Cairo University

Faculty of Engineering
Cairo University

Big Data Project Final Document

Team 5

Team members

Name	Section	B.N.
Aya Adel Hassan	1	17
Dina Alaa Ahmed	1	25
Dai Alaa Hassan	1	28
Nerdeen Ahmad Shawqi	2	28

i. Problem Description

Human resources has been using analytics for years. However, the collection, processing and analysis of data has been largely manual, and given the nature of human resources dynamics and HR KPIs, the approach has been constraining HR. The goal is to try to use predictive and descriptive analytics in identifying the employees most likely to get promoted.

ii. Project Pipeline

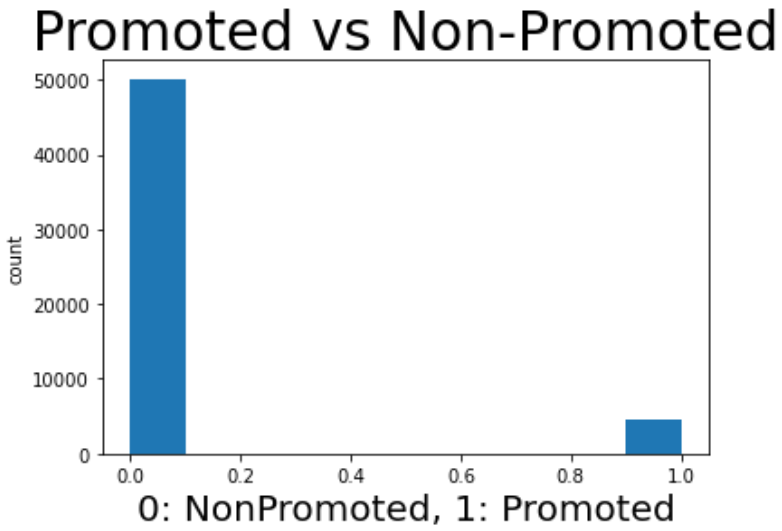
1. Load dataset and required libraries
2. Basic dataset statistics + conclusions
    1) Shape of the dataset
    2) Viewing some rows from the dataset
    3) Description of the numerical attributes
    4) Description of the categorical attributes
3. Checking for null values in columns
4. Checking for columns with only one value
5. Checking for duplicate rows
6. Dropping the 'employee_id' column as it's not important in the following steps
7. Data Visualization
8. Descriptive Analysis (Association Rule Mining)
9. Feature Selection
10. Data Preprocessing
11. Training Models

iii. Analysis and Solution of the problem

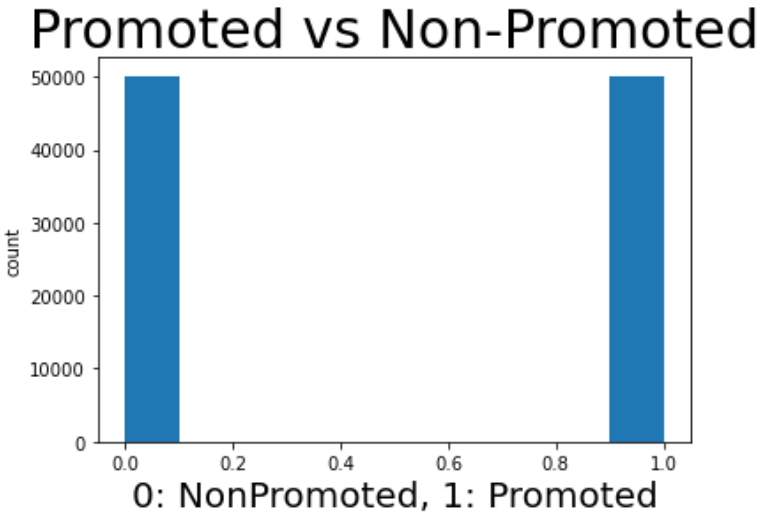
1. Data preprocessing

1) Handling categorical data using one hot encoding.
2) Handling unbalanced classes using SMOTE.

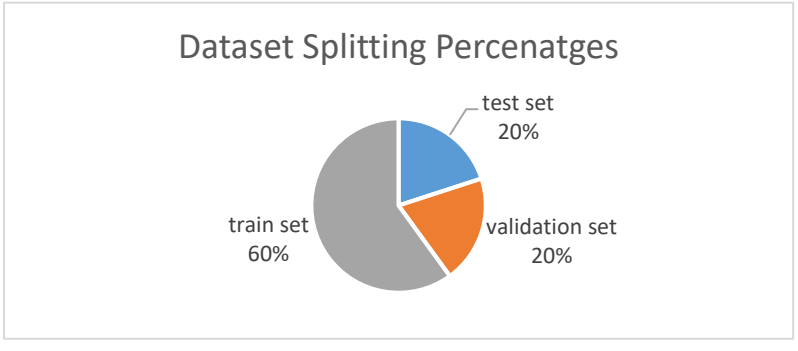
Target column **BEFORE** SMOTE



Target column **AFTER** SMOTE



3) Splitting data into train-test-validation:



4) Normalize data values:

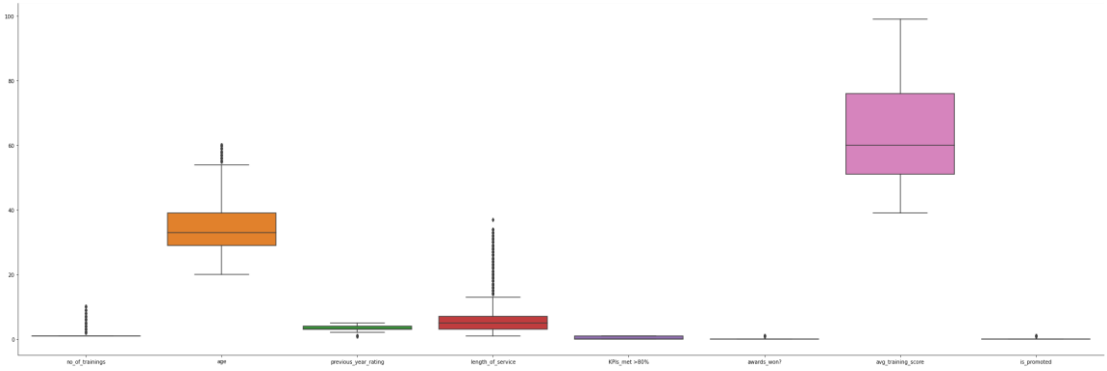
Normalization of the train and test set using the following formula

$$set = \frac{set - \min(set)}{\max(set) - \min(set)}$$

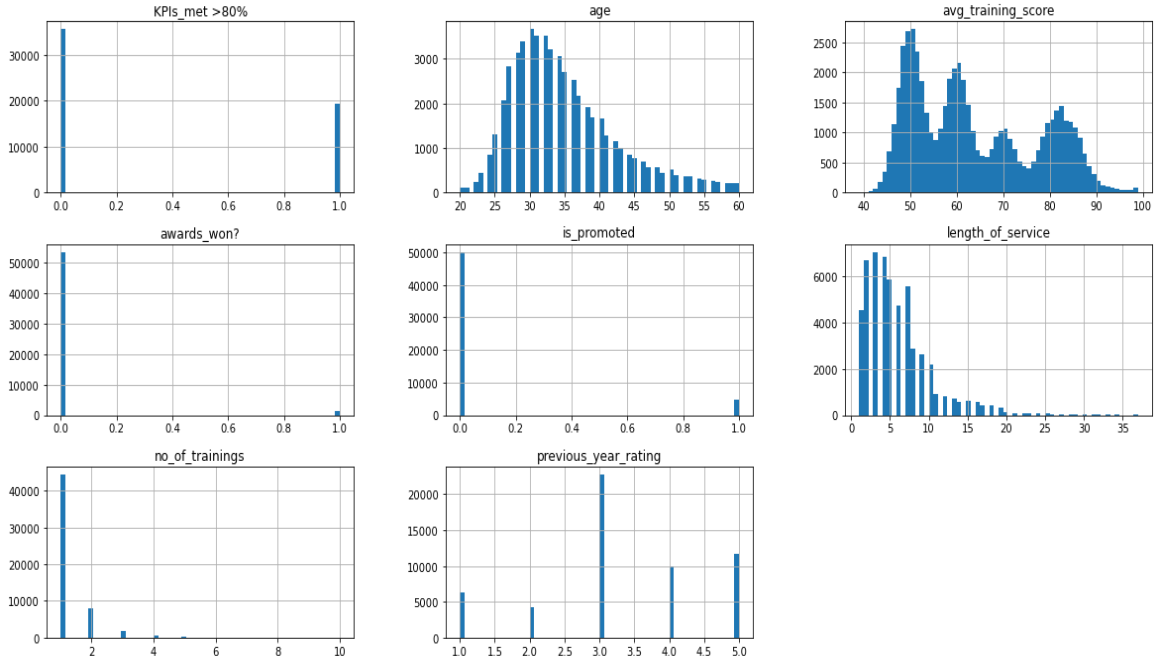
2. Data Visualization

1) Uni-Variate:

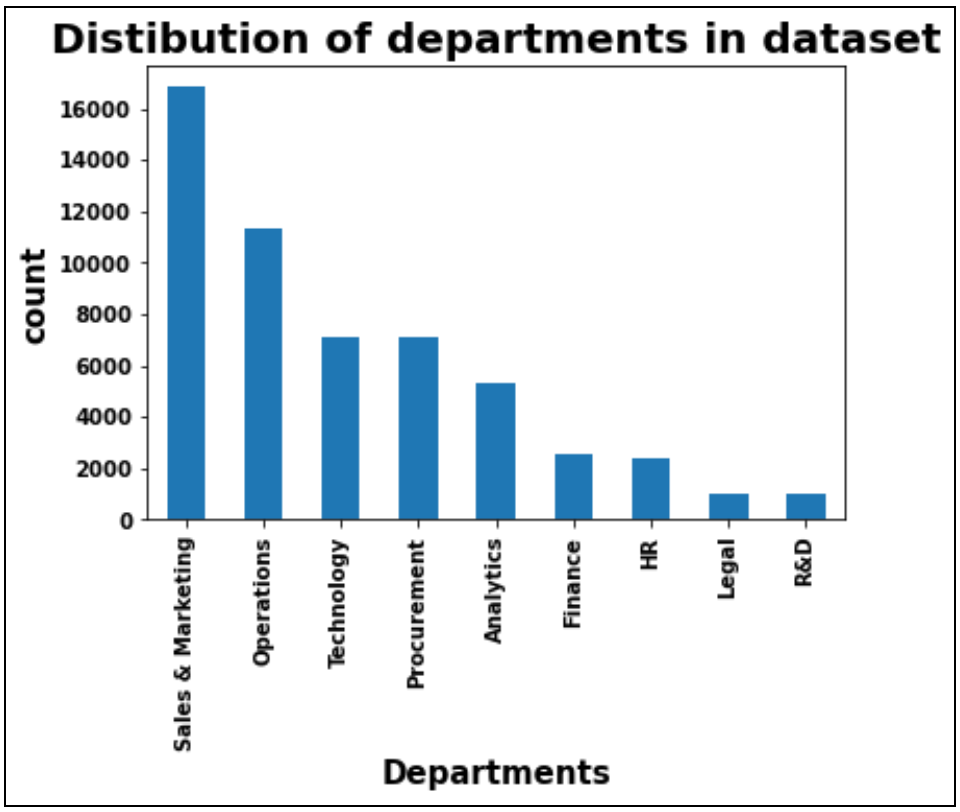
a) Boxplots (**Unsuccessful**) -> because features have different scales.

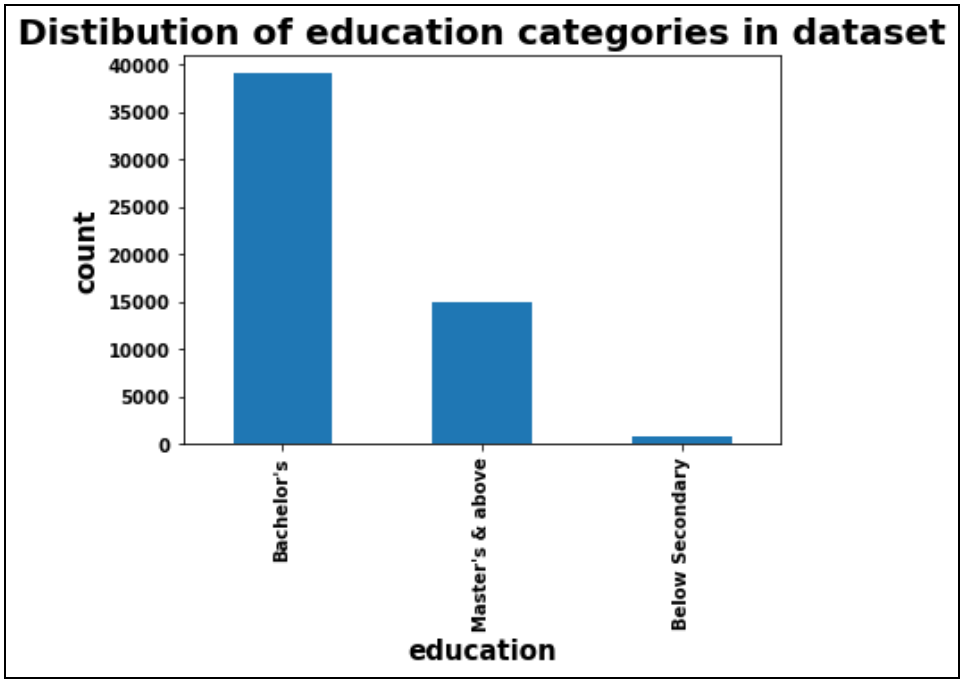


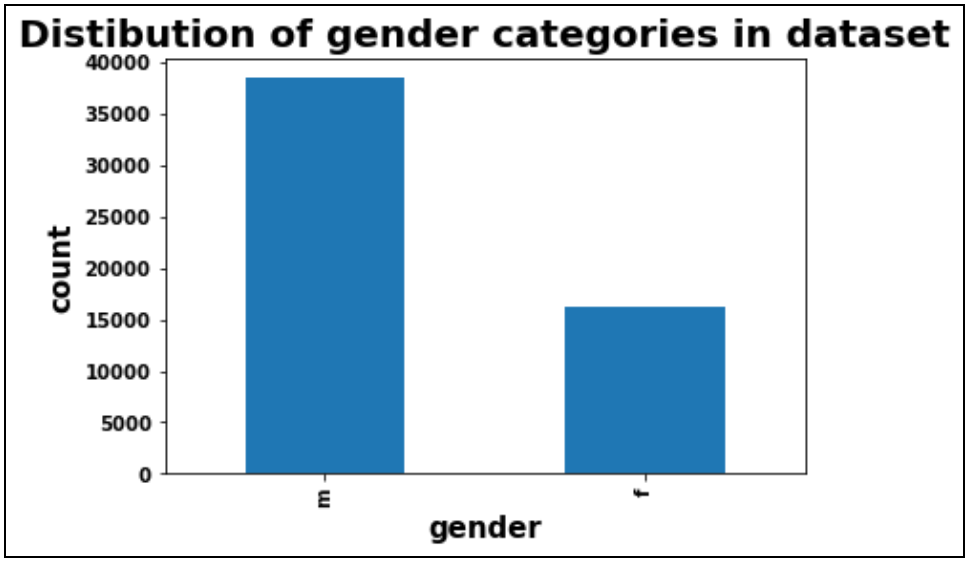
b) Histogram distributions for numerical columns.

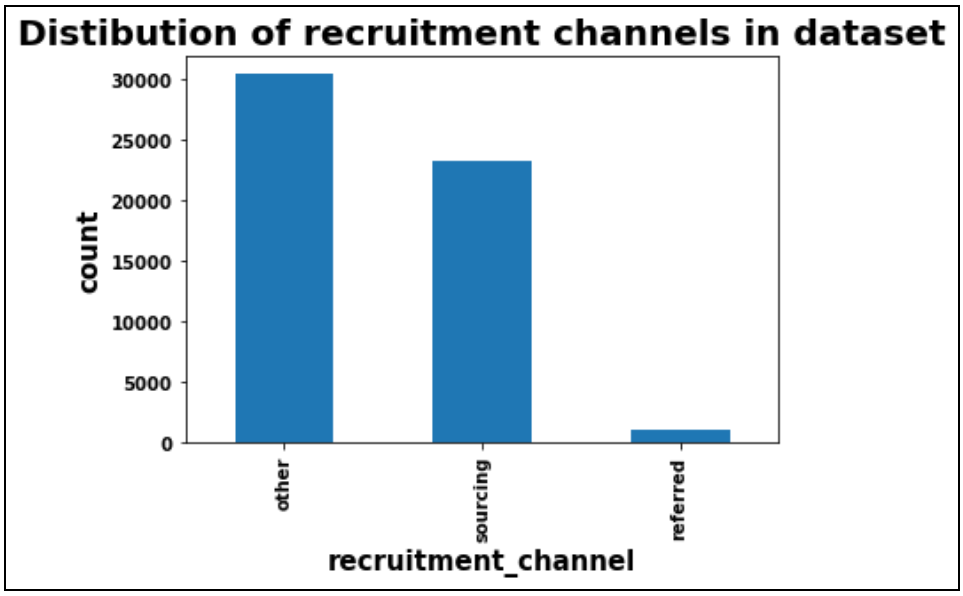


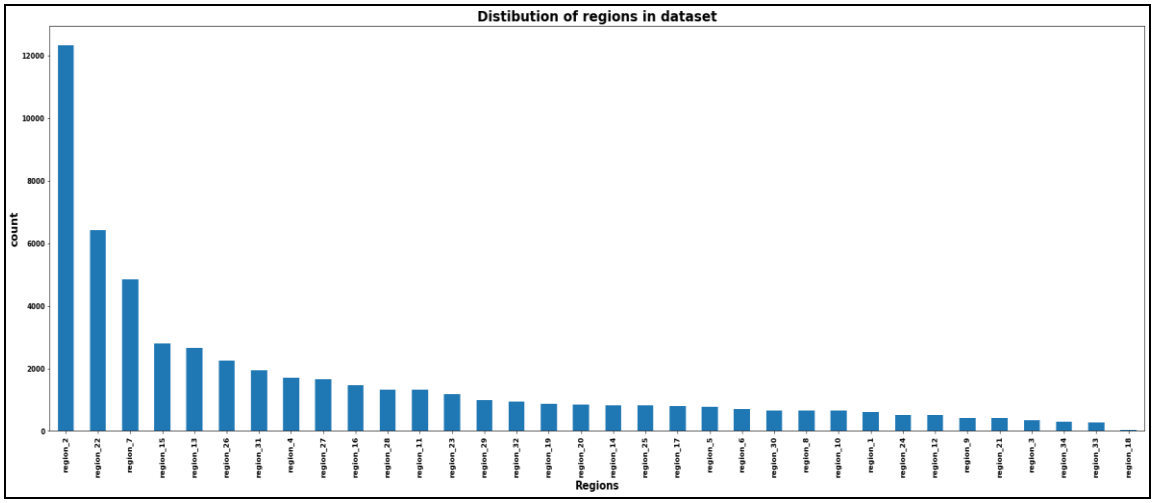
c) Bar Plots for categorical columns.



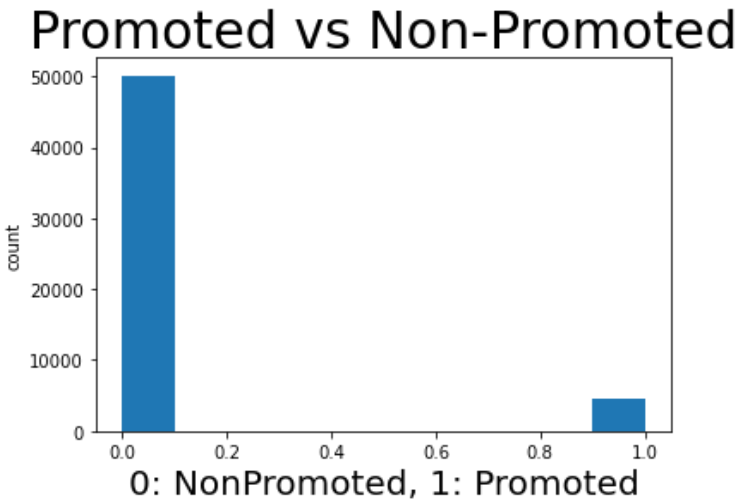



Distibution of gender categories in dataset
40000
35000
30000
25000
20000
15000
10000
5000
0
count
m
f
gender


Distibution of recruitment channels in dataset
30000
25000
20000
15000
10000
5000
0
count
other
sourcing
referred
recruitment_channel

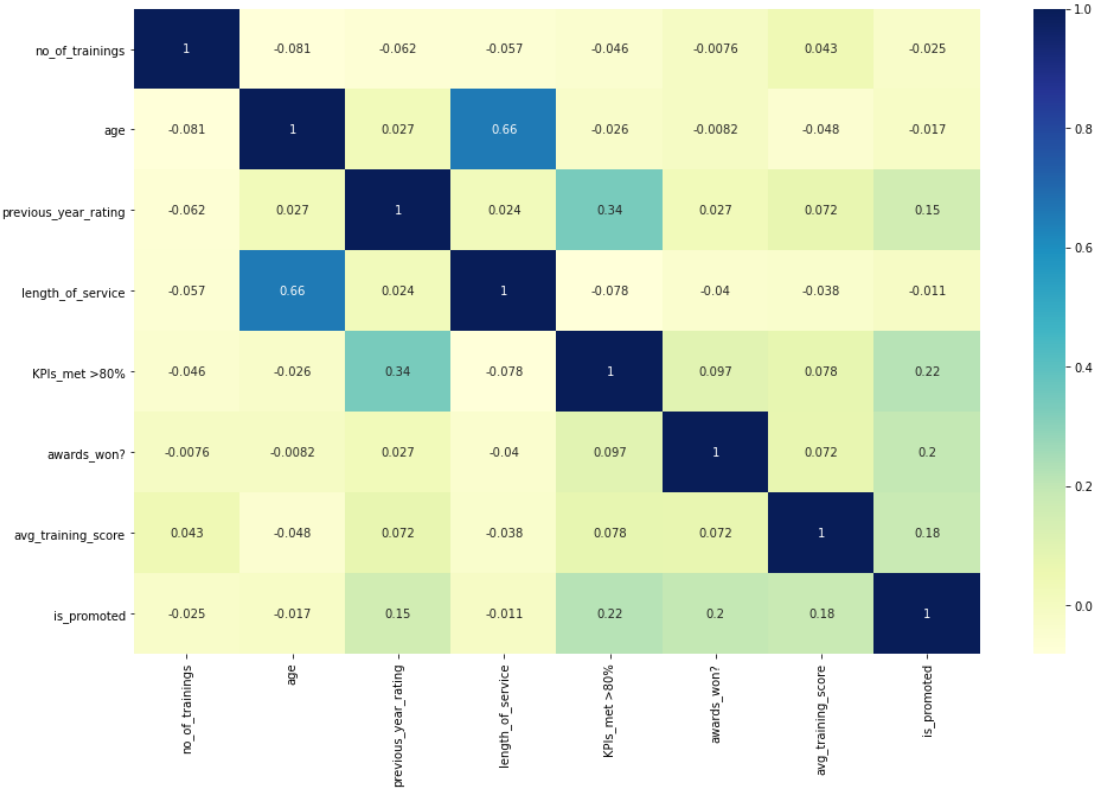

Distibution of regions in dataset
count
Regions

d) Bar Plot for the target column that shows there is a clear unbalancing problem in classes.

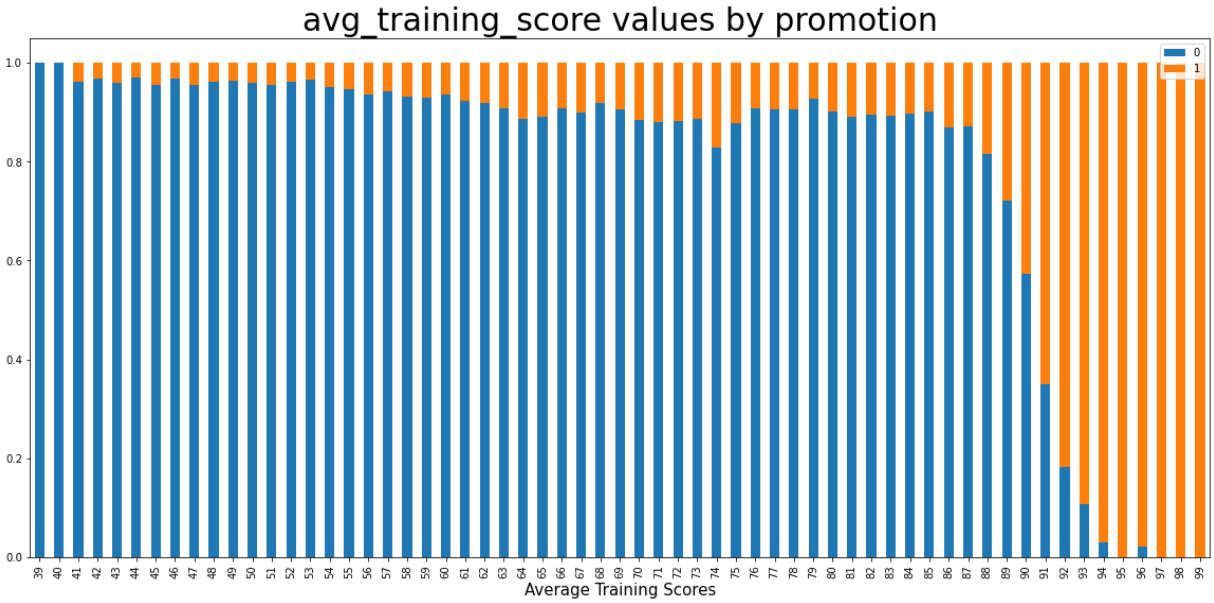


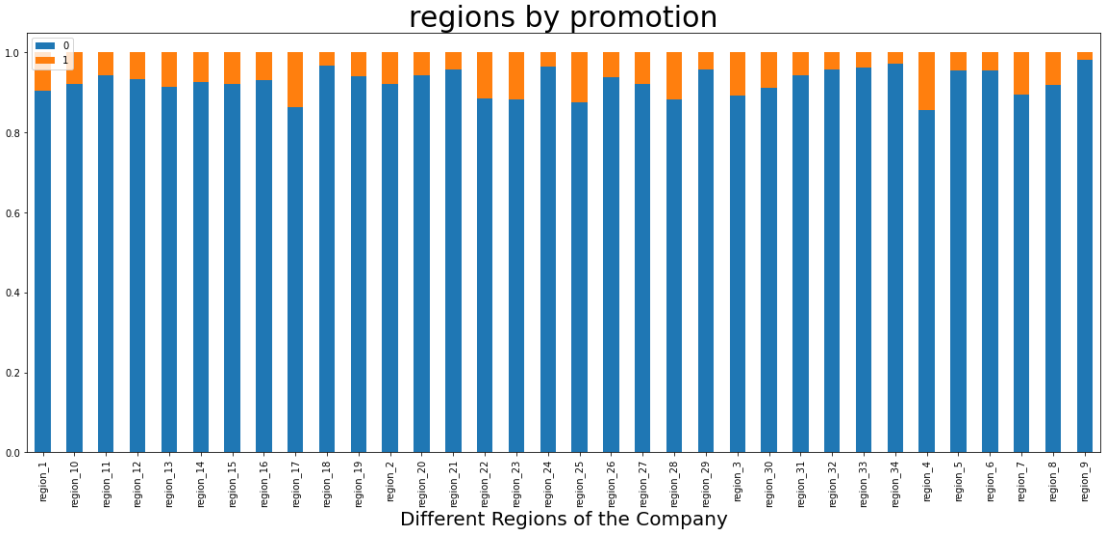
2) Multi-Variate:

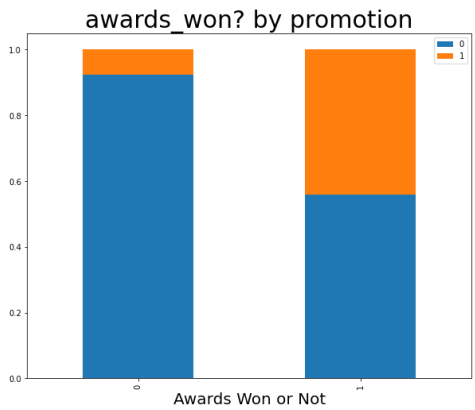
a) Heatmap to show correlation between columns.

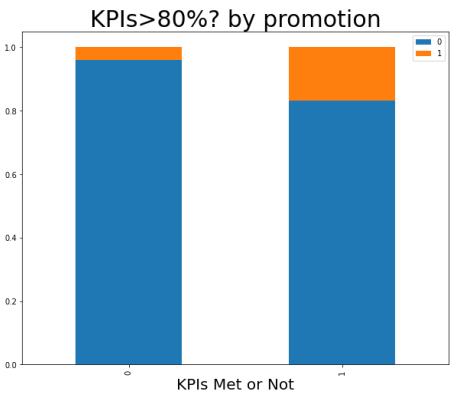


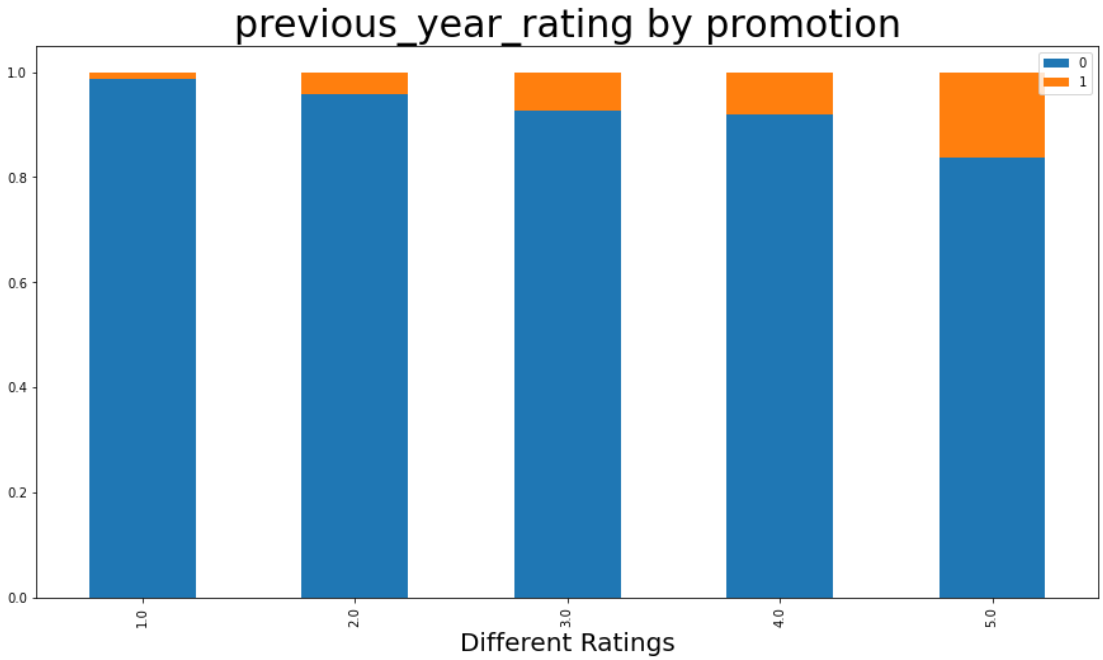
b) Crosstab plots between the target and features to show their effect on the target.

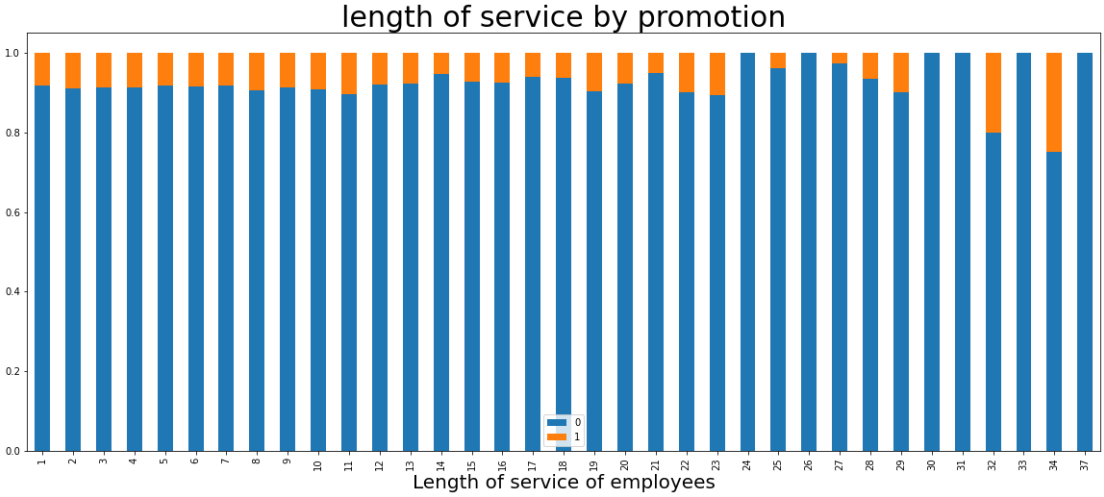


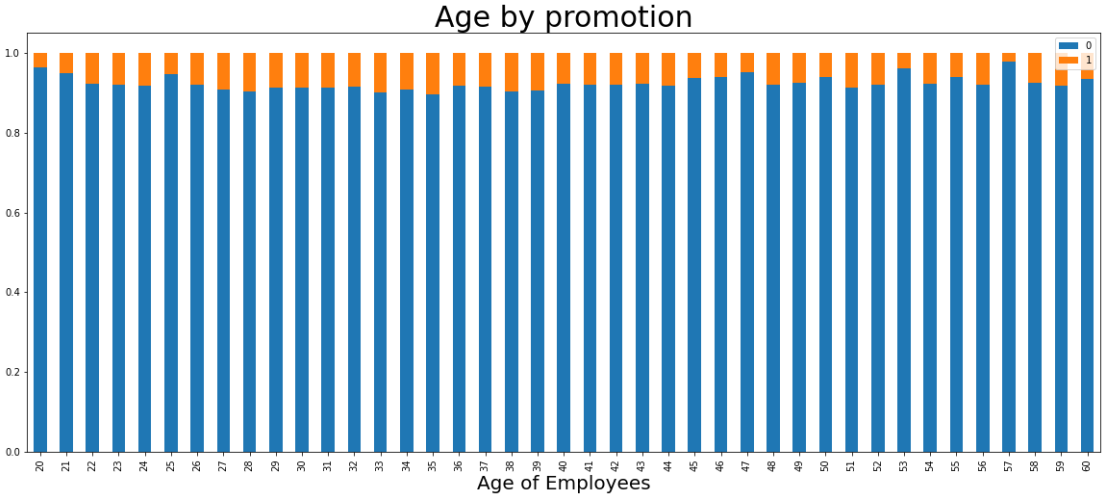


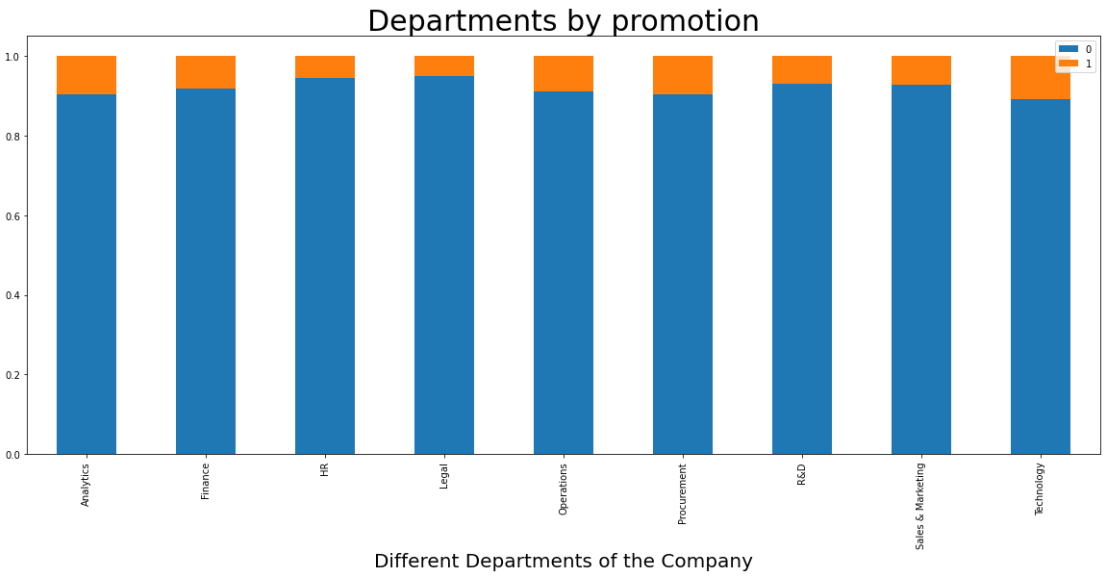


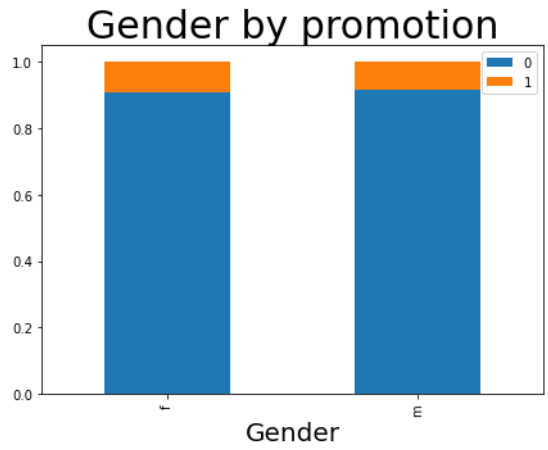


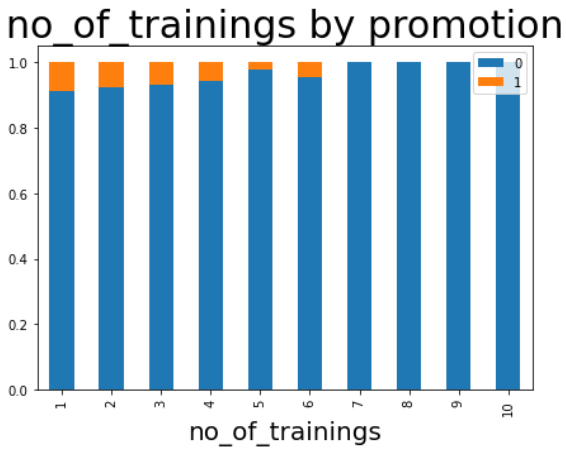

previous_year_rating by promotion
0
1
1.0
0.8
0.6
0.4
0.2
0.0
1.0
2.0
3.0
4.0
5.0
Different Ratings

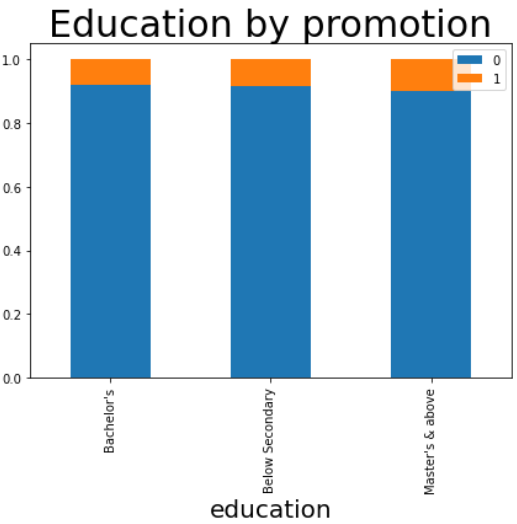

length of service by promotion
0
1
Length of service of employees

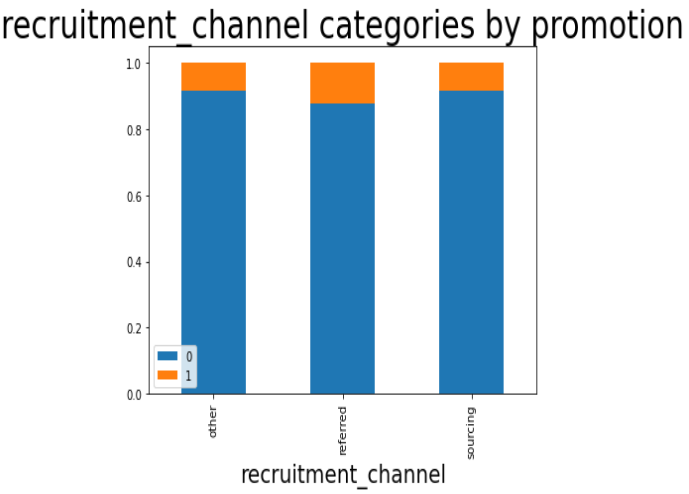

Age by promotion
0
1
Age of Employees


Departments by promotion
0
1
1.0
0.8
0.6
0.4
0.2
0.0
Analytics
Finance
HR
Legal
Operations
Procurement
R&D
Sales & Marketing
Technology
Different Departments of the Company


Gender by promotion
0
1
1.0
0.8
0.6
0.4
0.2
0.0
f
m
Gender


no_of_trainings by promotion
0
1
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
no_of_trainings


Education by promotion
0
1
1.0
0.8
0.6
0.4
0.2
0.0
Bachelor's
Below Secondary
Master's & above
education


recruitment_channel categories by promotion
0
1
1.0
0.8
0.6
0.4
0.2
0.0
other
referred
sourcing
recruitment_channel

3. Extracting insights from data

1) From statistical analysis

a) The distribution of data for features 'no_of_trainings', 'age', 'length_of_service', 'avg_training_score' looks normal because it's clear that the mean and the median are close enough.
b) Data majority in feature 'gender' is male with frequency 38496.
c) Data majority in feature 'department' is Sales&Marketing with frequency 16840.
d) Data majority in feature 'education' is Bachelor's with frequency 36669.
e) Data majority in feature 'region' is region_2 with frequency 12343.
f) Data majority in feature 'recruitment_channel' is other with frequency 30446.

2) From Visualizations

a) There is a clear unbalancing problem in classes in the target column.
b) **From heatmap:** Each variable has correlation = 1 with itself as expected.
c) **From heatmap:** There is somehow high correlation between 'age' and 'length_of_service', because older employees are more likely to have been working in the company for longer time.
d) **From heatmap:** 'KPIs_met>80%' is somehow related to 'previous_year_rating' as high rated employees are more likely to have their KPIs met the condition.
e) **From crosstab:** It's clear that the higher the 'avg_training_score', the more likely the employee is to be promoted.

f) **From crosstab:** Some regions have more promoted employees than other regions like region_4 & region_17.
g) **From crosstab:** If the employee has won awards, they are more likely to be promoted.
h) **From crosstab:** KPIs met have large effect on the promotion of the employee.
i) **From crosstab:** The higher the rating of the employee, the more likely he/she is to be promoted.
j) **From crosstab:** No clear pattern to determine the effect of 'length_of_service' on the promotion of the employee.
k) **From crosstab:** Middle-aged employees have higher chances of promotion.
l) **From crosstab:** Some departments have more promoted employees such as Technology department.
m) **From crosstab:** Gender doesn't affect the promotion of the employee.
n) **From crosstab:** The less trainings the employee has, the more likely he/she is to be promoted.
o) **From crosstab:** Education doesn't affect the promotion of the employee.
p) **From crosstab:** Employees who were referred to the company are more likely to be promoted.

3) From Descriptive analysis (Association rule mining)

Because of the unbalancing in data the insights are not good, but some rules make sense such as:

1. When average training score is low, the employee is not promoted with support = 0.2103 and confidence = 0.961
2. When the KPIs are not met, the employee is not promoted with support = 0.6223 and confidence = 0.96

3. When average training score is high, the employee is promoted with support = 0.0435 and confidence = 0.142
4. When the KPIs are met, the employee is promoted with support = 0.0595 and confidence = 0.1691

And so on as we deduced from the data visualizations. Note that the support of the rules that has promoted is yes is very low because of the unbalancing problem.

4. Feature Selection:

We used the insights extracted from data analysis and the Forward Selection method and found the best set of features which are:

- department
- region
- recruitment_channel
- age
- length_of_service
- no_of_trainings
- previous_year_rating
- KPIs_met >80%
- awards_won?
- avg_training_score

5. Model/ Classifier training:

We tried out multiple classifiers, but the one that performed the best is:

Random Forests (RF), It was trained on the train dataset with the following hyper parameters:

- n_estimators = 100

iv. Results and Evaluation:

- On Train set

	Class 0	Class 1
precision	1.0	1.0
recall	1.0	1.0
f1-score	1.0	1.0
accuracy	0.9996	

- On Validation set

	Class 0	Class 1
precision	0.95	0.97
recall	0.97	0.95
f1-score	0.96	0.96
accuracy	0.9621	

- On Test Set (final evaluation)

	Class 0	Class 1
precision	0.95	0.97
recall	0.97	0.95
f1-score	0.96	0.96
accuracy	0.9582	

Even though there is a clear overfitting in the model evaluation, but the model performed the best out of the models we tried, why? Because:

1. It has the highest test set accuracy (better generalization).
2. Internally, it makes feature selection as it takes the majority vote of many decision trees each of them has different sets of features.
3. It's stable, i.e. doesn't get affected by noise.

v. Unsuccessful Trials

- In model training: We tried out different models but all got less accuracies on the test set than the random forests.
  - SVM (accuracy = 0.935)
  - KNN (accuracy = 0.933)
  - Naïve Bayes (accuracy = 0.75)
  - Logistic Regression (accuracy = 0.92)
  - Decision Trees (accuracy = 0.932)
- In Feature Selection: We tried to remove unimportant features based on the data analysis only but the combining it with the forward selection method got us better accuracies.

vi. Enhancements and Future Work

- More data Visualizations and analysis.
- Try out different feature selection methods.
- Try out different learning models.
- Tuning the parameters of the tried out models to avoid problems like overfitting and under fitting.

vii. Map-Reduce

1. Technology used

- Hadoop distributed system in the pseudo distributed mode.

2. Model used

- KNN.

3. Methodology

- Put the entire dataset in HDFS.
- Write one code file for the mapper.
- Write another code file for the reducer.
- Run Hadoop to predict the class of the given input features.