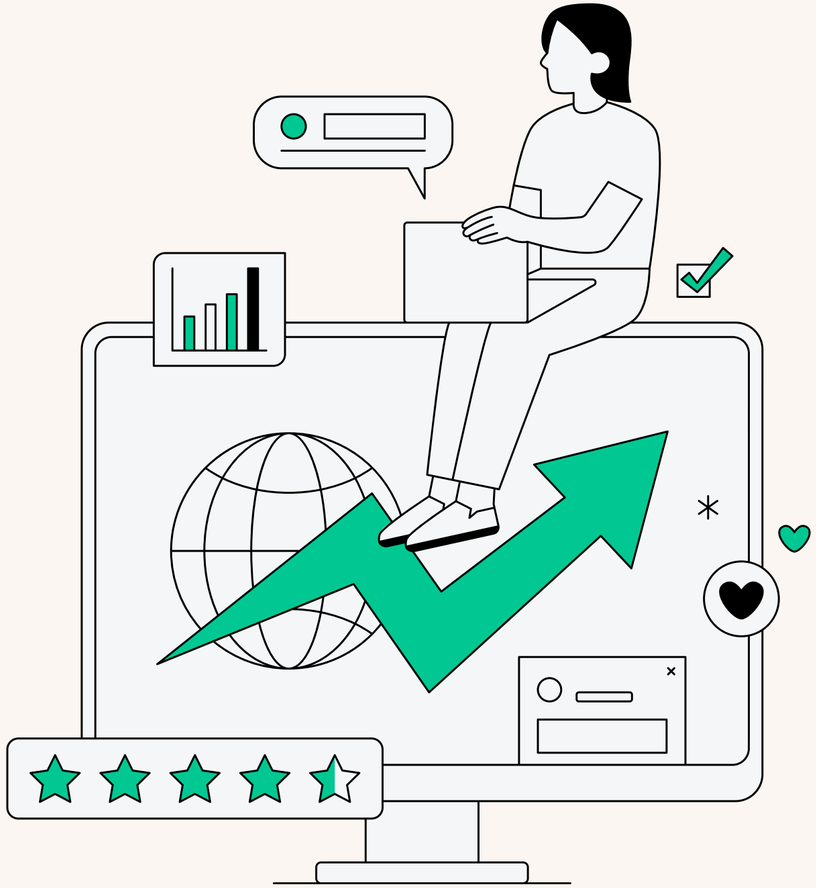
SC1015 Mini Project

Predicting Employee Promotion using Machine Learning

ECDS Team 3:

Kum Zhi Yan Marenda
Raksha Ramachandran
Back SunJin



Dataset from Kaggle

https://www.kaggle.com/datasets/arashnic/hr-ana?select=train.csv

	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won
0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	8	
1	65141	Operations	region_22	Bachelor's	m	other	1	30	5.0	4	
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	7	
3	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10	
4	48945	Technology	region_26	Bachelor's	m	other	1	45	3.0	2	
...	...	...	...	...	...	...	...	...	...	...	.
54803	3030	Technology	region_14	Bachelor's	m	sourcing	1	48	3.0	17	
54804	74592	Operations	region_27	Master's & above	f	other	1	37	2.0	6	
54805	13918	Analytics	region_1	Bachelor's	m	other	1	27	5.0	3	
54806	13614	Sales & Marketing	region_9	NaN	m	sourcing	1	29	1.0	2	
54807	51526	HR	region_22	Bachelor's	m	other	1	27	1.0	5	

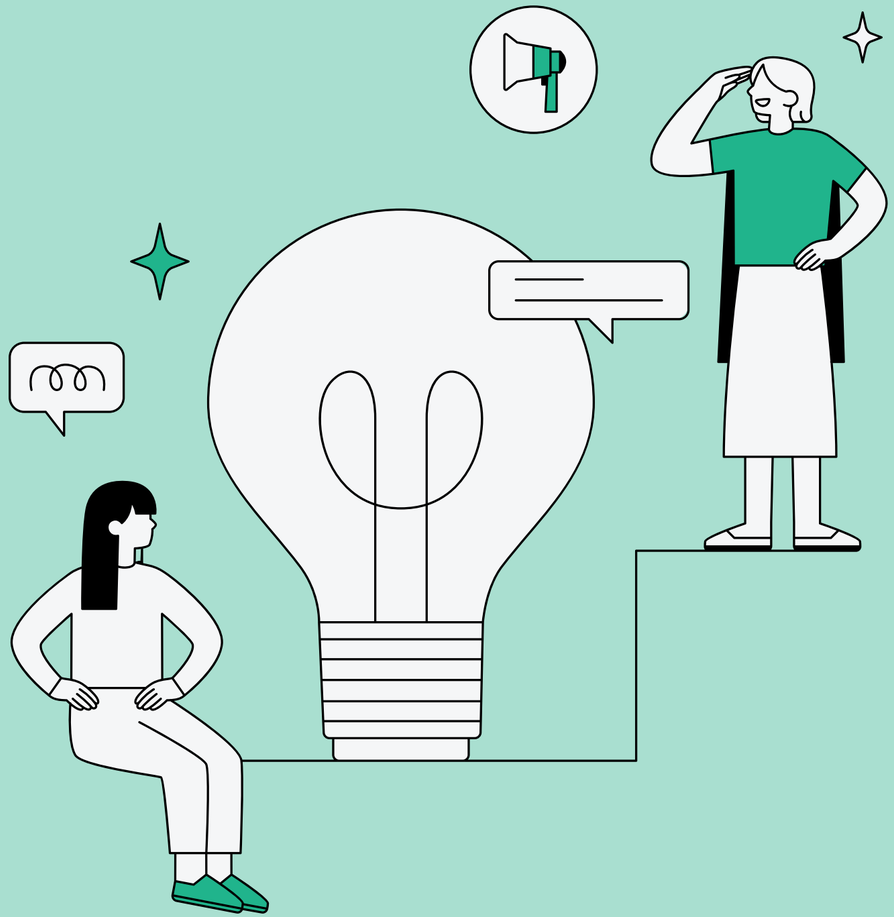
54808 rows × 13 columns

Background

Promotions are key milestones in an employee's career — affecting morale, salary, and career path

Thousands of employees are eligible for each promotion cycle.

67% of employees view performance evaluations as unfair, highlighting a lack of trust in promotion processes.



How can HR teams fairly identify the high-potential employees?

How can employees understand where they stand and how they can improve?

How do different variables affect *whether an employee is promoted* in different departments?

Why different departments?

We assume that different departments may have different criteria for promotion. We decided to analyse the top 3 departments – Sales & Marketing, Operations and Technology.

department	
Sales & Marketing	16840
Operations	11348
Technology	7138
Procurement	7138
Analytics	5352
Finance	2536
HR	2418
Legal	1039
R&D	999

What are the different variables?

<u>Categorical variables:</u>

region
education
gender
recruitment_channel
previous_year_rating
awards_won?
is_promoted

<u>Continuous variables:</u>

age
no_of_trainings
length_of_service
avg_training_score

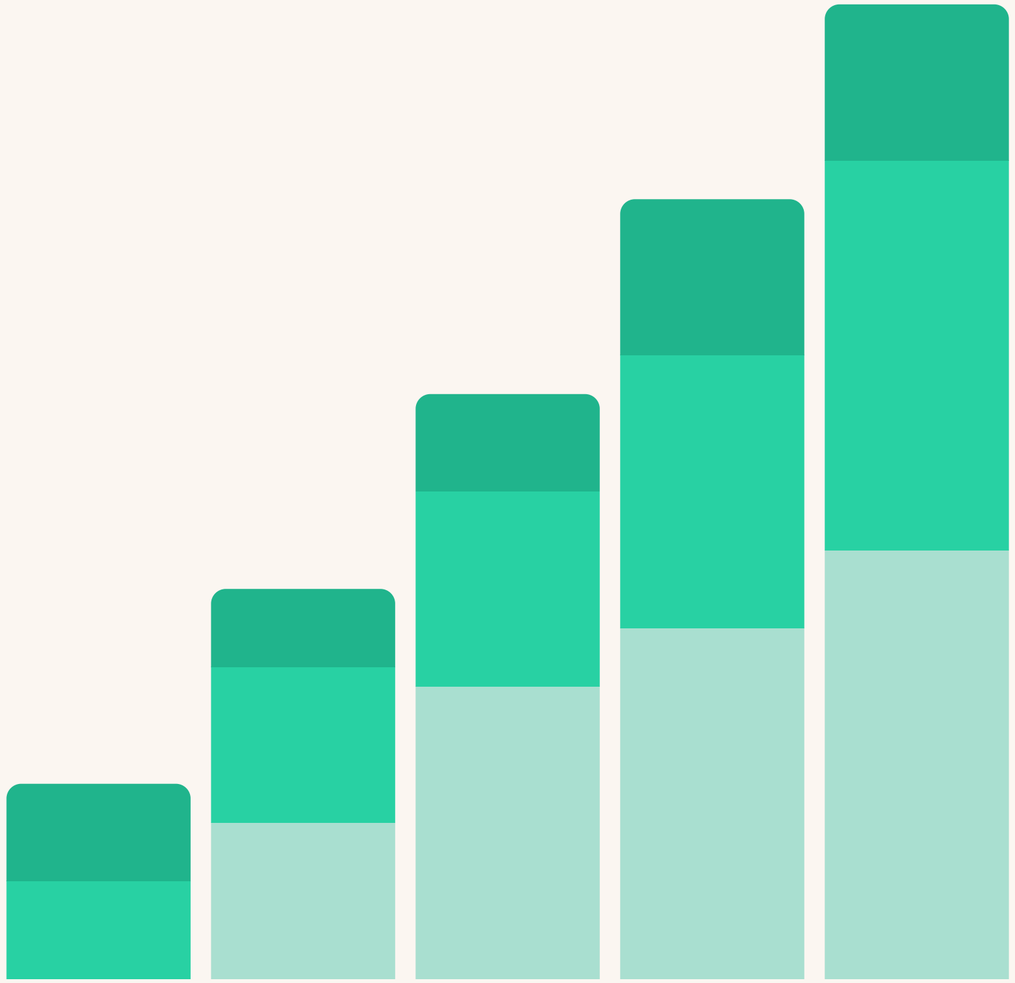
Exploratory Data Analysis

For categorical data:

- Bar Plot
- Mosiac Plot
- Cramér's V values

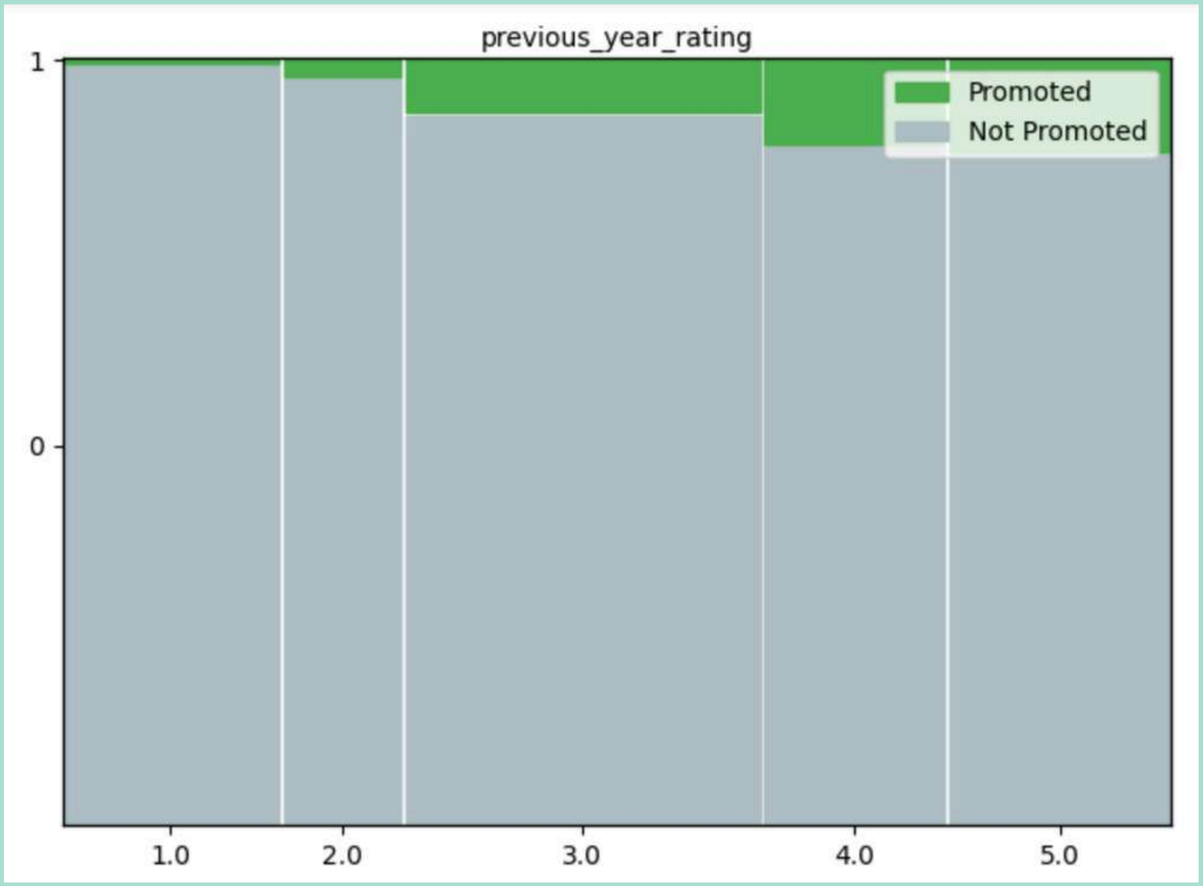
For continuous data:

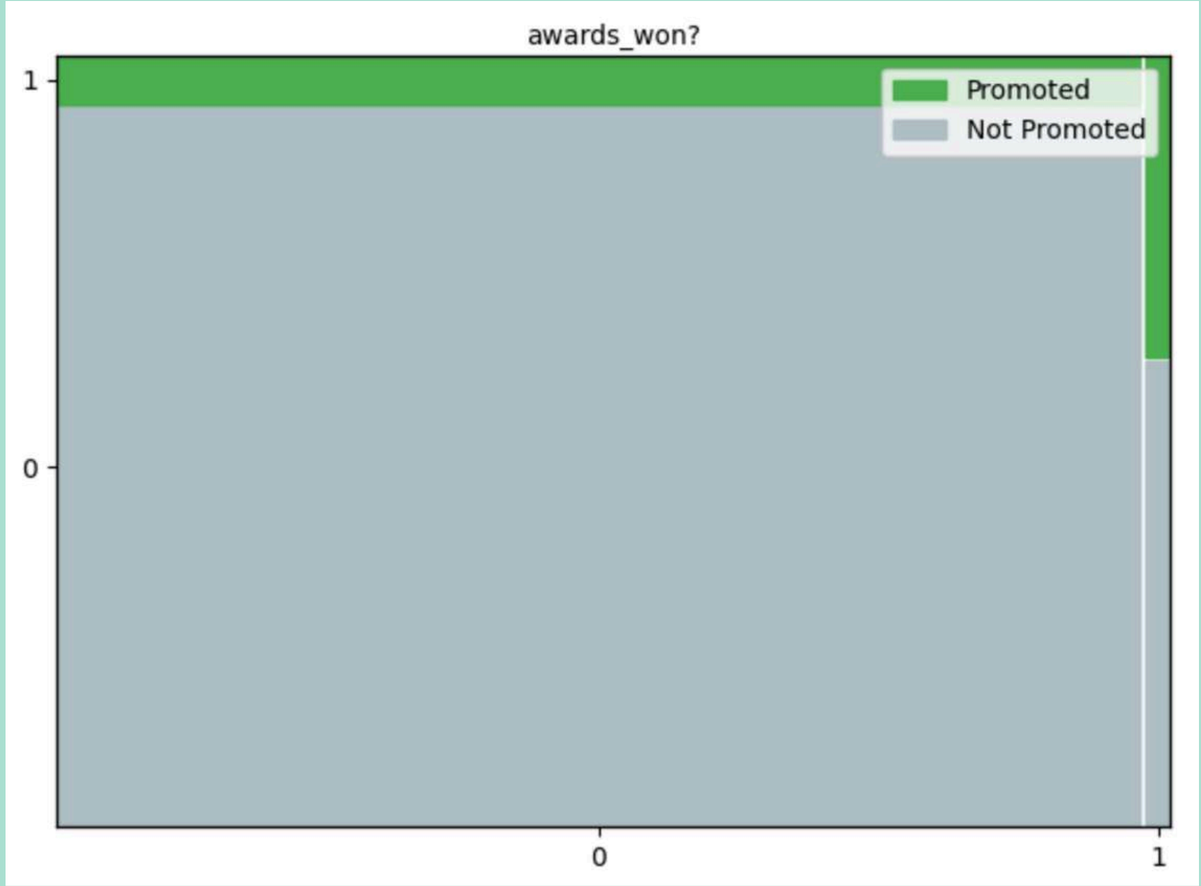
- Box Plot
- Correlation Coefficient



We performed EDA for each of the 3 departments individually.
Taking the *Sales & Marketing* as an example in the following slides:

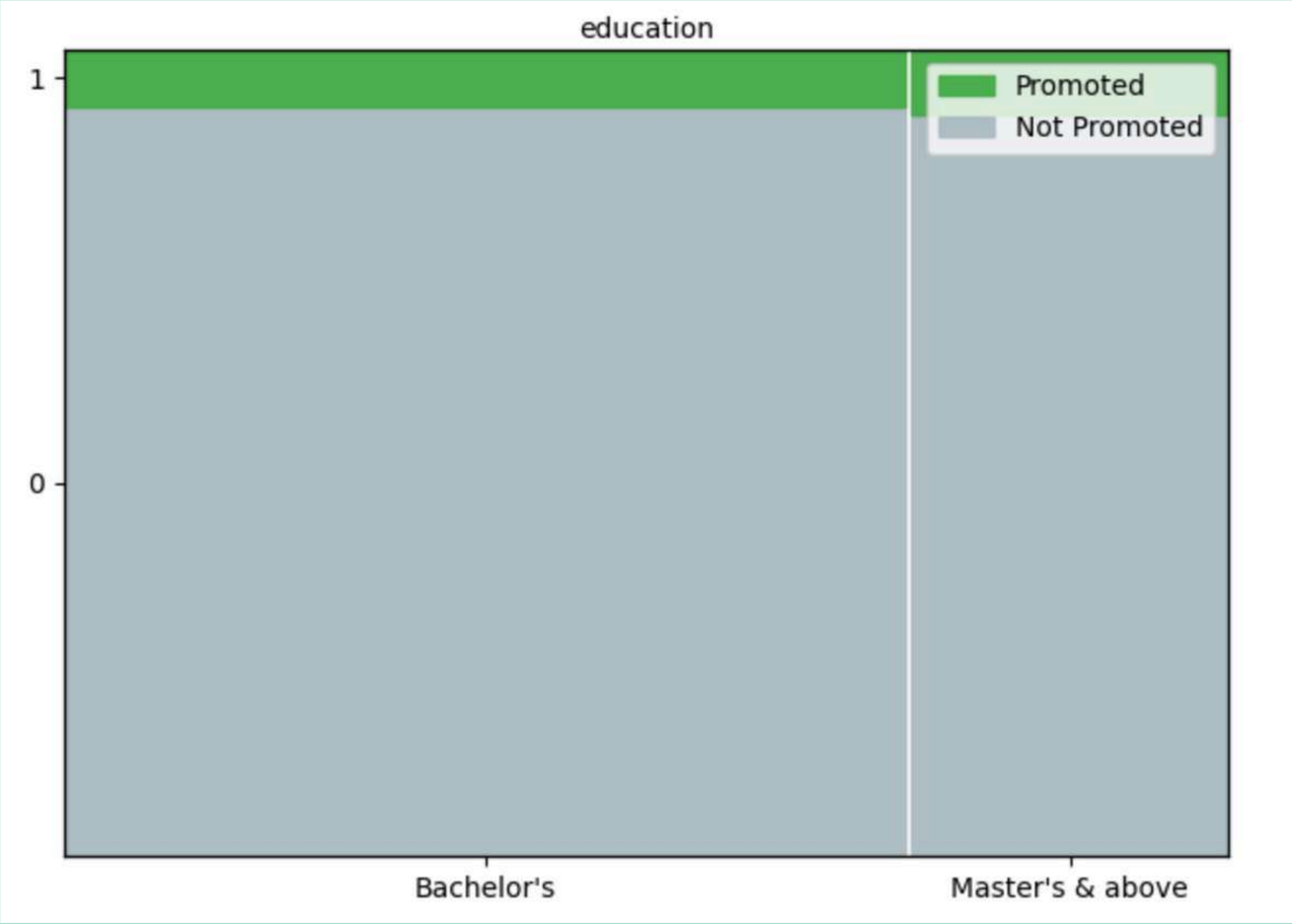
Categorial variables





Only ***previous_year_rating and awards_won?*** seem to have a relationship with is_promoted

Categorial variables



Other factors such as education for example do not show a clear relationship with is_promoted

Categorial variables

Cramér's V values: measures the strength of association between 2 categorial variables. Values:

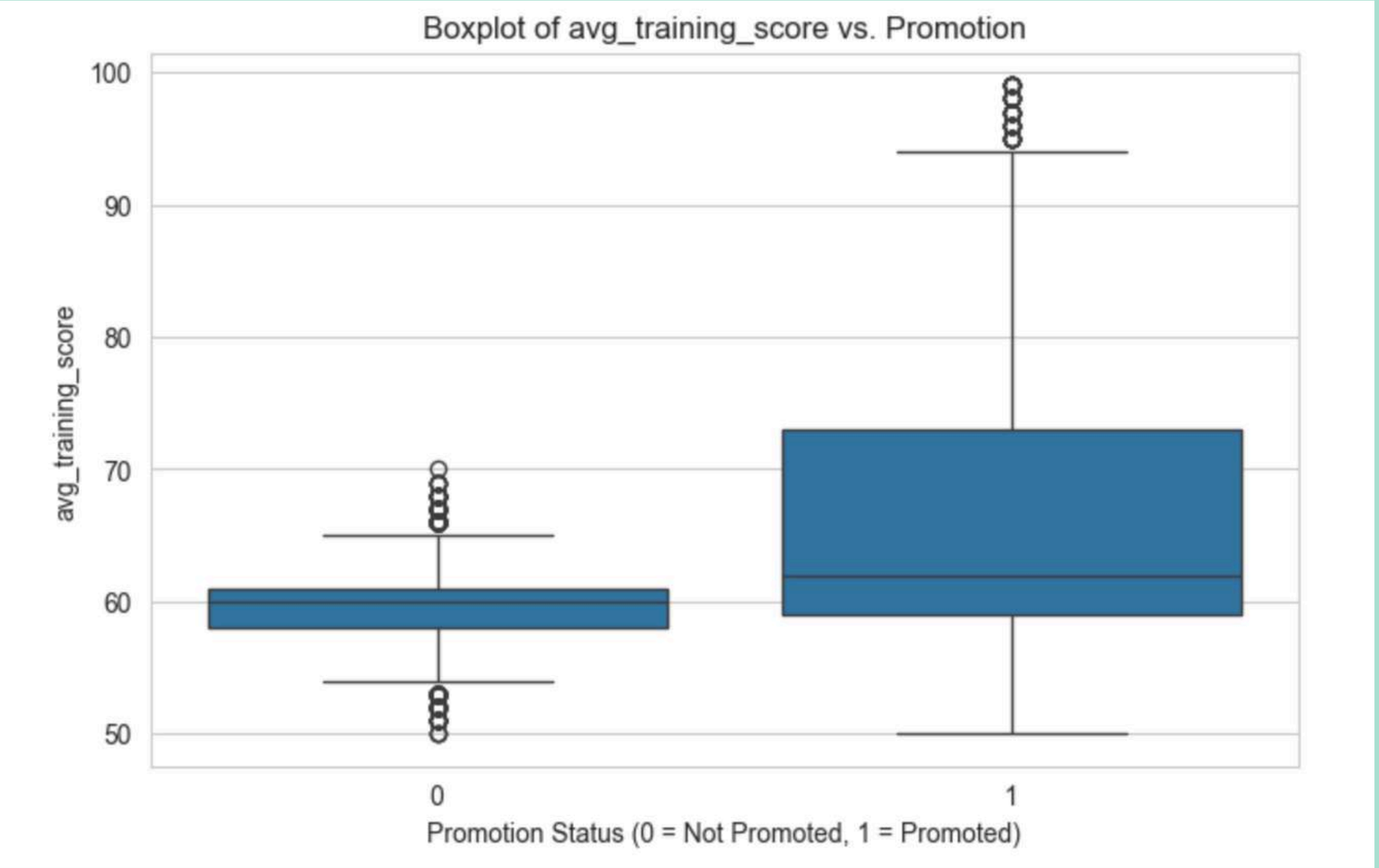
- Closer to 0 → very weak or no association
- ~0.1 to 0.3 → small to moderate association
- More than 0.3 → strong association

Previous_year_rating, awards_won? and region have a moderate association with is_promoted, and will be useful for training the model.

```
Cramér's V values with is_promoted:

previous_year_rating      : 0.1411
education                 : 0.0485
gender                    : 0.0124
recruitment_channel       : 0.0039
awards_won?               : 0.2067
region                    : 0.1219
```

Continuous variables



From the box plots, only avg_training_score seems to be somewhat correlated to promotion where a higher number correlates to a higher chance of being promoted.

The relationship of other continuous variables with is_promoted is not obvious from the box plots and violin plot.

Continuous variables

	no_of_trainings	age	length_of_service	avg_training_score	is_promoted
no_of_trainings	1.000000	-0.055069	-0.033504	0.018226	0.043493
age	-0.055069	1.000000	0.626715	0.025140	0.013982
length_of_service	-0.033504	0.626715	1.000000	0.024048	0.013368
avg_training_score	0.018226	0.025140	0.024048	1.000000	0.457027
is_promoted	0.043493	0.013982	0.013368	0.457027	1.000000

Only ***avg_training_score*** seems to have a decent correlation with is_promoted, and thus it will be useful for training the model.

How do different variables such as previous_year_rating, awards_won?, region and avg_training_score affect *whether an employee is promoted* in each of the different departments, Sales & Marketing, Operations and Technology?

Cleaning of Data

- Use SMOTE-NC to oversample the training data to balance the proportion of is_promoted
- Remove columns with low correlation values with is_promoted
- Replace NA & null values with 0 for previous_year_rating
- Change type float to string for categorical variables
- Encode categorical variables to convert them into numerical values. This is required to utilise for our models for machine learning later.

Machine Learning

01.

Decision Tree

(classifies our defined categorical variable (avg_training_score) using predictors)

02.

Random Forest

(combining multiple decision trees to produce more accurate predictions)

03.

Cat Boost

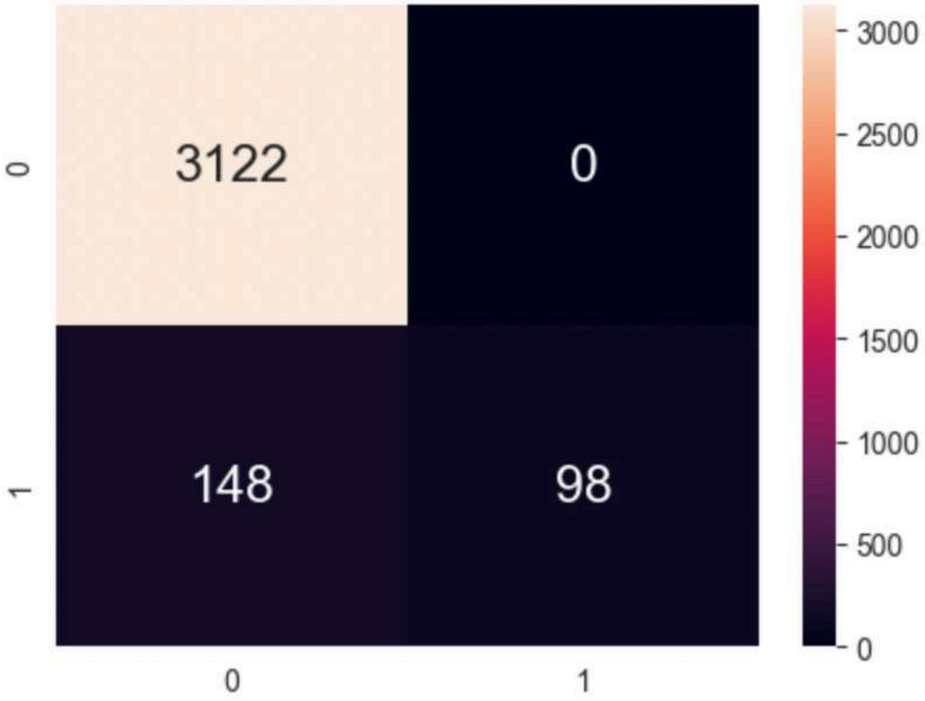
(handles categorical values without any encoding needed)

Hyperparameter tuning

(Sales and Marketing, Decision Tree, Test Data Set)

Before tuning

```
Classification Accuracy for train       : 0.8305077968812475
Classification Accuracy for test        : 0.9560570071258907
F1-Score        : 0.5697674418604651
Recall  : 0.3983739837398374
Time taken      : 0.040663719177246094
```



After tuning

```
Time taken to search for hyperparameters:       12.173465967178345
Best Hyperparameters: {'criterion': 'entropy', 'max_depth': 20, 'min_samples_split': 12}
Classification Accuracy for train       : 0.8272690923630548
Classification Accuracy for test        : 0.9560570071258907
F1-Score        : 0.5697674418604651
Recall  : 0.3983739837398374
```

Decision Tree
(Sales and Marketing)

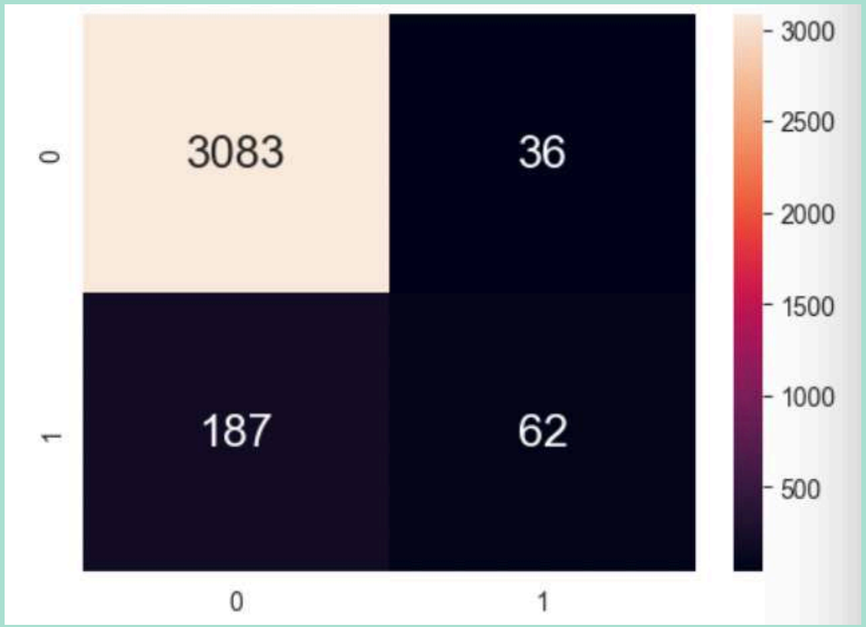
Goodness of Fit of Model: Train Data Set
Classification Accuracy: 0.831
True Negative Rate: 0.816
True Positive Rate: 0.847
False Negative Rate: 0.153
False Positive Rate: 0.184

Goodness of Fit of Model: Test Data Set
Classification Accuracy: 0.947
True Negative Rate: 0.988
True Positive Rate: 0.422
False Negative Rate: 0.578
False Positive Rate: 0.012

Random Forest
(Sales and Marketing)

Goodness of Fit of Model: Train Data Set
Classification Accuracy: 0.831
True Negative Rate: 0.811
True Positive Rate: 0.851
False Negative Rate: 0.149
False Positive Rate: 0.189

Goodness of Fit of Mode: Test Data Set
Classification Accuracy: 0.934
True Negative Rate: 0.988
True Positive Rate: 0.249
False Negative Rate: 0.751
False Positive Rate: 0.012



Cat Boost
(Sales and Marketing)

Goodness of Fit of Model: Train Data Set
Classification Accuracy: 0.815
True Negative Rate: 0.772
True Positive Rate: 0.857
False Negative Rate: 0.143
False Positive Rate: 0.228

Goodness of Fit of Model: Test Data Set
Classification Accuracy: 0.765
True Negative Rate: 0.771
True Positive Rate: 0.695
False Negative Rate: 0.305
False Positive Rate: 0.229

What we learned

Using different statistical models to find the relevant variables that affect our predictor

Implementing various machine learning functions

- Decision Tree
- Random Forest
- Cat Boost

Outcomes of project

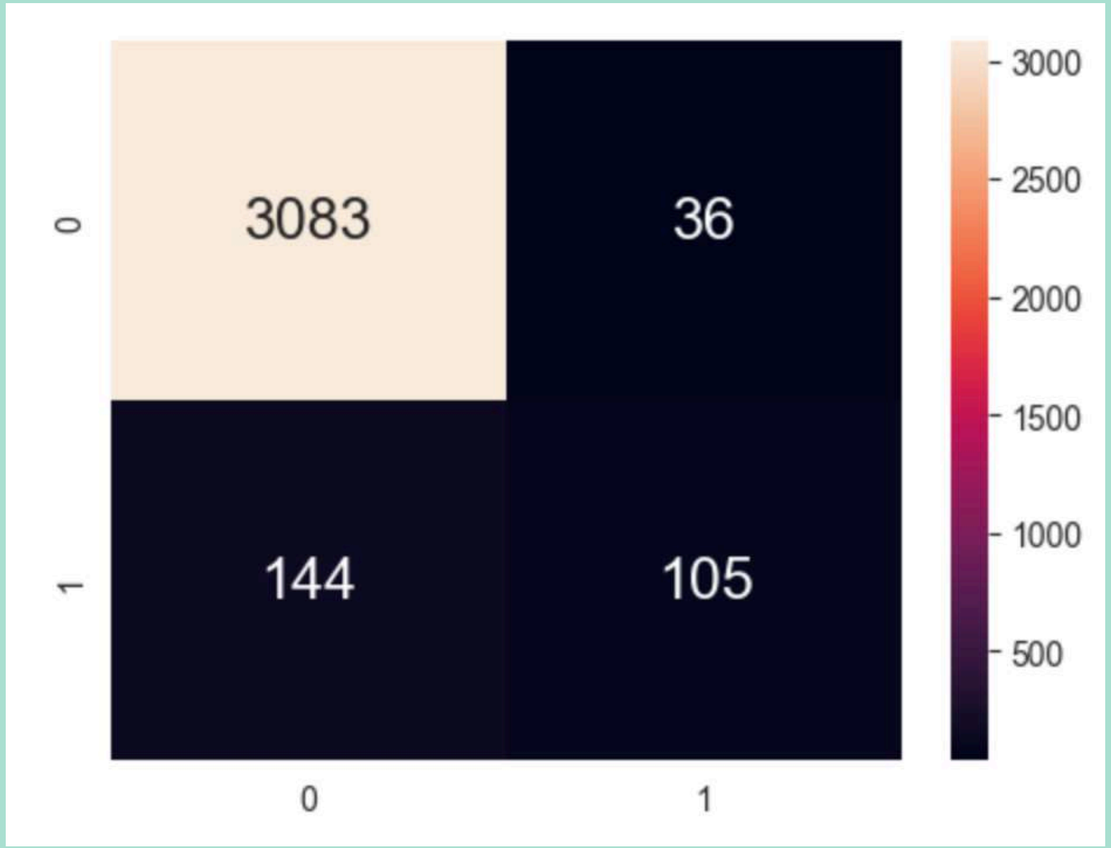
Through our analysis, employees can:

- understand which aspect of their resume they should build on
- adjust their performance to increase their likelihood of being promoted

Data Analysis

The classification accuracy for the test data is even higher than train data for our Decision Tree and Random Forest. Why?

Test Data (Decision Tree)



- (TN + FP)/ Total = 91.3%
- Looking at the Test Data, out of the workers with previous rating of 0, those who were not promoted made up 91.8%

Since the percentages are simliar, this shows that the previous_year_rating variable gives a straightforward result

Conclusion (data driven insights)

Decision Tree and Random Forest have a higher classification accuracy (over 90%) as compared to Catboost (average of 79%)

Decision Tree vs. Random Forest

- Provide similar statistics
- Accuracy is generally above 90% on the test data set
- F1-Score ranges from about 0.35-0.6

Cat Boost

- Lower classification accuracy
- Lower F1-Score
- Takes more time but can automatically handle categorical data without any encoding

Decision Tree

```
Classification Accuracy for train       : 0.8305077968812475
Classification Accuracy for test        : 0.9560570071258907
F1-Score        : 0.5697674418604651
Recall  : 0.3983739837398374
Time taken      : 0.040663719177246094
```

Random Forest

```
Classification accuracy for train       : 0.83046781287485
Classification accuracy for test        : 0.9569477434679335
F1-Score        : 0.5845272206303725
Recall  : 0.4146341463414634
Time taken      : 1.1758198738098145
```

CatBoost

```
Fitted: True
Classification accuracy for train: 0.810875649740104
Classification accuracy for test: 0.7568289786223278
F1-Score        : 0.2884448305821025
Recall  : 0.6747967479674797
Time taken      : 56.9407913684845
```

Conclusion (data driven insights)

Possible Improvements

- General predictors seem to be the same across the 3 departments (sales and marketing, operations, and technology). Therefore, it seems possible to implement these models to other departments as well.
- However, the models seem to be a better predictor for larger data sets as seen by the metrics presented from the Sales and Marketing department being the best out of the 3 departments.

Since the predictors are the same across all departments, aggregating all the departments into 1 dataset may provide better accuracy and prediction.

Thank you very much!