HR Analytics Promotion Recommendation

Madhuri Dilliker
Siri Sirur Prakash
Siva Deepthi Gajula

Agenda

- Dataset Description
- Objective
- Exploratory Data Analysis
- Data Cleaning/Preprocessing
- Feature Engineering
- Models Used
- Results/Comparison
- Challenges
- Future Work

Dataset Description

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

Dataset Overview

Data Source : Analytics Vidhya

https://datahack.analyticsvidhya.com/contest/wns-analytics-hackathon-2018-1/

Data Dictionary:

Attributes : 14
Observations: 54808

	employee_id <int>	department <chr>	region <chr>	education <chr>	gender <chr>	recruitment_channel <chr>	no_of_trainings <int>	age <int>	previous_year_rating <int>	length_of_service <int>	KPIs_met..80. <int>	awards_won. <int>	avg_training_score <int>	is_promoted <int>
1	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5	8	1	0	49	0
2	65141	Operations	region_22	Bachelor's	m	other	1	30	5	4	0	0	60	0
3	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3	7	0	0	50	0
4	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1	10	0	0	50	0
5	48945	Technology	region_26	Bachelor's	m	other	1	45	3	2	0	0	73	0
6	58896	Analytics	region_2	Bachelor's	m	sourcing	2	31	3	7	0	0	85	0

Objective

To predict whether an employee will be recommended for promotion based on various factors given in the data set.

Exploratory Data Analysis

- Identifying the categorical and numerical variables
- Analyzing the target variable
- Analyzing the categorical variable with target variable
- Analyzing the numerical variable
- Analyzing the relationship between variables

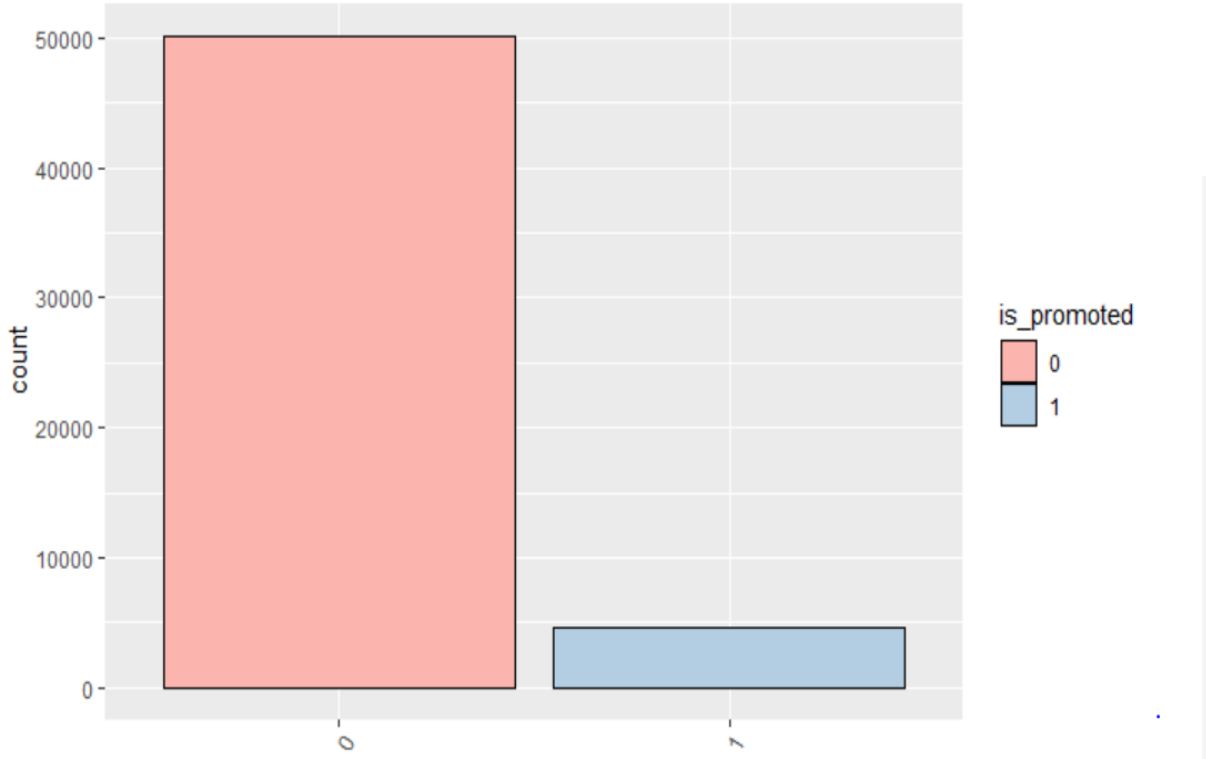
Exploratory Data Analysis

Step 1: Identifying the categorical and numerical variables

```
Classes ‘data.table’ and 'data.frame':  54808 obs. of  14 variables:
 $ employee_id         : int  65438 65141 7513 2542 48945 58896 20379 16290 73202 28911 ...
 $ department          : chr  "Sales & Marketing" "Operations" "Sales & Marketing" "Sales &
Marketing" ...
 $ region              : chr  "region_7" "region_22" "region_19" "region_23" ...
 $ education           : chr  "Master's & above" "Bachelor's" "Bachelor's" "Bachelor's" ...
 $ gender              : chr  "f" "m" "m" "m" ...
 $ recruitment_channel : chr  "sourcing" "other" "sourcing" "other" ...
 $ no_of_trainings     : int  1 1 1 2 1 2 1 1 1 1 ...
 $ age                 : int  35 30 34 39 45 31 31 33 28 32 ...
 $ previous_year_rating: int  5 5 3 1 3 3 3 3 4 5 ...
 $ length_of_service   : int  8 4 7 10 2 7 5 6 5 5 ...
 $ KPIs_met            : int  1 0 0 0 0 0 0 0 0 1 ...
 $ awards_won          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ avg_training_score  : int  49 60 50 50 73 85 59 63 83 54 ...
 $ is_promoted         : int  0 0 0 0 0 0 0 0 0 0 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

Exploratory Data Analysis

Step 2: Analyzing the target variable



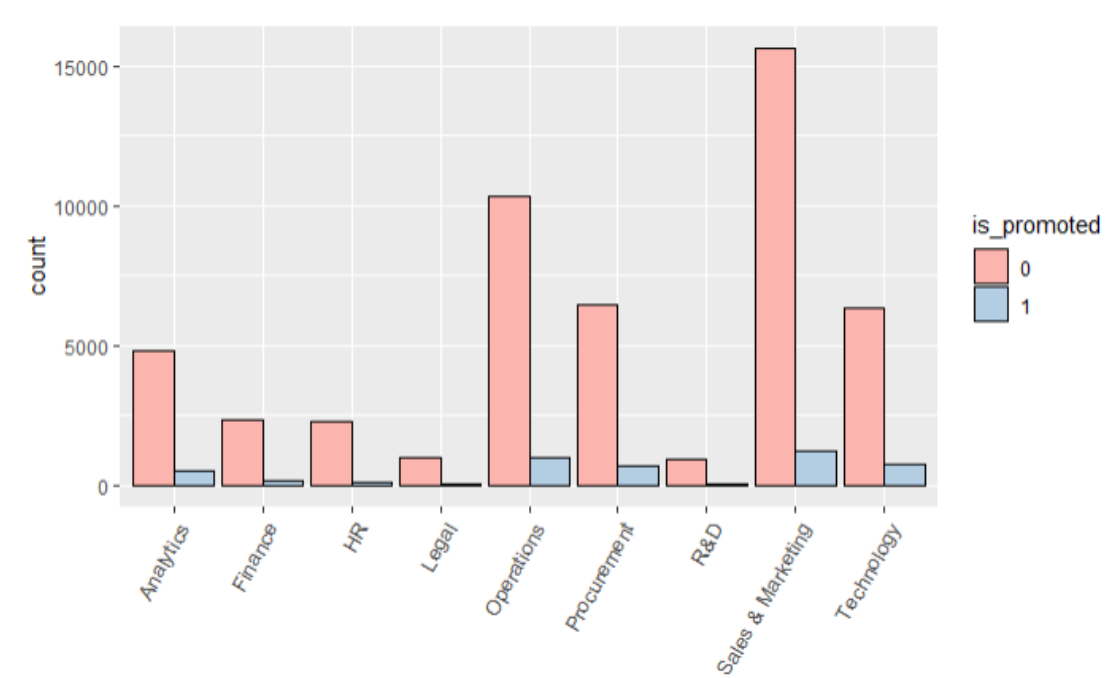

```
> prop.table(table(cat_hr_analytics$is_promoted))

          0           1
0.91482995 0.08517005
```

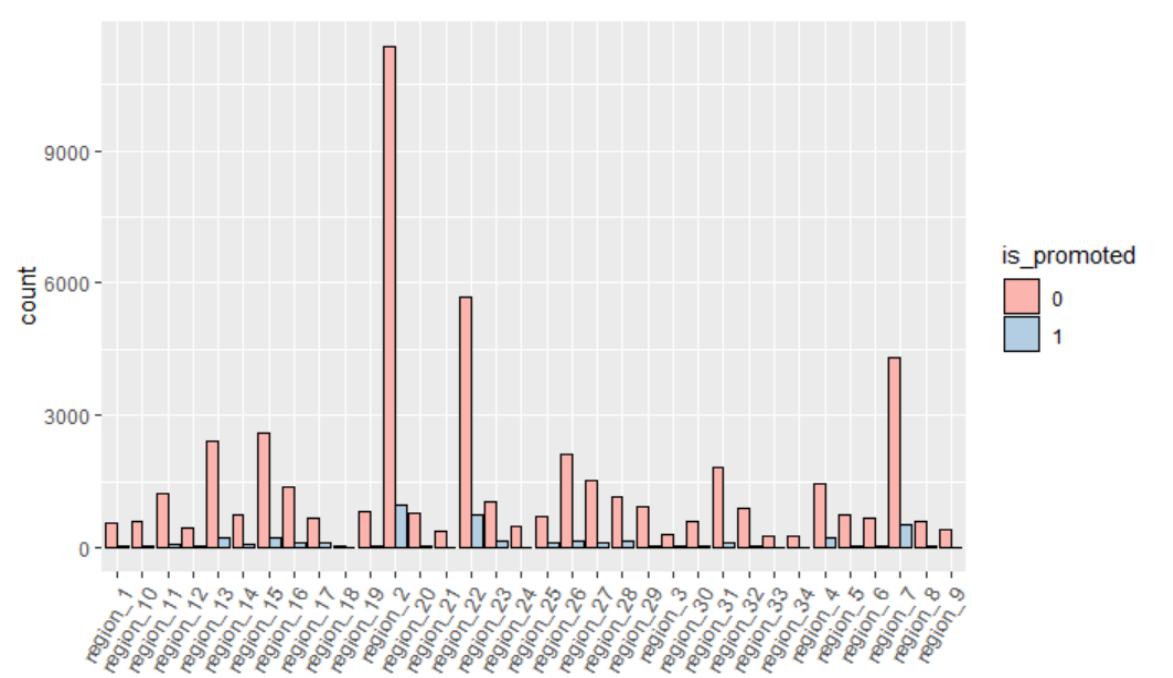
- This indicates a huge imbalance in the target variable for classification.
- This issue needs to be addressed before modeling.

Exploratory Data Analysis

Department



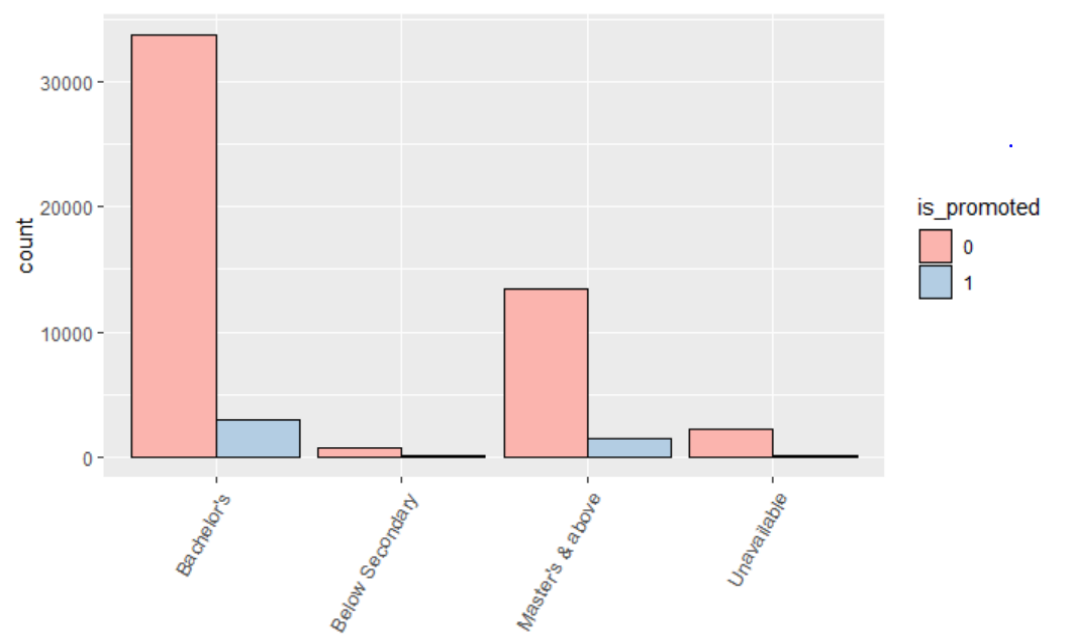
Region



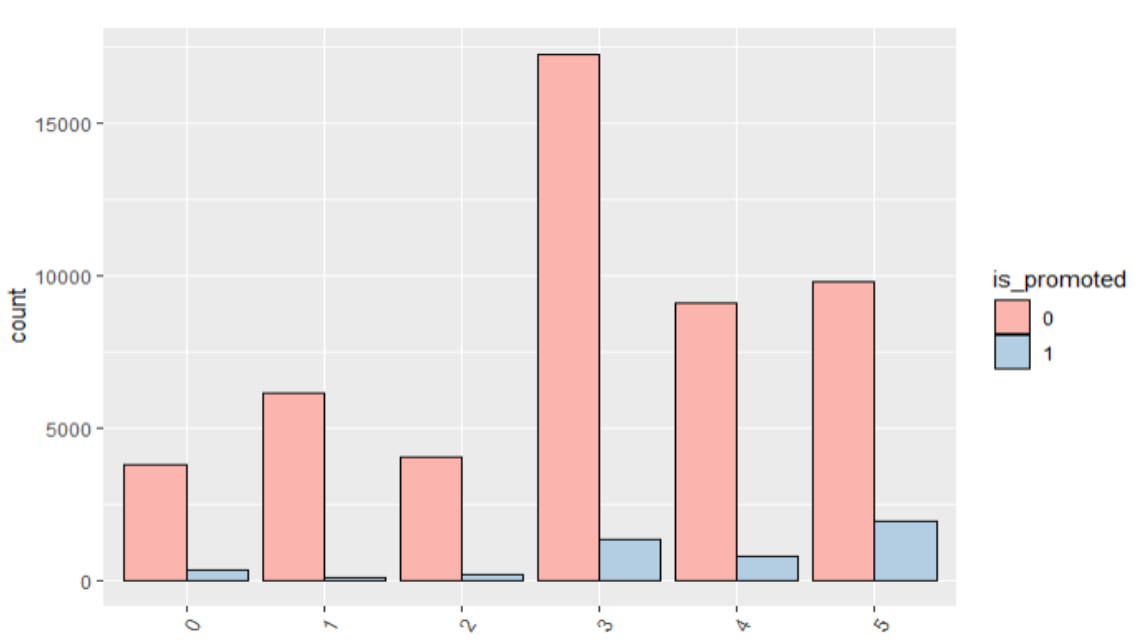
- Department:
  I. We observe that even though Sales & Marketing department is big, employees recommended are very few.
  II. In other departments like Analytics, Operations, Technology, Procurement employee recommendation is relatively good
- Region : Even though Region 19 has more employees, promotion recommendation is comparatively low

Exploratory Data Analysis

Education



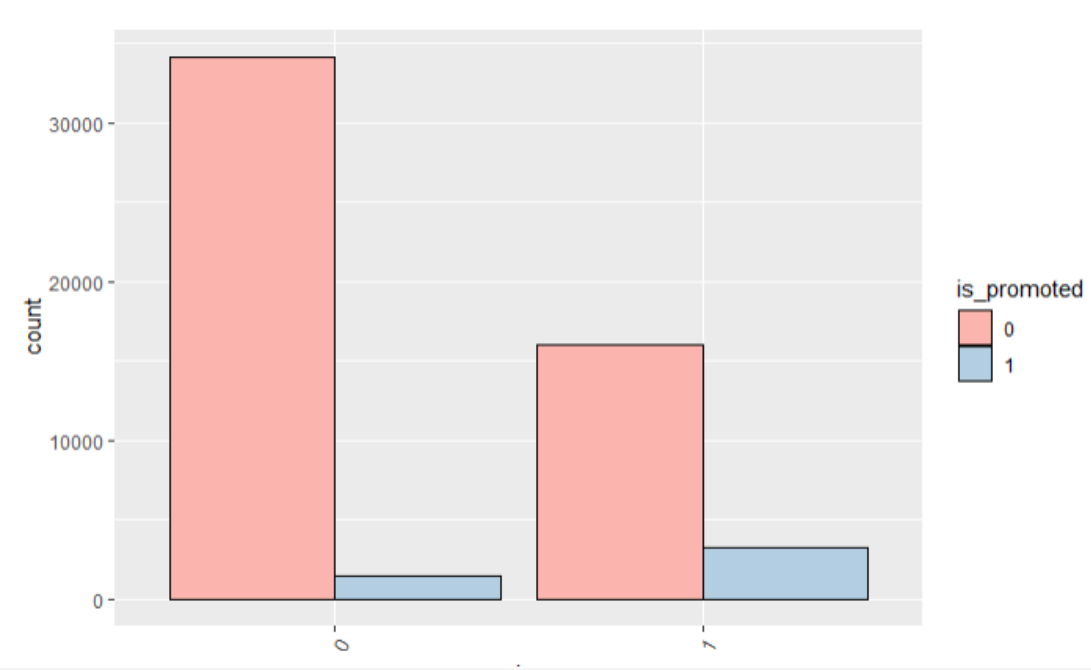
Previous Year Rating



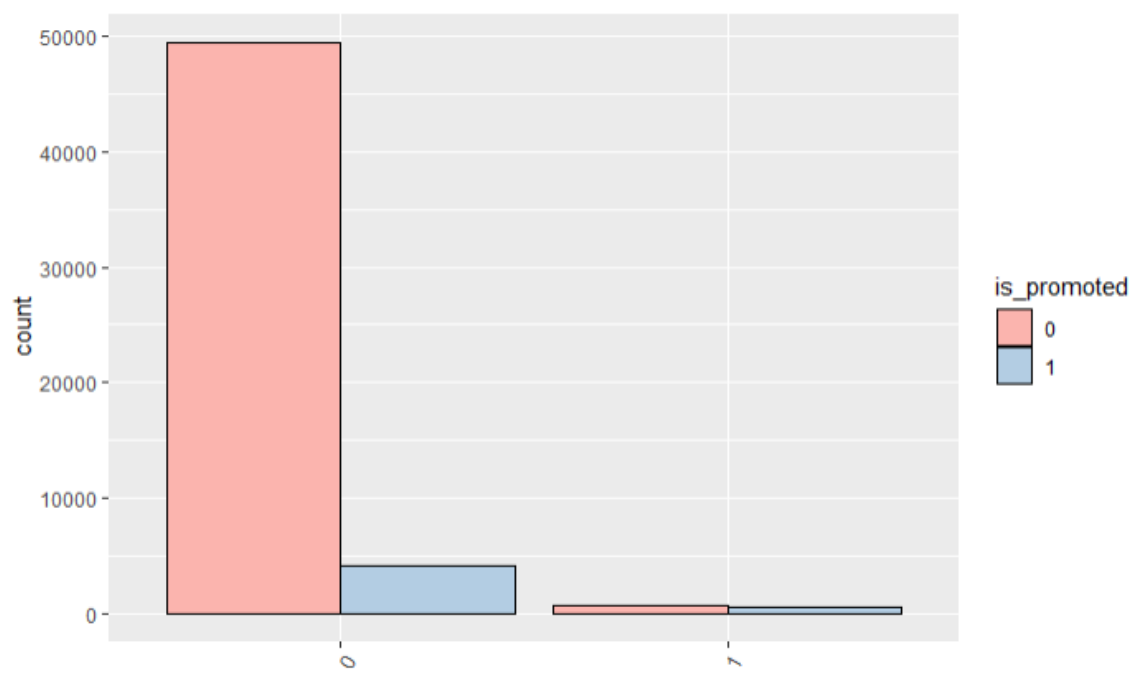
- Education: People who are recommended for promotion mostly holds a Master's & above
- Previous Year Rating: Employees with previous year rating of 5 have fair amount of chance of being recommended for promotion

Exploratory Data Analysis

KPI's met > 80%



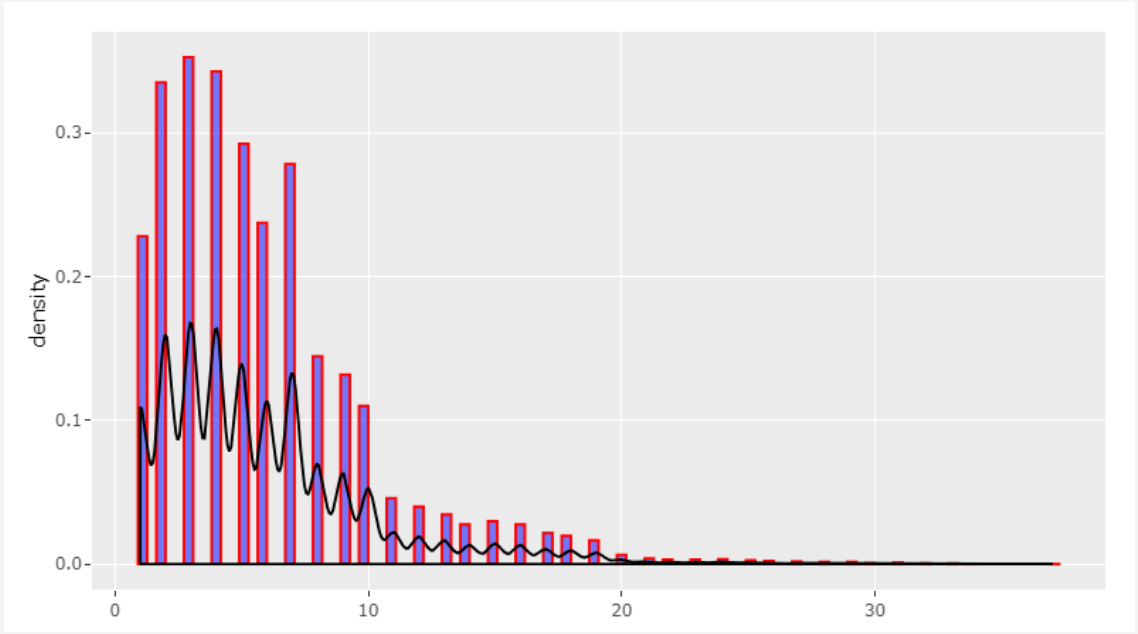
Awards Won



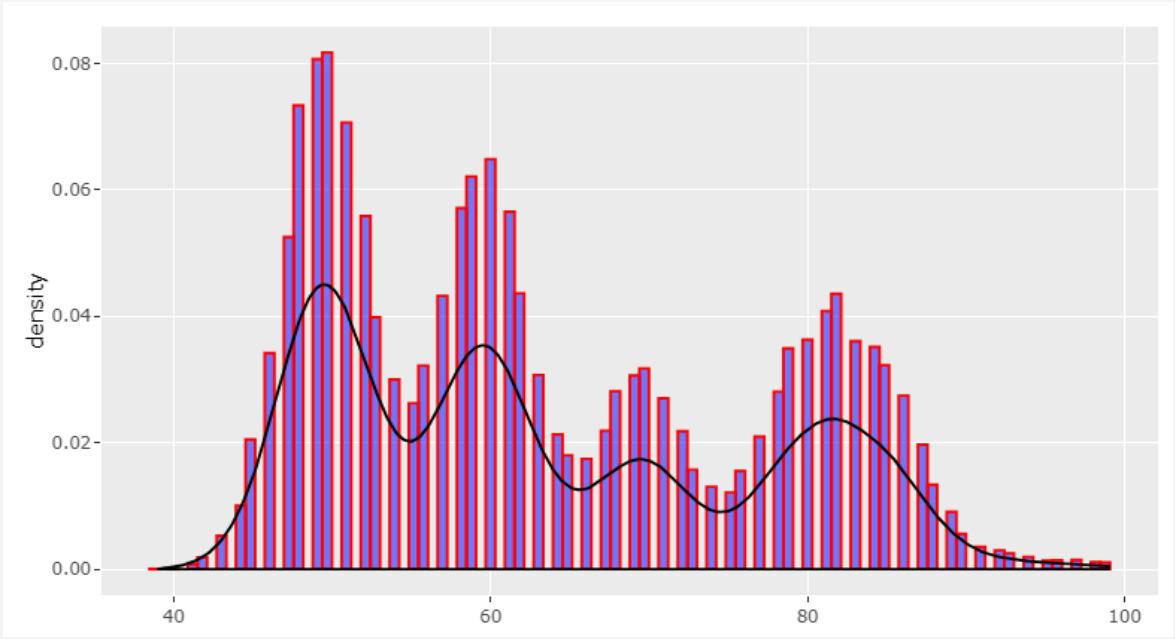
- KPI's met >80%: Employees who have KPI's greater than 80 have higher chances of being recommended for promotion
- Awards Won: People who have not won more awards are not likely to be recommended for promotion

Exploratory Data Analysis

Length of Service



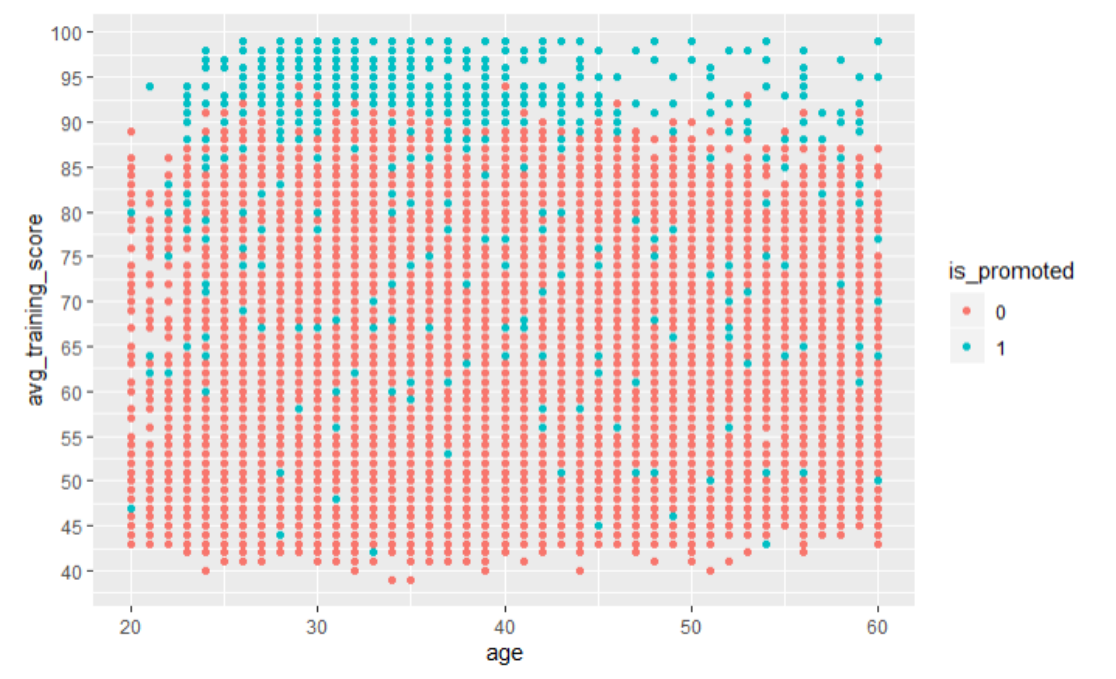
Avg Training Score



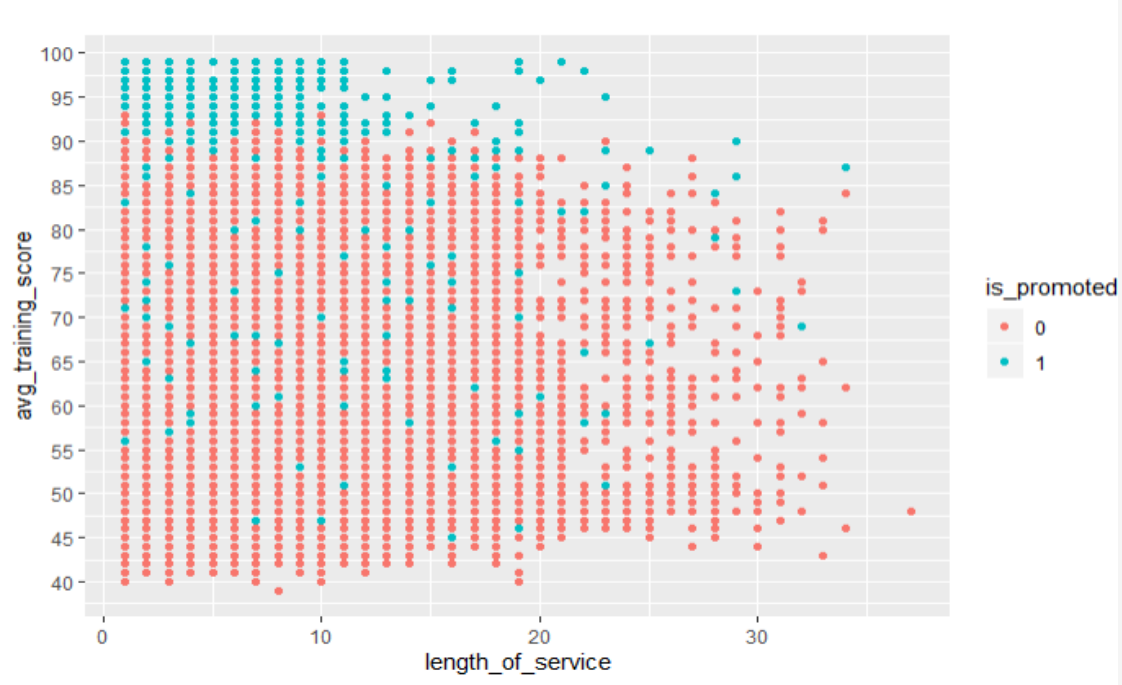
- Length of Service: Distribution shows a right skewness
- Avg Training Score: Skewness is not evident in the distribution

Exploratory Data Analysis

Avg Training Score vs Age



Length of Service vs Avg Training Score



- Avg_training_score vs Age: Higher the average score across all ages, are highly recommended for promotion
- Length of Service vs Avg Training Score : Employees with length of service of 1 to 11 years along with high average training scores are highly recommended for promotion

Data Cleaning/Data Preprocessing

- Observed NA values in previous_year_rating and education attributes.
- Replaced the NA values in the previous_year_rating with 0, since the length of the service of those employees was only one year.

```
> relation<-hr_analytics[is.na(previous_year_rating),.(length_of_service,previous_year_rating)]
> unique(relation$length_of_service)
[1] 1
```

- Added the NA as a factor level for the education attribute.

```
> levels(hr_analytics$education)
[1] "Bachelor's"       "Below Secondary"  "Master's & above" NA
```

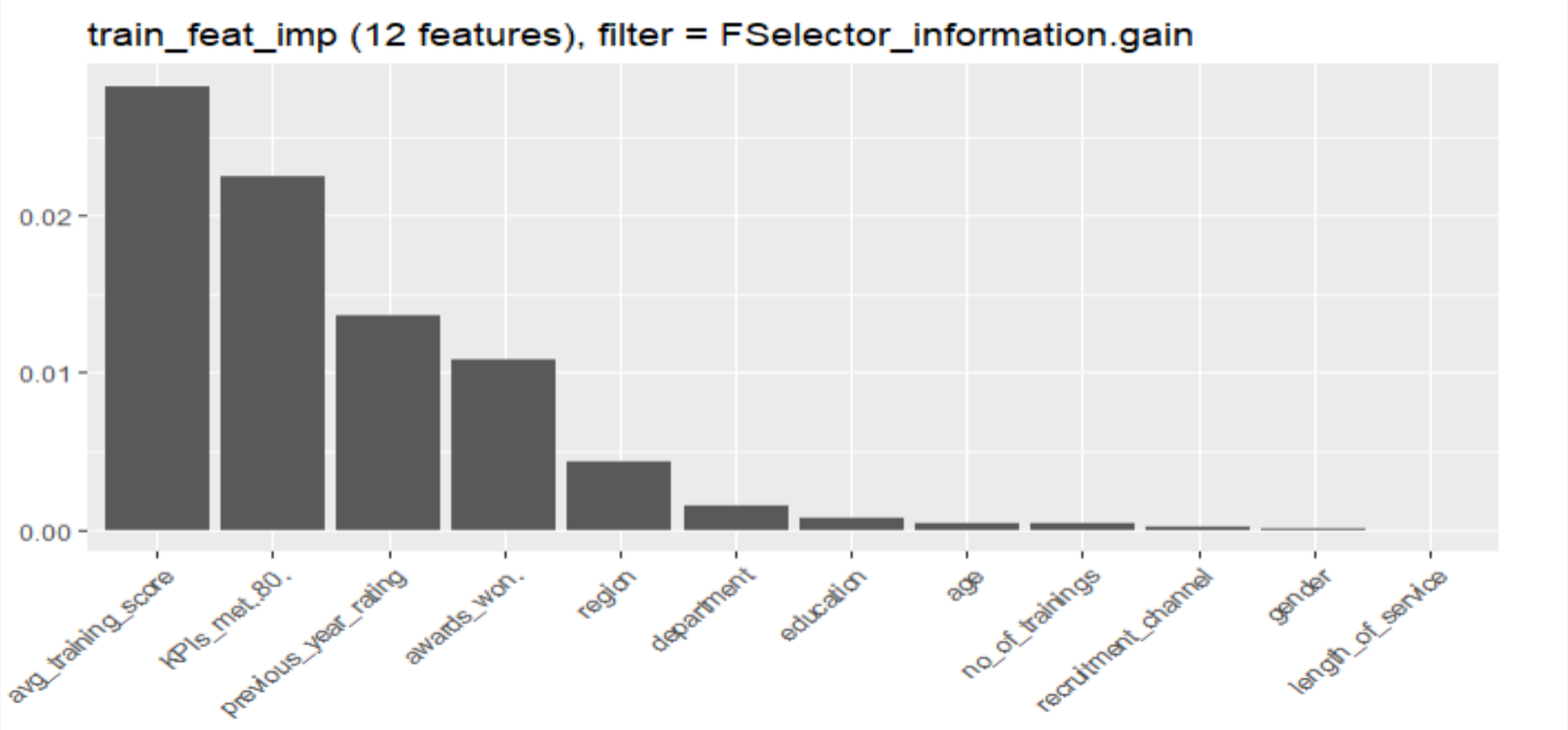
- No significant correlation observed between the attributes
- Converted Age, Department, Region, No: of Trainings, Previous Year Rating, KPI's >80%, Awards Won, Region, is promoted, Education, Recruitment Channel, Gender to factors.

Feature Engineering

- Converted age into bins E.g.: 20-30,30-40,40-50,50-60
- Converted NA level in the education attribute to ‘Unavailable’ as the rows with NA’s were being omitted during logistic regression
- Employee ID’s attribute was not considered for modelling as it was a nominal factor.
- Handling the imbalanced data by applying synthetic resampling techniques as ROSE and SMOTE.
- Split the data into train and validation set in the ratio 75:25

Feature Selection

Variable importance chart before applying various models on the data



Models Used

ROSE

- Decision Tree(CART/RPART)
- Random Forest
- Logistic Regression

SMOTE

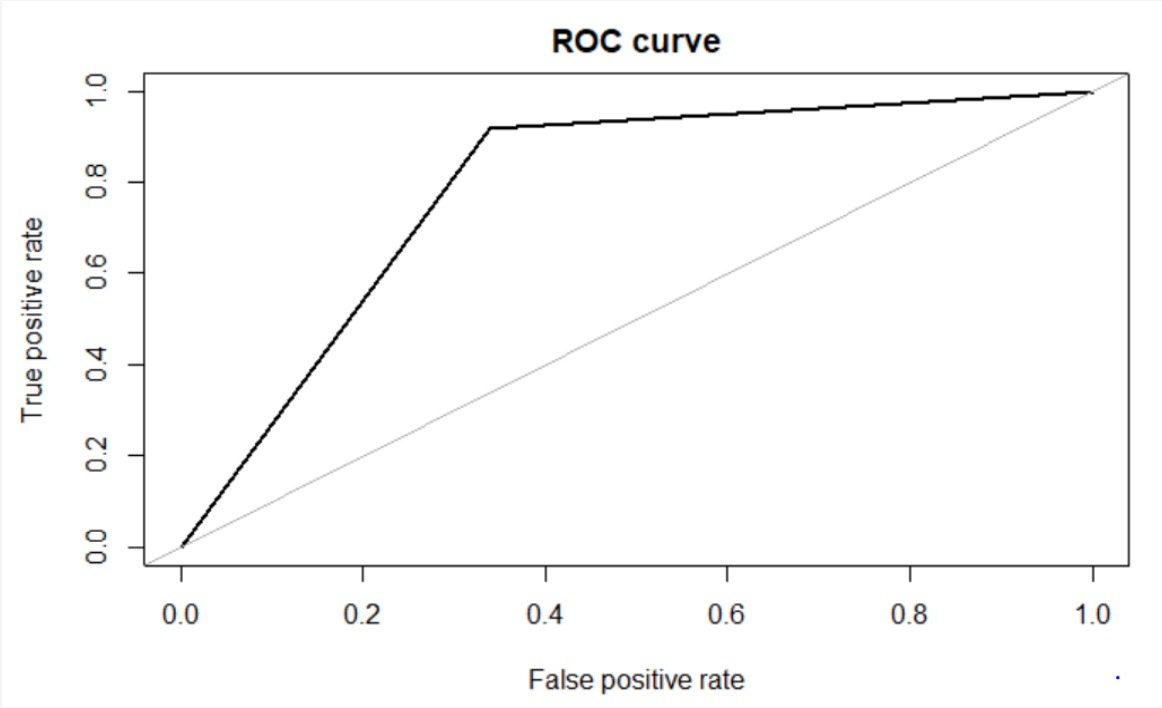
- Decision Tree(CART/RPART)
- Random Forest
- Logistic Regression

ROSE-Decision Tree (CART/RPART)

Accuracy is 68.22%

AUC:0.789

	Predicted 0	Predicted 1
Actual 0	8283 TN	96 FP
Actual 1	4258 FN	1065 TP

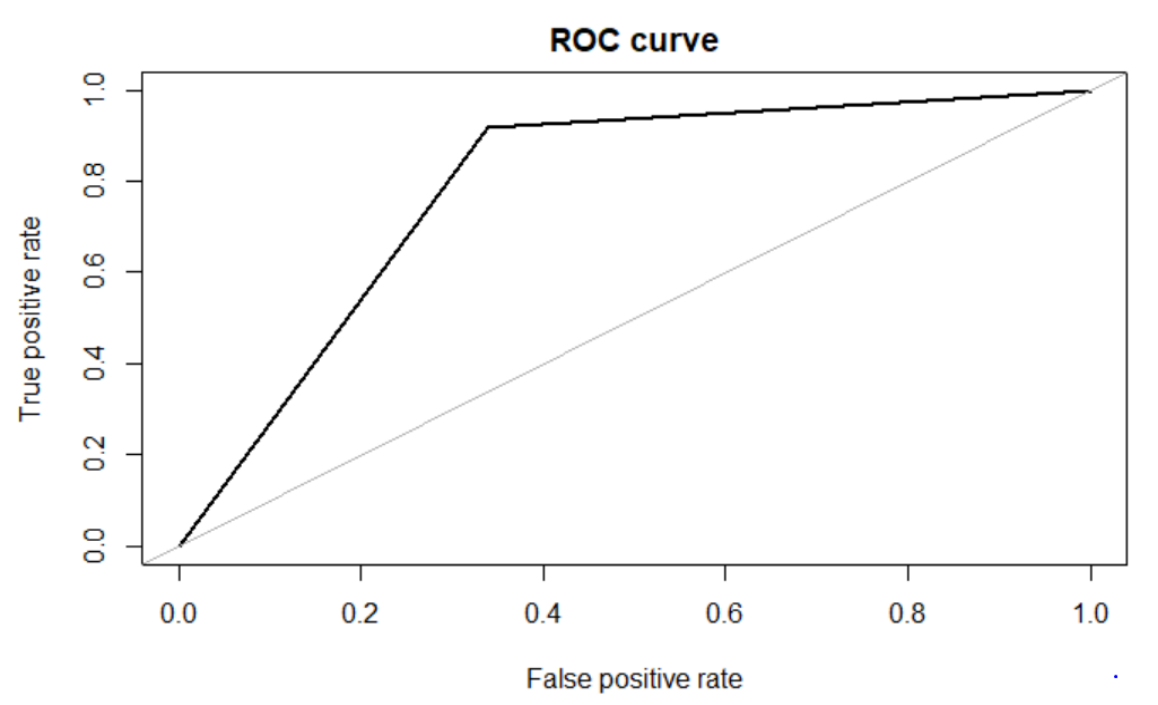


ROSE-Random Forest

Accuracy is 82.36%

AUC:0.785

	Predicted 0	Predicted 1
Actual 0	10428 TN	303 FP
Actual 1	2113 FN	858 TP

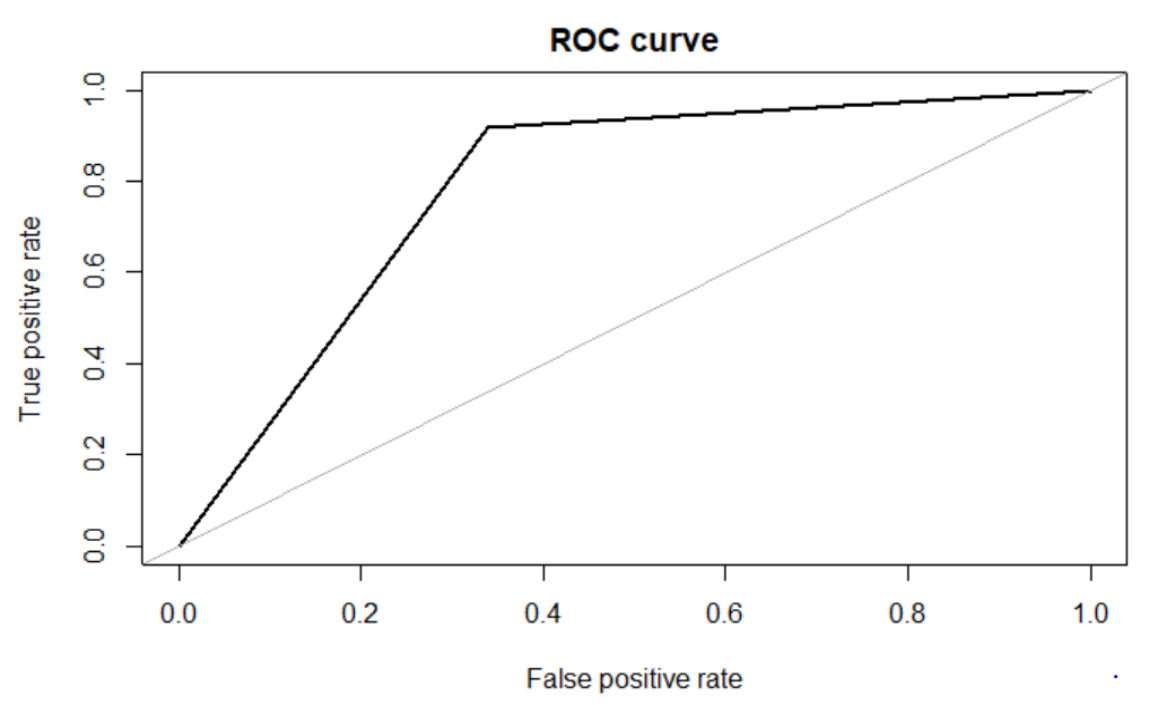


ROSE-Logistic Regression

Accuracy is 78.15%

AUC:0.87

	Predicted 0	Predicted 1
Actual 0	9799 TN	252 FP
Actual 1	2742 FN	909 TP

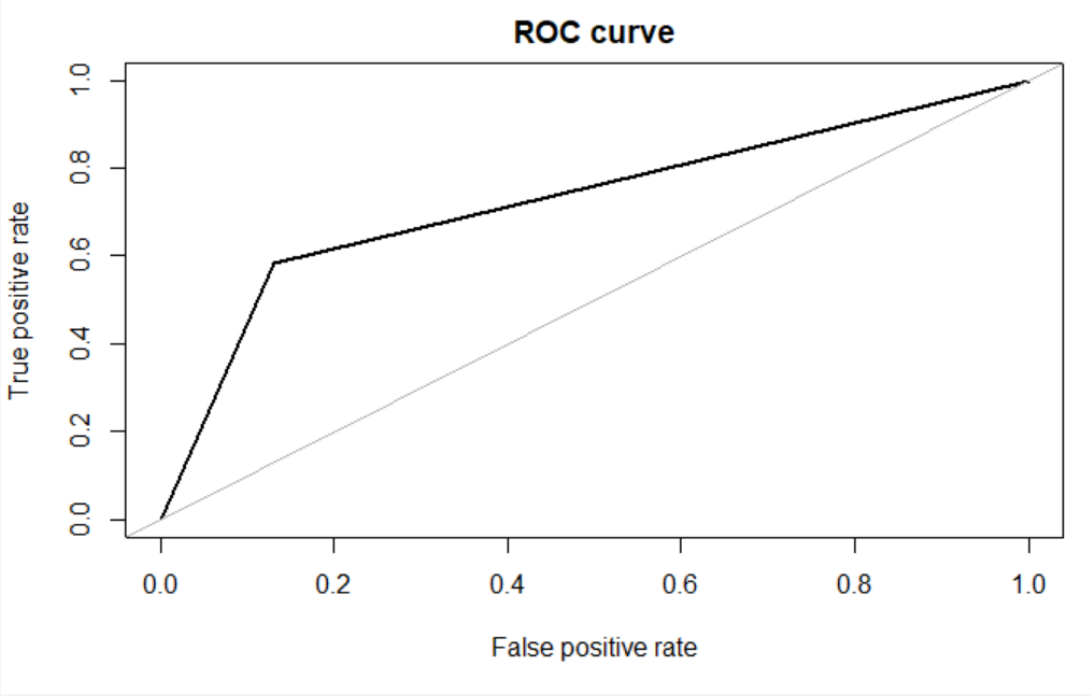


SMOTE – Decision tree (CART/RPART)

Accuracy is 84.6%

AUC:0.72

	Predicted 0	Predicted 1
Actual 0	10926 TN	479 FP
Actual 1	1629 FN	667 TP

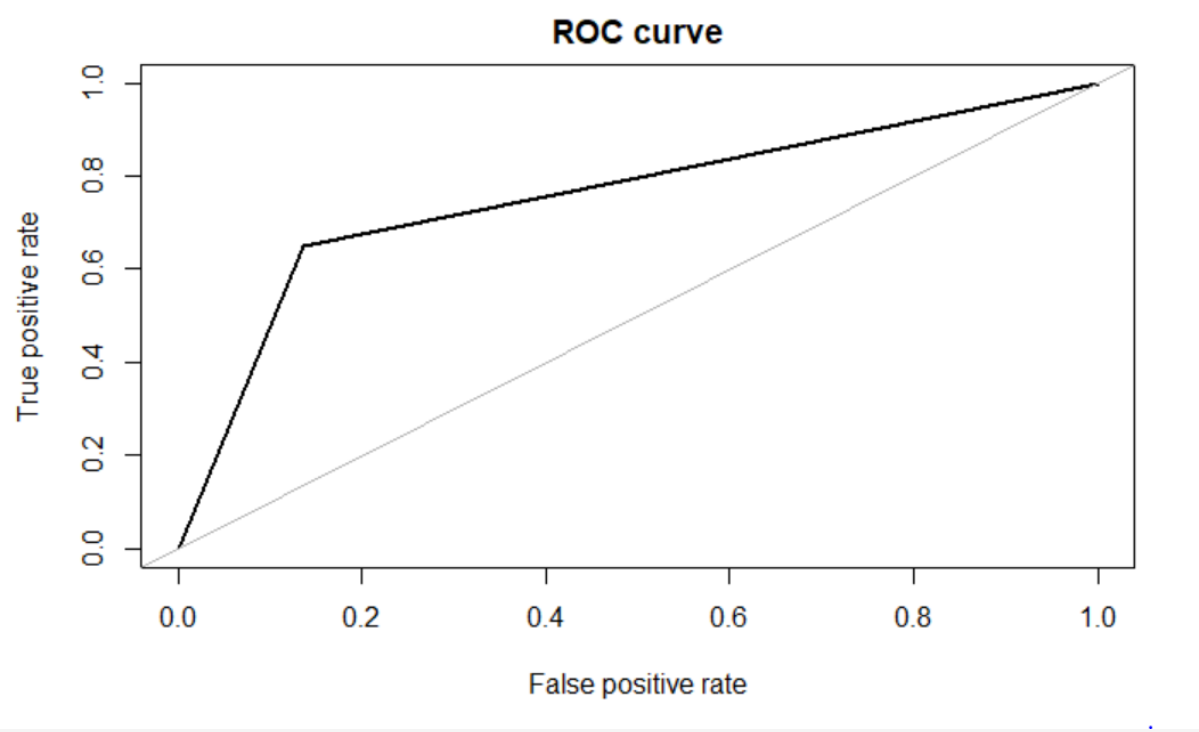


SMOTE – Random Forest

Accuracy is 84.6%

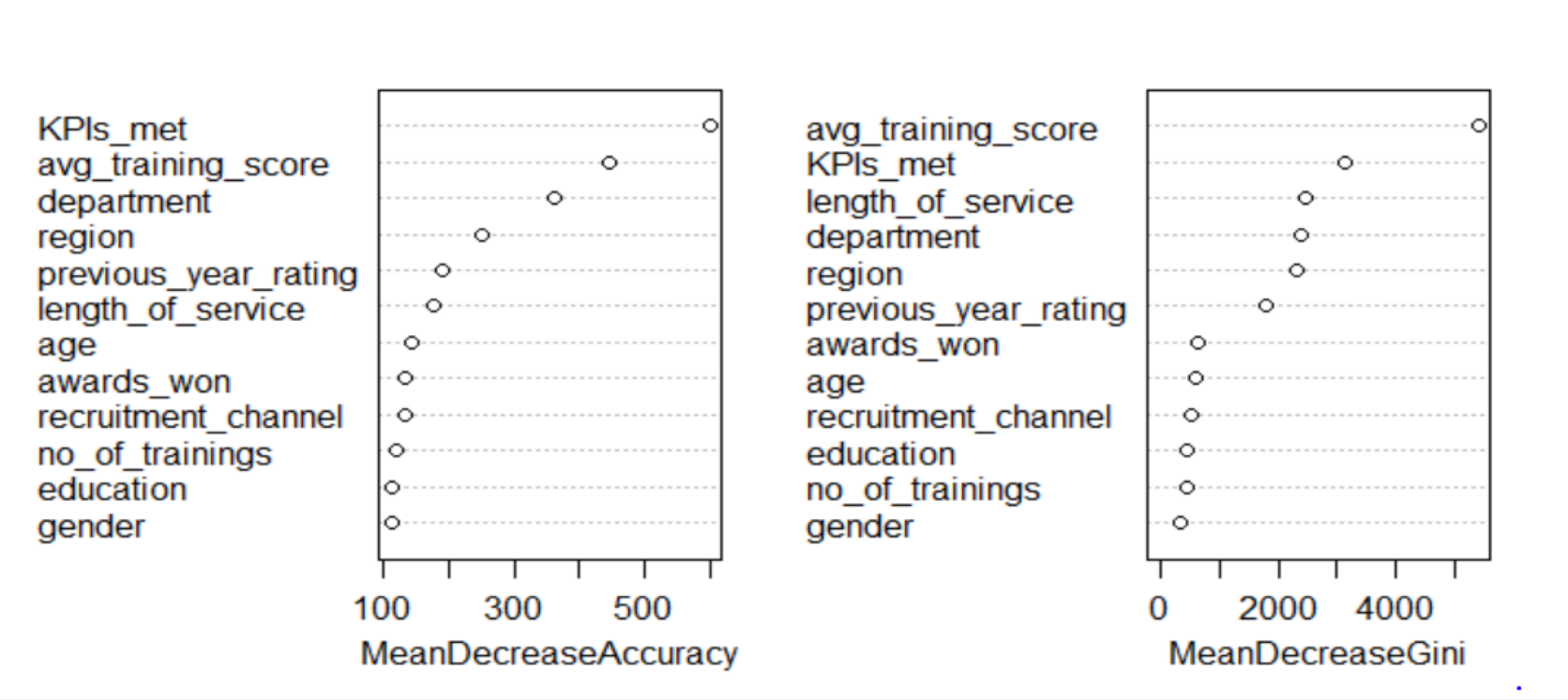
AUC:0.75

	Predicted 0	Predicted 1
Actual 0	10848 TN	401 FP
Actual 1	1708 FN	745 TP



SMOTE – Random Forest

- Variable Importance Plot

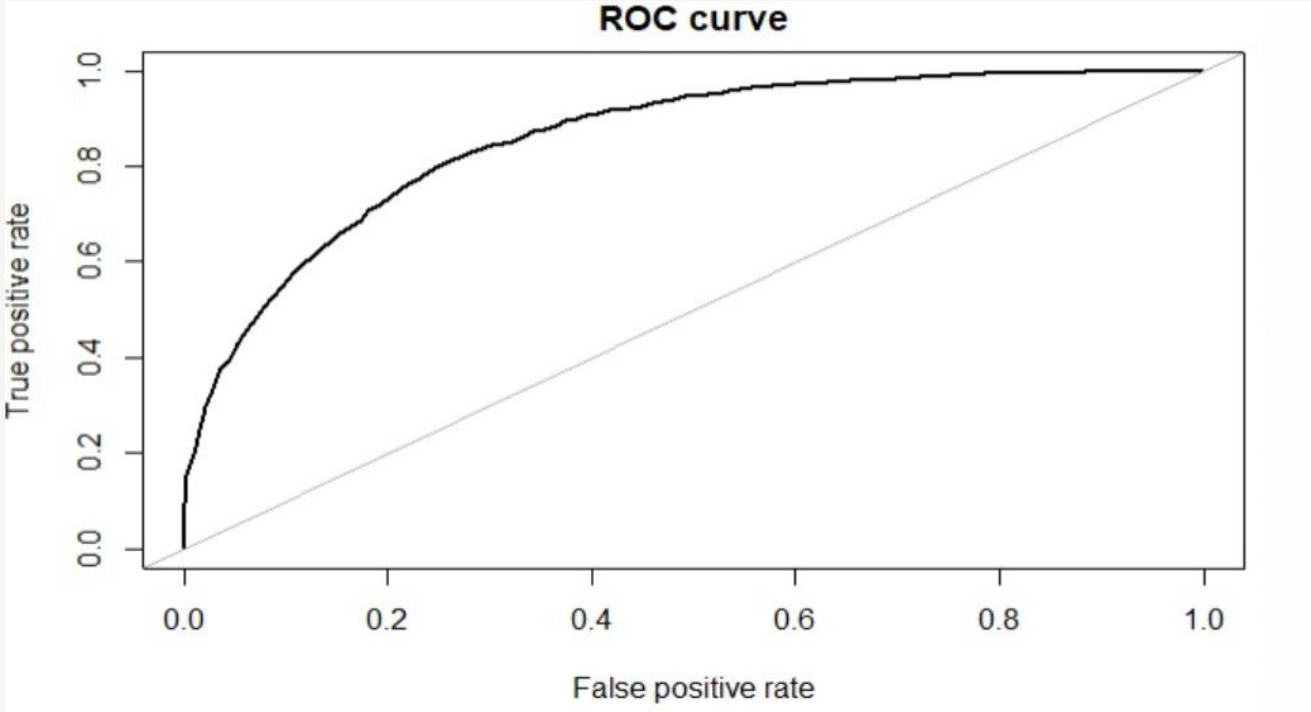


SMOTE – Logistic Regression

Accuracy is 88.29%

AUC:0.857

	Predicted 0	Predicted 1
Actual 0	11493 TN	577 FP
Actual 1	1028 FN	604



Results Comparison

Model	Accuracy (%)	AUC
ROSE – Decision tree (RPART)	68.22	0.789
ROSE – Random Forest	82.36	0.785
ROSE – Logistic Regression	78.15	0.87
SMOTE – Decision tree (RPART)	84.6	0.72
SMOTE – Random Forest	84.6	0.75
SMOTE – Logistic Regression	88.29	0.857

Challenges

- High Imbalance in the data set
- Only 2 Quantitative attributes

Future Work

- Creating new relevant features
- Using other resampling techniques to overcome the imbalance
- Other Machine Learning models like Clustering.

Thank You