HR CASE STUDY ANALYSIS

Name of Student (1): Nikhil Shah (B2020092)

Name of Student (2): Sam John (B2020107)

Name of Student (3): Souvik Das (B2020118)

Name of Student (4): Varsha Satpathy (B2020122)

**Case Illustration:**

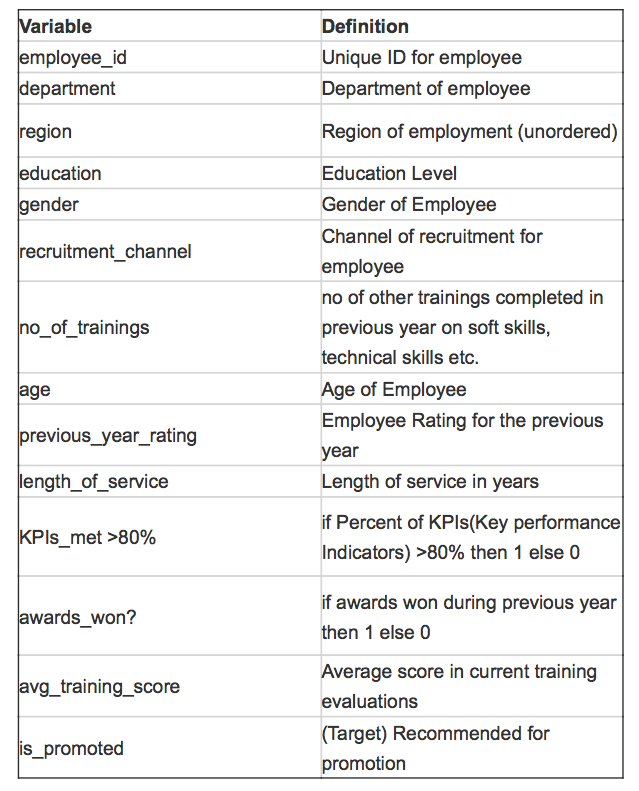
Your client is a large MNC and they have 9 broad verticals across the organisation. One of the problems your client is facing is around identifying the right people for promotion (only for manager position and below) and prepare them in time. Currently the process, they are following is:

They first identify a set of employees based on recommendations/ past performance

Selected employees go through the separate training and evaluation program for each vertical. These programs are based on the required skill of each vertical

At the end of the program, based on various factors such as training performance, KPI completion (only employees with KPIs completed greater than 60% are considered) etc., employee gets promotion

For above mentioned process, the final promotions are only announced after the evaluation and this leads to delay in transition to their new roles. Hence, company wants to design some model which help in identifying the eligible candidates at a particular checkpoint so that they can expedite the entire promotion cycle.



1. **Describe the problem and dataset.**

**Problem Statement:** Identifying the right people who should be promoted.

The main factors for decision making are:

**KPI’s completion (only employees with KPIs completed greater than 60% are considered)**

**Training performance and score (Average Training Score)**

**Previous year rating**

Mainly on the basis of these factors we will build a model to determine which employees should be promoted and which employees should not.

We are given two sets of datasets.

One is historical dataset (promotion\_tr) where it was already given which employees were promoted based on certain factors (such as above mentioned), on the basis of which we have to build a model and also train the model within the given dataset.

The other dataset is the new dataset (promotion\_ts) where after building our model we have to determine which employees should get promoted.

1. **List the variable as per your understanding of the case which will helpful in decision making for promotion:**

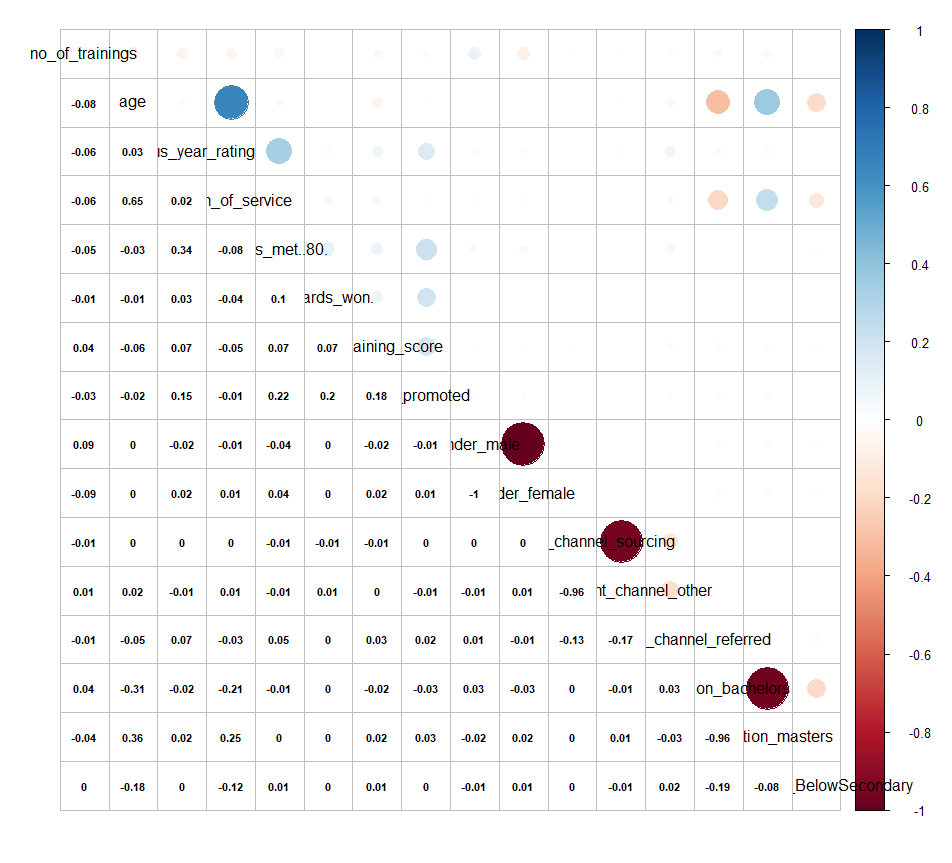
**Answer:** The variables which can be useful for in decision making for promotion are

* no\_of\_trainings
* age
* previous\_year\_rating
* length\_of\_service
* KPIs\_met..80.
* awards\_won.
* avg\_training\_score
* education
* gender

Leaving “Department” and “Region” variables from consideration into building the model as suggested.

1. **Find correlation of "is\_promoted" variable with all the other variables in the dataset.**

**Answer:** The correlation of “is\_promoted” with all the other variables are shown in the diagram below.



We considered all the variables except department and region from the given datasets as suggested.

1. **Develop a Logistic regression Model and discuss the results?**

**Answer:**

**First developing a Logistic Regression Model**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Sets Of Variables** | **Residual Deviance** | **AIC** |
| model 1 | no of trainings | 21611 | 21615 |
| model 2 | no of trainigs+age | 21593 | 21599 |
| model 3 | no of trainings+age+\_previous year rating | 20672 | 20680 |
| model 4 | no of trainings+age+\_previous year rating +length of service | 20670 | 20680 |
| model 5 | no of trainings+age+\_previous year rating +length of service+KPI met 80 | 19571 | 19583 |
| model 6 | no of trainings+age+\_previous year rating +length of service+KPI met 80+awards\_won | 18975 | 18989 |
| model 7 | no of trainings+age+\_previous year rating +length of service+KPI met 80+awards\_won+avg training score | 18053 | 18069 |
| model 8 | no of trainings+age+\_previous year rating +length of service+KPI met 80+awards\_won+avg training score+education | 18025 | 18045 |
| model 9 | no of trainings+age+\_previous year rating +length of service+KPI met 80+awards\_won+avg training score+education+recruitment channel | 18023 | 18047 |
| model 10 | no of trainings+age+\_previous year rating +length of service+KPI met 80+awards\_won+avg training score+education+recruitment channel+gender | 18021 | 18047 |
| **model 11** | **model10 - recruitment** | **18022** | **18044** |
| model 12 | model11 - age | 18043 | 18063 |
| model 13 | model12+age- no of training | 18044 | 18064 |
| model 14 | model 13 - age+no of training -gender | 18046 | 18064 |
| model 15 | model14- education below secondary | 18046 | 18064 |
| model 16 | model15- length of service | 18046 | 18062 |

After building many models, the best model that have come up is hereby **model 11,** shown below with the certain sets of independent variables predicting the dependent variable “is\_promoted”.

glm(formula = is\_promoted ~ no\_of\_trainings + age + previous\_year\_rating +

length\_of\_service + KPIs\_met..80. + awards\_won. + avg\_training\_score +

education\_bachelors + education\_masters + education\_BelowSecondary +

gender\_male + gender\_female, family = binomial, data = HR\_TR)

Null deviance: 21638 on 36678 degrees of freedom

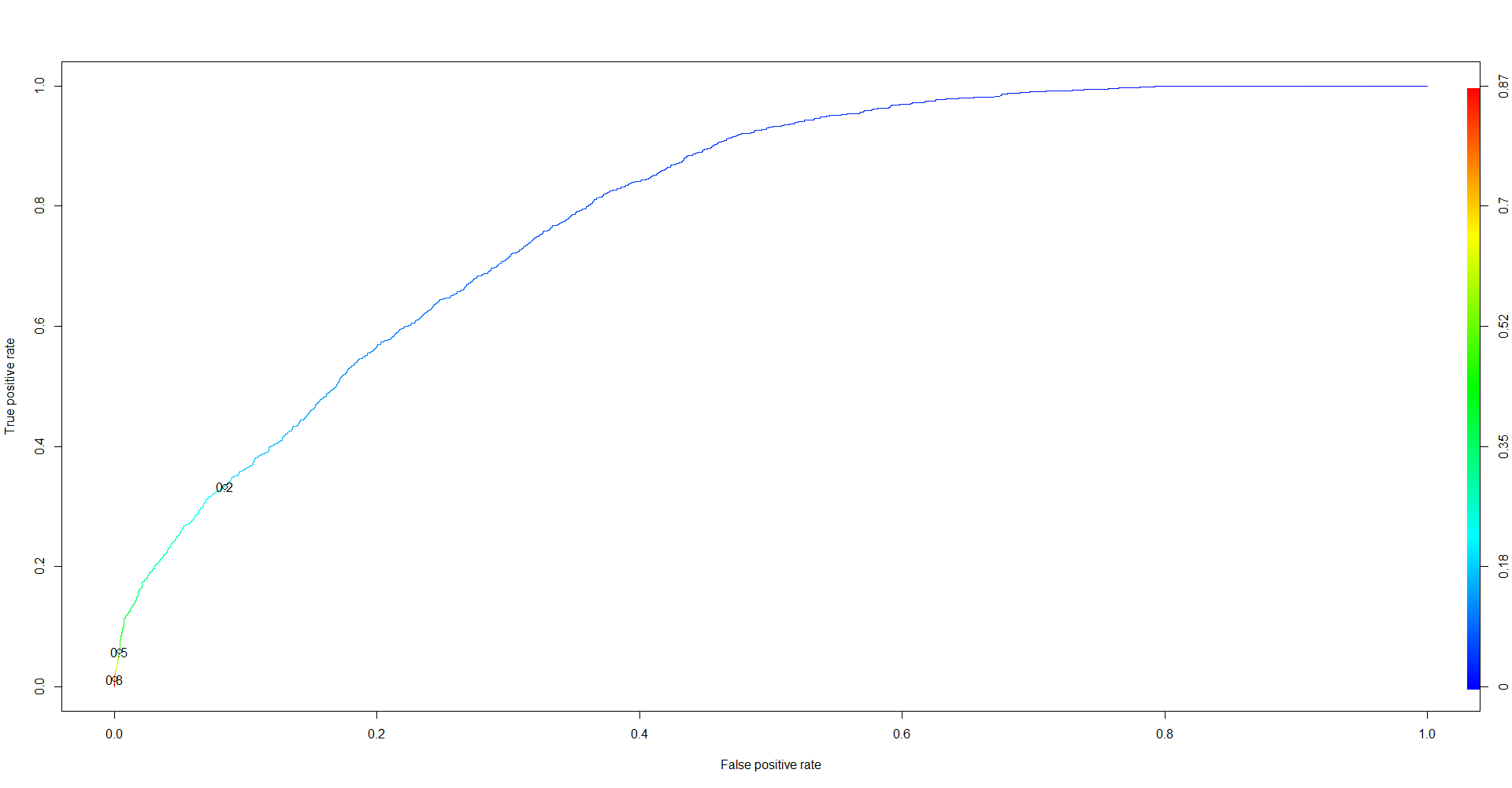
Residual deviance: 18022 on 36668 degrees of freedom

AIC: 18044

The model 11 is giving **the least AIC (= 18044) and Residual Deviance (= 18022)** among all the other models thus making it the best model for prediction.

1. **With the help of ROCR Curve build the confusion matrix using different threshold values. Based on the CF Matrix you have built, write the best threshold value suitable for the dataset and also accuracy of the best fit model.**

**Answer:** **The ROCR curve** is hereby given by;



Now, checking the confusion matrix with different threshold values to determine the best threshold value.

|  |  |  |
| --- | --- | --- |
| **Threshold Value (>)** | **Confusion Matrix** | **Accuracy (%)** |
| > 0.2 | PredictedValue Actualvalue FALSE TRUE  0 13149 1207  1 910 454 | 86.53% |
| > 0.3 | PredictedValue Actualvalue FALSE TRUE  0 13899 457  1 1086 278 | 90.18% |
| > 0.4 | PredictedValue Actualvalue FALSE TRUE  0 14255 101  1 1220 144 | 91.59% |
| > 0.45 | PredictedValue Actualvalue FALSE TRUE  0 14289 67  1 1252 112 | 91.6% |
| > 0.5 | PredictedValue Actualvalue FALSE TRUE  0 14303 53  1 1284 80 | 91.49% |
| > 0.6 | PredictedValue Actualvalue FALSE TRUE  0 14324 32  1 1309 55 | 91.46% |

From the table, the threshold value **0.45 is giving the best accuracy** for predicting, so we are taking 0.45 threshold value for our model predicting.

1. **Develop a Decision Tree using same set of variables and same training dataset. Draw the tree or write the rules?**

**Answer:** Building a Decision Tree using the same sorts of variables and same training dataset

model20 <- rpart(is\_promoted ~ no\_of\_trainings+age+previous\_year\_rating+length\_of\_service

+KPIs\_met..80.+awards\_won.+avg\_training\_score+education\_bachelors

+education\_masters+education\_BelowSecondary

+gender\_male+gender\_female,

data=HR\_TR,

method = "class")

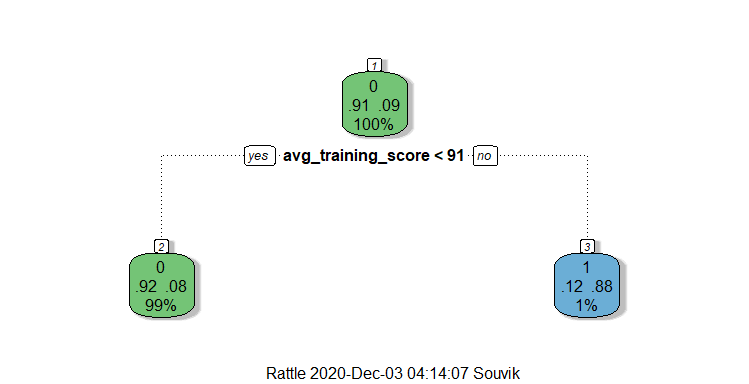
**Output:**

1) root 36679 3182 0 (0.91324736 0.08675264)

2) avg\_training\_score< 90.5 36267 2820 0 (0.92224336 0.07775664) \*

3) avg\_training\_score>=90.5 412 50 1 (0.12135922 0.87864078) \*

**Decision Tree Diagram:**



**Confusion Matrix and Statistics**

pred2 0 1

0 14340 1232

1 16 132

Accuracy : 0.9206

95% CI : (0.9163, 0.9248)

No Information Rate : 0.9132

P-Value [Acc > NIR] : 0.0004632

The model is giving 92.06% accuracy.

1. **Use “Information Gain” and “Gini Index” as splitting criteria to build Decision Tree and write the confusion matrix for both. Also discuss which splitting criteria you will choose for this dataset.**

**Answer:**

**Splitting Criteria: Information Gain**

model18 <- rpart(is\_promoted ~ no\_of\_trainings+age+previous\_year\_rating+length\_of\_service

+KPIs\_met..80.+awards\_won.+avg\_training\_score+education\_bachelors

+education\_masters+education\_BelowSecondary

+gender\_male+gender\_female,

data=HR\_TR,

method = "class",

parms = list(split = "information"))

**Output:**

1) root 36679 3182 0 (0.91324736 0.08675264)

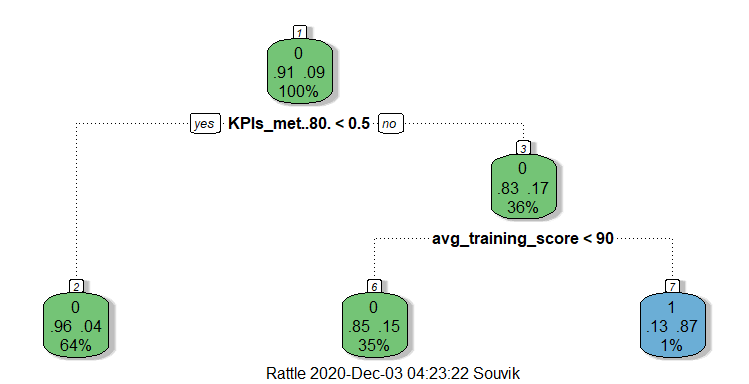
2) KPIs\_met..80.< 0.5 23589 938 0 (0.96023570 0.03976430) \*

3) KPIs\_met..80.>=0.5 13090 2244 0 (0.82857143 0.17142857)

6) avg\_training\_score< 89.5 12772 1968 0 (0.84591293 0.15408707) \*

7) avg\_training\_score>=89.5 318 42 1 (0.13207547 0.86792453) \*

**Decision Tree Diagram:**



**Confusion Matrix and Statistics**

pred 0 1

0 14341 1267

1 15 97

Accuracy : 0.9184

95% CI : (0.9141, 0.9227)

No Information Rate : 0.9132

P-Value [Acc > NIR] : 0.009993

The Decision Tree model with “Information Gain” as splitting criteria is giving **91.84% accuracy**.

**Splitting Criteria: Gini Index**

model19 <- rpart(is\_promoted ~ no\_of\_trainings+age+previous\_year\_rating+length\_of\_service

+KPIs\_met..80.+awards\_won.+avg\_training\_score+education\_bachelors

+education\_masters+education\_BelowSecondary

+gender\_male+gender\_female,

data=HR\_TR,

method = "class",

parms = list(split = "gini"))

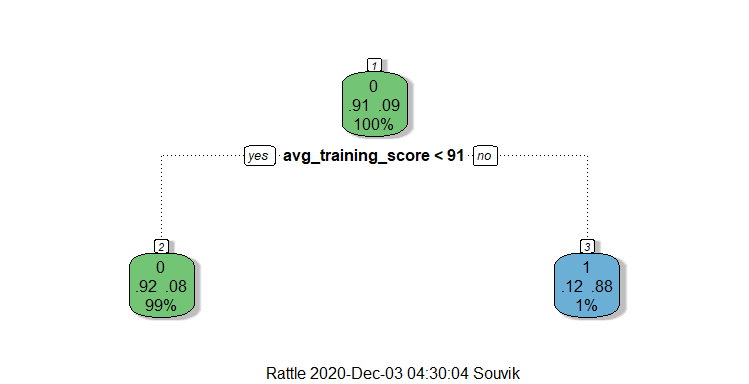
**Output:**

1) root 36679 3182 0 (0.91324736 0.08675264)

2) avg\_training\_score< 90.5 36267 2820 0 (0.92224336 0.07775664) \*

3) avg\_training\_score>=90.5 412 50 1 (0.12135922 0.87864078) \*

**Decision Tree Diagram:**



**Confusion Matrix and Statistics**

pred0 0 1

0 14340 1232

1 16 132

Accuracy : 0.9206

95% CI : (0.9163, 0.9248)

No Information Rate : 0.9132

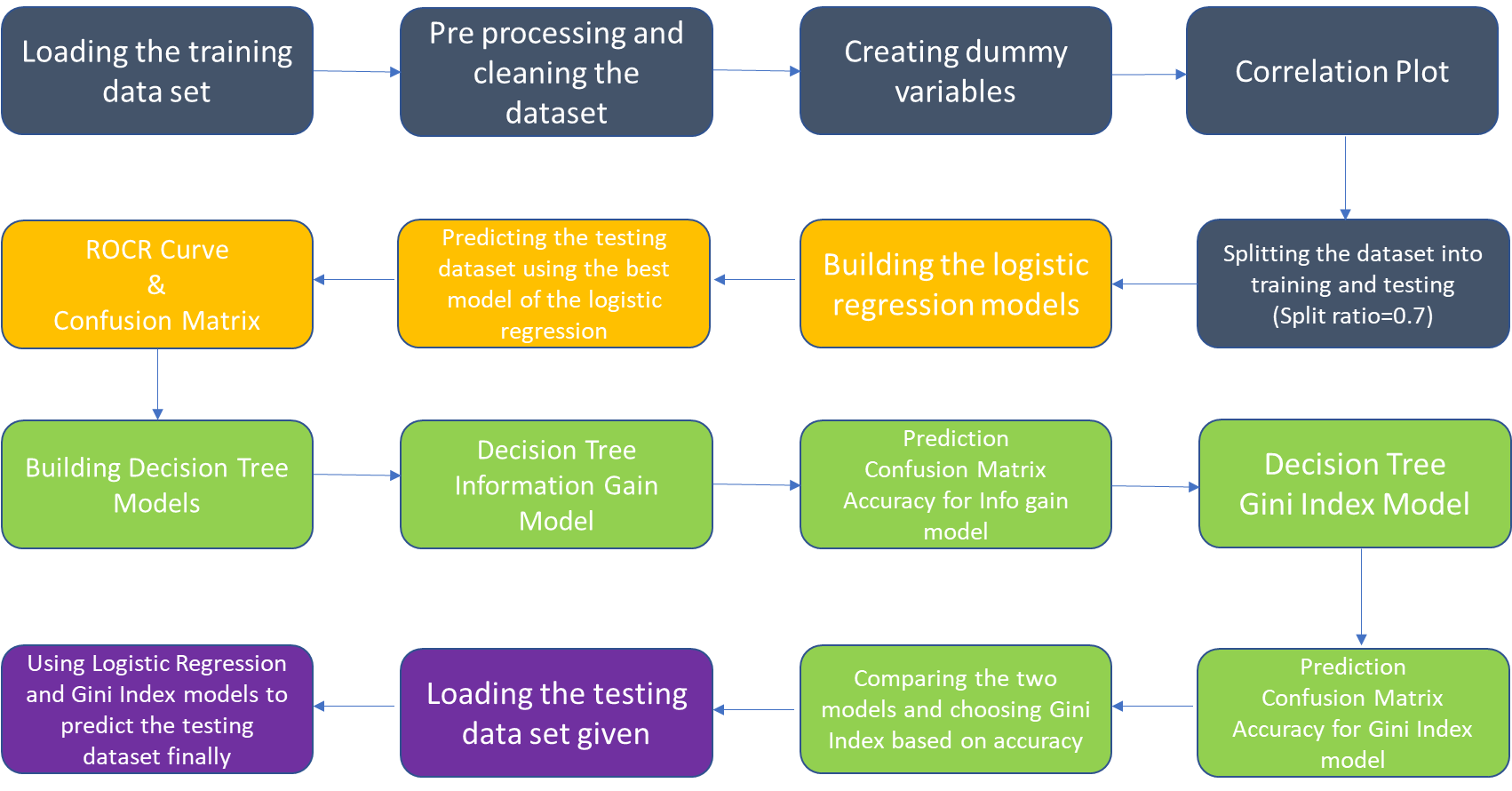
P-Value [Acc > NIR] : 0.0004632

The Decision Tree model with “Gini Index” as splitting criteria is giving **92.06% accuracy**.

So, among the two splitting criteria models **“Gini Index”** is giving a **slightly higher accuracy** than **“Information gain”** in predicting the dataset, so we will consider the decision tree model with splitting criteria as **“Gini Index”** for predicting this dataset as well will take this model predicting the given testing data set.

1. **Draw the process flow diagram. Also, using your best model predict the final number of employees that are being promoted in the given testing dataset?**

**Answer: Process flow diagram**



Among the three models we built, Gini Index model is giving the best accuracy but for comparing reason we have done prediction with Gini index model and Logistic Regression model.

The final number of employees that are being promoted is **259.**

(by using the **Logistic Regression Model** on the **testing data set**)

The final number of employees that are being promoted is **231**.

(by using the **Decision Tree Gini Index Model** on the **testing data set**)

**In the end we will take the prediction of Gini Index model into consideration for promoting employees.**