PREDICTING PROMOTABILITY

PROBLEM OVERVIEW

Accurately determining whether an employee is ready for a promotion is important to both the employee and the employer.

Understanding Potential	Employers can proactively look to start training and developing employees in key business areas and improve employee retention
Professional Development	Leaders can use findings as a tool to develop employee reports' strengths and areas of improvement
Reduce Bias	Standardizing promotion qualities can reduce implicit human bias in selecting employees for promotion

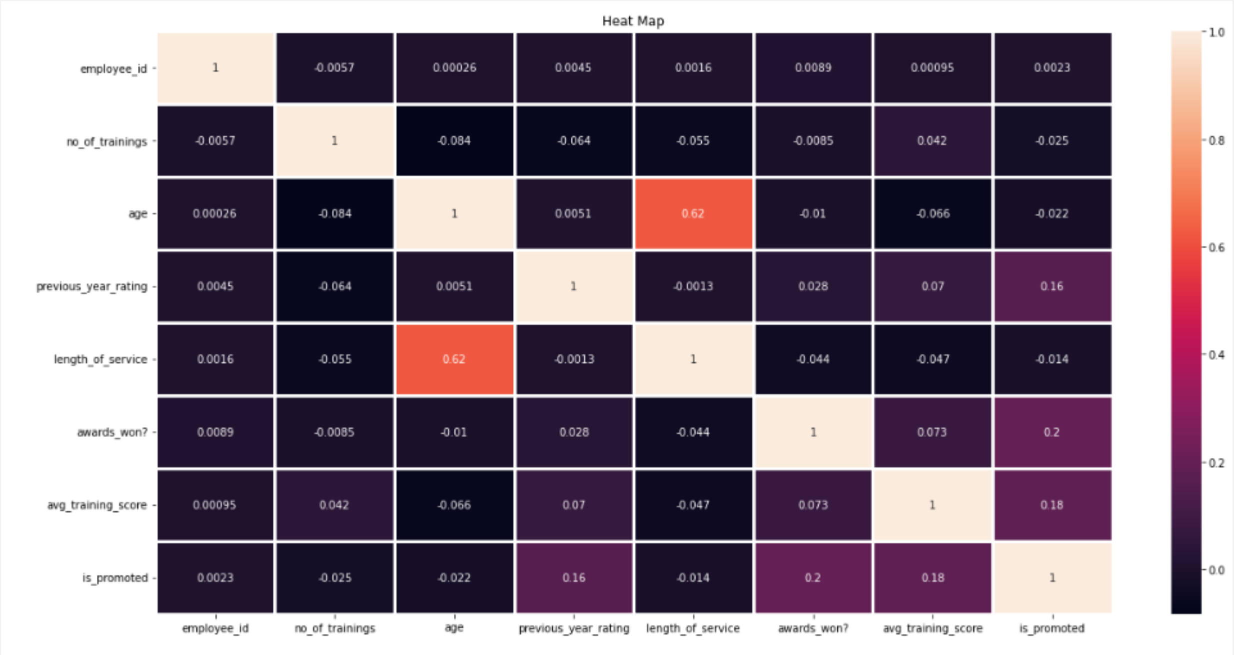
ABOUT THE DATA

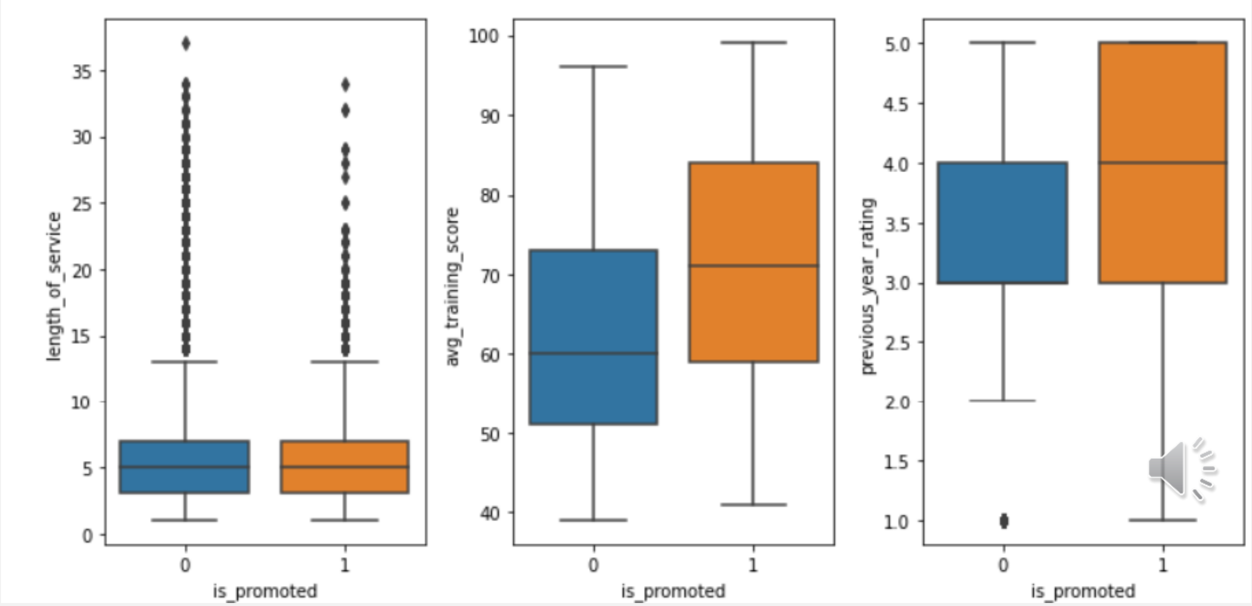
- Using fabricated dataset from Kaggle that mimics a real-world organization that is looking to identify the right candidates for promotion
- Includes employee workforce data and basic demographic data

•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
•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

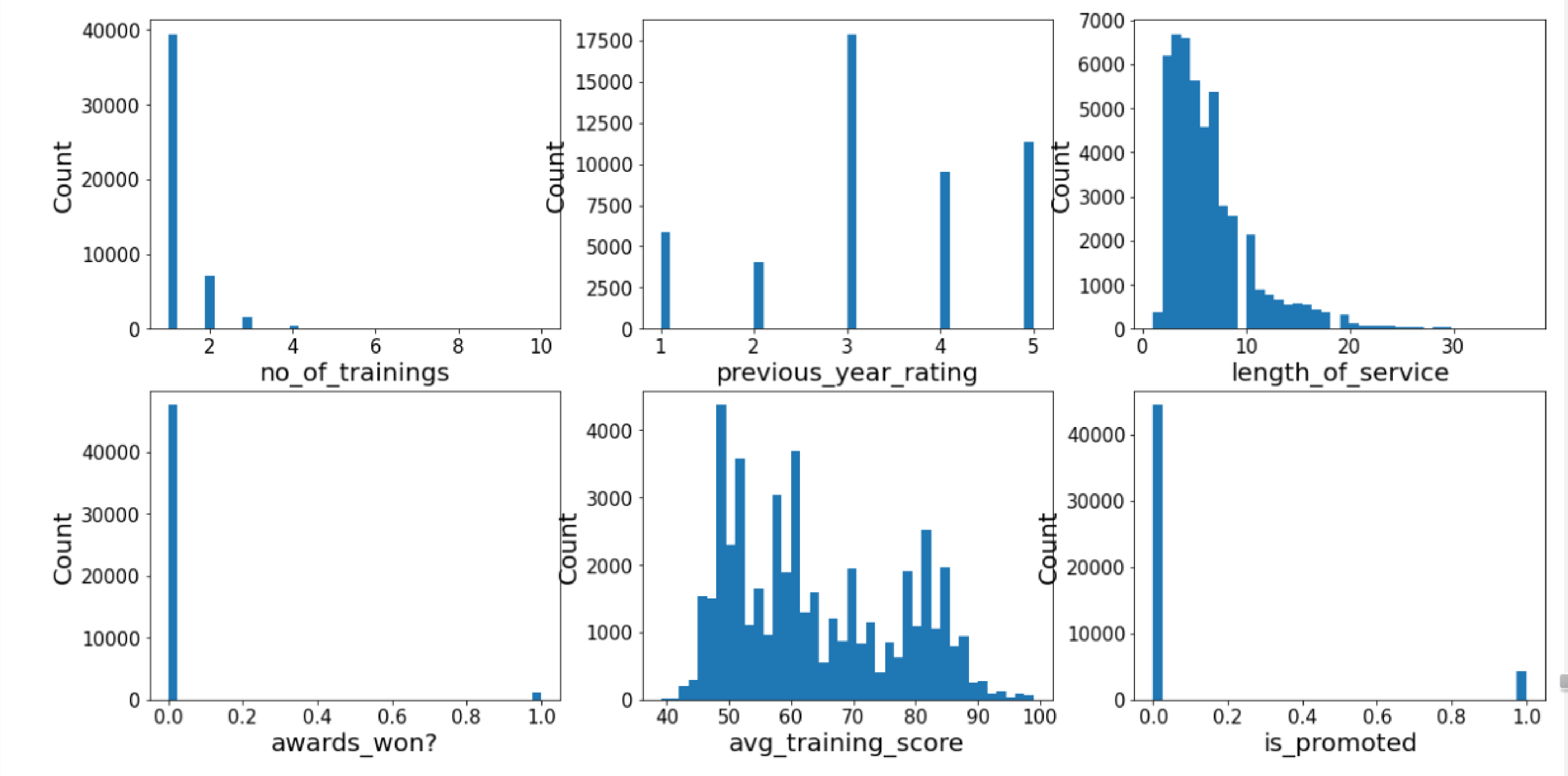
DATA CLEANSING AND PREPARATION

- Validated dataset for any NaN/Null values (no adjustments needed)
- Descriptive analysis was performed in order to better understand the basic statistics of the attributes (distribution, mean, median, min, and max)
- Evaluated the target field
  - Addressed class imbalance by adding class weights to penalize the minority class for misclassification

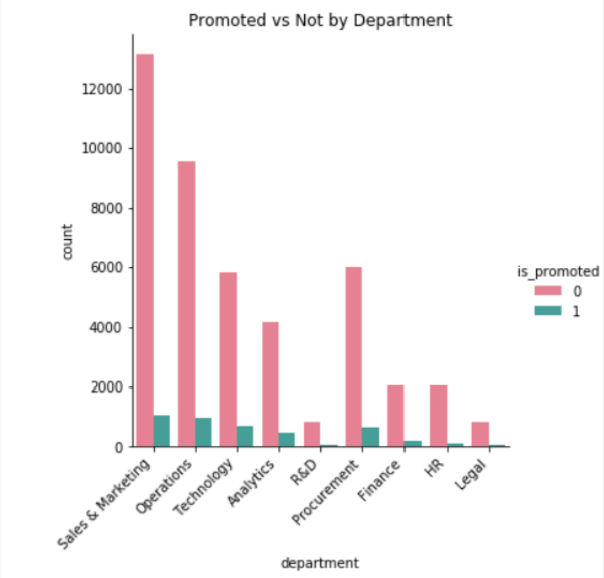


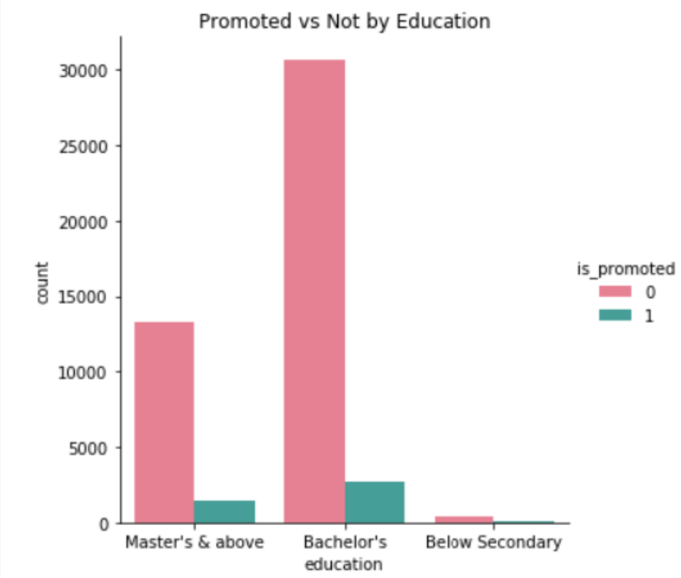


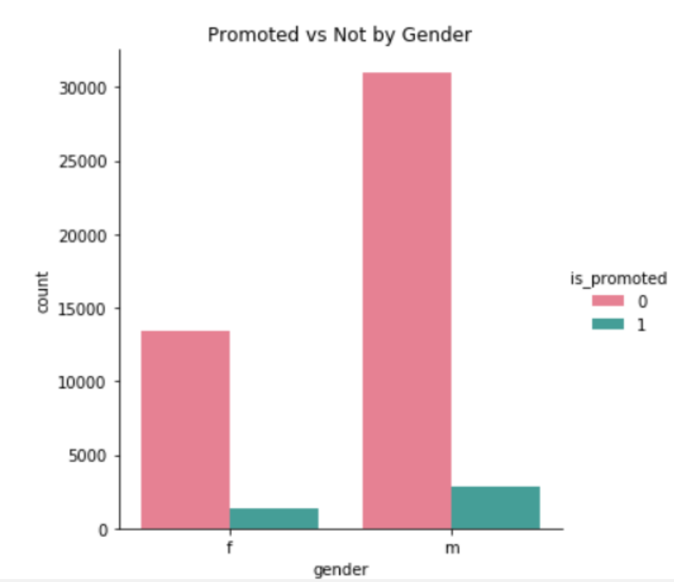
NUMERICAL FEATURES

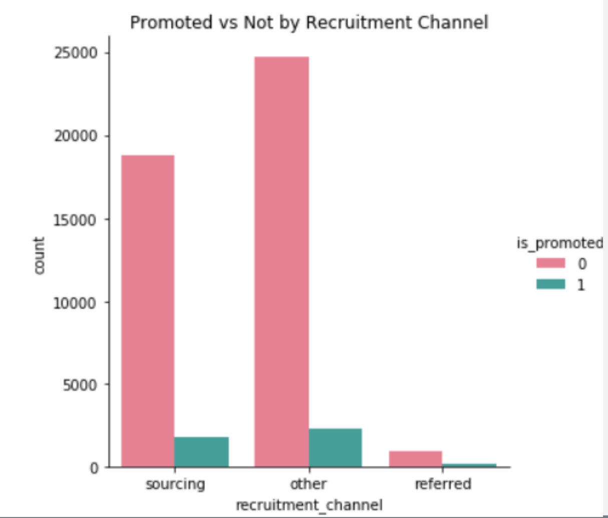


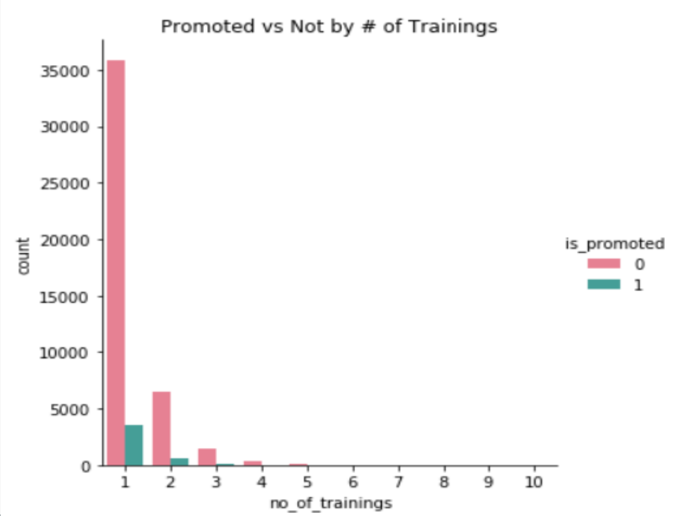
PROMOTED COMPARISONS

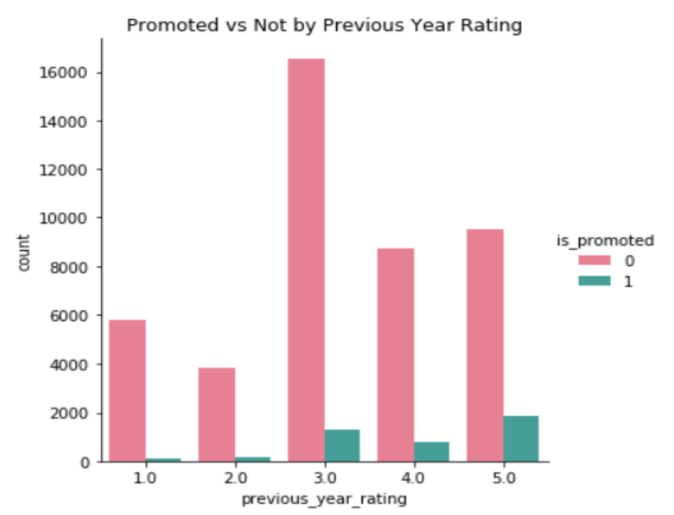






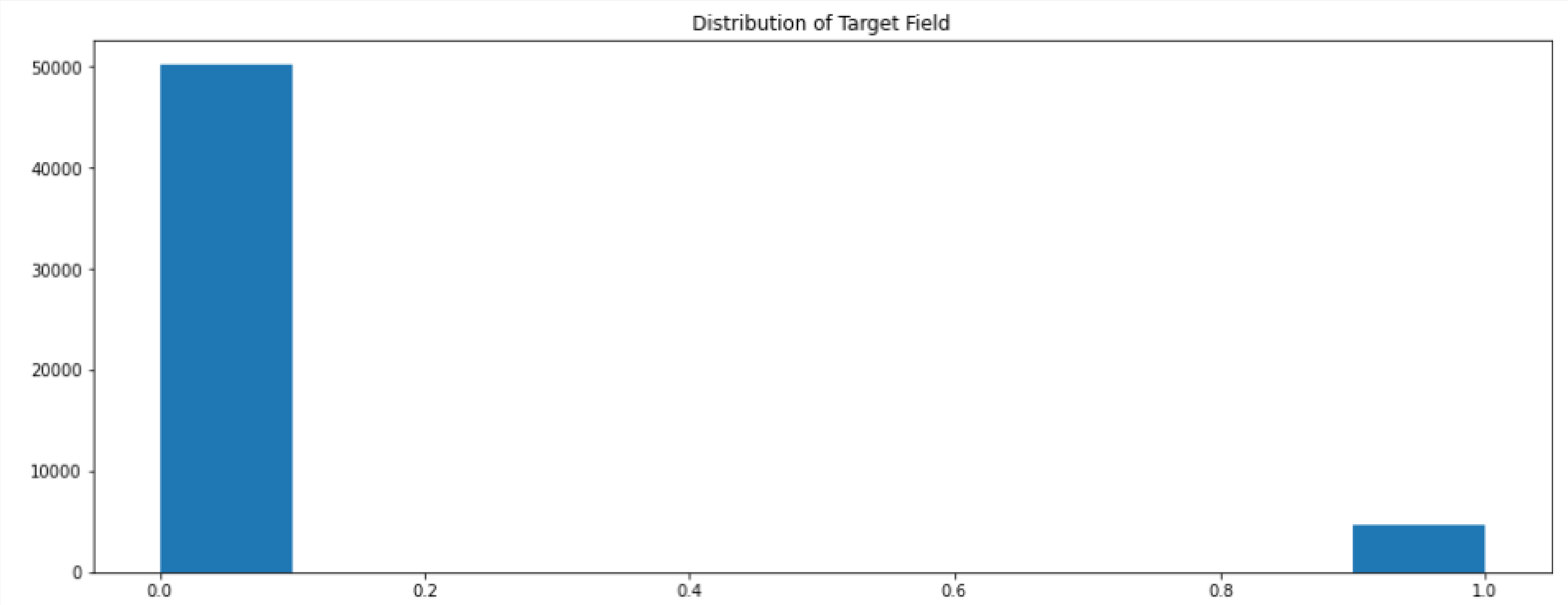






DEFINING THE TARGET

- Binary Indicator- “Is_Promoted”
- Higher density of non-promoted vs promoted (~9%), which could cause the model to skew towards a negative result
- Adjusted parameter “class weight” to lessen the majority value’s influence on the model



MODEL SELECTION + BUILD

Logistic Regression Model

- Ease of application
- Interpretability
- Well-suited for our binary target

Base Model

```
## add class weight to address 90/10 promoted split
lr = LogisticRegression(class_weight={0:0.1,1:0.9})
lr.fit(X_train,y_train)
```

```
LogisticRegression(class_weight={0: 0.1, 1: 0.9})
```

```
## set base model
base_model = lr
y_pred_base_model = base_model.predict(X_test)
pred_prob = base_model.predict_proba(X_test)
```

MODEL EVALUATION

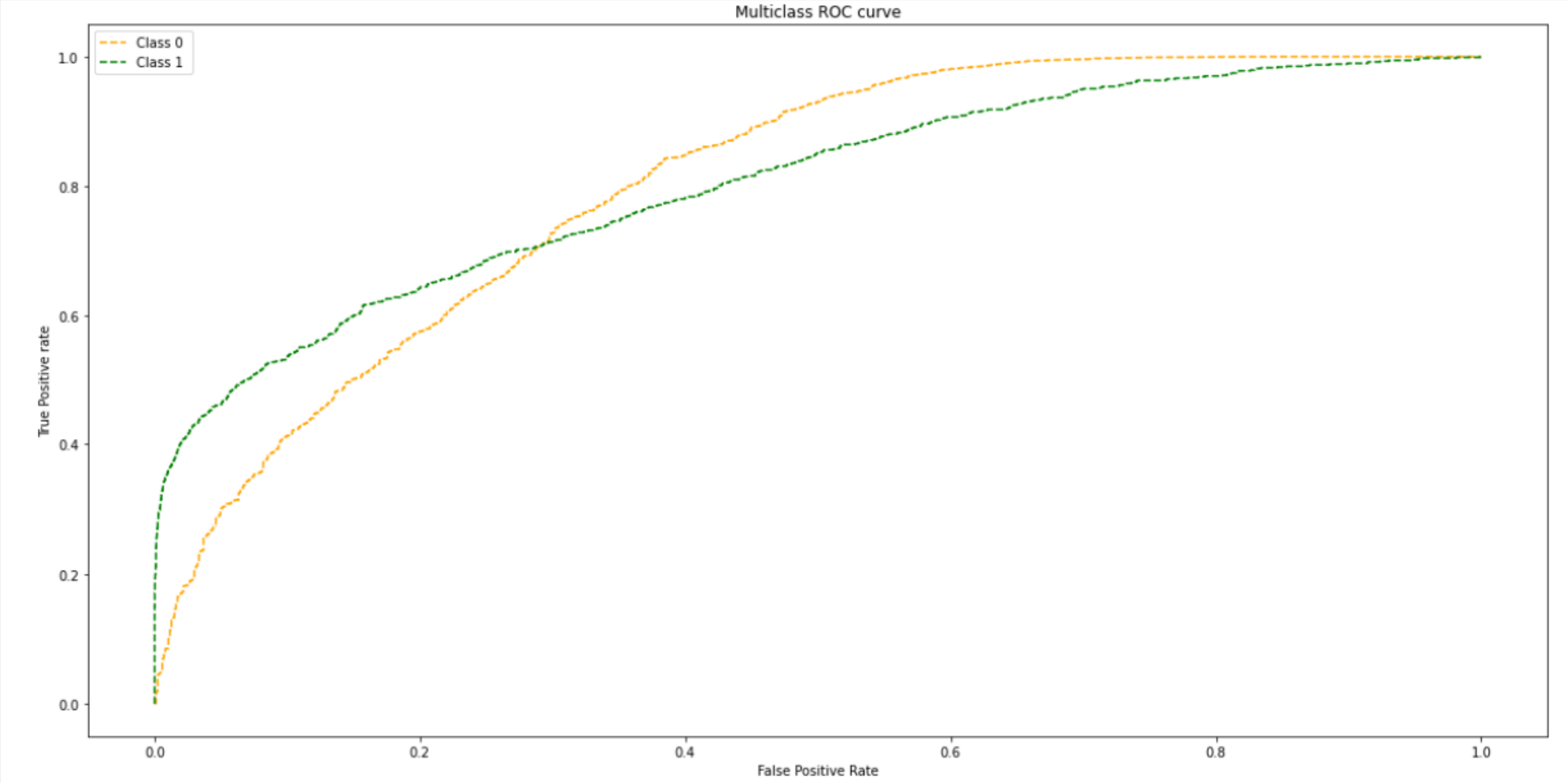
Confusion Matrix

7190	1675
320	547

- True Positive (Upper Left): 7190, which indicates where the model correctly predicted the positive class
- False Positive (Upper Right): 1675, which indicates where the model incorrectly predicted the positive class when it was actually negative (type 1 error)
- False Negative (Lower Left): 320, which indicates where the model incorrectly predicted the negative class when it was actually positive (type 2 error)
- True Negative (Lower Right): 547, which indicates where the model correctly predicted a negative class.

- **Model Accuracy Score:** .80

ROC CURVE



Multiclass ROC curve
Class 0
Class 1
True Positive rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

CHANGES TO FIELDS

FEATURE ENGINEERING

- Created two new columns from original dataframe from performance and productivity fields
  - 'sum_metric' combined 'awards_won?' and 'previous_year_rating' fields
  - 'total_score' = 'avg_training_score' and 'no_of_trainings'

DROPPED COLUMN

- Dropped 'region'

```python
## combine 'awards_won' and 'previous_year_rating'
## combine 'avg_training_score' and 'no_of_trainings'

#Creating a sum metric column
df['sum_metric'] = df['awards_won?']+ df['previous_year_rating']

# creating a total score column
df['total_score'] = df['avg_training_score'] * df['no_of_trainings']
```

```python
df['region'].unique()
```

```
array(['region_7', 'region_22', 'region_19', 'region_23', 'region_26',
       'region_2', 'region_20', 'region_34', 'region_1', 'region_4',
       'region_29', 'region_31', 'region_15', 'region_14', 'region_11',
       'region_5', 'region_28', 'region_17', 'region_13', 'region_16',
       'region_25', 'region_10', 'region_27', 'region_30', 'region_12',
       'region_21', 'region_32', 'region_6', 'region_33', 'region_8',
       'region_24', 'region_3', 'region_9', 'region_18'], dtype=object)
```

2ND MODEL EVALUATION

Confusion Matrix

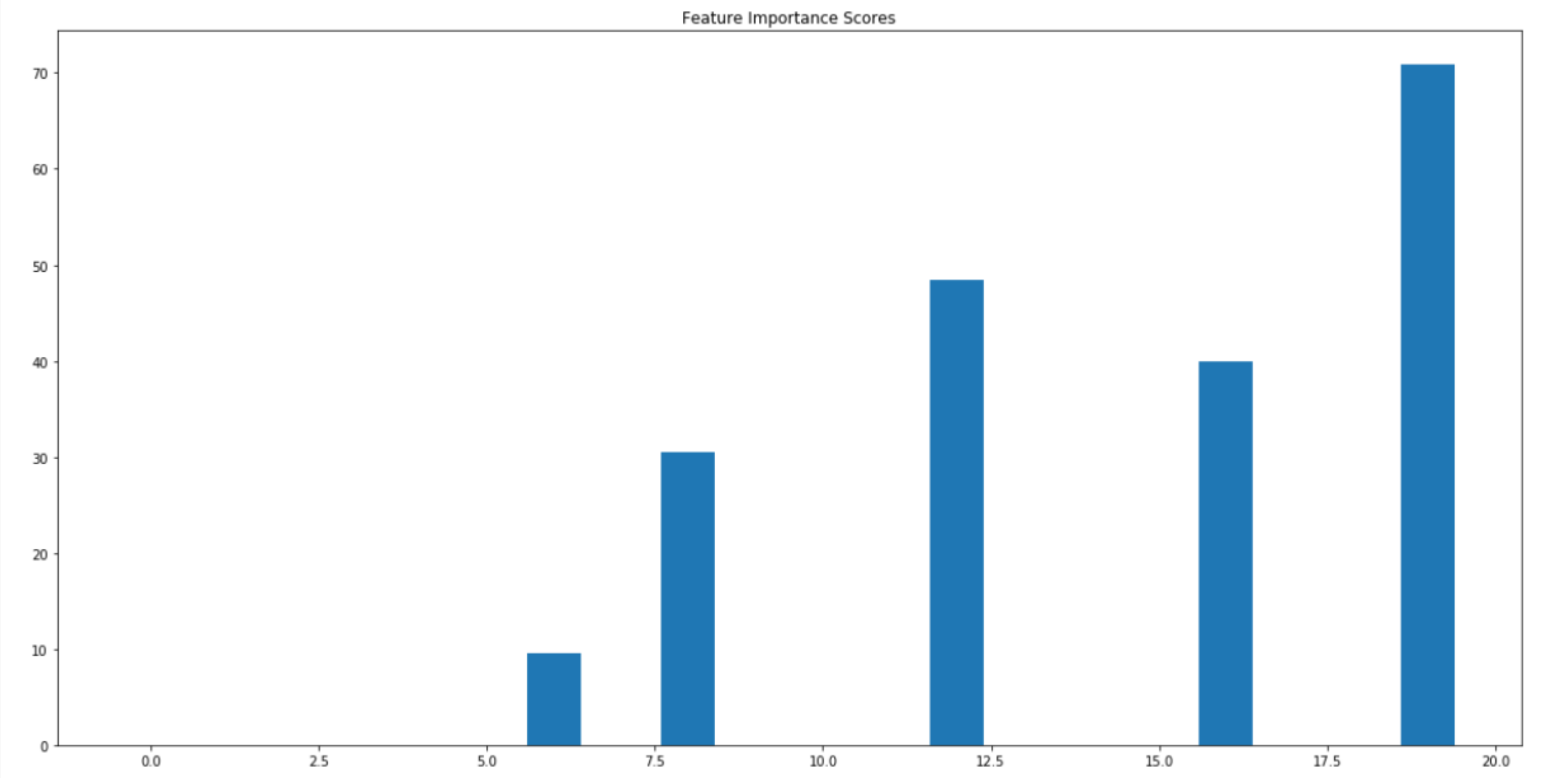
7190	**1675**
321	**546**

- True Positive (Upper Left): 7190, which indicates where the model correctly predicted the positive class
- False Positive (Upper Right): 1675, which indicates where the model incorrectly predicted the positive class when it was actually negative (type 1 error)
- False Negative (Lower Left): 321, which indicates where the model incorrectly predicted the negative class when it was actually positive (type 2 error)
- True Negative (Lower Right): 546, which indicates where the model correctly predicted a negative class.

- **Model Accuracy Score:** .79 (-.01 from 1st model)

FEATURE EVALUATION FOR IMPACT

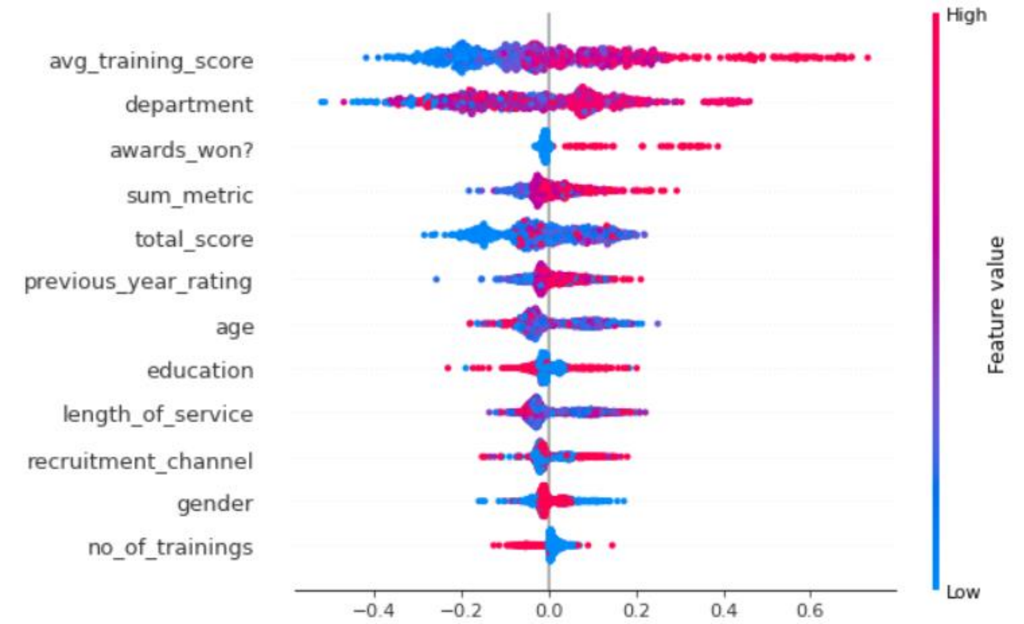
```
Feature: 0, Score: 0.00000
Feature: 1, Score: -0.00000
Feature: 2, Score: -0.00000
Feature: 3, Score: 0.00000
Feature: 4, Score: -0.00000
Feature: 5, Score: 0.00000
Feature: 6, Score: 9.61372
Feature: 7, Score: 0.00000
Feature: 8, Score: 30.51944
Feature: 9, Score: 0.00000
Feature: 10, Score: 0.00000
Feature: 11, Score: 0.00000
Feature: 12, Score: 48.38204
Feature: 13, Score: 0.00000
Feature: 14, Score: -0.00000
Feature: 15, Score: -0.00000
Feature: 16, Score: 39.99774
Feature: 17, Score: -0.00000
Feature: 18, Score: -0.00000
Feature: 19, Score: 70.86224
```



- Department had a high impact on predicting an employee's promotion
- Average Training Score and Length of Service followed

RESULTS

- Nominal difference in model performance
  - 2nd model interpreted two employees differently than the 1st mode
- Features had a strong difference of impact on predicting an employee's promotion
  - Workforce details like department and length of service had a noticeable impact
  - Demographic details such as age also contributed



ETHICAL IMPLICATIONS AND LIMITATIONS

- Dataset required is sensitive in nature; project data was fabricated due to difficulties procuring real-world data
- Ethical data collection and analysis pertinent to prevent bias, manipulation, or influence
- Project understands that there are limits to how well the data can portray people and their actions, only meant to guide and inform businesses
- There are many factors that influence whether an employee is qualified for a promotion; tested features were limited in availability and do not wholly represent impact for promotions

POTENTIAL FUTURE WORK

Expand Sample Dataset

- Dataset was fabricated in order to mimic a real-world company
- Future workstreams could expand to different companies to test the impact on department on promoted employees
- Additional workforce and demographic details could be tested

Different Models

- Final model results did not surpass 80%
- Future work could evolve this project's model's accuracy score or delve into alternative model options such as Random Forest Classifier

REFERENCES

Mobius. (2021). *HR Analytics: Employee Promotion Data*. Retrieved from Kaggle: https://www.kaggle.com/datasets/arashnic/hr-ana?select=train.csv