HR CASE – EMPLOYEE PROMOTION
PROJECT ANALYSIS REPORT

Sarthak Agarwal
25 Years | +91-9123723044 | sarthak.agarwal20b@gim.ac.in
https://www.linkedin.com/in/sarthak-agarwal20

Case Illustration

Your client is a large MNC, and they have 9 broad verticals across the organization. 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. Theseprograms 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 thisleads 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.

Variable	Definition
employee_id	Unique ID for employee
department	Department of employee
region	Region of employment (unordered)
education	Education Level
gender	Gender of Employee
recruitment_channel	Channel of recruitment for employee
no_of_trainings	no of other trainings completed in previous year on soft skills, technical skills etc.
age	Age of Employee
previous_year_rating	Employee Rating for the previous year
length_of_service	Length of service in years
KPIs_met >80%	if Percent of KPIs(Key performance Indicators) >80% then 1 else 0
awards_won?	if awards won during previous year then 1 else 0
avg_training_score	Average score in current training evaluations
is_promoted	(Target) Recommended for promotion

CASE SOLUTION

1. Describe the problem and dataset.

The problem is regarding a large MNC which is facing issues identifying the right people who are eligible for promotion for the manager level and below. The current process is a lengthy and rigorous one in which shortlisted employees based on past performance and recommendations are made to go through separate training and evaluations for each vertical, based on the required skillset. And at the end of the process, only those employees get promoted who have KPI completion rate greater than 60%. But the promotion decision is made only after the evaluation and this leads to delay in the transition into new roles.
We have to build a model which can predict the promotion of employees and hence boost the promotion cycle faster.

We had been given a training dataset which contains the employee details who were promoted and not promoted and using that we have to build a model and check its accuracy. If the model is accurate enough, we can use it to predict the promotions of the employees given in the testing dataset.

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

After the Deep Dive Exploratory Data Analysis of the "promotion_tr" Dataset we found few interesting facts about the variables in the dataset and their effect on the is_promoted variable,

a) The Number of Trainings an employee is getting has no relation with respect to its promotion.
After filtering the employee who has taken trainings more than 4 wrt their promotion we get to know only 4 employees out of 172.

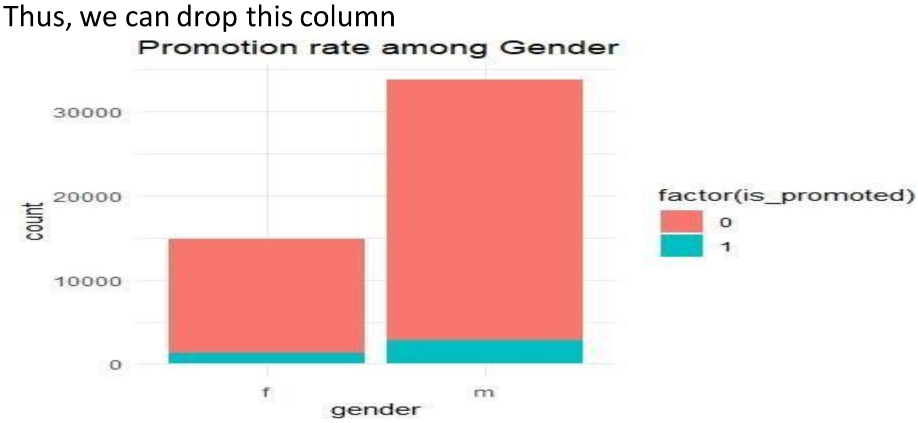
is_promoted	**count**
0	168
1	4

Thus, we can drop this column

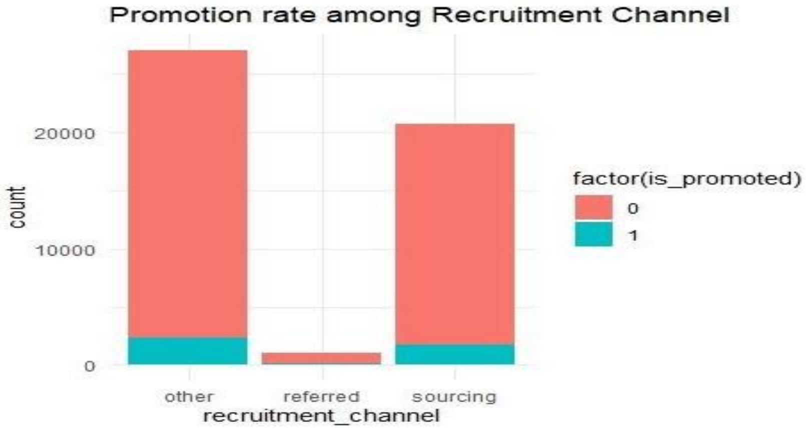
b) Gender has no role w.r.t. to the employee's promotion
After filtering the employee based on the gender and the promotion, we got the following result

is_promoted	**gender**	**count**
0	f	13445
0	m	30983
1	f	1363
1	m	2869

Thus, we can drop this column



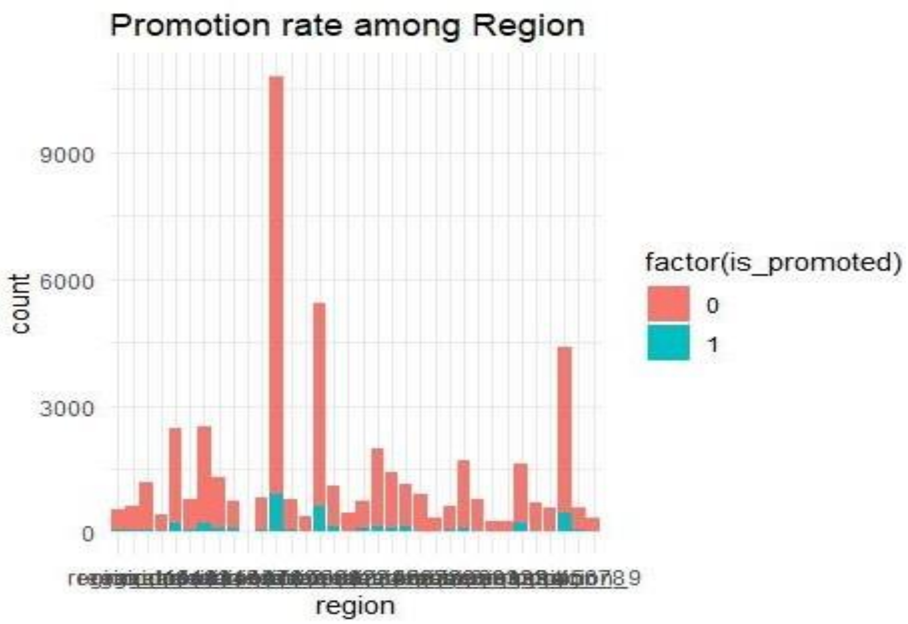
c) Recruitment Channel also does not play a major role in individuals' promotion.



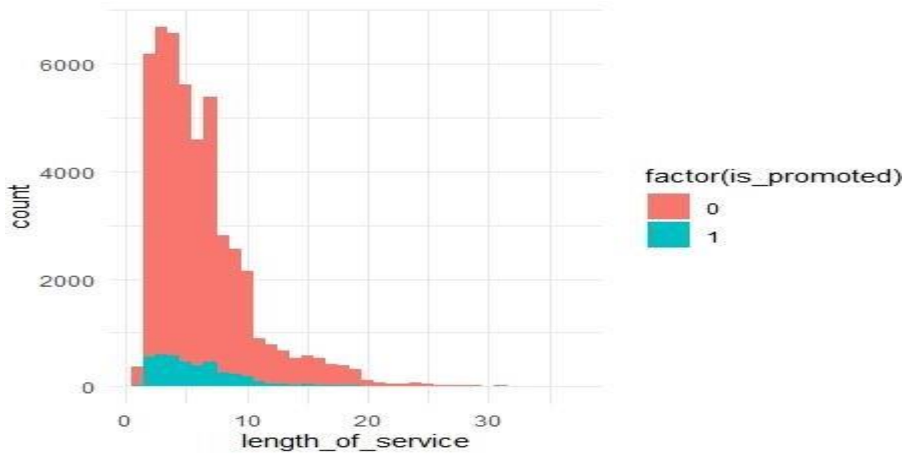
Thus, we can drop this column

d) Employee_ID has no relevance wrt the modelling purpose as it is just for the reference purpose.

e) Region can also be dropped off as it has not important for the model



f) Length of service is also not helpful/relevant wrt the promotion of an employee,

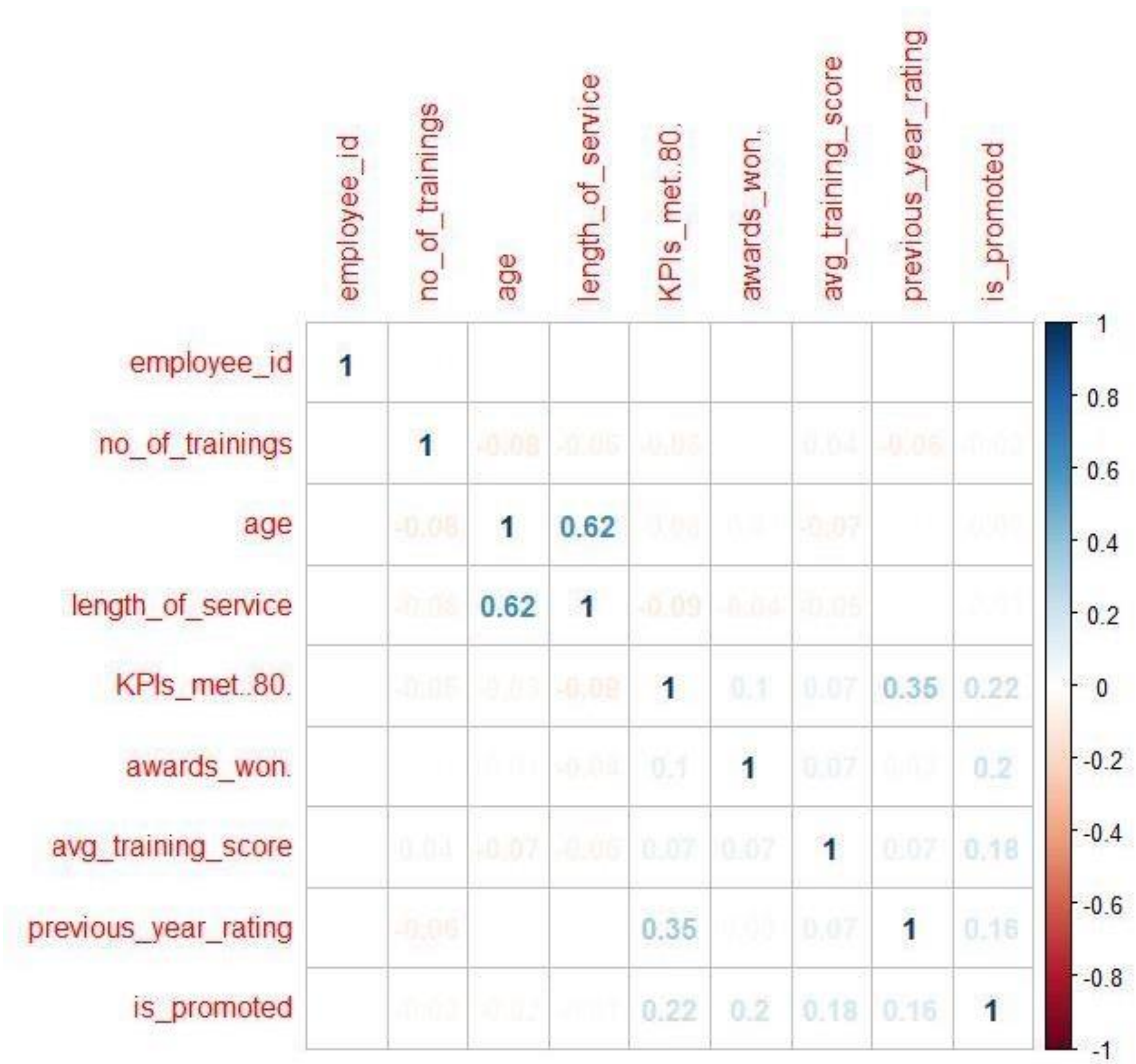


As we can see length of service is positively skewed or skewed toward right, we can say the promotion does not depends on the length of service an employee is serving for. Thus, we can drop it off.

Now coming on to the helpful or relevant variables in the dataset, are

a) KPI met 80
b) Award won
c) Previous year Rating
d) Average Training Score
e) Department
f) Education
g) Age

3. Find correlation of "is_promoted" variable with all the other variables in the dataset.

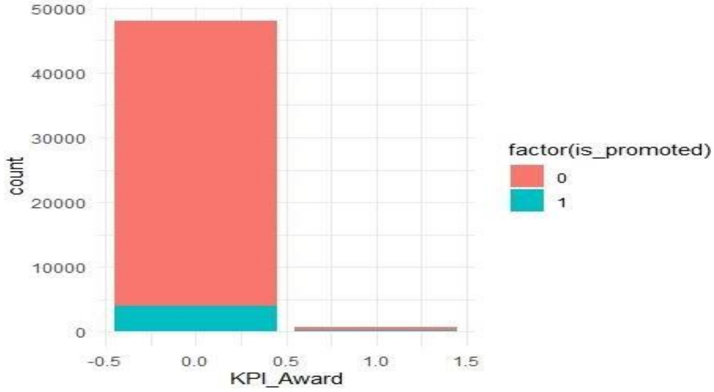


We see from the correlation plot that almost all the variables are not directly correlated with promotion rate. The variables with the highest correlation (0.22 and 0.20 respectively) is whether the KPI is met and whether an award was won.

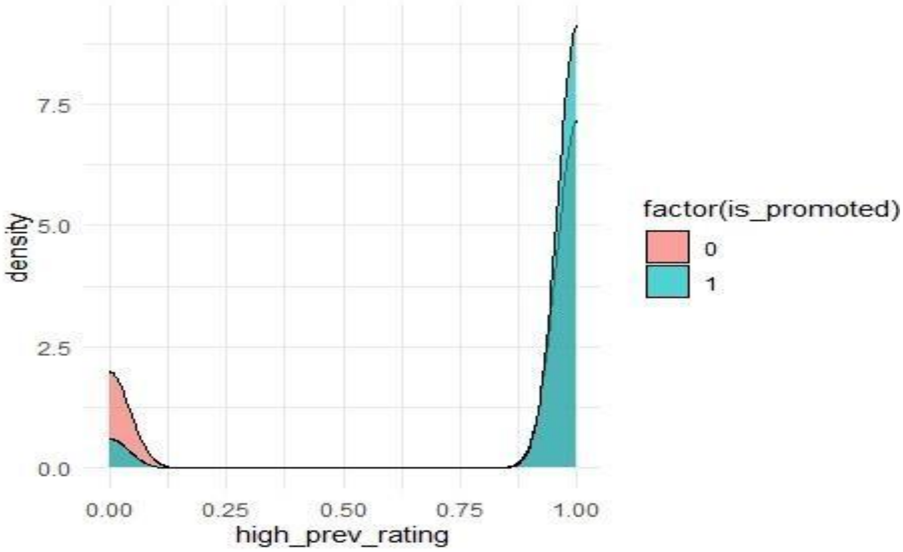
CASE SOLUTION

After looking at the correlation between the variables wrt to the is_promoted (Response Variable) we did some feature Engineering.

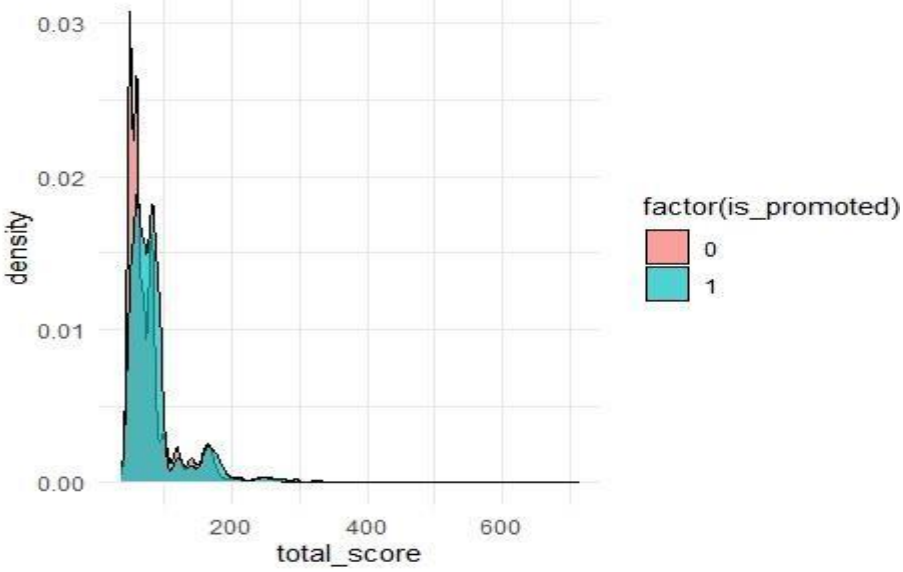
a) We have added another variable that states if an employee has won both awards and met KPI- **KPI_Award**. (Exhibit 3.1)



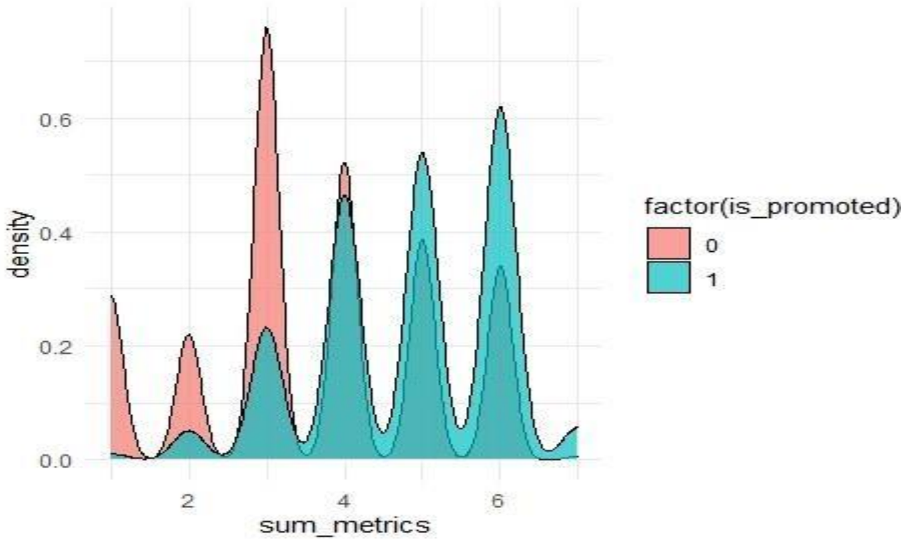
b) We created two columns based on the previous year rating that is,
High previous year Rating: if previous year rating is greater than equal to 3.
Low previous year Rating: otherwise



c) **Total Score** Column which is derived from the avg_training_score and No_of_trainings That is **avg_training_score * No_of_trainings.**



d) **Metrics of sum** column, which is derived from previous year rating, award won and KPI_met_80, that is **previous year rating + award won + KPI_met_80.**



4. Develop a Logistic regression Model and discuss the results?

Formed a Logistic Regression model taking following relevant variables and performing the one hot on the categorical variables,

```
> colnames(promotion_training_final_Ltr)
 [1] "age"                          "KPIs_met..80."                "awards_won."
 [4] "avg_training_score"           "is_promoted"                  "KPI_Award"
 [7] "sum_metrics"                  "total_score"                  "high_prev_rating"
[10] "low_prev_rating"              "educationBachelor's"          "educationBelow Secondary"
[13] "educationMaster's & above"    "departmentAnalytics"          "departmentFinance"
[16] "departmentHR"                 "departmentLegal"              "departmentOperations"
[19] "departmentProcurement"        "departmentR&D"                "departmentSales & Marketing"
[22] "departmentTechnology"
> |
```

Performing the splitting of promotion_tr data set into 70:30 ratio of training and testing dataset in order to avoid overfitting which can occur if we do the train and test on the entire dataset.

```
> model1_logi <- glm(is_promoted~.,data = promotion_training_final_Ltr,family = binomial)
> summary(model1_logi)

Call:
glm(formula = is_promoted ~ ., family = binomial, data = promotion_training_final_Ltr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6647  -0.3676  -0.2009  -0.1224   3.3008

Coefficients: (3 not defined because of singularities)
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -2.835e+01  5.620e-01 -50.449  < 2e-16 ***
age                            -1.709e-02  3.566e-03  -4.791 1.66e-06 ***
KPIs_met..80.1                  1.779e+00  6.915e-02  25.733  < 2e-16 ***
awards_won.1                    2.628e+00  1.951e-01  13.473  < 2e-16 ***
avg_training_score              3.077e-01  6.504e-03  47.312  < 2e-16 ***
KPI_Award1                     -1.830e+00  2.281e-01  -8.026 1.01e-15 ***
sum_metrics                     1.824e-01  2.811e-02   6.487 8.78e-11 ***
total_score                    -1.144e-03  6.089e-04  -1.878   0.0604 .
high_prev_rating1               4.512e-01  1.123e-01   4.020 5.82e-05 ***
low_prev_rating1                       NA         NA      NA       NA
`educationBachelor's`1         -2.431e-01  5.269e-02  -4.613 3.98e-06 ***
`educationBelow Secondary`1    -2.446e-01  2.399e-01  -1.020   0.3078
`educationMaster's & above`1           NA         NA      NA       NA
departmentAnalytics1           -1.674e+00  9.318e-02 -17.962  < 2e-16 ***
departmentFinance1              5.411e+00  1.764e-01  30.673  < 2e-16 ***
departmentHR1                   8.227e+00  2.414e-01  34.076  < 2e-16 ***
departmentLegal1                5.335e+00  2.481e-01  21.500  < 2e-16 ***
departmentOperations1           5.655e+00  1.461e-01  38.702  < 2e-16 ***
```

CASE SOLUTION

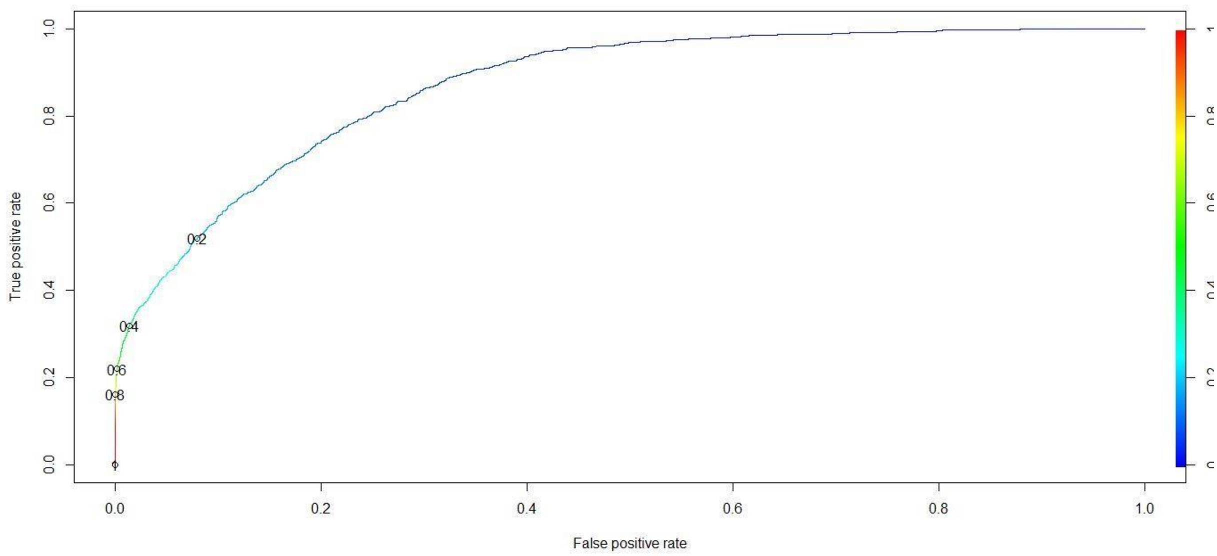
```
departmentAnalytics1          -1.674e+00  9.318e-02 -17.962  < 2e-16 ***
departmentFinance1             5.411e+00  1.764e-01  30.673  < 2e-16 ***
departmentHR1                  8.227e+00  2.414e-01  34.076  < 2e-16 ***
departmentLegal1               5.335e+00  2.481e-01  21.500  < 2e-16 ***
departmentOperations1          5.655e+00  1.461e-01  38.702  < 2e-16 ***
departmentProcurement1         2.726e+00  1.035e-01  26.328  < 2e-16 ***
`departmentR&D`1              -2.182e+00  1.833e-01 -11.903  < 2e-16 ***
`departmentSales & Marketing`1 8.797e+00  2.052e-01  42.874  < 2e-16 ***
departmentTechnology1                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20127  on 34061  degrees of freedom
Residual deviance: 13612  on 34043  degrees of freedom
AIC: 13650

Number of Fisher Scoring iterations: 7

> 
```

After performing lot of feature engineering and removal of irrelevant columns we get the model with least AIC and Residual Deviance.

Variable	Residual Deviance	AIC
Age	20113	20117
Age + KPI_met_80	18506	18512
Age + KPI_met_80+ Awards_won.	17967	17975
age+KPIs_met..80.+awards_won.+avg_training_score	17135	17145
age+KPIs_met..80.+awards_won.+avg_training_score+KPI_Award	16944	16956
age+KPIs_met..80.+awards_won.+avg_training_score +KPI_Award+total_score	16924	16938
age+KPIs_met..80.+awards_won.+avg_training_score +KPI_Award+total_score+sum_metrics	16661	16667
age+KPIs_met..80.+awards_won.+avg_training_score +KPI_Award+total_score+sum_metrics+ high_prev_rating	16650	16668
age+KPIs_met..80.+awards_won.+avg_training_score +KPI_Award+total_score+sum_metrics+ high_prev_rating + departmentAnalytics	16391	16411
age+KPIs_met..80.+awards_won.+avg_training_score +KPI_Award+total_score+sum_metrics+ high_prev_rating+ departmentAnalytics+ departmentFinance	16391	16413
age+KPIs_met..80.+awards_won.+avg_training_score +KPI_Award+total_score+sum_metrics+ high_prev_rating+ departmentAnalytics+ departmentFinance+ departmentHR	16388	16412
age+KPIs_met..80.+awards_won.+avg_training_score +KPI_Award+total_score+sum_metrics+ high_prev_rating+ departmentAnalytics+ departmentFinance+ departmentHR+departmentLegal	16387	16413
age+KPIs_met..80.+awards_won.+avg_training_score +KPI_Award+total_score+sum_metrics+ high_prev_rating+ departmentAnalytics+ departmentFinance+ departmentHR+ departmentLegal+ departmentOperations	16359	16387
age+KPIs_met..80.+awards_won.+avg_training_score +KPI_Award+total_score+sum_metrics+ high_prev_rating+ departmentAnalytics+ departmentFinance+ departmentHR+ departmentLegal+ departmentOperations+departmentProcurement+ departmentTechnology	15874	15906
age+KPIs_met..80.+awards_won.+avg_training_score +KPI_Award+total_score+sum_metrics+ high_prev_rating+ departmentAnalytics+ departmentFinance+ departmentHR+ departmentLegal+ departmentOperations+departmentProcurement+ departmentTechnology+departmentSales & Marketing+educationBachelor's+educationBelow Secondary	13612	13650

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

Performing prediction on the 30% of training dataset which we have sliced and
Forming ROCR Curve on the predicted values from the above model to find the threshold value,



After forming the ROCR curve we observed, between 0.8 - 0.4 our threshold value lies as the TPR is maximum and FPR is minimum thus the requirement for better accuracy.

#Accuracy testing Using Confusion Matrix - (Threshold = 0.8)

```
> table(promotion_training_final_Lts$is_promoted,promotion_training_final_Lts$predicted)

       0     1
  0 13328     0
  1  1065   205
> misClasificError <- mean(promotion_training_final_Lts$predicted != promotion_training_final_Lts$is_pro
moted)
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.927044800657624"
> 
```

#Accuracy testing Using Confusion Matrix - (Threshold Value = 0.7)

```
> table(promotion_training_final_Lts$is_promoted,promotion_training_final_Lts$predicted)

       0     1
  0 13321     7
  1  1038   232
>
> #Accuracy
> misClasificError <- mean(promotion_training_final_Lts$predicted != promotion_training_final_Lts$is_pr
moted)
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.9284148513495"
> 
```

#Accuracy testing Using Confusion Matrix - (Threshold = 0.6)

```
> table(promotion_training_final_Lts$is_promoted,promotion_training_final_Lts$predicted)

       0     1
  0 13304    24
  1   991   279
>
> #Accuracy
> misClasificError <- mean(promotion_training_final_Lts$predicted != promotion_training_final_Lts$is_
moted)
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.930469927387313"
> 
```

#Accuracy testing Using Confusion Matrix - (Threshold = 0.5)

```
> table(promotion_training_final_Lts$is_promoted,promotion_training_final_Lts$predicted)

        0     1
  0 13258    70
  1   942   328
>
> #Accuracy
> misClasificError <- mean(promotion_training_final_Lts$predicted != promotion_training_final_Lts$is_promoted)
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.930675434991095"
```

#Accuracy testing Using Confusion Matrix - (Threshold = 0.4)

```
> table(promotion_training_final_Lts$is_promoted,promotion_training_final_Lts$predicted)

        0     1
  0 13149   179
  1   864   406
>
> #Accuracy
> misClasificError <- mean(promotion_training_final_Lts$predicted != promotion_training_final_Lts$is_promoted)
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.92855185641868"
```

Threshold	Accuracy
0.8	92.70%
0.7	92.84%
0.6	93.04%
0.5	93.06%
0.4	92.85%

The Threshold Value from the ROCR is 0.5 giving Accuracy of 93.06% for the best fit model.

Confusion Matrix

	0	1
0	13258	70
1	942	348

Result after applying the above Logistic Regression Model on the promotion_ts to predict the promotion of an employee we get,

0	1
2037	422

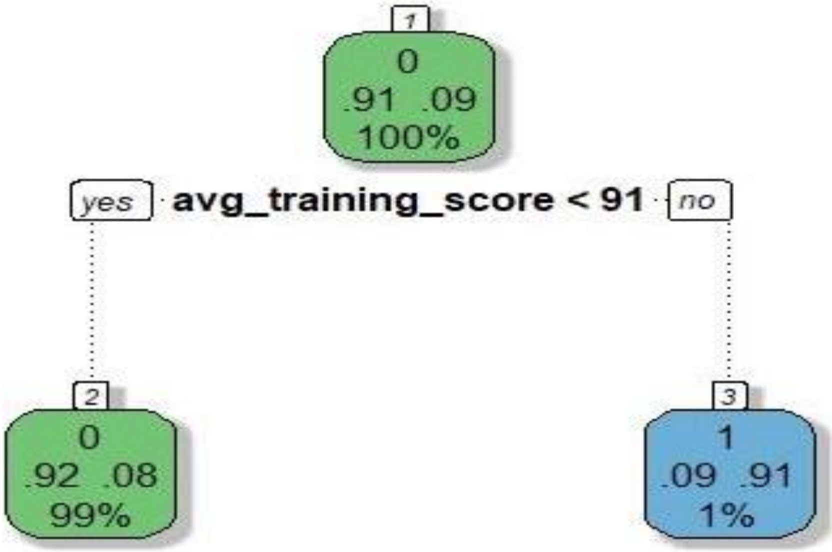
Means total of 432 employees get to promote as per the historical data and its modelling.

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

Forming Decision Tree on the below mentioned set of relevant variables,

```
> colnames(promotion_training_final_Dtr)
 [1] "age"                      "KPIs_met..80."          "awards_won."
 [4] "avg_training_score"       "is_promoted"            "KPI_Award"
 [7] "sum_metrics"              "total_score"            "high_prev_rating"
[10] "low_prev_rating"          "educationBachelor's"    "educationBelow Secondary"
[13] "educationMaster's & above" "departmentAnalytics"   "departmentFinance"
[16] "departmentHR"             "departmentLegal"        "departmentOperations"
[19] "departmentProcurement"    "departmentR&D"          "departmentSales & Marketing"
[22] "departmentTechnology"
> 
> model1_DT <- rpart(is_promoted~., data = promotion_training_final_Dtr, method = "class")
> model1_DT
n= 34062

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 34062 2962 0 (0.91304093 0.08695907)
  2) avg_training_score< 90.5 33703 2635 0 (0.92181705 0.07818295) *
  3) avg_training_score>=90.5 359   32 1 (0.08913649 0.91086351) *
```

Visualization of Decision Tree,



Confusion Matrix and Statistics

```
          Reference
Prediction     0     1
         0 13300  1140
         1    28   130
               Accuracy : 0.92
                 95% CI : (0.9155, 0.9243)
    No Information Rate : 0.913
    P-Value [Acc > NIR] : 0.00129
                  Kappa : 0.166
```

Kappa : 0.166

```
 Mcnemar's Test P-Value : < 2e-16
            Sensitivity : 0.9979
            Specificity : 0.1024
         Pos Pred Value : 0.9211
         Neg Pred Value : 0.8228
             Prevalence : 0.9130
         Detection Rate : 0.9111
   Detection Prevalence : 0.9892
      Balanced Accuracy : 0.5501
       'Positive' Class : 0
```

7. 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.

When Using Gini Index as the splitting criteria,

```
          Reference
Prediction     0     1
         0 13300  1140
         1    28  130
               Accuracy : 0.92
                 95% CI : (0.9155, 0.9243)
    No Information Rate : 0.913
    P-Value [Acc > NIR] : 0.00129
                  Kappa : 0.166
Mcnemar's Test P-Value : < 2e-16
```

Decision Tree,

When Using Information Gain as the splitting criteria,

Confusion Matrix and Statistics

```
          Reference
Prediction     0     1
         0 13300  1140
         1    28  130
               Accuracy : 0.92
                 95% CI : (0.9155, 0.9243)
    No Information Rate : 0.913
    P-Value [Acc > NIR] : 0.00129
                  Kappa : 0.166
 Mcnemar's Test P-Value : < 2e-16
```

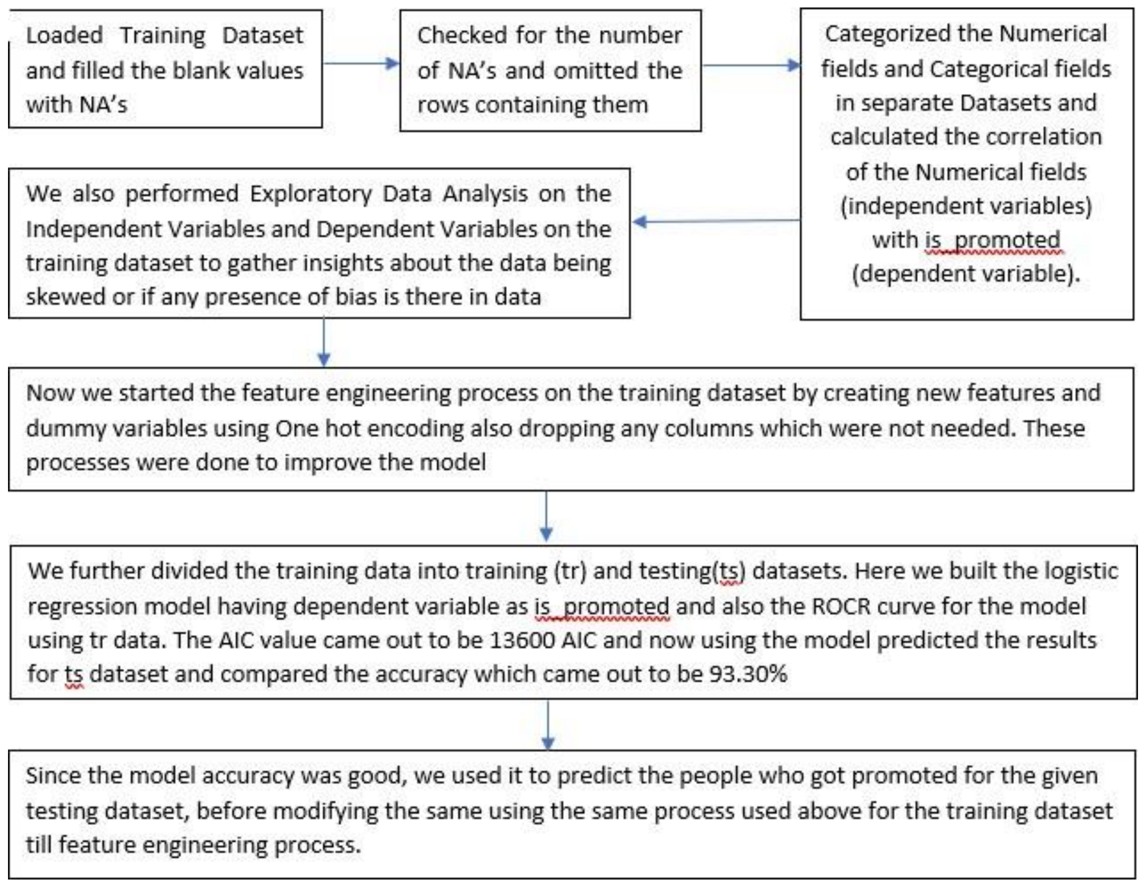
CASE SOLUTION

R Studio by Default takes "Gini Index" as the splitting criteria for Decision Tree.
Since the accuracy is same for Model irrespective of the splitting criteria, we use any one for the splitting purpose.
Result after applying the above Decision Tree Model on the promotion_ts to predict the promotion of an employee we get,

1 1
20603 216
Means total of 216 employees get to promote as per the historical data and its modelling.

8. 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?

Process Flow Diagram:



Best Model: Logistic Regression giving an accuracy of 93.06%
Result after applying the best Logistic Regression Model on the promotion_ts to predict the promotion of an employee we get,

0 1
2037 422
Means total of **432 employees get to promote as per the historical data and its modelling.**