is with research director followed by managers and health care representative.

3.2.2- Box plot of monthly income vs Attrition: -

On analysing the box plot, it is observed that the mean value of employee who leaves the job have have monthly income 5100 units while the people who leave the job have monthly income of around 2700 units.
There are certain outliers which are visible in the box plot. Most of the people who don't leave the job have a very high monthly income. The spread of outliers of employees who leaves job is quite a lot which creates certain unpredictivity.

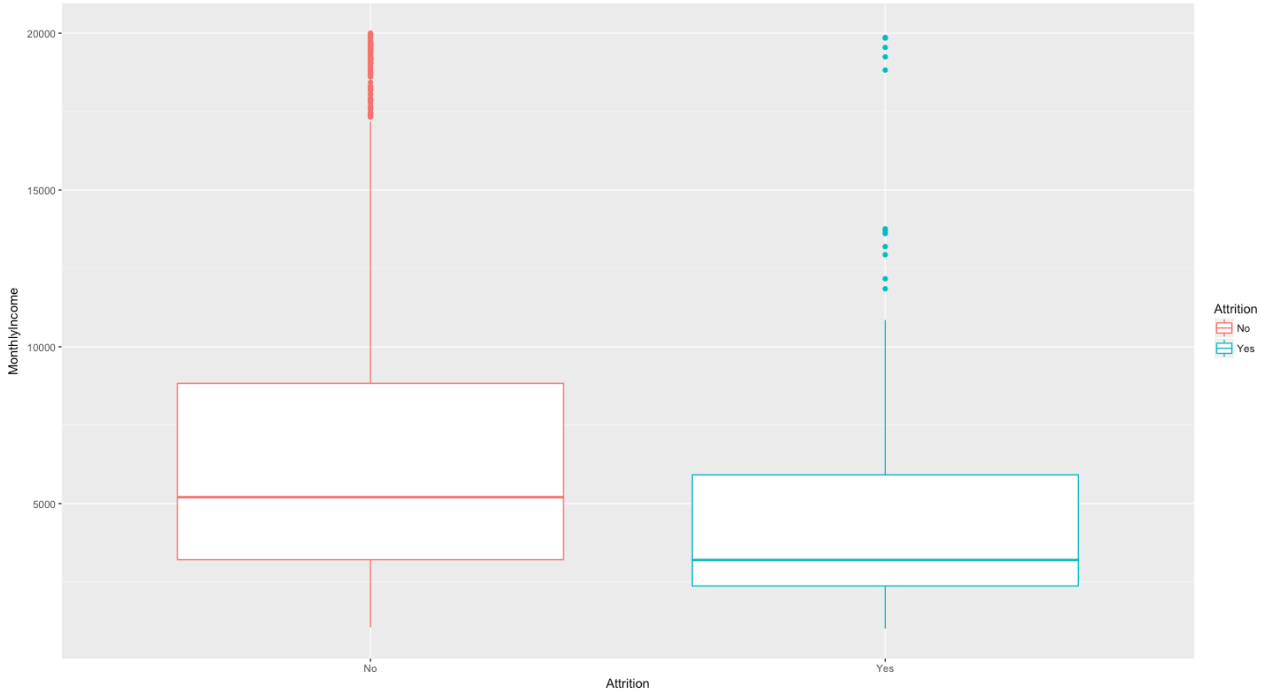

Fig 4- Box plot of Attrition rate with the monthly income

3.2.3- Box plot of monthly income vs Attrition: -

Below is the box plot of job role vs Monthly income spreaded over attrition.
Managers and research director have the high monthly income while sales representative and research scientist have lower monthly income. One thing to notice is that the mean value of monthly income of research director has a huge difference. For certain roles monthly income has no significance. Mean value of monthly income for manufacturing director is quite same.

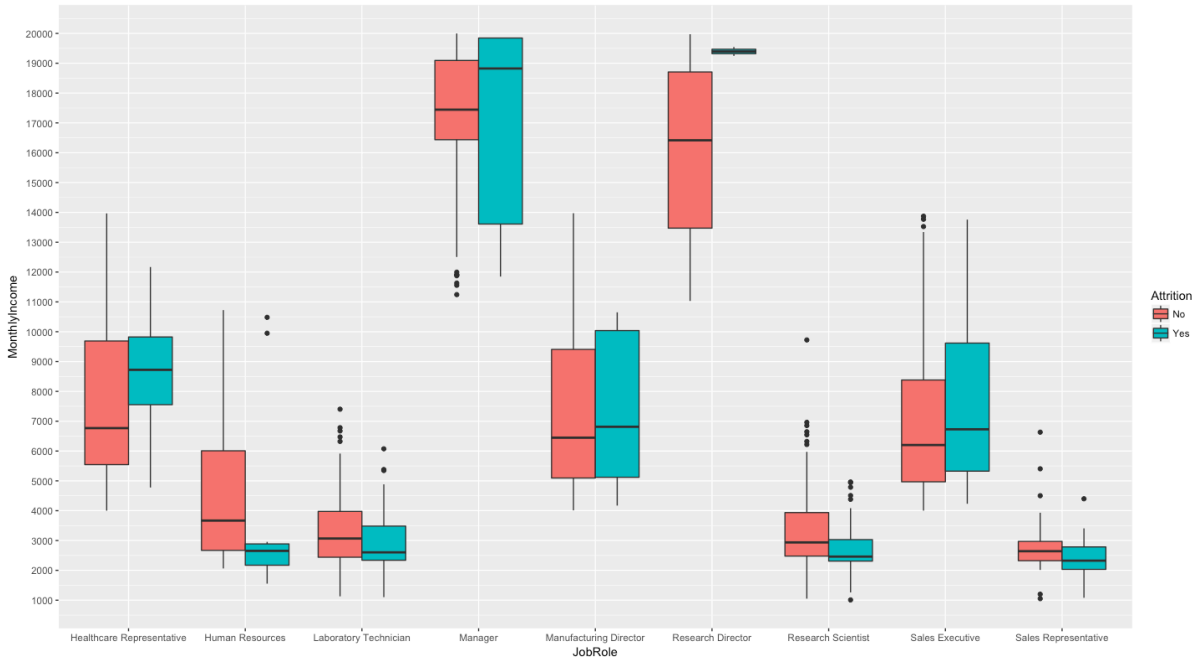

Fig 5- Box plot of Attrition rate with the monthly income vs Job Role

3.2.3- Scatter plot of monthly income vs total working year spreaded over attrition: -

The spread of attrition is quite high especially for attrition.

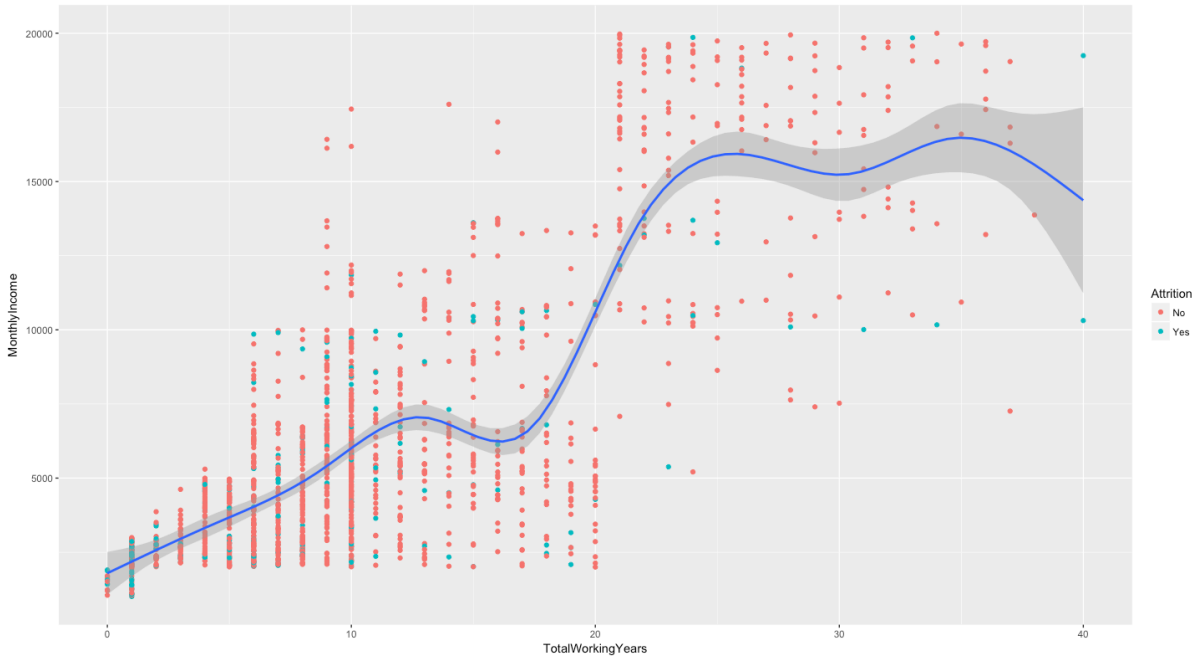

Fig 6- Scatter plot of Attrition rate with the monthly income vs Total working year

4- Single classification trees: -

The single tree is grown by splitting the the data into 70/30. 70% of the data is used for training and the rest is for testing. To check the working of the model confusion matrix built to confusion table.
It has 381 true negative and 23 false positive. Also it has 21 true positive and 37 false negative.
The accuracy achieved by this tree is 87.01%.
The sensitivity of the table is 91.15% and specificity is 47.73%, the detection rate of the trees is 82.47%.
The kappa value of the tree is 0.34 which indicates fair agreement.
Analysing the tree it is visible that the first split is done on the overtime and subsequent splits are done on total working hour and monthly income.
The trees works pretty well for the dataset and can be easily used for predictions,

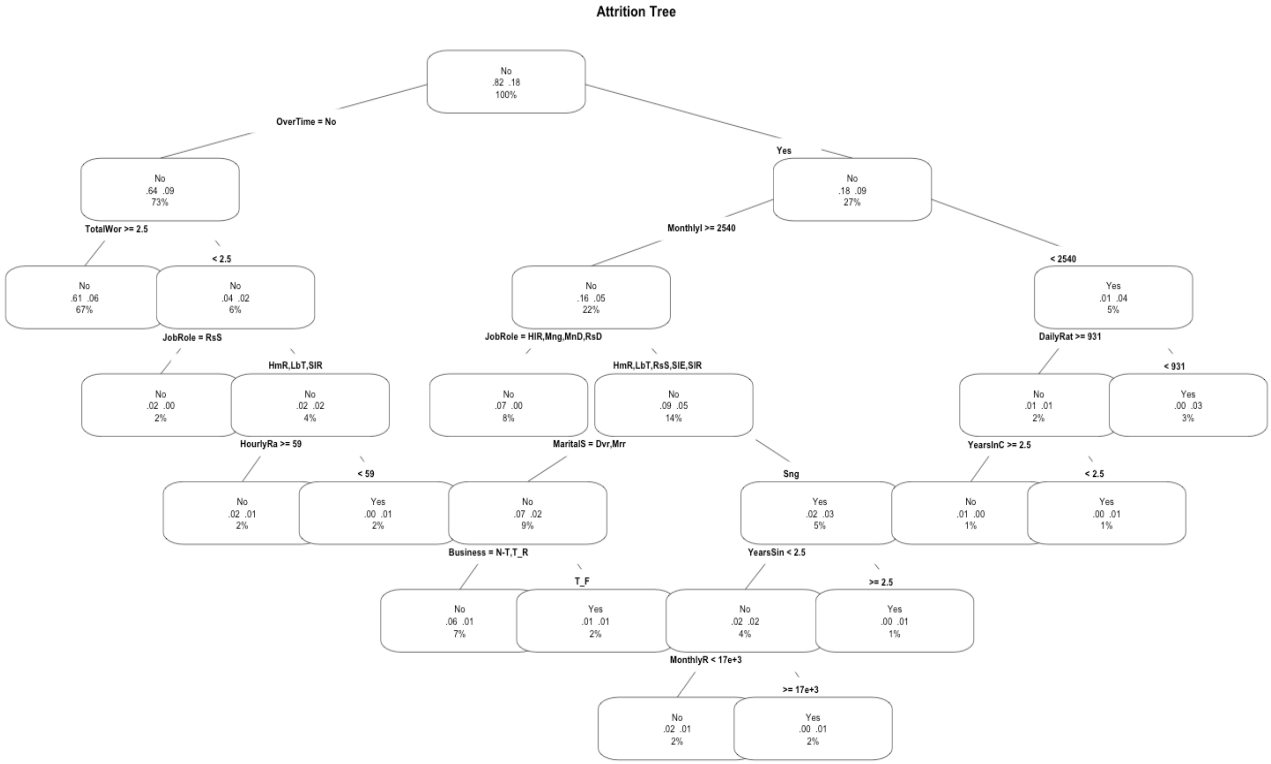

Fig 7-Single classification tree for attrition

4-Ensemble Techniques: -

4.1: - Bagging Ensemble technique: -

Bagging (stands for Bootstrap Aggregating) is a way to decrease the variance of your prediction by generating additional data for training from your original dataset using combinations with repetitions to produce multisets of the same cardinality/size as your original data. By increasing the size of your training set you can't improve the model predictive force, but just decrease the variance, narrowly tuning the prediction to expected outcome.
In this analysis bagging does the classification trees with 25 bootstrap replications. Confusion matrix is made to analyse the working of the model.

The accuracy of the model is coming out to be 84.69%, with 1212 true negative, 21 false positive, 33 true positive and 202 false positive.

The sensitivity of the model is 61.11% and specificity of 85.59%.
The kappa value of the model is 0.17 which indicates good agreement.

4.2 Stacking ensemble technique: -

Stacking is a similar to boosting: you also apply several models to your original data. The difference here is, however, that you don't have just an empirical formula for your weight function, rather you introduce a meta-level and use another model/approach to estimate the input together with outputs of every model to estimate the weights or, in other words, to determine what models perform well and what badly given these input data.

In the stacking two layers are made. The first layer contains n number of model techniques to first train the model on train data set and then predict the classes on the test data set.
A new data frame is build using these predictions and attrition class from the original data set is added to this data frame.
Now the level 1or meta level model technique is chosen and trained on this new data frame and predict using the test data set.
Hence the model are stacked over each other to make the predictions. This whole box of stacked model takes the input from the dataset and gives the output.

In this analysis the three models chosen are decision tree, Logistic regression, KNN regression as the levelo-0 models. We can choose more than 3 models for this layer but for the sake of computational cost it is better to choose 3 only. For level-1 layer log regression is chosen.

For level-0 layer the decision tree gives the accuracy of 85.48% and sensitivity of 98.12% and 16.12% specificity. It has 366 true negative and 11 true positive. The kappa value is coming out to be 0.2 which indicates the fair agreement.

For level- 0 layer the log regression gives the accuracy of 86.85% and sensitivity of 96.25% and specificity of 35.29%. It has 359 true negative and 24 true negative. The kappa value is coming out to be 0.384 which indicates excellent agreement.

For level-0 layer the KNN regression gives the accuracy of 84.35% and sensitivity of 98.9% and specificity of 4.0%. It has 369 true positive and 3 true negative. The kappa value is coming out to be 0.05 which indicates poor agreement.

The new data frame created has 441 observations and 4 variables.

The level-1 or meta level is trained on this data frame.
The level-1 layer has the accuracy of 86.25% and the sensitivity of 96.25% and specificity of 35.25%. It has 359 true negative and 24 true positive and the kappa value of 0.38 which shows the excellent agreement.

Thus the final accuracy of the ensemble is 86.25% which is satisfactory.
There would have been a significant increment of the accuracy if the model chosen at level-0 were not related. All the models are giving the same accuracy and almost same predictions which could not show subsequent increment increment of accuracy.

5- Comparison of models and conclusions: -

Total three models were used for predicting the target variables in this analysis: -

1- Single regression tree
2- Bagging
3- Stacking

The accuracy achieved in the single regression tree is very low giving the unpredictable nature and the hypothesis taken in the beginning that the target variable can be predicted without any financial attribute variable. Below tables shows the achieved accuracy of the model: -

Model	Accuracy
Single regression tree	87.65%
Bagging	84.69%
Stacking	86.25%

Table 2 Accuracy comparison of different models

The accuracy is increased a bit in stacking but it is almost same in all the techniques used. Single regression tree predicts very well all alone.