Business Report: Employee Promotion Prediction Using Machine Learning

Extended Project Report

Submitted to

By

Subhadeep Seal

In Partial Fulfillment of PDP-DSBA

CONTENTS

1. PROBLEM STATEMENT

1.1. Context

Employee Promotion means the ascension of an employee to higher ranks, this aspect of the job is what drives employees the most. The ultimate reward for dedication and loyalty towards an organization and the HR team plays an important role in handling all these promotion tasks based on ratings and other attributes available.

The HR team in JMD company stored data on the promotion cycle last year, which consists of details of all the employees in the company working last year and also if they got promoted or not, but every time this process gets delayed due to so many details available for each employee - it gets difficult to compare and decide.

1.2. Problem Definition

The Human Resources (HR) department at **JMD Company** is responsible for conducting employee performance appraisals and determining promotions. However, due to the **large volume of employee data** and the **complexity of comparing various employee attributes**, the promotion decision process often becomes delayed and inefficient.

To address this issue, the HR team aims to **leverage historical employee data** from the previous appraisal cycle and apply **machine learning techniques** to automate and **predict which employees are likely to be promoted** in the upcoming cycle. The goal is to develop a reliable and interpretable model that can help prioritize eligible candidates, making the appraisal process **faster, fairer, and data-driven.**

This project involves:

- Understanding the factors that influence promotion decisions.
- Preparing and analyzing historical employee data.
- Building and comparing machine learning models.
- Improving prediction accuracy through sampling techniques and hyper-parameter tuning.
- Delivering actionable recommendations to the HR team.

1.3. Objective

You as a data scientist at JMD Company need to come up with a model that will help the HR team to predict if a person is eligible for promotion or not.

1. Explore and visualize the dataset.
2. Build a classification model to predict if the customer has a higher probability of getting a promotion.
3. Optimize the model using appropriate techniques.
4. Generate a set of insights and recommendations that will help the company.

1.4. Data Description

The detailed data dictionary is given below.

Data Dictionary

- employee_id: Unique ID for the 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 the 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 the previous year then 1 else 0
- avg_ training_ score: Average score in current training evaluations
- is_promoted: (Target) Recommended for promotion

2.DATA OVERVIEW

- We will view the first 5 & last 5 rows of the dataset.

ex	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won	avg_training_score	is_promoted
0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	8	0	49.0	0
1	65141	Operations	region_22	Bachelor's	m	other	1	30	5.0	4	0	60.0	0
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	7	0	50.0	0
3	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10	0	50.0	0
4	48945	Technology	region_26	Bachelor's	m	other	1	45	3.0	2	0	73.0	0
index	**employee_id**	**department**	**region**	**education**	**gender**	**recruitment_channel**	**no_of_trainings**	**age**	**previous_year_rating**	**length_of_service**	**awards_won**	**avg_training_score**	**is_promoted**
54803	3030	Technology	region_14	Bachelor's	m	sourcing	1	48	3.0	17	0	78.0	0
54804	74592	Operations	region_27	Master's & above	f	other	1	37	2.0	6	0	56.0	0
54805	13918	Analytics	region_1	Bachelor's	m	other	1	27	5.0	3	0	79.0	0
54806	13614	Sales & Marketing	region_9	NaN	m	sourcing	1	29	1.0	2	0	NaN	0
54807	51526	HR	region_22	Bachelor's	m	other	1	27	1.0	5	0	49.0	0

Table 1: First 5 & last 5 rows of the dataset

2.1. Shape of the Dataset

- The dataset contains 54808 rows & 13 columns.

2.2. Check the type of data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54808 entries, 0 to 54807
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   employee_id           54808 non-null  int64  
 1   department            54808 non-null  object 
 2   region                54808 non-null  object 
 3   education             52399 non-null  object 
 4   gender                54808 non-null  object 
 5   recruitment_channel   54808 non-null  object 
 6   no_of_trainings       54808 non-null  int64  
 7   age                   54808 non-null  int64  
 8   previous_year_rating  50684 non-null  float64
 9   length_of_service     54808 non-null  int64  
 10  awards_won            54808 non-null  int64  
 11  avg_training_score    52248 non-null  float64
 12  is_promoted           54808 non-null  int64  
dtypes: float64(2), int64(6), object(5)
memory usage: 5.4+ MB
```

Table 2: Data types

There are 5 object data types, 6 integer data types, and 2 float data type in the dataset. All these features could be good predictors for Promotion eligibility.

2.3. Check for missing values

	0
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	2560
is_promoted	0

dtype: int64

Table 3: Missing Values

- There are 2409 missing values in education, 4124 missing values in previous_year_rating & 2560 missing values avg_training_score the dataset.

2.4. Statistical summary of the dataset

	count	mean	std	min	25%	50%	75%	max
employee_id	54808.000	39195.831	22586.581	1.000	19669.750	39225.500	58730.500	78298.000
no_of_trainings	54808.000	1.253	0.609	1.000	1.000	1.000	1.000	10.000
age	54808.000	34.804	7.660	20.000	29.000	33.000	39.000	60.000
previous_year_rating	50684.000	3.329	1.260	1.000	3.000	3.000	4.000	5.000
length_of_service	54808.000	5.866	4.265	1.000	3.000	5.000	7.000	37.000
awards_won	54808.000	0.023	0.150	0.000	0.000	0.000	0.000	1.000
avg_training_score	52248.000	63.712	13.522	39.000	51.000	60.000	77.000	99.000
is_promoted	54808.000	0.085	0.279	0.000	0.000	0.000	0.000	1.000

Table 4: Statistical summary

- In the above table we can see the counts, mean, standard deviation, minimum value and maximum value of numerical features.

2.5. Observations

- **Department Distribution**: The dataset includes multiple departments, with a notable presence of Sales & Marketing, Operations, and Technology.
- **Education Level**: Most employees have at least a Bachelor's degree, with some holding Master's degrees or higher.
- **Gender Representation**: The dataset includes both male and female employees, but the gender distribution may be imbalanced, with some entries missing gender data.
- **Training and Development**: Employees have attended varying numbers of training sessions, and their average training scores differ, indicating different levels of engagement in training.
- **Promotion Status**: The 'is_promoted' column indicates that very few employees were promoted.
- This dataset can be used for various analyses related to employee performance, promotion trends, and organizational development strategies.
- ID column consists of unique IDs which will not add value to the modeling & hence this column is dropped for further analysis.

3.EXPLORATORY DATA ANALYSIS (EDA)

3.1. Univariate Analysis

- Revealed distributions of no_of_trainings, age, length_of_service, Average_training_score, Department, Education, gender, recruitment_channel, previous_year_rating, awards_won, region & is_promoted. Labeled Bar plots & Histogram-Box plots for each distribution are as follows:

Observations on No. of Trainings:

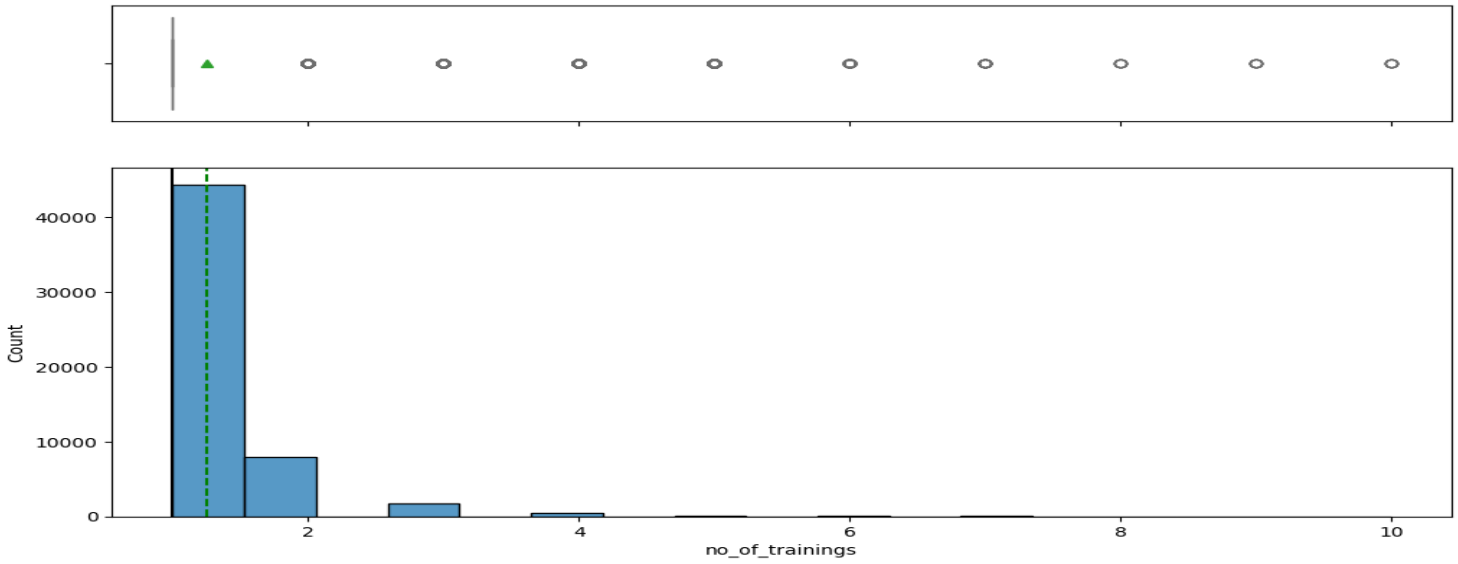

Fig-1

- There are 9 outliers present in the distribution of no_of_trainings.
- Most employees have done one training session & few have more than 2.

Let's see the distribution of age of employee

Observations on Age:

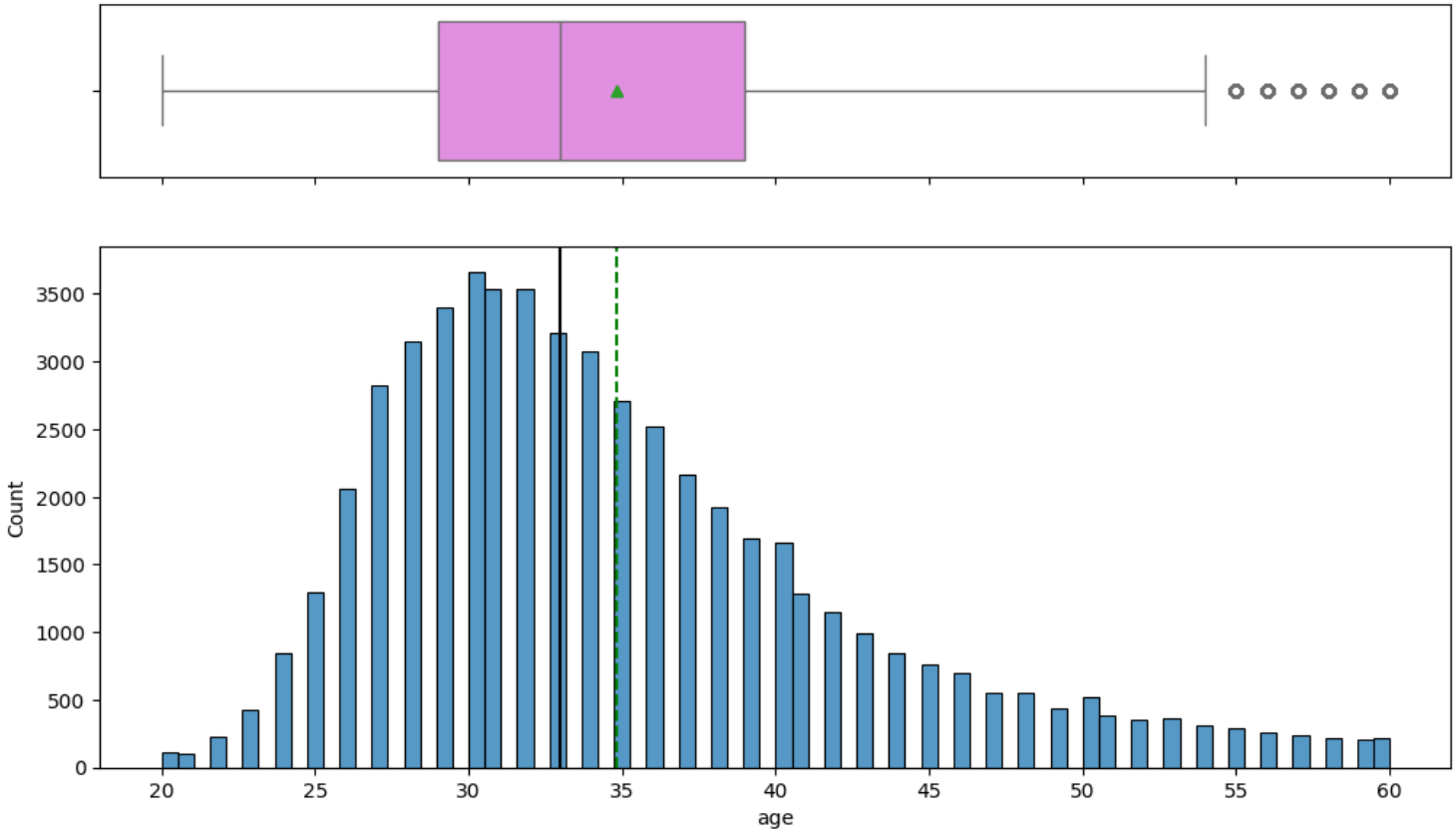

Fig-2

- There are 6 outliers present in the distribution of age.
- Age is normally distributed with slightly skewed towards right and peak is observed around 30–35 years.

Observations on length of service:

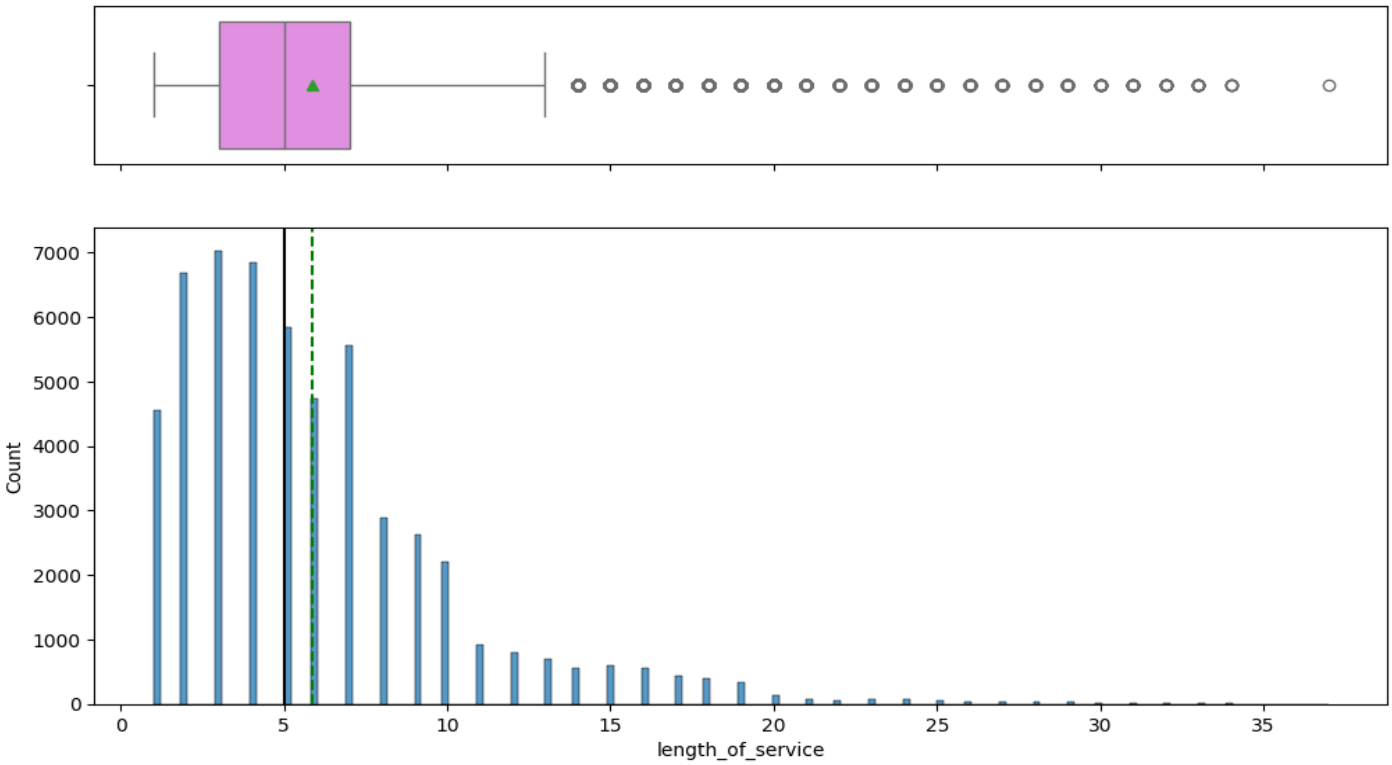

Fig-3

- Maximum length of service is observed between 0 to 5 years.
- There are few outliers present in the distribution of length of service.

Let's see the distribution of average training score of employee

Observations on Average Training Score:

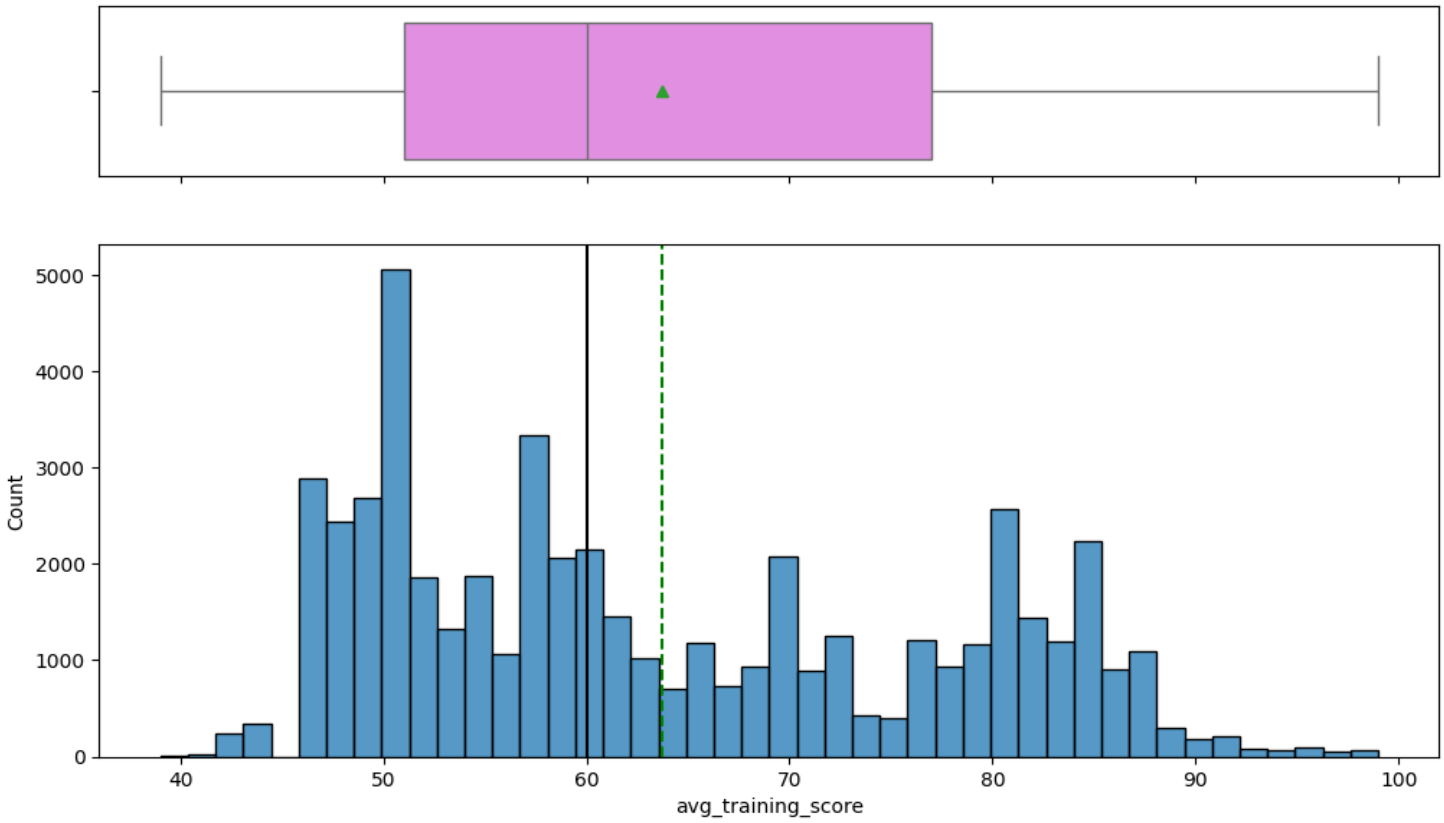

Fig-4

- Distribution of Average Training Score is skewed towards right.
- Highest training score is observed as 50.

Observations on Department:

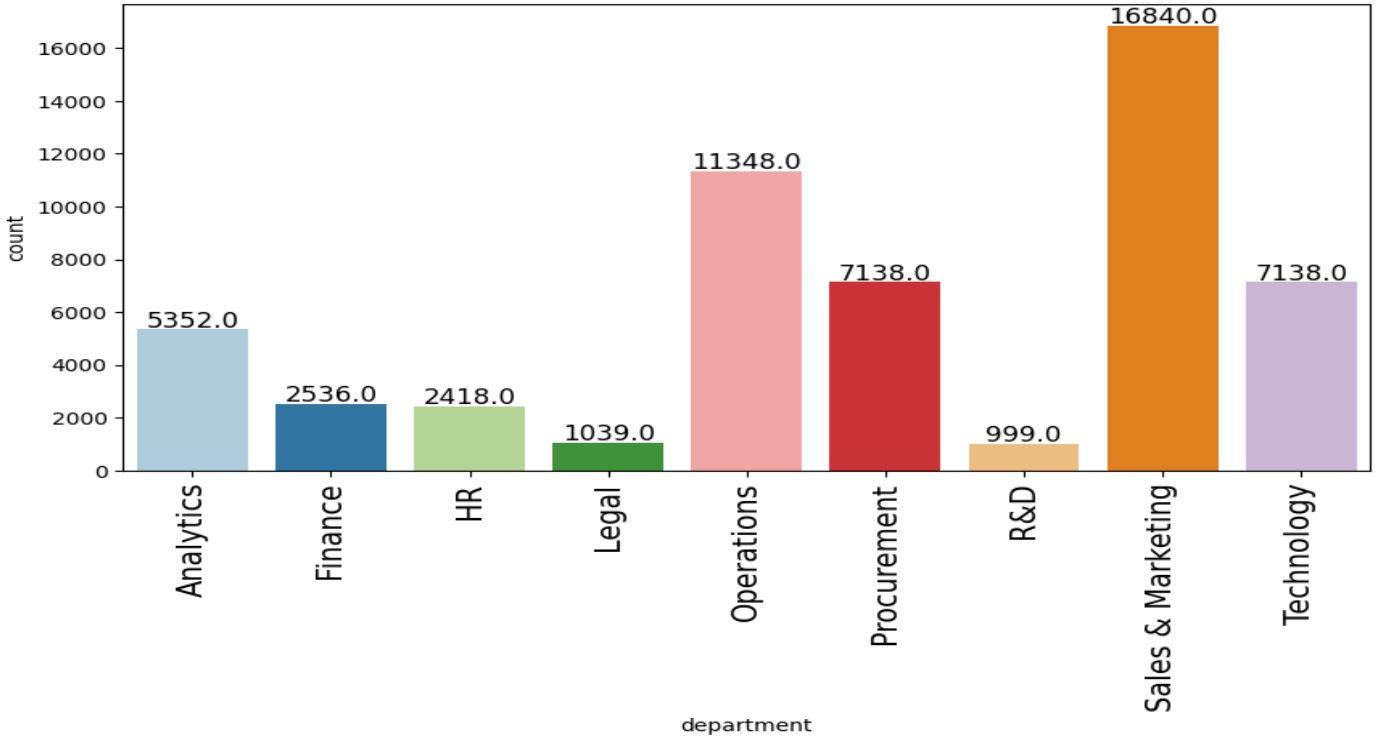

Fig-5

- Sales & Marketing has accounted for maximum number of promotions.

Observations on Education:

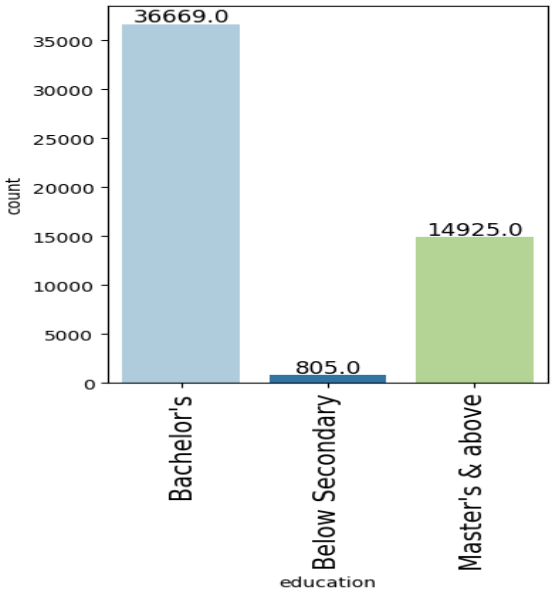

Fig-6

- Most employees are Bachelor's Degree Holders.

Observations on gender:

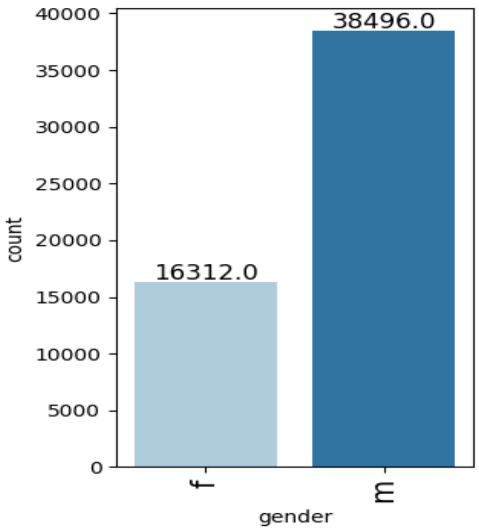

Fig-7

- Most employees who got promoted are males.

Observations on Recruitment Channel:

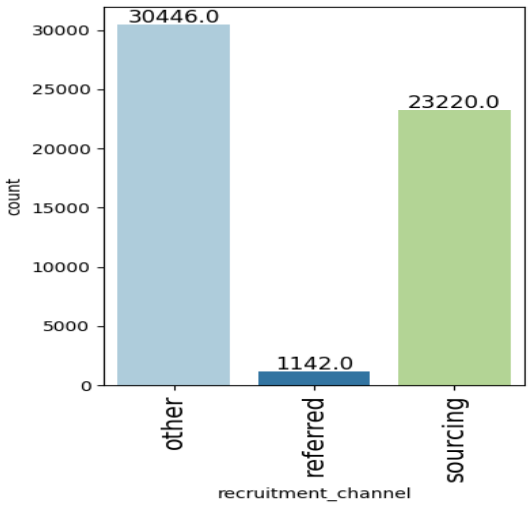

Fig-8

- Most of the employees got selected based on other recruitment channels.

Observations previous_year_rating:

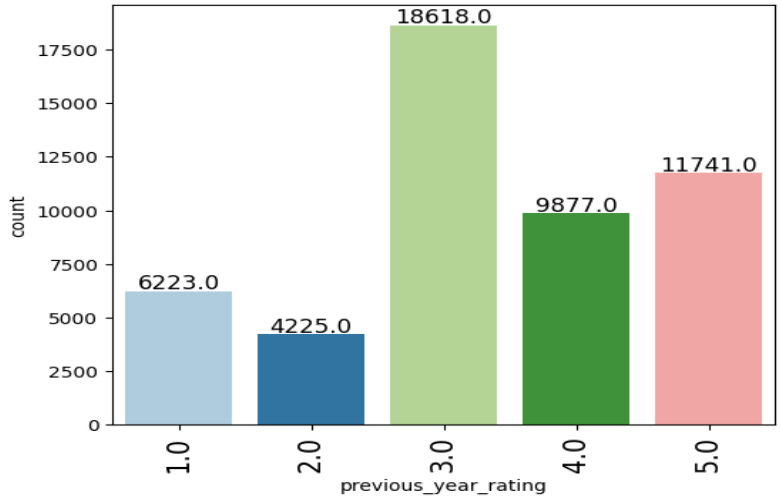

Fig-9

- Maximum rating observed is 3.0.

Observations on Awards Won:

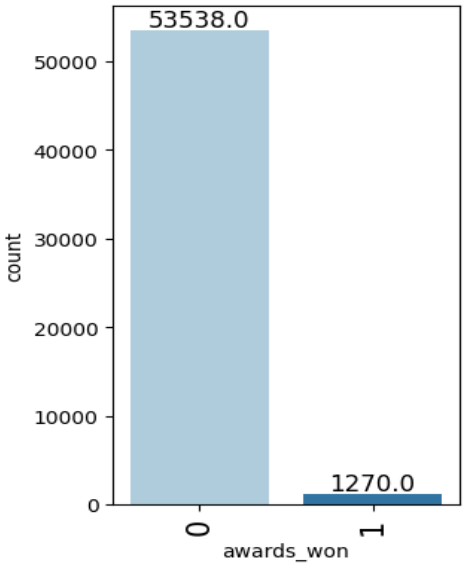

Fig-10

- Very few employees have won awards.

Observations on Region:

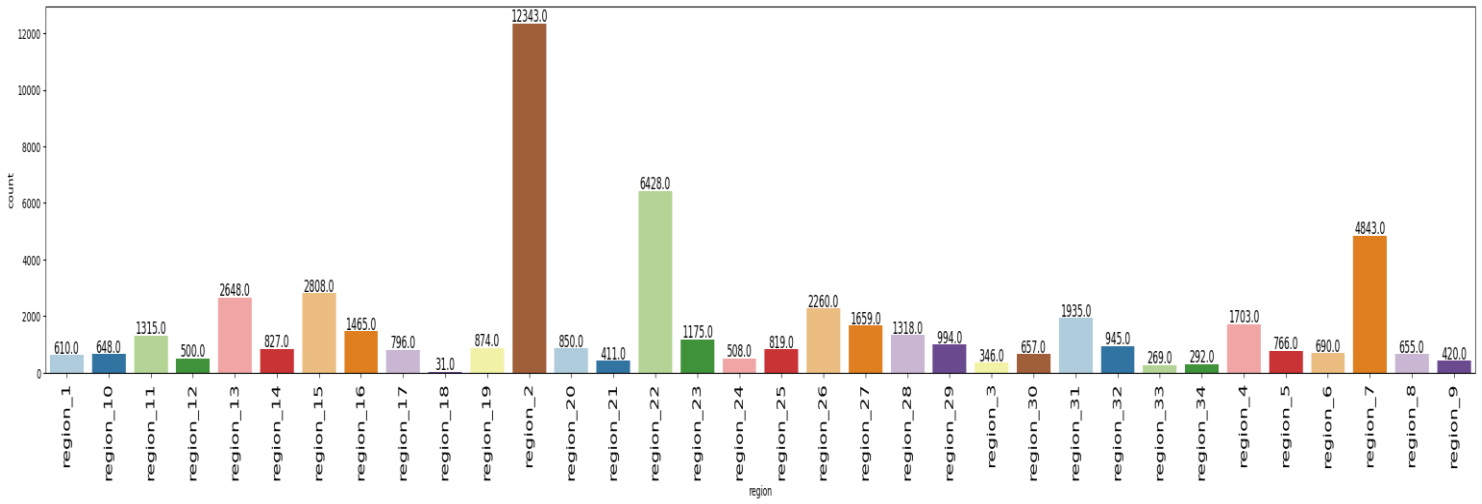

Fig-11

- Maximum records of employment are from region 2.

Observations on Target Variable:

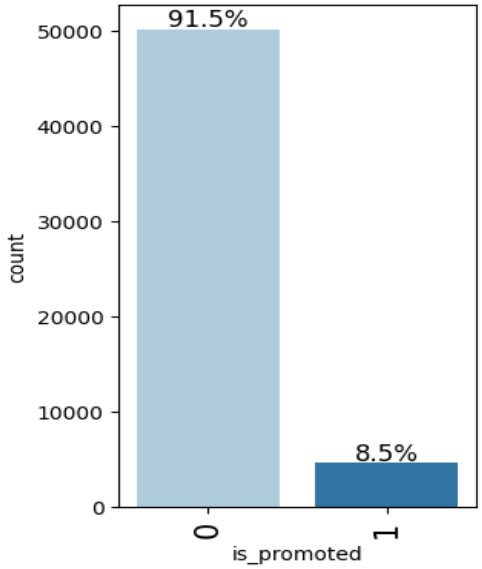

Fig-12

- Only ~8.5% of employees were promoted (high class imbalance).

3.2. Bivariate Analysis

- Used stacked bar-plot, distribution plots & correlation heatmaps to identify relationships with target variable 'is_promoted'. Plots for each distribution are as follows:

Target variable vs Age:

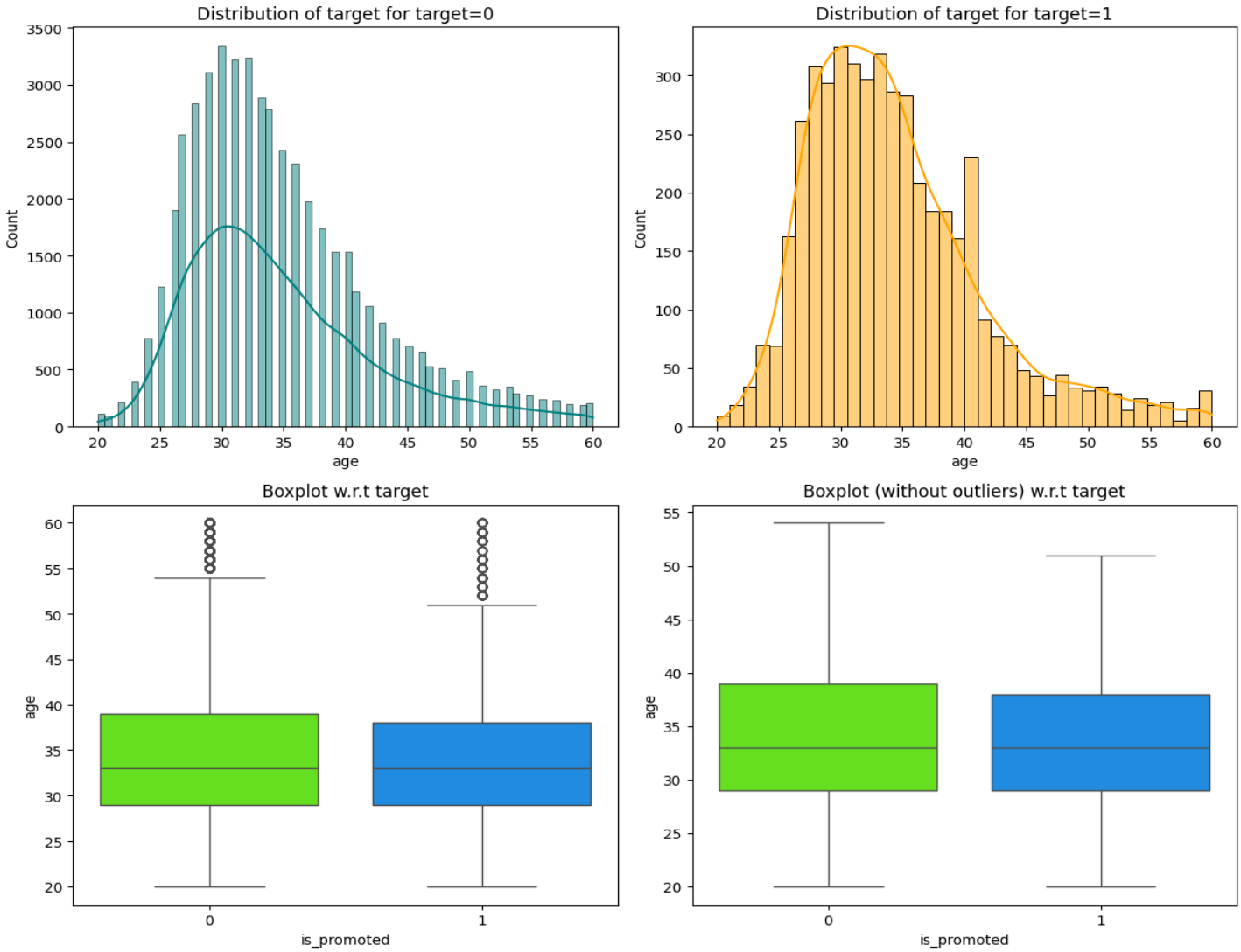

Fig-13

- There is a peak in promotion of employees between ages 25 to 35.

Let's see the change in length of service (length_of_service) vary by the employee's promotion status (is_promoted)?

Target variable vs Length of Service:

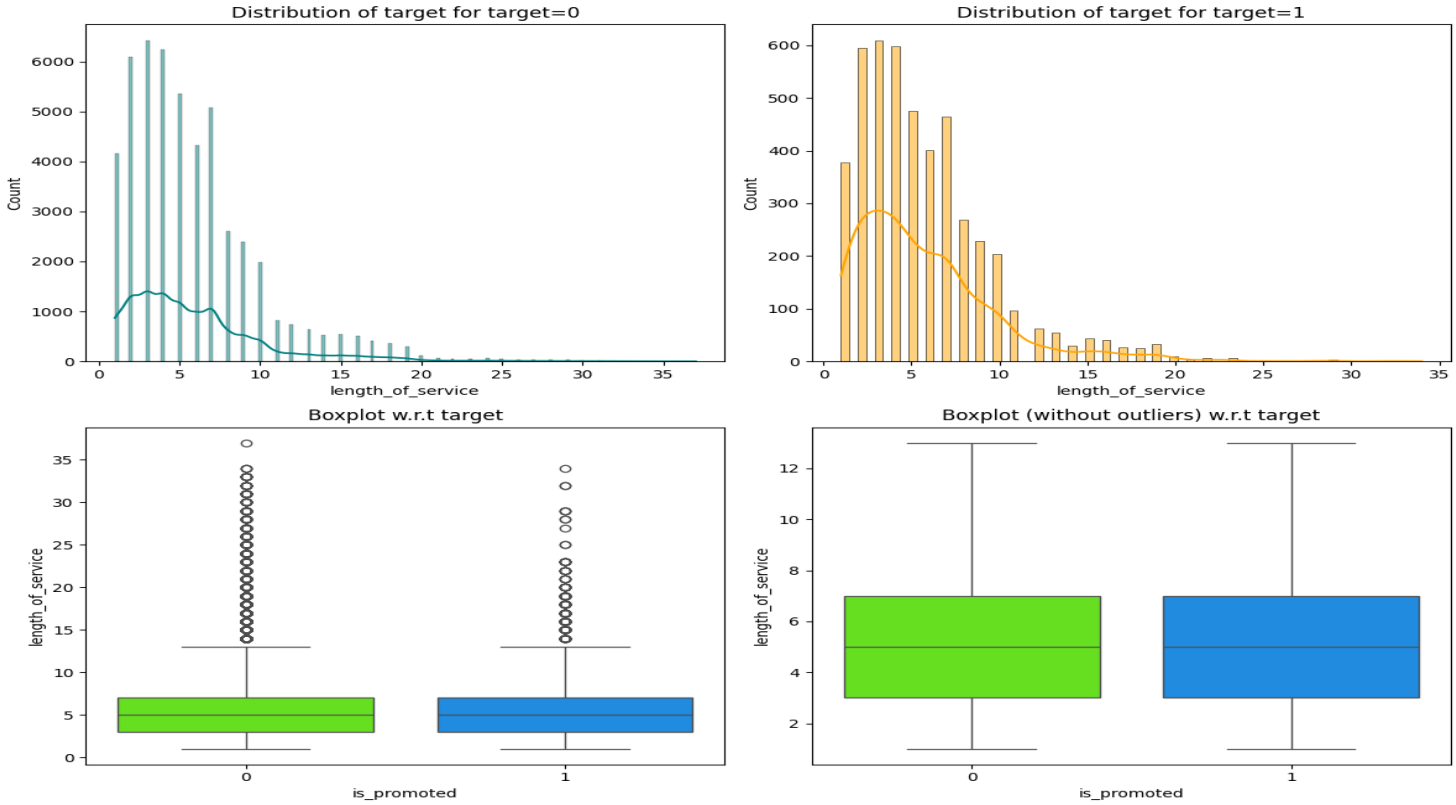

Fig-14

- Promotion likelihood increases slightly with **moderate tenure (5–10 years)**.
- Very **short (<2 years)** or **very long (>15 years)** service durations show **lower promotion rates**, possibly due to lack of experience or nearing retirement.

Target variable vs Average Training Score

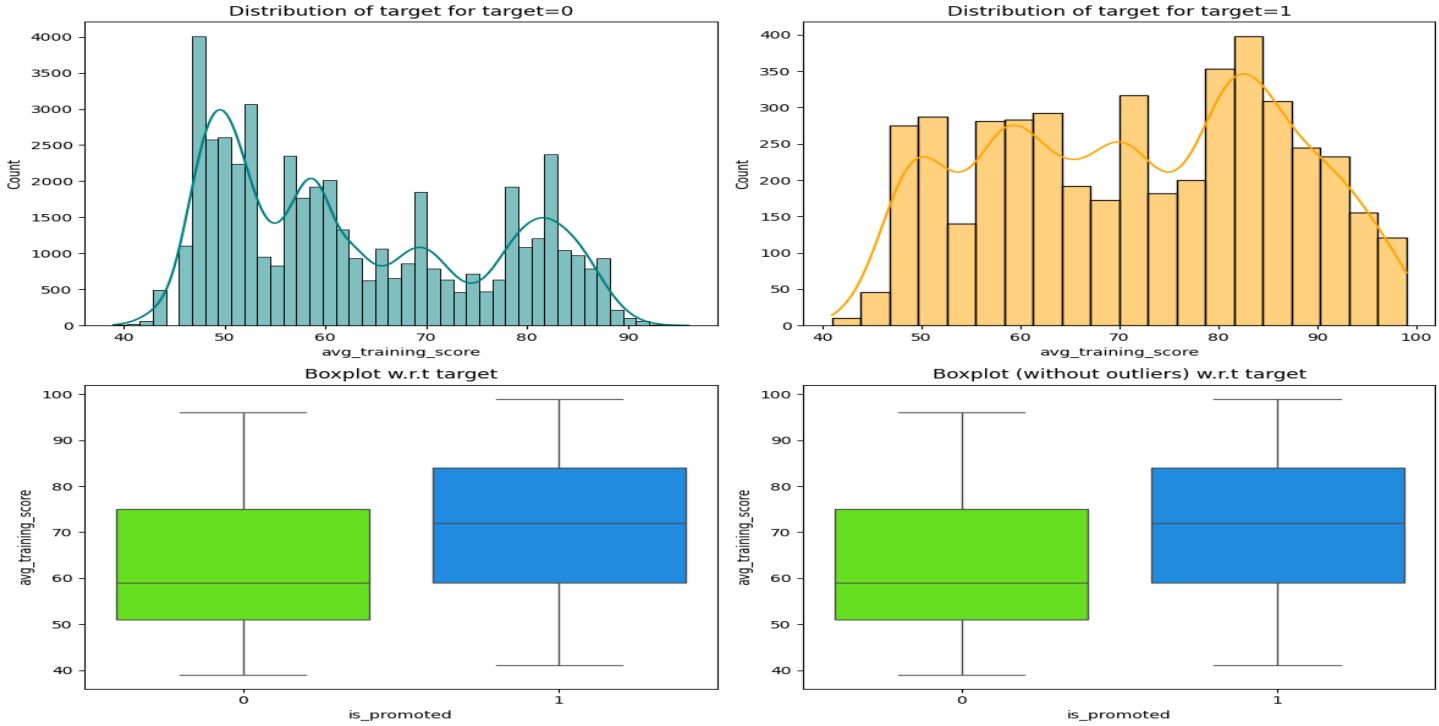

Fig-15

- Employees with **training scores above 80** show a **notable increase in promotion rates**.
- Those with low training scores (below 50) rarely get promoted.

Target variable vs Department:

is_promoted	0	1	All
department			
All	50140	4668	54808
Sales & Marketing	15627	1213	16840
Operations	10325	1023	11348
Technology	6370	768	7138
Procurement	6450	688	7138
Analytics	4840	512	5352
Finance	2330	206	2536
HR	2282	136	2418
R&D	930	69	999
Legal	986	53	1039

Table-5

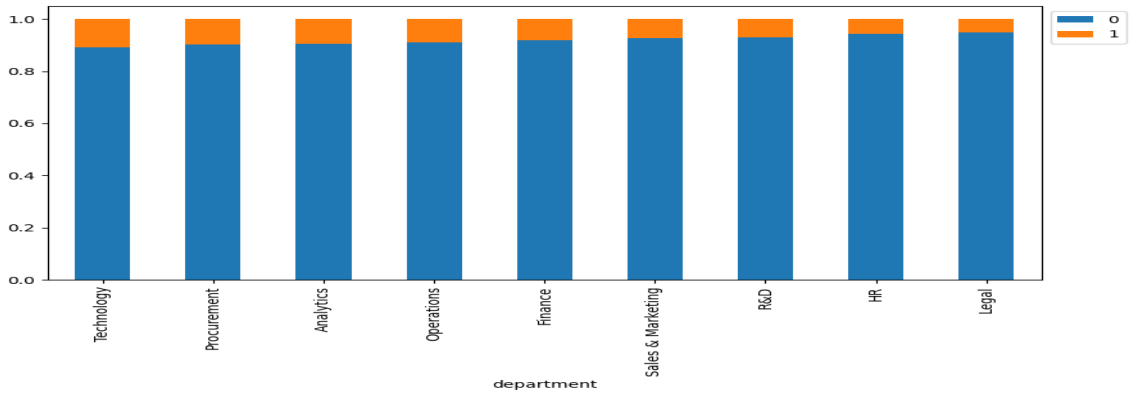

Fig-16

- **Sales & Marketing** and **Operations** are departments which have **higher promotion rates**.
- Departments like **Legal**, **R&D** and **HR** show **lower promotion frequencies**.

Target variable vs Region:

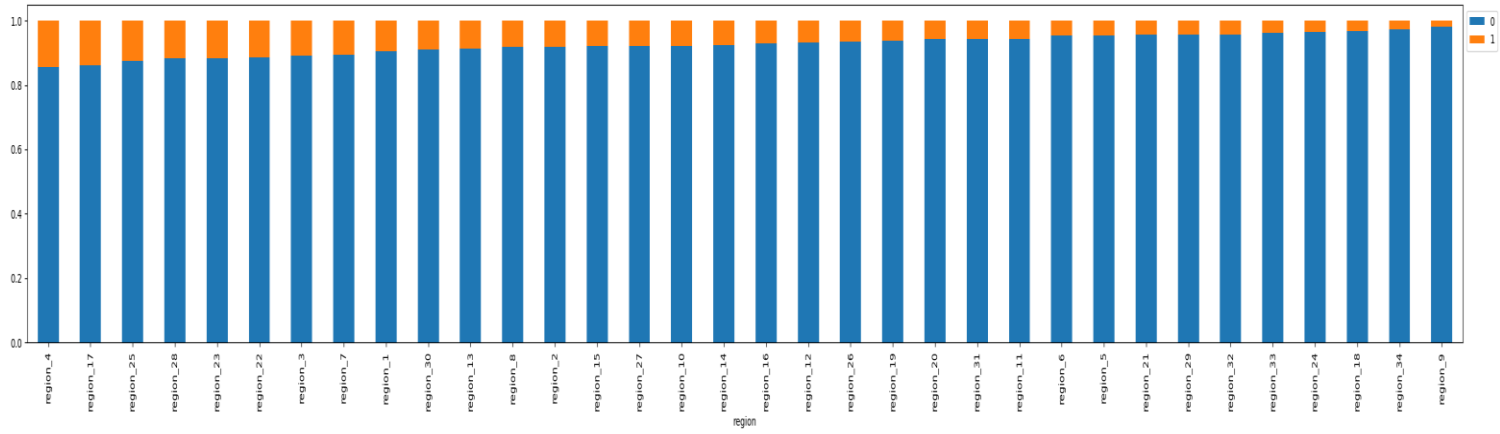

Fig-17

- There are **regional differences** in promotion rates, but these may reflect **departmental and demographic concentrations** rather than regional policy bias.

Target variable vs Education:

is_promoted	0	1	All
education			
All	47853	4546	52399
Bachelor's	33661	3008	36669
Master's & above	13454	1471	14925
Below Secondary	738	67	805

Table-6

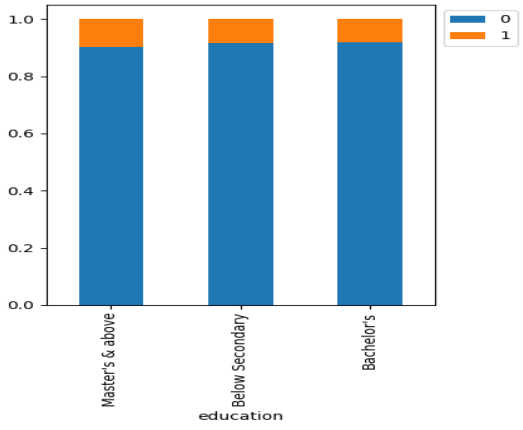

Fig-18

- **Bachelor's degree holders** have a **slightly higher promotion rate** than those with only a **Master**'s.
- However, the **difference isn't very large**, suggesting that promotions are not heavily biased toward educational qualification.

Target variable vs Gender:

is_promoted	0	1	All
gender			
All	50140	4668	54808
m	35295	3201	38496
f	14845	1467	16312

Table-7

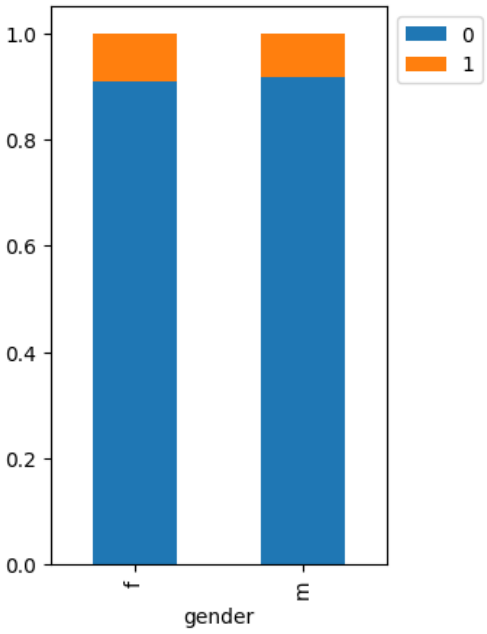

Fig-19

- Males are more likely to get promoted than females.

Target variable vs Recruitment Channels:

is_promoted	0	1	All
recruitment_channel			
All	50140	4668	54808
other	27890	2556	30446
sourcing	21246	1974	23220
referred	1004	138	1142

Table-8

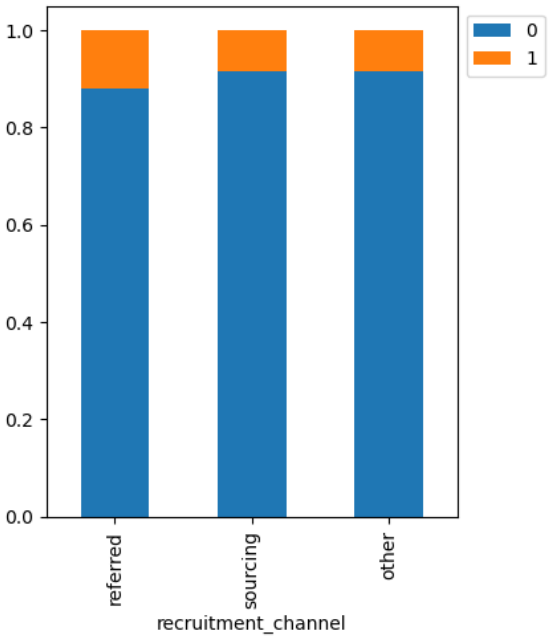

Fig-20

- Employees recruited via the **'other'** channel have **higher promotion rates** than those from **'sourcing'** or **'referred'** channels.
- Could indicate stronger performance from formally sourced candidates.

Let's see the previous rating (previous_year_rating) vary by the employee's promotion status (is_promoted)

Target variable vs previous_year_rating:

is_promoted	0	1	All
previous_year_rating			
All	46355	4329	50684
5.000	9820	1921	11741
3.000	17263	1355	18618
4.000	9093	784	9877
2.000	4044	181	4225
1.000	6135	88	6223

Table-9

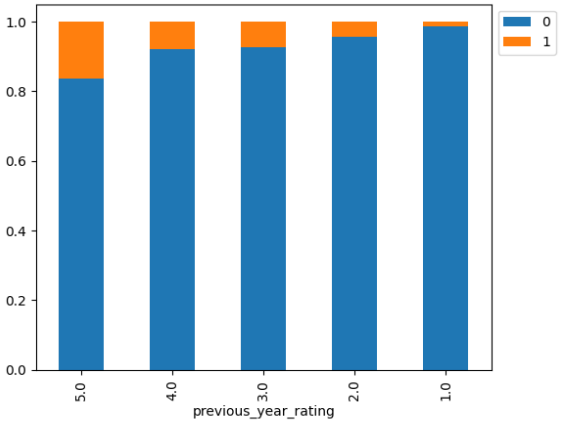

Fig-21

- Employees with **higher ratings (3.0 and 5.0)** have a **much higher promotion rate** compared to those with ratings of 1.0 or 2.0.
- Very few employees with a rating of 1.0 were promoted.
- Rating appears to be **one of the strongest predictors** for promotion.

Target variable vs awards_won:

is_promoted	0	1	All
awards_won			
All	50140	4668	54808
0	49429	4109	53538
1	711	559	1270

Table-10

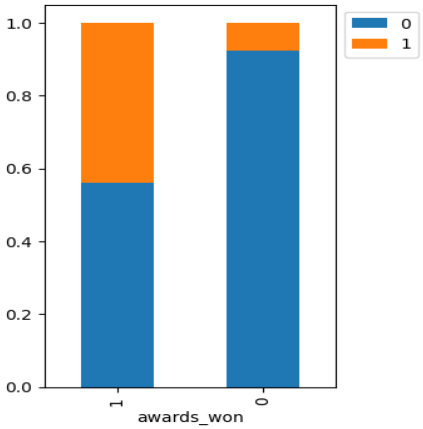

Fig-22

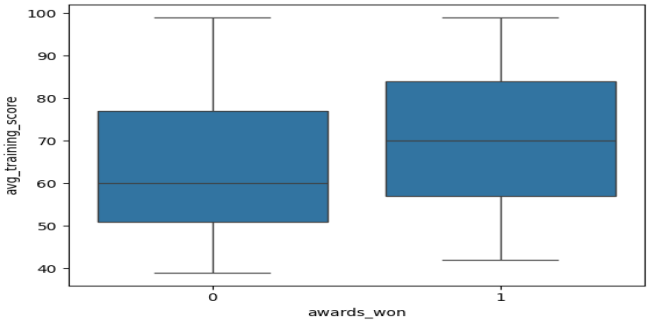

Fig-23

- Employees who won awards have a much higher likelihood of being promoted compared to those who did not receive awards.

Correlation Heatmap:

Fig-24

Insights from Correlation Heatmap

The heatmap and correlation values with the target variable is_promoted reveal the following:

✓ *Positive Correlations:*

1. **awards_won (0.20)**:
   - Strongest positive correlation with promotion.
   - Suggests award-winning employees are more likely to be promoted.
2. **avg_training_score (0.18)**:
   - Higher training scores are linked to a higher probability of promotion.
   - Indicates training quality/performance is valued in promotions.
3. **previous_year_rating (0.16)**:
   - Employees with higher performance ratings in the previous year tend to be promoted more.
   - Reflects the importance of recent job performance.

✗ *Negligible or Negative Correlations:*

4. **length_of_service (-0.01)**:
   - Virtually it has no correlation, indicating that tenure alone doesn't drive promotion.
5. **age (-0.02)**:
   - Slightly negative correlation; age is not a significant promotion factor.
6. **no_of_trainings (-0.02)**:
   - Counterintuitive: More training does not necessarily lead to promotions.
   - Could imply quality matters more than quantity.

Conclusion:

- Top Predictors**: awards_won, avg_training_score, and previous_year_rating.**
- Not Useful Alone**: age, tenure, and number of trainings may not be strong independent predictors.**
- Actionable Insight**: Encourage award programs and performance-focused training to influence promotions.**

4. DATA PREPROCESSING

4.1. Missing Value Treatment:

- **Education:** Imputed with mode (most common category).
- **previous_year_rating:** Imputed using median.
- **avg_training_score:** Imputed using median.

Checking that no column has missing values in train, validation and test sets

```
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
dtype: int64
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
dtype: int64
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
dtype: int64
```

Table-10

4.2. Feature Engineering:

- Derived features such as "seniority level", "training intensity", or "promotion probability category".

4.3. Encoding:

- Label encoding for binary categorical features.
- One-hot encoding for multi-class features like department, region, etc.

4.4. Train-Test Split:

- 80-20 & 75-25 stratified split to maintain class balance for train-test & validation set.

5. MODEL BUILDING

5.1. Model evaluation criterion

Model can make wrong predictions as:

- Predicting an employee should get promoted when he/she should not get promoted.
- Predicting an employee should not get promoted when he/she should get promoted.

Which case is more important?

- Both cases are important here as not promoting a deserving employee might lead to less productivity and the company might lose a good employee which affects the company's growth. Further, giving promotion to a non-deserving employee would lead to loss of monetary resources and giving such employee higher responsibility might again affect the company's growth.

How to reduce this loss i.e. need to reduce False Negatives as well as False Positives?

- Bank would want F1-score to be maximized, as both classes are important here. Hence, the focus should be on increasing the F1-score rather than focusing on just one metric i.e. Recall or Precision.

5.2. Model Building-original data

Models Used:

1. **Bagging**
2. **Decision Tree**
3. **Random Forest**
4. **Gradient Boosting**
5. **XGBoost**

```
Cross-Validation Cost:

Bagging: 0.7068560427147808
Decision Tree: 0.6492655176387899
Random Forest: 0.6772424198845639
Gradient Boosting: 0.6968180072640978
XGBoost: 0.7259915450319645

Validation Performance:

Bagging: 0.33771929824561403
Decision Tree: 0.39035087719298245
Random Forest: 0.24561403508771928
Gradient Boosting: 0.25
XGBoost: 0.3201754385964912
```

Performance:

- Based on F1- Score metrics Bagging & XGBoost outperformed than models like Decision Tree and Random Forest.
- XGBoost has the best F1-Score (~0.72) on the original imbalanced dataset.

Insight: Class imbalance caused biased predictions toward the majority class (non-promoted).

Plotting box-plots for CV scores of all models defined above:

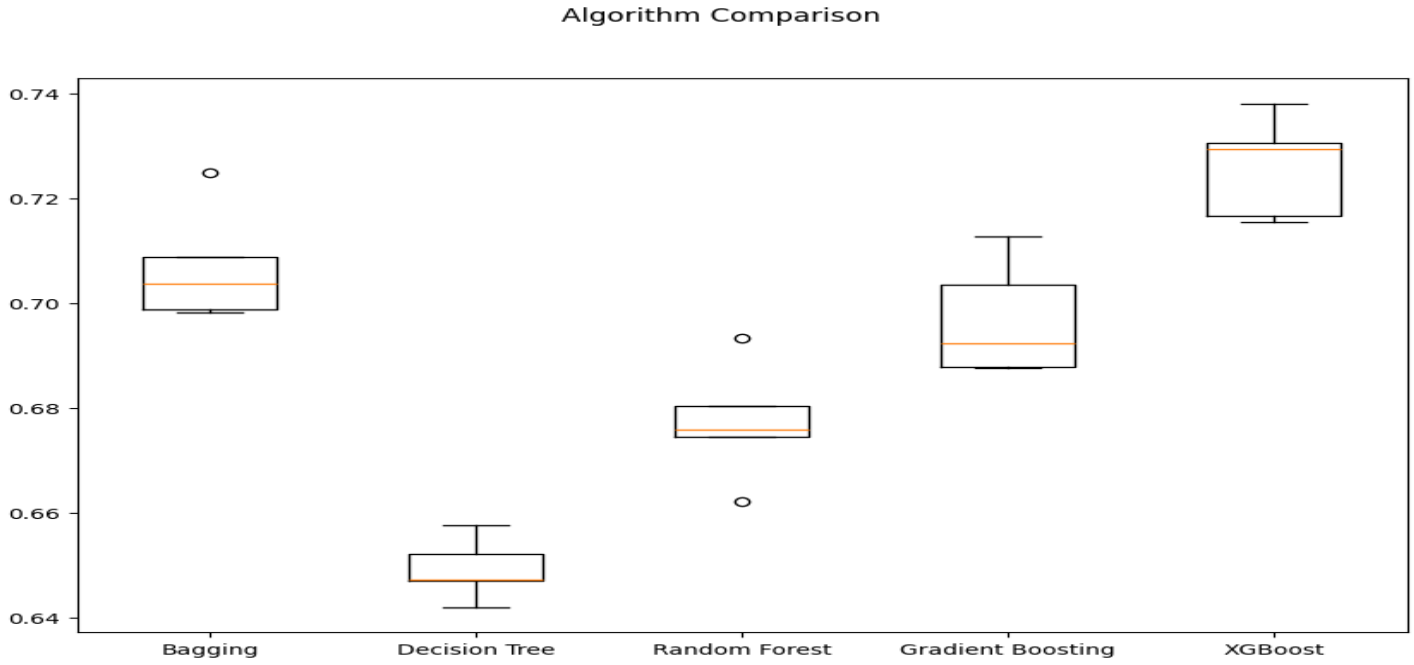

Fig-25

5.3. Model Building-Oversampled data

Technique used: SMOTE (Synthetic Minority Oversampling Technique)

Models Used: Same as above

```
Cross-Validation Cost:

Bagging: 0.9501723540415414
Decision Tree: 0.9278538453444046
Random Forest: 0.9608850251873754
Gradient Boosting: 0.8816740226552275
XGBoost: 0.9409470927488123

Validation Performance:

Bagging: 0.2807017543859649
Decision Tree: 0.3508771929824561
Random Forest: 0.2719298245614035
Gradient Boosting: 0.5175438596491229
XGBoost: 0.39035087719298245
```

Performance:

- Random Forest and Bagging had the best F1-score (~0.96).
- The model could now correctly identify more promoted candidates.

Insight: Oversampling effectively helped balance the model's sensitivity toward both classes.

Plotting box-plots for CV scores of all models defined above:

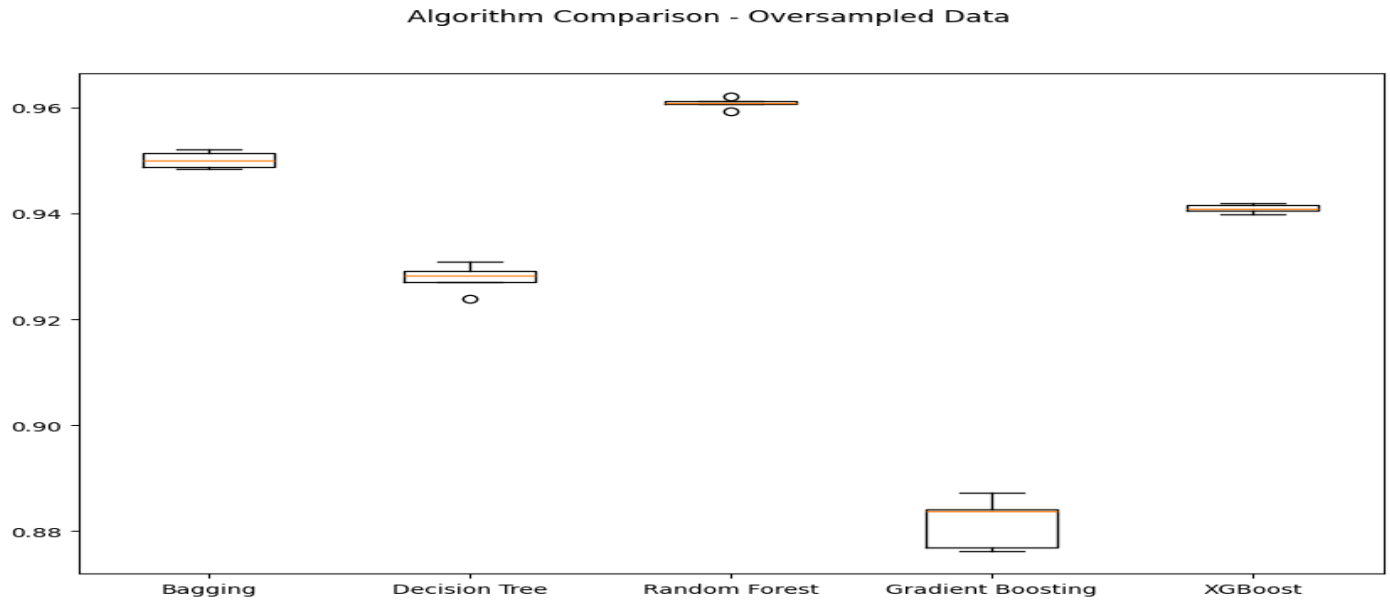

Fig-26

5.4. Model Building – Under-sampled Data

Technique: Random under-sampling of the majority class to match the size of the minority class.

Models Used: Same as above

```
Cross-Validation Cost:

Bagging: 0.6891217987163694
Decision Tree: 0.6496104414468349
Random Forest: 0.7080043951870578
Gradient Boosting: 0.7173828174825883
XGBoost: 0.7023025029406976

Validation Performance:

Bagging: 0.6666666666666666
Decision Tree: 0.6798245614035088
Random Forest: 0.7017543859649122
Gradient Boosting: 0.6403508771929824
XGBoost: 0.6666666666666666
```

Performance:

- Balanced F1-scores (~0.71).
- Random Forest and Gradient Boosting were most stable.

Insight: Under-sampling helps fast model training and understanding feature influence but sacrifices predictive power on real data.

Plotting box-plots for CV scores of all models defined above:

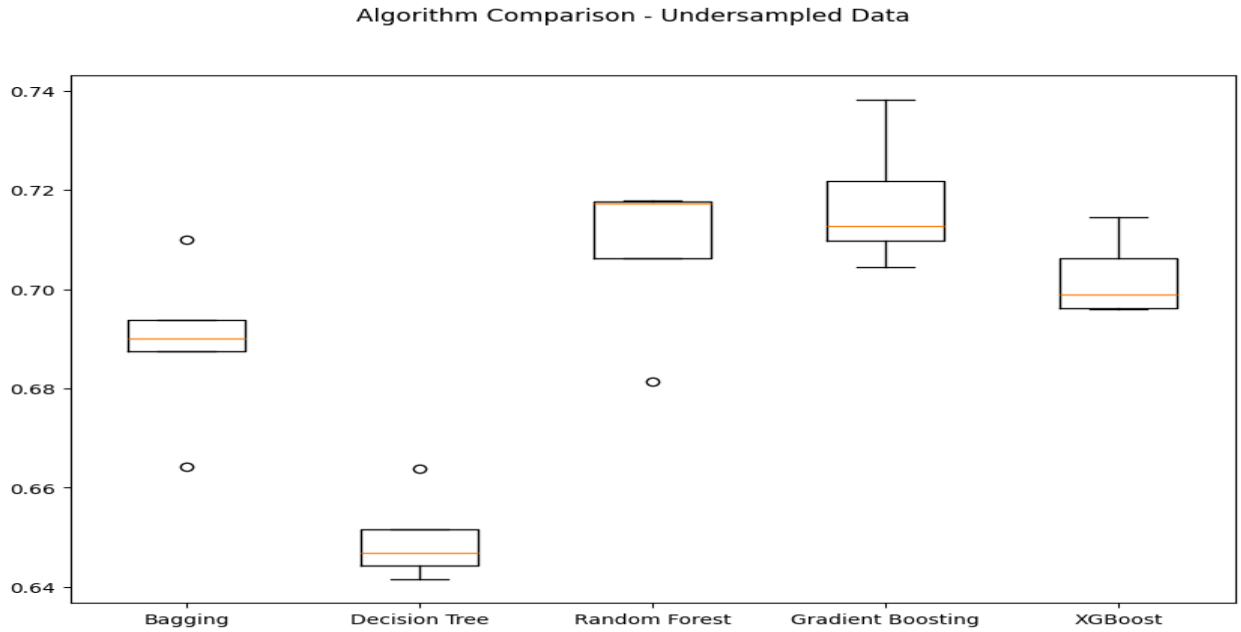

Fig-26

6. MODEL PERFORMANCE IMPROVEMENT

6.1. Hyper-parameter tuning

Tuned models:

- Adaboost using under-sampled data
- Adaboost using Original data
- Gradient boosting using under-sampled data
- Gradient boosting using Original data

Approach:

- RandomizedSearchCV
- Metric: F1-score

Result:

- Adaboost using under-sampled data

Performance on train set:

	Accuracy	Recall	Precision	F1
0	0.724	0.648	0.765	0.723

Table-11

Performance on validation set:

	Accuracy	Recall	Precision	F1
0	0.781	0.654	0.222	0.600

Table-12

- Adaboost using Original data

Performance on train set:

	Accuracy	Recall	Precision	F1
0	0.928	0.193	0.817	0.637

Table-13

Performance on validation set:

	Accuracy	Recall	Precision	F1
0	0.928	0.175	0.816	0.625

Table-14

- Gradient boosting using under-sampled data

Performance on train set:

	Accuracy	Recall	Precision	F1
0	0.796	0.664	0.244	0.618

Table-15

Performance on validation set:

	Accuracy	Recall	Precision	F1
0	0.798	0.702	0.248	0.623

Table-16

- Gradient boosting using Original data

Performance on train set:

	Accuracy	Recall	Precision	F1
0	0.941	0.321	0.955	0.724

Table-17

Performance on validation set:

	Accuracy	Recall	Precision	F1
0	0.940	0.289	0.971	0.707

Table-18

- All four models showed performance improvement.
- Gradient boosting under Original data achieved the best AUC and F1 scores on validation set.

7. MODEL PERFORMANCE COMPARISON & FINAL MODEL SELECTION

Training performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Original data	AdaBoost trained with Undersampled data	AdaBoost trained with Original data
Accuracy	0.796	0.941	0.724	0.928
Recall	0.664	0.321	0.648	0.193
Precision	0.244	0.955	0.765	0.817
F1	0.618	0.724	0.723	0.637

Table-19

Validation performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Original data	AdaBoost trained with Undersampled data	AdaBoost trained with Original data
Accuracy	0.798	0.940	0.781	0.928
Recall	0.702	0.289	0.654	0.175
Precision	0.248	0.971	0.222	0.816
F1	0.623	0.707	0.600	0.625

Table-20

Observations:

- All models benefit from sampling and tuning.
- **Gradient Boosting** is consistently the best performer for F1-score and AUC.
- Adaboost is a strong alternative.

Checking the performance of the best model on the test data:

	Accuracy	Recall	Precision	F1
0	0.794	0.648	0.244	0.616

Table-21

7.1. Final Model Selection

Gradient Boosting selected for best balance and interpretability. The final model is **Gradient Boosting** trained on under-sampled data with hyper-parameter tuning, providing the best balance of F1- Score, recall, precision, and computational efficiency.

7.2. Feature Importance

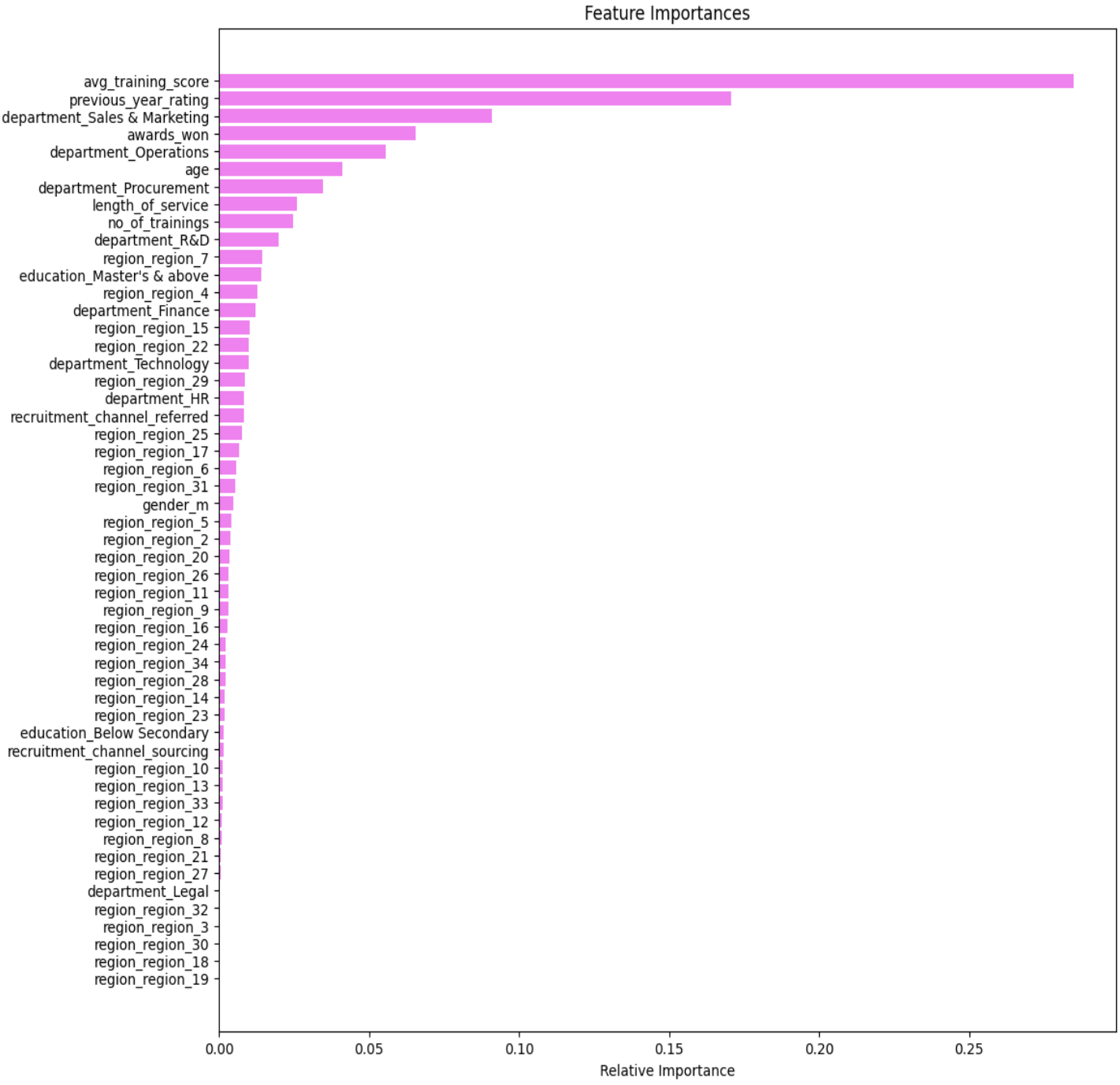

Fig-27

8. ACTIONABLE INSIGHTS & RECOMMENDATIONS

8.1. Actionable Insights

- Encourage award programs and performance-focused training to influence promotions.

8.2. Recommendations

- Focus on tracking and recognizing awards.
- Maintain accurate performance ratings.
- High training scores strongly indicate promotion readiness.
- Monitor promotion probabilities in a dashboard.

8.3. Conclusion

The machine learning model will enhance fairness, transparency, and HR efficiency at JMD Company by supporting evidence-based promotion decisions.

- **Top Predictors**: awards_won, avg_training_score, and previous_year_rating.
- **Not Useful Alone:** age, tenure, and number of trainings may not be strong independent predictors.