**Section 2- Data modeling** - (Data set: Employee Test data No. 1.csv) (45 points)

1. As mentioned in section 1 above, the company would like to predict which valuable employee (key talent) will leave next.

Using predictive model(s) of your choice, please build and compare different predictive models. Please provide details of each technique tested (e.g. decision tree learning, logistic regression, etc.). Select a model of choice and explain why. (30 points)

**Data Preprocessing:**

Before performing any modeling, the data must preprocess. Firstly, the independent and dependent variables are created.

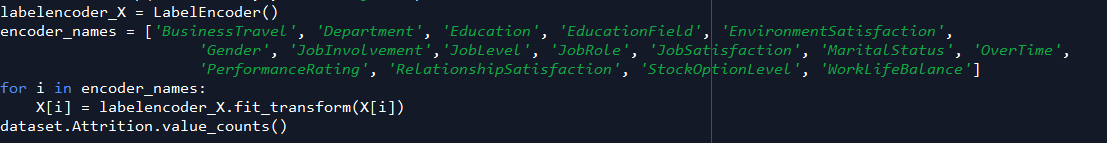
Creation of the Independent variable (X).



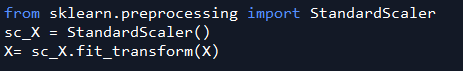
Creation of dependent variable(y).



A label encoder is applied to X to deal with the categorical data in it.



Perform feature scaling on the Independent variable.

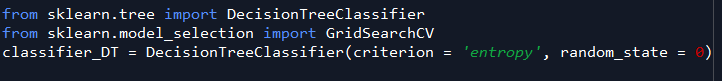


Split the X and y variable into training and test set

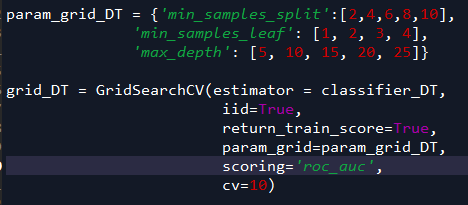


**Decision Tree:**

The libraries are imported.



The model is optimized using grid search.



The optimized model is fitted on the training set of the X and y variables.



The best estimator instance is used on the fitted optimized model



The best estimator model is then fitted to the training set of X and y.



The test set of X is the predicted. This will produce a result showing the probabilities of both an employee leaving or not.

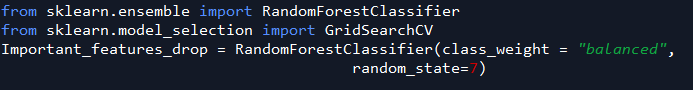


But we want the probabilities of an employee leaving.

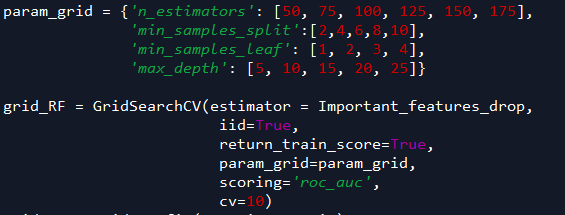


**Random Forest regression:**

The necessary libraries are first imported.



The model is optimized using grid search.



The optimized model is fitted on the training set of the X and y variables.



The best estimator instance is used on the fitted optimized model



The best estimator model is then fitted to the training set of X and y



The test set of X is the predicted. This will produce a result showing the probabilities of both an employee leaving or not.

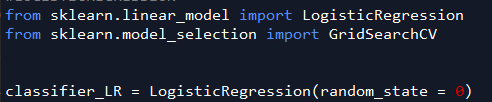


But we want the probabilities of an employee leaving.

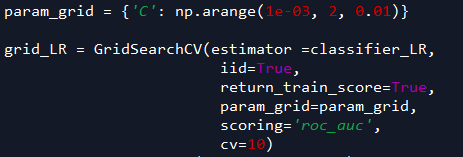


**Logistic Regression:**

The necessary libraries are imported.



The model is optimized using grid search.



The optimized model is fitted on the training set of the X and y variables.



The best estimator instance is used on the fitted optimized model



The best estimator model is then fitted to the training set of X and y. This will produce a result showing the probabilities of both an employee leaving or not.



The test set of X is the predicted. This will produce a result showing the probabilities of both an employee leaving or not.



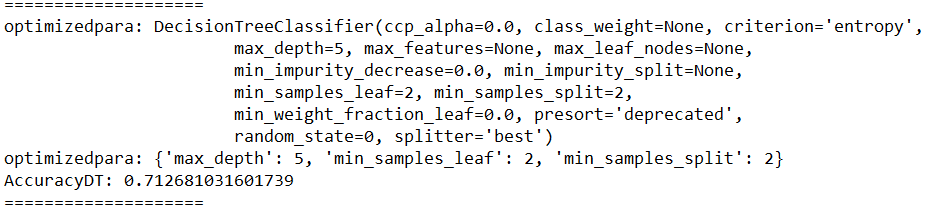
But we want the probabilities of an employee leaving.



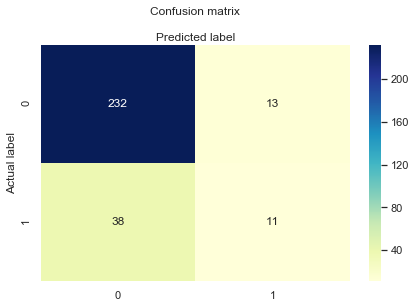
**Best Model:**

To find the best model of the three, we must do some analysis. Three techniques were used to determine which was the best model of all, there are; Accuracy, Confusion matrix and ROC curve.

**Decision Tree:**

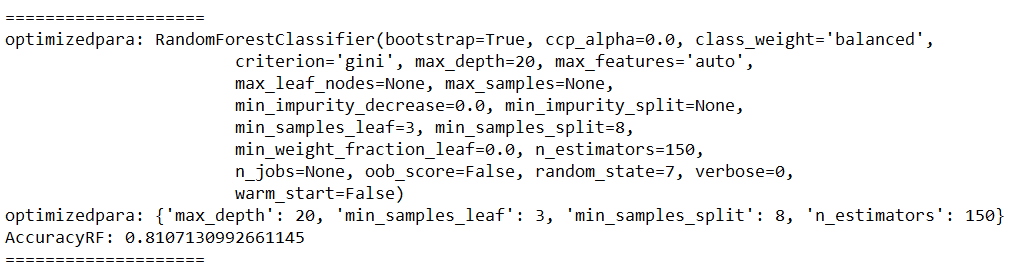


Decision tree showed an accuracy of 71%

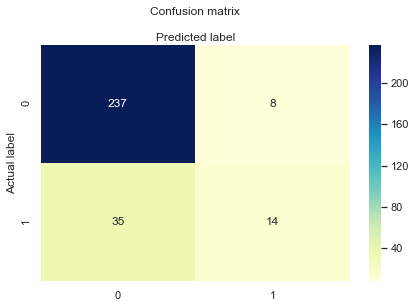


The confusion matrix tells us that the random forest had 232+ 11 predictions correctly and 13 + 38 predictions incorrectly. Leading to the accuracy we got above; 71%

**Random Forest:**

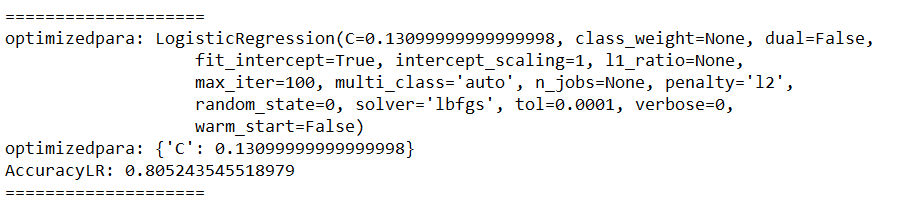


The random forest had an accuracy or best score of 81%

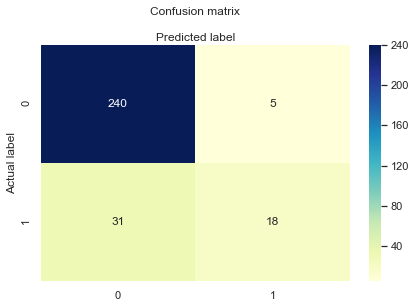


The confusion matrix tells us that the random forest had 237 + 14 predictions correctly and 8 + 35 predictions incorrectly.

**Logistic Regression:**

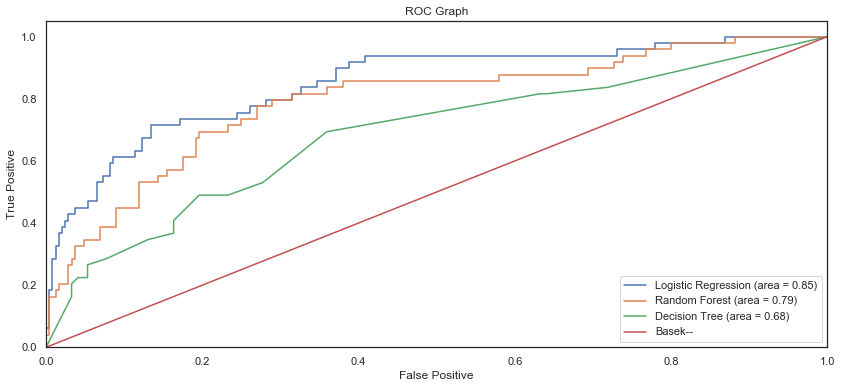


Logistic regression showed an accuracy of 80%



The confusion matrix tells us that the random forest had 240 + 18 predictions correctly and 5+ 31 predictions incorrectly

We plot the ROC curve,



From our analysis, we see that logistic Regression performed the best of all three models having an AUC score of 0.85.

**BLOCK DIAGRAM**

Block Diagram of modelling approach between *“company would like to predict which valuable employee (key talent) will leave”*  Vs *“company would like to predict which valuable employee (key talent) will leave* ***next****”*

