PREDICTING STAFF PROMOTION USING THE EMPLOYEE PROMOTION DATASET

BY

AKOLADE BETTY ABIMBOLA

Course: CIS7031 – Programming for Data Analysis – 20 credits

Programming-driven analysis of a real-world dataset

May, 2022

Executive Summary

This report gives the analysis of the Employee Promotion dataset and the result obtained from the machine learning models (binary classifier) used to predict whether a staff was promoted or not.

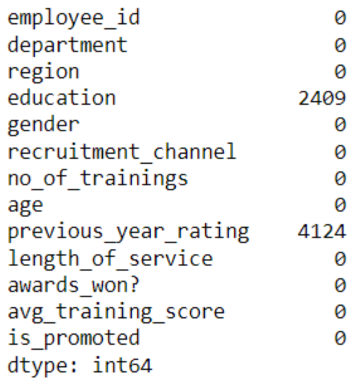
The dataset consists of 54,808 ad 23,490 train and test records respectively, having 12 features aside the target feature. After performing exploratory data analysis, data visualization, data cleaning, data processing and feature engineering, I was able to identify the features that are highly important for predicting the outcome of a staff being promoted. These features were further used in several machine learning binary classifiers to predict the probability of a staff being promoted. The catboost machine learning binary classifier proved to be the best by predicting the promotion of a staff with an accuracy of 94%, and having precision score of 94%.

The most important features in the dataset are:

- No_of_trainings: the number of trainings attended by the staff.
- previous_year_rating: the rating by the employer in the previous year.
- awards_won: an affirmation if the staff has ever won an award.
- avg_training_score: the average score obtained by the staff based on trainings attended.
- Overall_performance: the aggregate performance of the staff.
- region: the region where the staff falls.
- department: the department of the staff
- year_of_experience: the number of years of experience of the staff.

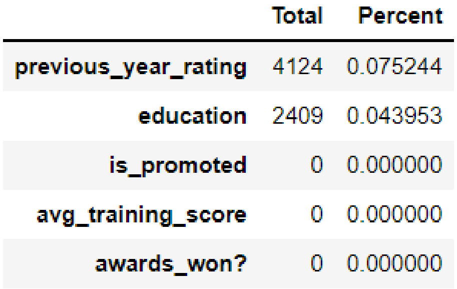
Exploratory Data Analysis

Firstly, I checked for the presence of missing values. This is important and needs to be addressed if any, because while some of the machine learning algorithms do not know how to handle missing data (values), others perform poorly in the presence of missing values.

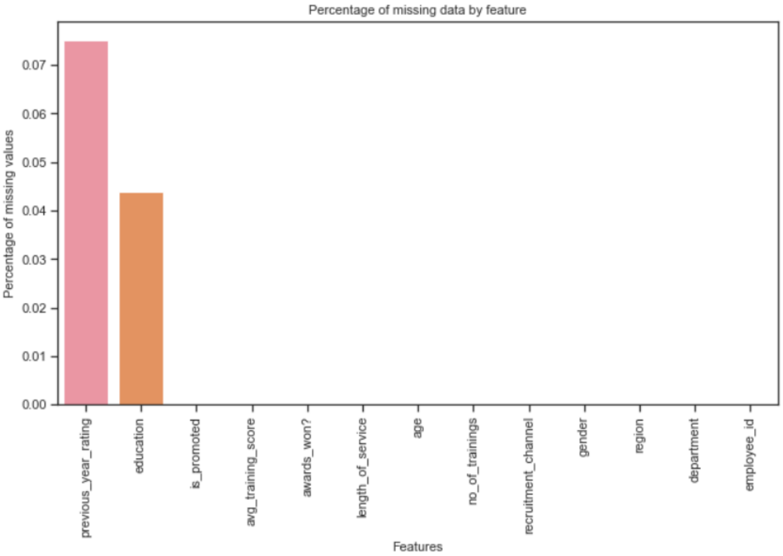


```
employee_id              0
department               0
region                   0
education             2409
gender                   0
recruitment_channel      0
no_of_trainings          0
age                      0
previous_year_rating  4124
length_of_service        0
awards_won?              0
avg_training_score       0
is_promoted              0
dtype: int64
```

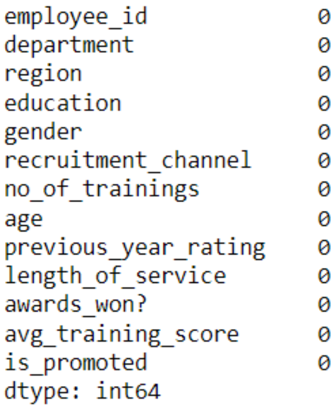
We can see that while the features education and previous_year_rating have 2,409 and 4,124 missing values respectively, the other features do not have missing values



	Total	Percent
previous_year_rating	4124	0.075244
education	2409	0.043953
is_promoted	0	0.000000
avg_training_score	0	0.000000
awards_won?	0	0.000000

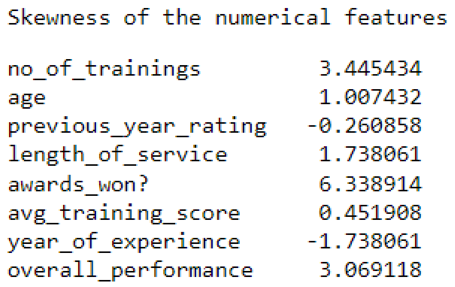


Bar charts were created to show the percentage of missing data by feature. The table and the bar chart show the percentage of missing values from each feature. One of the most important features previous_year_rating has 4,124 missing values, which accounts for 7% of the entire dataset, this is a big leap. We can see how dropping observations (rows) affect the shape of the dataset. Hence, we want to try to retain as much data as possible.



```
employee_id             0
department              0
region                  0
education               0
gender                  0
recruitment_channel     0
no_of_trainings         0
age                     0
previous_year_rating    0
length_of_service       0
awards_won?             0
avg_training_score      0
is_promoted             0
dtype: int64
```

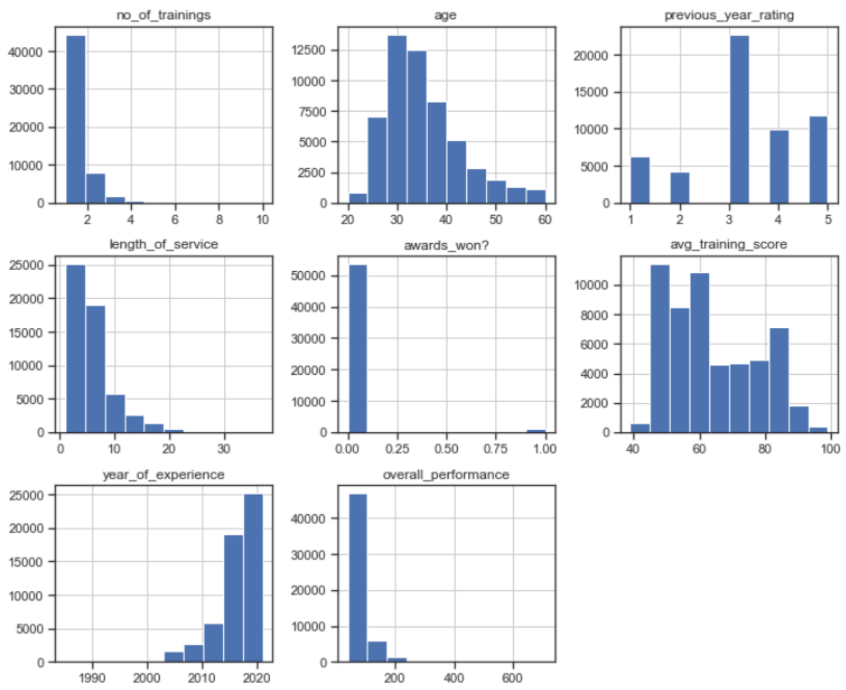
To resolve the problem of the missing values, I filled the categorical features missing values with the mode of each feature.



```
Skewness of the numerical features

no_of_trainings          3.445434
age                      1.007432
previous_year_rating    -0.260858
length_of_service        1.738061
awards_won?              6.338914
avg_training_score       0.451908
year_of_experience      -1.738061
overall_performance      3.069118
```

Histograms showing the skewness of the numerical features



Ideally, for a machine learning model to perform optimally, the dataset given must be close or follow a normal distribution. From the histograms and table, the numeric features no_of_trainings, awards won and overall_performance are highly skewed to the right, and this would affect the performance of our machine algorithm. To reduce the skewness, the affected features were scaled and encoded before applying the predictive model.

Now, all the numeric features almost follow the normal distribution. In addition, the categorical features were explored. These categories include:

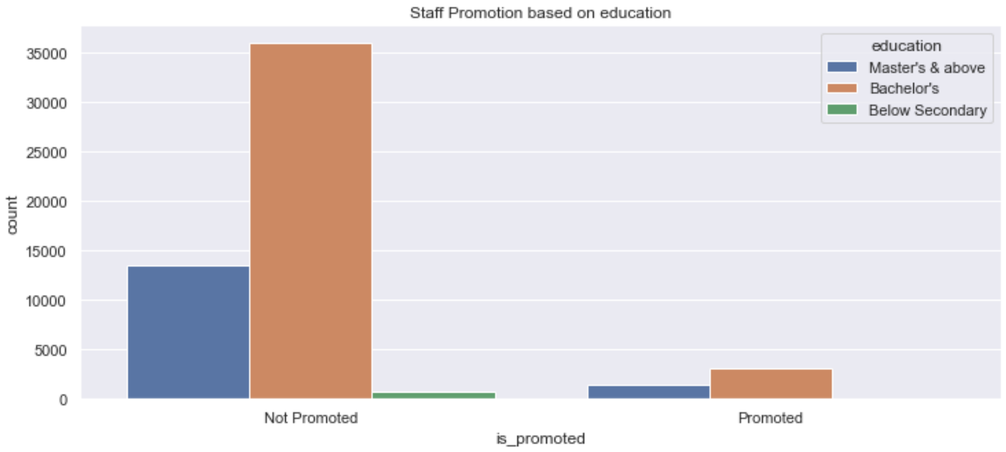
- recruitment_channel: sourcing, others, and referred
- education: Bachelor's, Below Secondary, and Master's & above
- gender: f—female, m-male
- region: region_1, region_2,….., region_34
- department: Analytics, Finance, HR, Legal, Operations, Procurement, R&D, Sales & Marketing, and Technology
- previous_year_rating: 1, 2, 3, 4 and 5
- awards: 1, 0

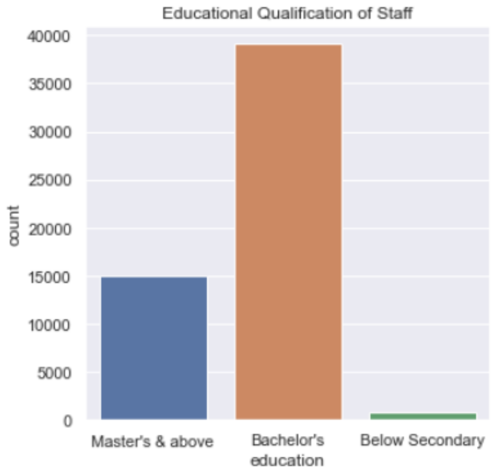
Bar charts were plotted to show the frequency of each of the categorical features, and we observed the following:

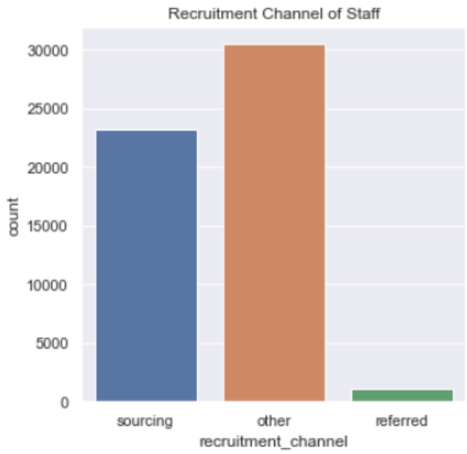
- Bachelor's education was the most common education among the staff, followed by 'Master's & above', while 'Below Secondary' was not common.
- The recruitment_channel, 'others' was the most common, followed by 'sourcing', and then 'referred'.

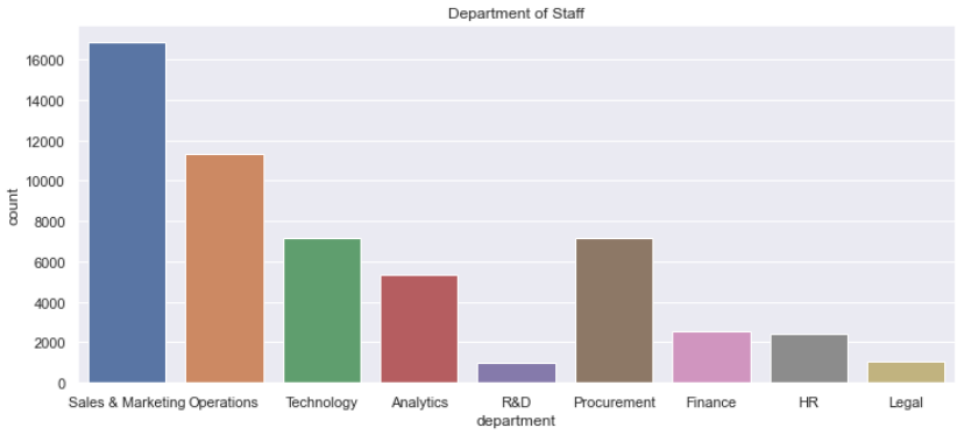
- Substantial percentage of the promoted staff are males. Based on the data, we can further say that preference was given to the males when it comes to promotion.
- Almost all the staff are in the Sales & Marketing department, while minimal number of staff are in the R&D department.
- Among the 34 regions, region_4 has the highest average number of promoted staff, while region_9 has the least average number of promoted staff
- Approximately 97.7% of the staff have never won an award while working for the company, while 2.3% have won an award.
- The number of staff that have received just a single training accounts for 80.9% of the total number of staff.
- 41.5%, 21.4%, 18%, 11.4%, and 7.7% of the staff had 3, 5, 4, 1, and 2 years previous year trainings respectively.

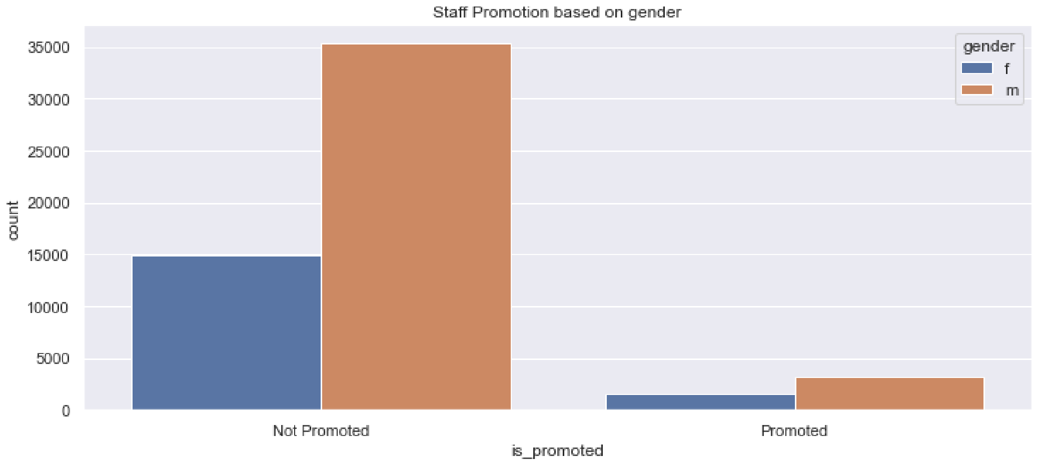
Based on the explanations given above, the graphical illustrations of the features in the dataset are further shown below

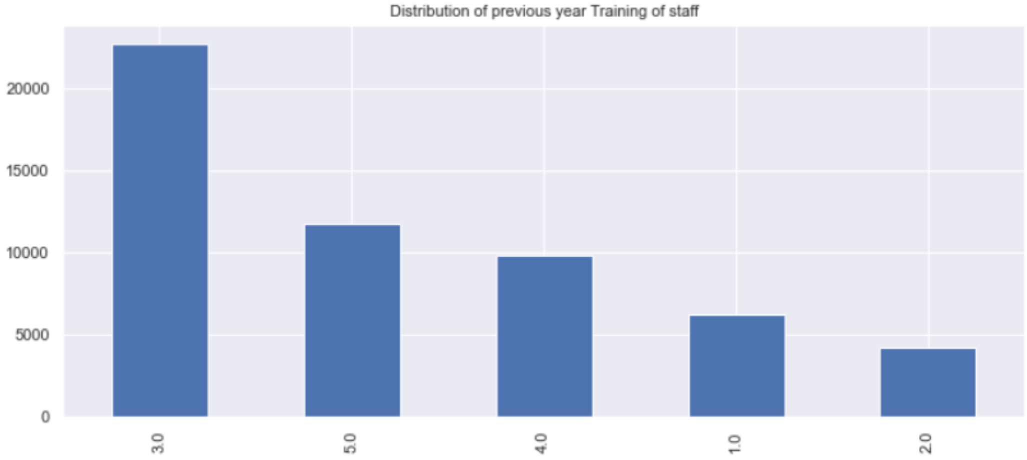

Staff Promotion based on education


Educational Qualification of Staff

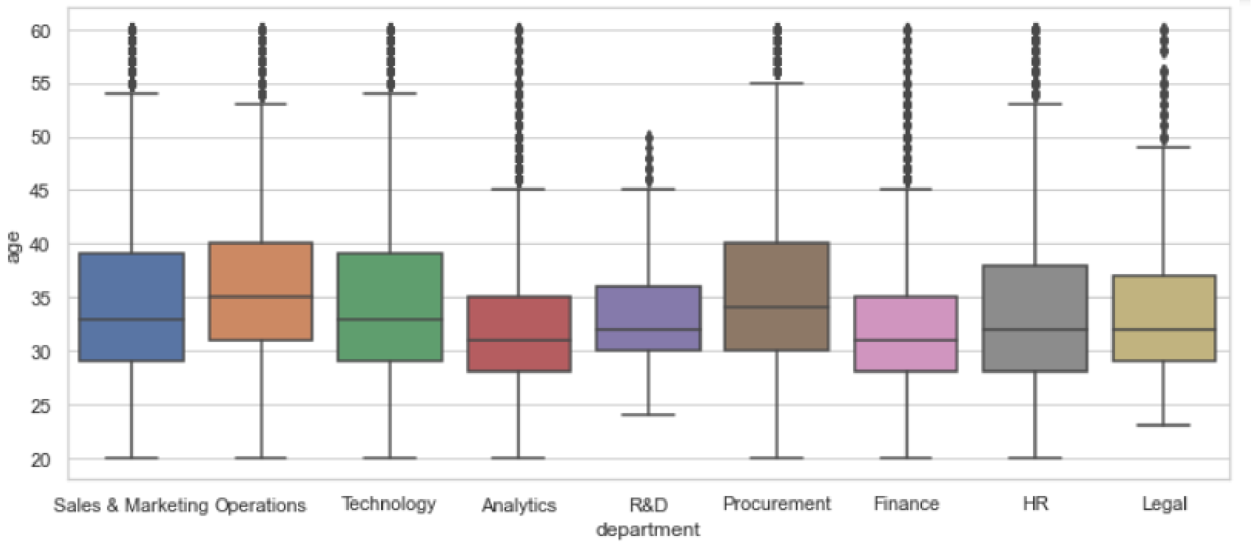

Recruitment Channel of Staff


Department of Staff

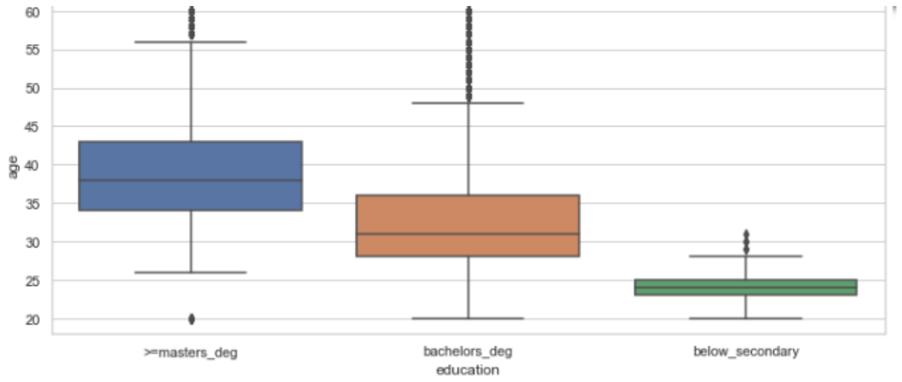


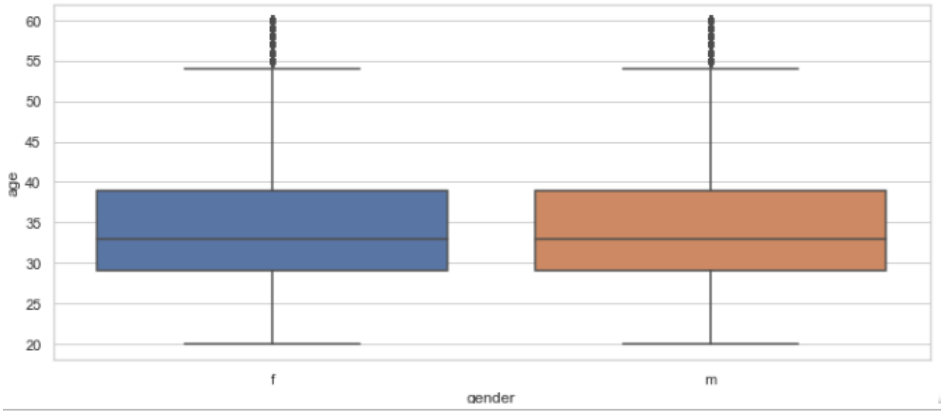


I further used the boxplot to establish the relationship between the age of the staff and some of the categorical features present in the dataset.



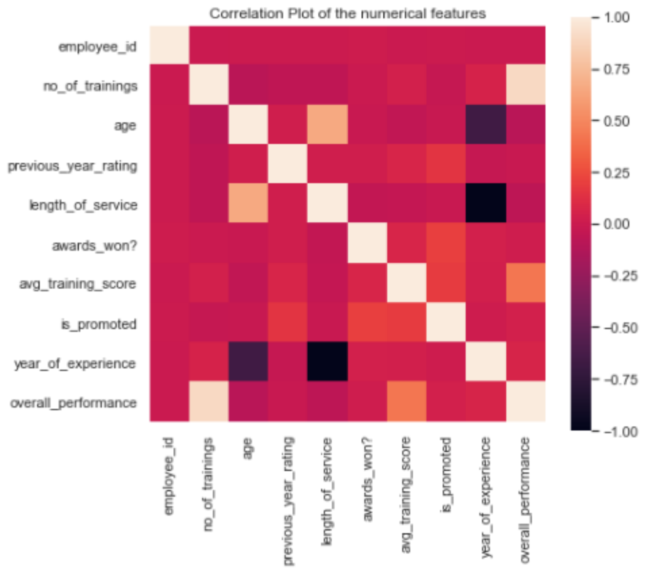
From the boxplot shown above, we can deduce that majority of the staff across the departments have an average of 33, maximum age range of 53, and a least age range of 20. There are also outliers (denoted by the dotted lines) in all the departments. Similar situation and observation is seen from the relationship between the age and other categorical features shown in the plots below

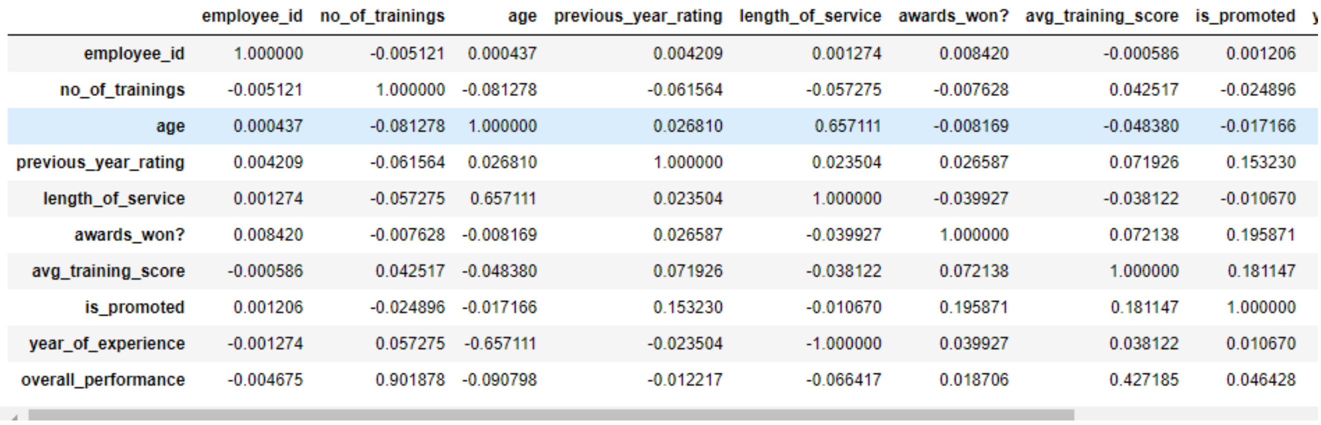




Correlation and Apparent Relationships

The relationship between the features in the dataset and the target variable (is_promoted) was plotted using the heatmap.





	employee_id	no_of_trainings	age	previous_year_rating	length_of_service	awards_won?	avg_training_score	is_promoted
employee_id	1.000000	-0.005121	0.000437	0.004209	0.001274	0.008420	-0.000586	0.001206
no_of_trainings	-0.005121	1.000000	-0.081278	-0.061564	-0.057275	-0.007628	0.042517	-0.024896
age	0.000437	-0.081278	1.000000	0.026810	0.657111	-0.008169	-0.048380	-0.017166
previous_year_rating	0.004209	-0.061564	0.026810	1.000000	0.023504	0.026587	0.071926	0.153230
length_of_service	0.001274	-0.057275	0.657111	0.023504	1.000000	-0.039927	-0.038122	-0.010670
awards_won?	0.008420	-0.007628	-0.008169	0.026587	-0.039927	1.000000	0.072138	0.195871
avg_training_score	-0.000586	0.042517	-0.048380	0.071926	-0.038122	0.072138	1.000000	0.181147
is_promoted	0.001206	-0.024896	-0.017166	0.153230	-0.010670	0.195871	0.181147	1.000000
year_of_experience	-0.001274	0.057275	-0.657111	-0.023504	-1.000000	0.039927	0.038122	0.010670
overall_performance	-0.004675	0.901878	-0.090798	-0.012217	-0.066417	0.018706	0.427185	0.046428

From the correlation table and heatmap plot, we can see that positively correlated features have correlation values close to 1, (e.g. as the age vs length of service features have a correlation value of 0. 65711, which implies as the age of the staff increases, the length_of_service also increases. In contrast, there is a negative correlation between the year_of_experience and the previous_year_rating. This means, as the year_of_experience decreases, the previous_year_rating increases (-0.023504).

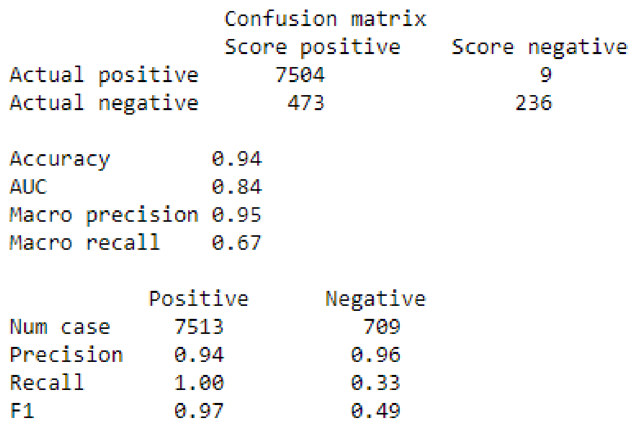
Feature Engineering

The education feature was regrouped due to the presence of the special characters i.e. & in the records, which was affecting the detection of those records by the machine learning algorithm.

Furthermore, using the existing features I created two new features/columns namely 'year_of_experience' and 'overall_performance'. The year_of_experience of a staff was calculated by subtracting his length_of_service from the current year, 2022. Likewise, the overall_performance of a staff was calculated by multiplying the 'no_of_trainings' by the 'avg_training_score'.

Predictive Analytics using Machine Learning Algorithms

After the analysis of the employee promotion, I tried to predict whether a staff will be promoted on no based on the historical data I had used to build a predictive model. High cardinality features i.e. region, length_of_service, avg_training_score, overall_performance, year_of_experience were encoded. Dummy variables were created for the categorical features, which lead to high increase in the size of the dataset. The dataset was then passed into scikit-learn standardScaler to scale the dataset. Scaling helps to resize the distribution of values to mean 0 and standard deviation 1. It makes all values have equal impact on the machine learning algorithm so as to avoid building a bias predictive model. After scaling, the dataset was split into train data and test data with 70% and 30% respectively. The train data was trained using several machine learning algorithms which included: *Logistic Regression, Gradient Boost Classifier, Ada Boost Classifier, KNN Classifier, xgboost, Random Forest Classifier and Decision Tree Classifier*. The accuracy obtained from these algorithms were below 94%, which were not acceptable. Although the Random Forest classifier, xgBoost classifier and gradient boost classifier had an accuracy of 94%, their AUC, precision and recall scores were low. Further research was made on the best algorithm for a dataset having high categorical features, the catboost algorithm proved to be a good choice for the dataset. This algorithm is good for dataset having many categorical features and high cardinality. It was observed that, unlike some algorithms that needed categorical features to be dummified and encoded, the catboost algorithm does not need this. The catboost algorithm was trained with 70% of the dataset, and later tested with the remaining 30% of the dataset. The following results were obtained:



```
                 Confusion matrix
                 Score positive    Score negative
Actual positive      7504               9
Actual negative       473             236

Accuracy        0.94
AUC             0.84
Macro precision 0.95
Macro recall    0.67

           Positive      Negative
Num case     7513          709
Precision    0.94          0.96
Recall       1.00          0.33
F1           0.97          0.49
```

Recommendation and Conclusion

The goal of this project was achieved, which was to maximize the accuracy of the classifier, and also obtain a high F1-score. I realized some form of nepotism or discrimination in our data. For example, a visual look at the education of the staff showed that the bachelor's degree holders are highly favored for promotion despite having staff with master and above degrees. The same thing applies for the staff's gender, almost all the staff given promotion were males. Although, my personal opinion is not needed for this project, but as a data scientist, if I find myself in this kind of organization, I will try to discuss with the management of the company on the importance of treating every person fairly, provided they meet the company's promotion requirements for the staff promotion.

Reference

- Dataset details: https://www.kaggle.com/datasets/arashnic/hr-ana