**Business Report: Employee Promotion Prediction Using Machine Learning**

Extended Project Report

Submitted to

By

**Subhadeep Seal**

In Partial Fulfillment of PDP-DSBA



# CONTENTS

|  |  |  |
| --- | --- | --- |
| SL.NO. | Title | **Page no.** |
| 1. | Problem Statement | 3 |
| 1.1. | Context | 3 |
| 1.2. | Problem Definition | 3 |
| 1.3. | Objective | 3 |
| 1.4. | Data Description-Data Dictionary | 3 |
| 2. | Data Overview | 4-5 |
| 2.1. | Shape of the dataset | 4 |
| 2.2. | Check the type of data | 4 |
| 2.3. | Check for missing values | 4 |
| 2.4. | Statistical summary of the dataset | 5 |
| 2.5. | Observations on various dataset | 5 |
| 3. | Exploratory Data Analysis (EDA) | 5-15 |
| 3.1. | Univariate Analysis | 5-10 |
| 3.2. | Bivariate Analysis | 10-15 |
| 4. | Data preprocessing | 16 |
| 4.1. | Missing Value Treatment | 16 |
| 4.2. | Feature Engineering | 16 |
| 4.3. | Encoding | 16 |
| 4.4. | Train-test split | 16 |
| 5. | Model Building | 17 |
| 5.1. | Model evaluation criterion | 17 |
| 5.2. | Model building-Original data | 17 |
| 5.3. | Model building-Oversampled data | 18 |
| 5.4. | Model building-Under-sampled data | 18-19 |
| 6. | Model Performance Improvement | 19-20 |
| 6.1. | Hyper-parameter tuning & observations | 19-20 |
| 7. | Model performance comparison and Final model selection | 20-21 |
| 7.1. | Final Model selection | 21 |
| 8. | Actionable Insights & Recommendations | 22 |
| 8.1. | Actionable Insights | 22 |
| 8.2. | Recommendations | 22 |
| 8.3. | Conclusion | 22 |

# 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.

# . 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****

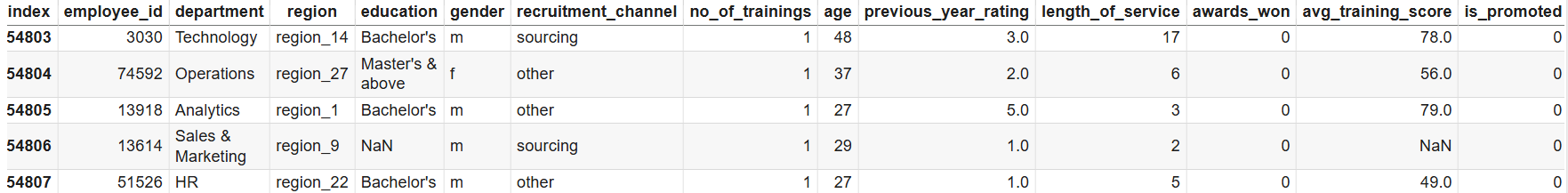
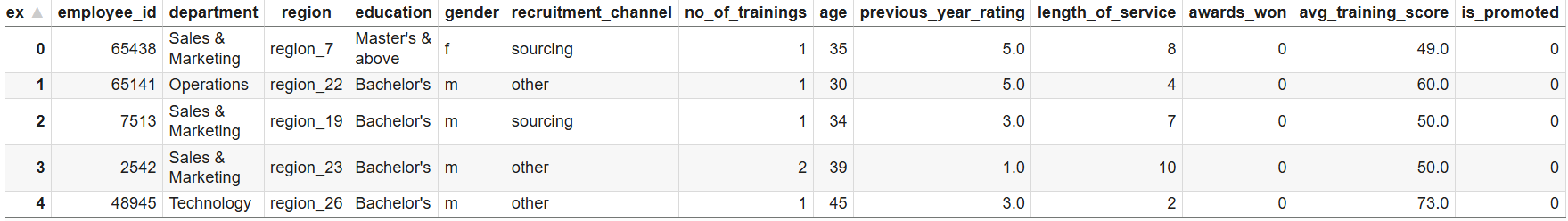
The detailed data dictionary is given below.

#### ****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

# DATA OVERVIEW

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

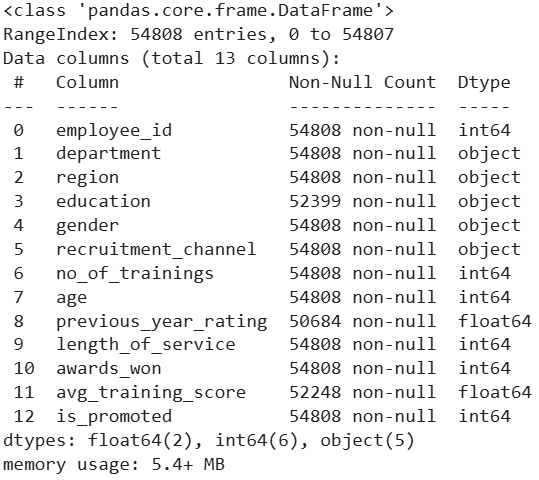


# ****Table 1: First 5 & last 5 rows of the dataset****

# ****2.1. Shape of the Dataset****

# **The dataset contains 54808 rows & 13 columns.**

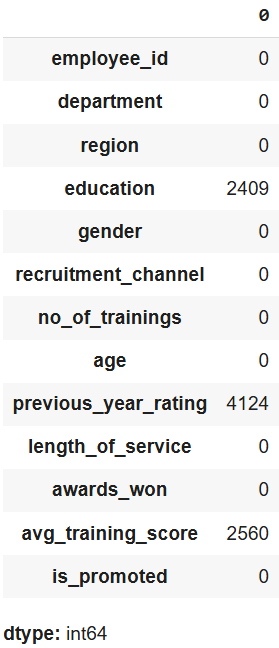
# ****2.2. Check the type of data****



**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.

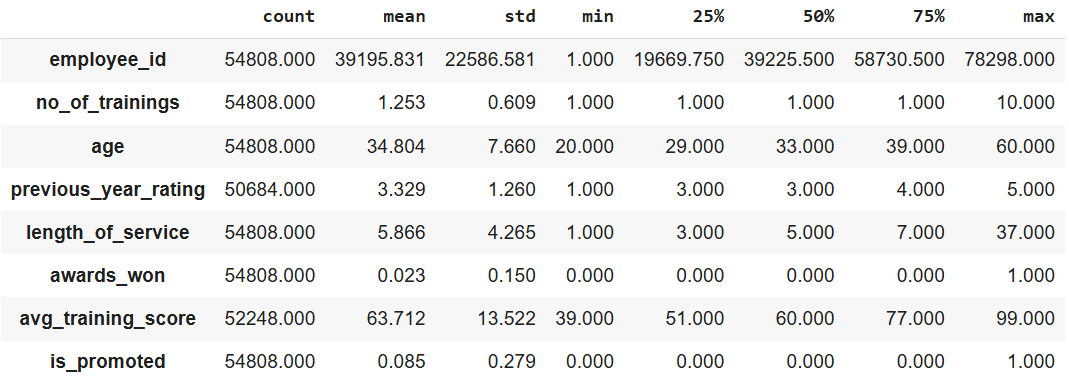
**2.3. Check for missing values**

****

**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**



**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**

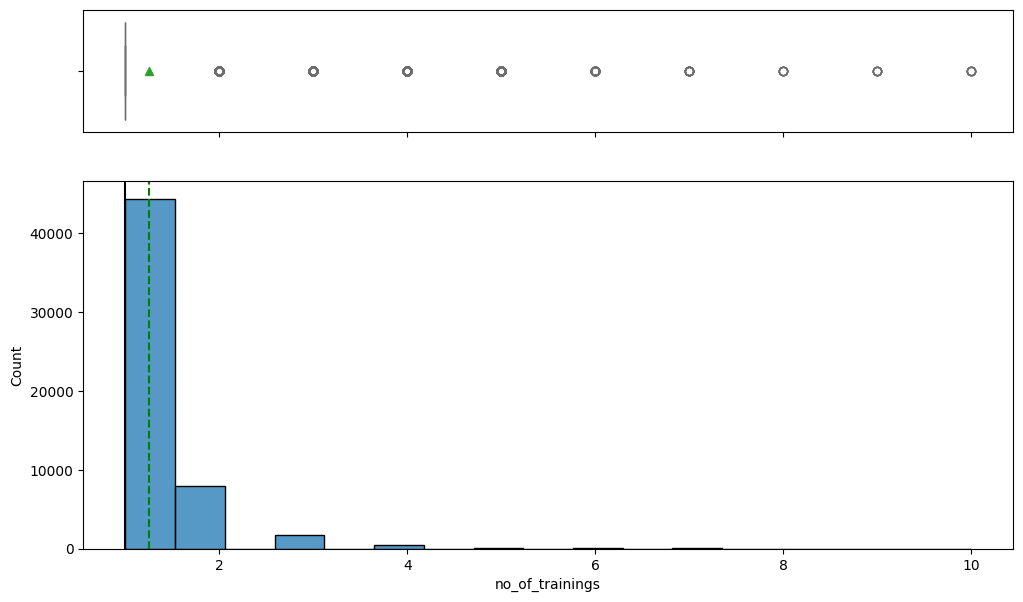
* **Department Distribution**: The dataset includes multiple departments, with a notable presence of Sales & Marketing, Operations, and Technology.
* **Education Level**: Most employees have at least a Bachelor's degree, with some holding Master's degrees or higher.
* **Gender Representation**: The dataset 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.

# 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:



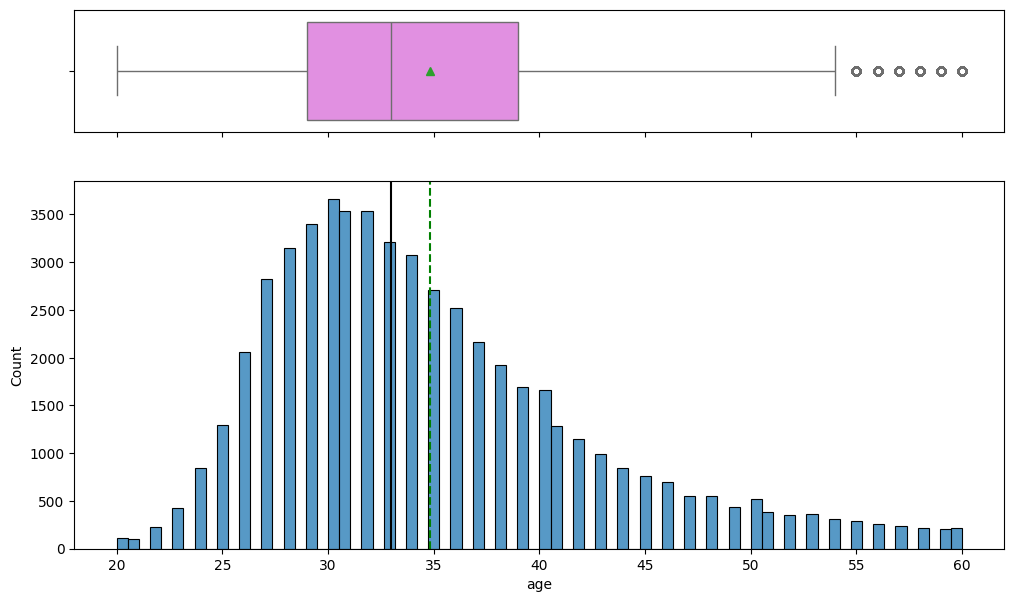
**Fig-1**

# 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:

# download (4).png

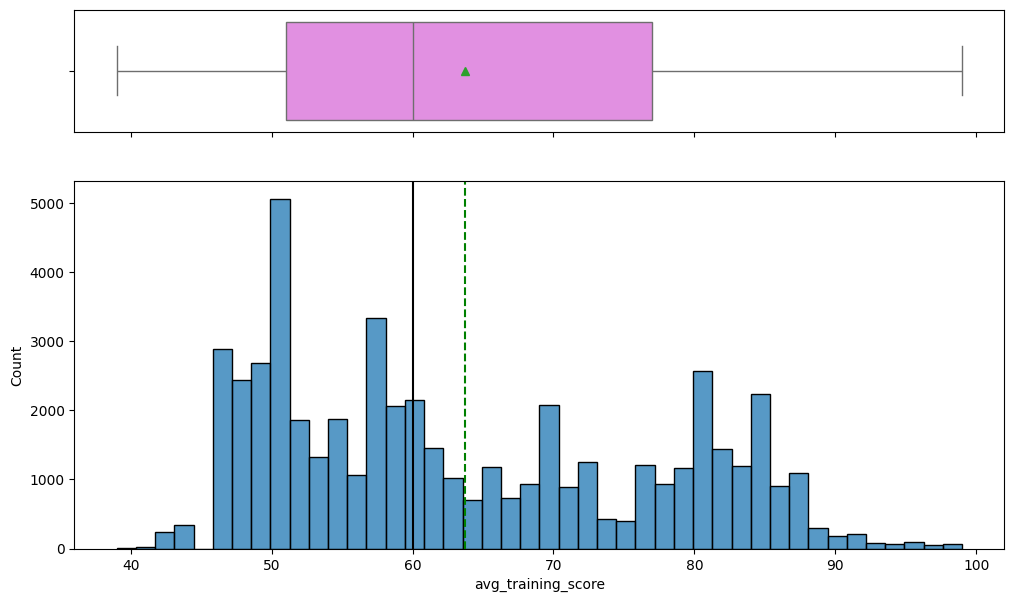
**Fig-3**

# 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*****

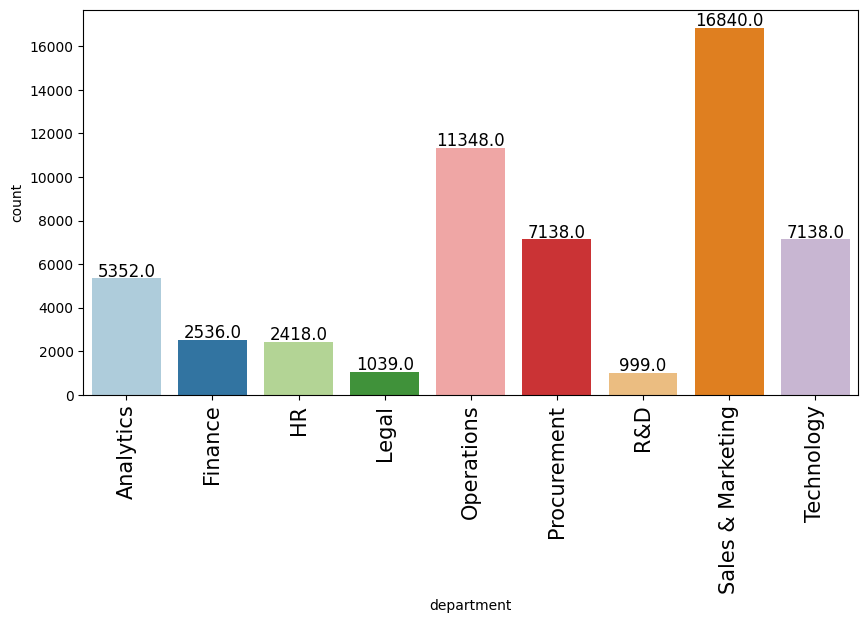
# Observations on Average Training Score:



**Fig-4**

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

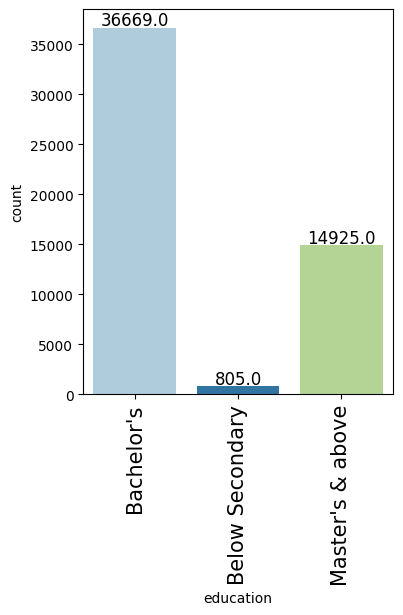
#### Observations on Department:



**Fig-5**

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

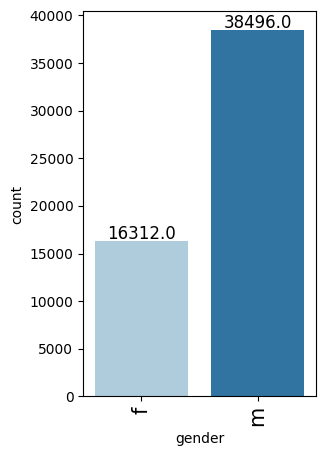
#### Observations on Education:



**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:**

#### download (9).png

**Fig-8**

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

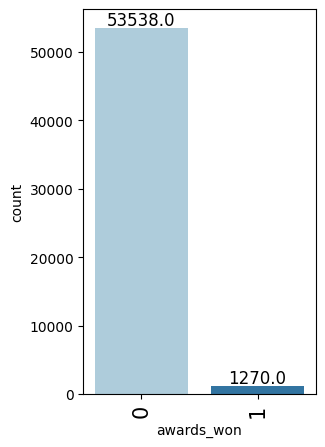
#### Observations previous\_year\_rating:

#### download (10).png

**Fig-9**

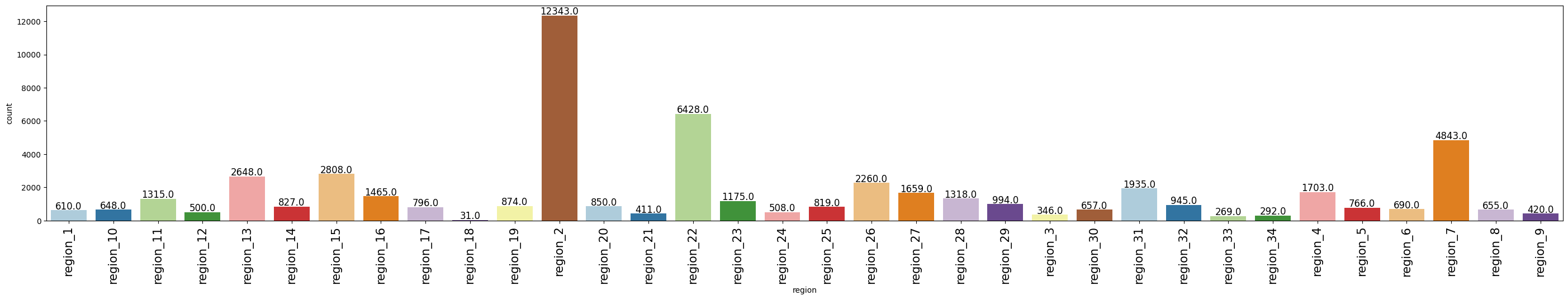
* Maximum rating observed is 3.0.

#### Observations on Awards Won:



**Fig-10**

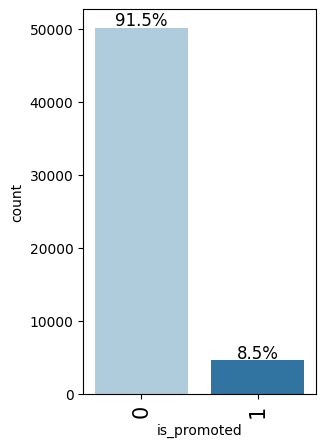
* Very few employees have won awards.

**Observations on Region:**

**Fig-11**

* Maximum records of employment are from region 2.

**Observations on Target Variable:**



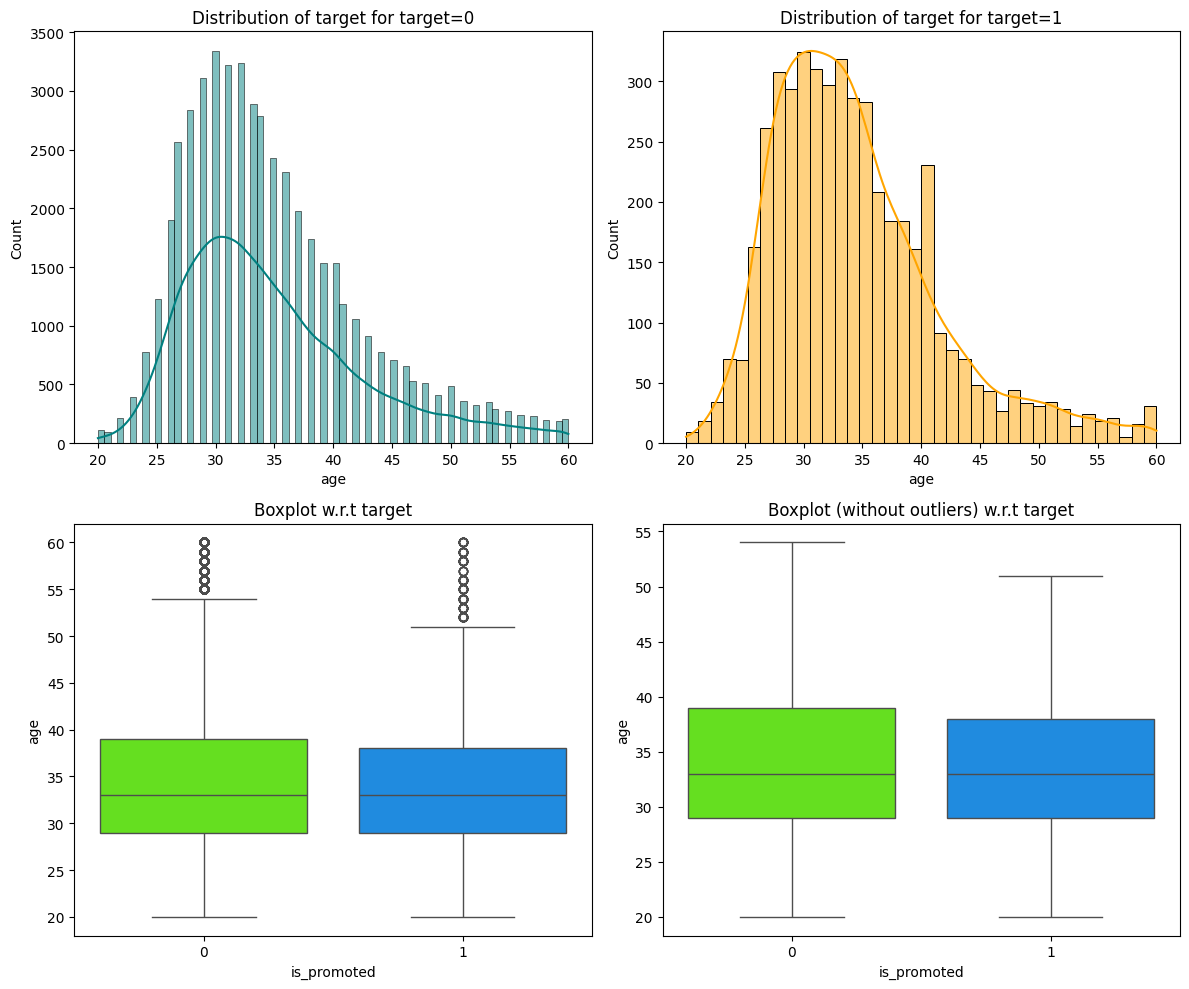
**Fig-12**

* 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**

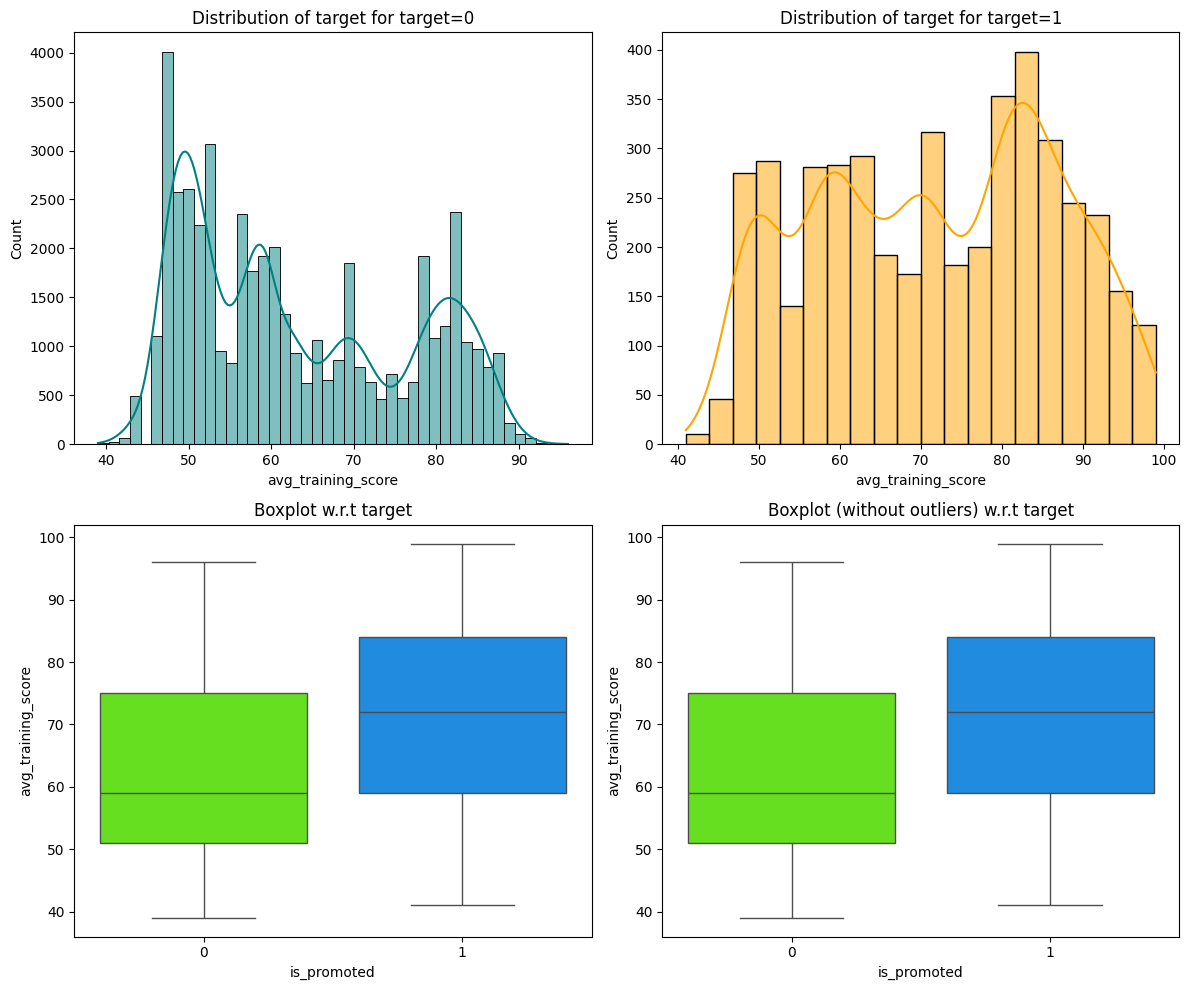
* There is a peak in promotion of employees between ages 25 to 35.

**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:download (17).pngFig-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.

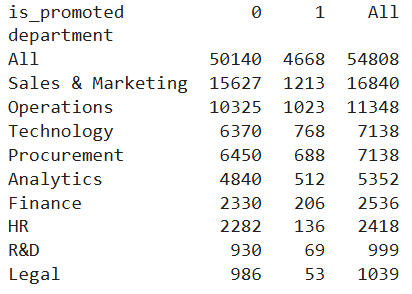
**Target variable vs Average Training Score**



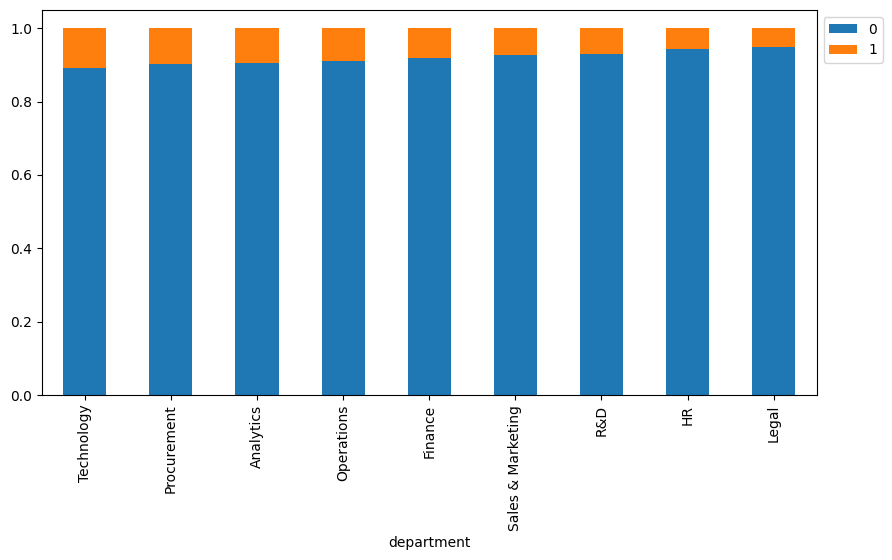
**Fig-15**

* 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:**



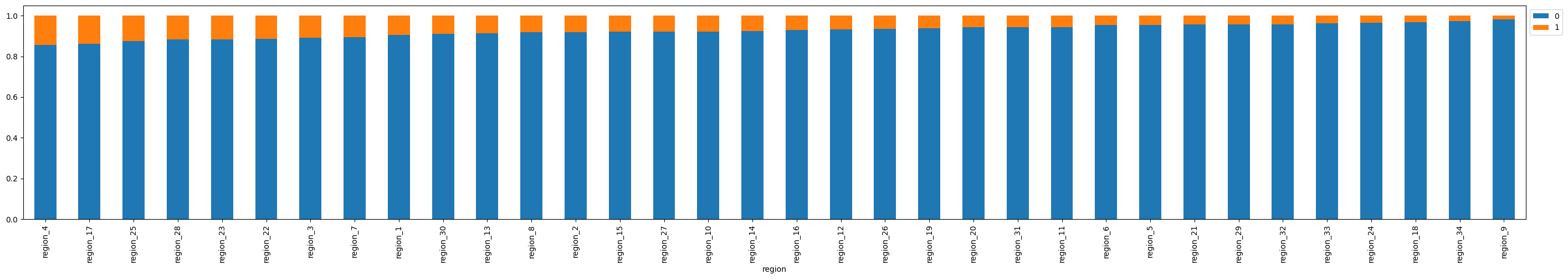
**Table-5**



**Fig-16**

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

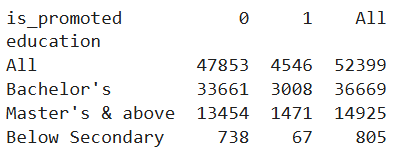
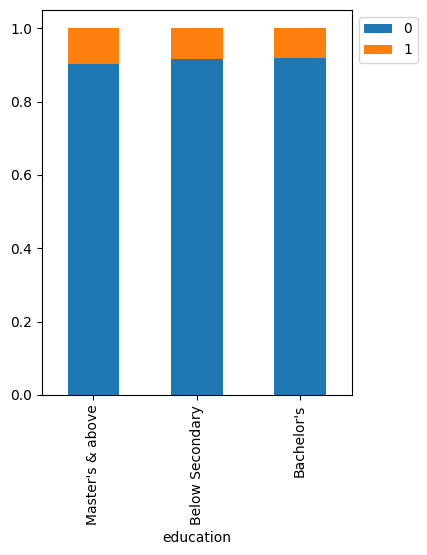
**Target variable vs Region:**

****

**Fig-17**

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

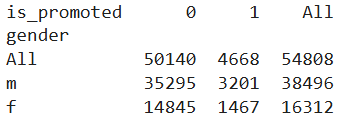
**Target variable vs Education:**

****

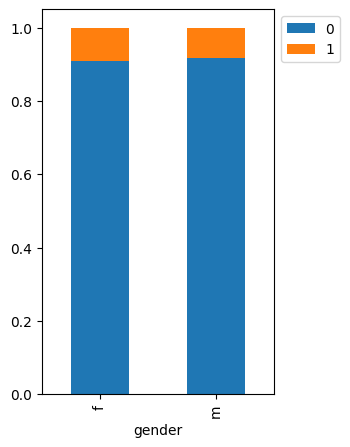
**Table-6 Fig-18**

* **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:**



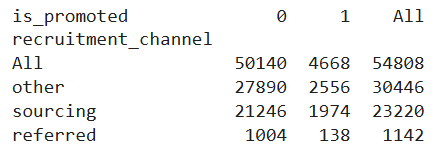
**Table-7**



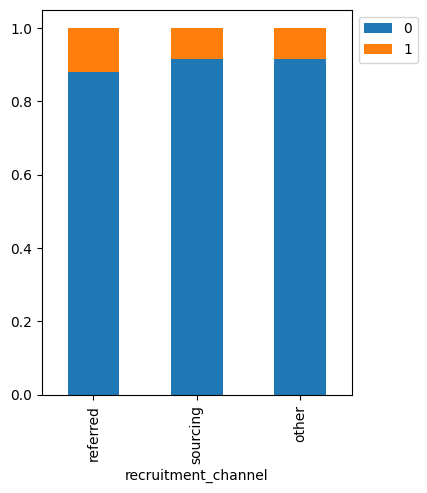
**Fig-19**

* Males are more likely to get promoted than females.

**Target variable vs Recruitment Channels:**

****

**Table-8**

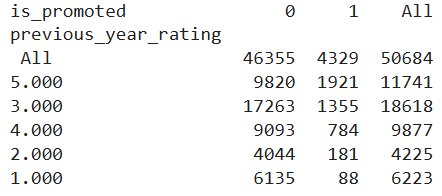
****

**Fig-20**

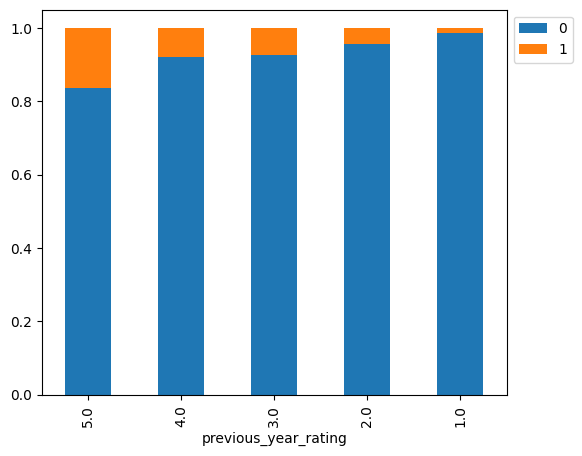
* Employees recruited via the **'other'** channel have **higher promotion rates** than those from **'sourcing'** or **'referred'** channels.
* Could indicate stronger performance from formally sourced candidates.

**Let's see the previous rating (previous\_year\_rating) vary by the employee's promotion status (is\_promoted)**

**Target variable vs previous\_year\_rating:**

****

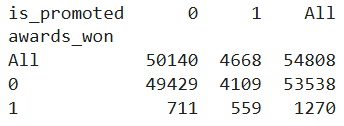
**Table-9**

****

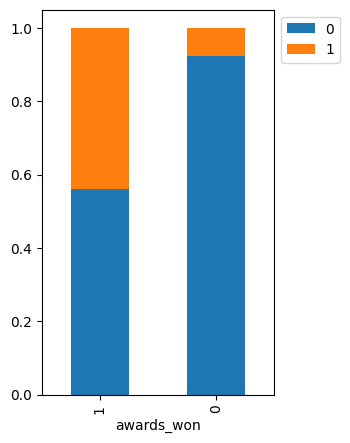
**Fig-21**

* 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.

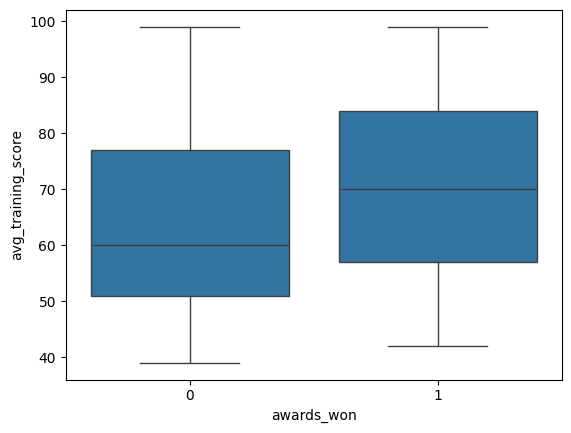
**Target variable vs awards\_won:**

****

**Table-10**

****

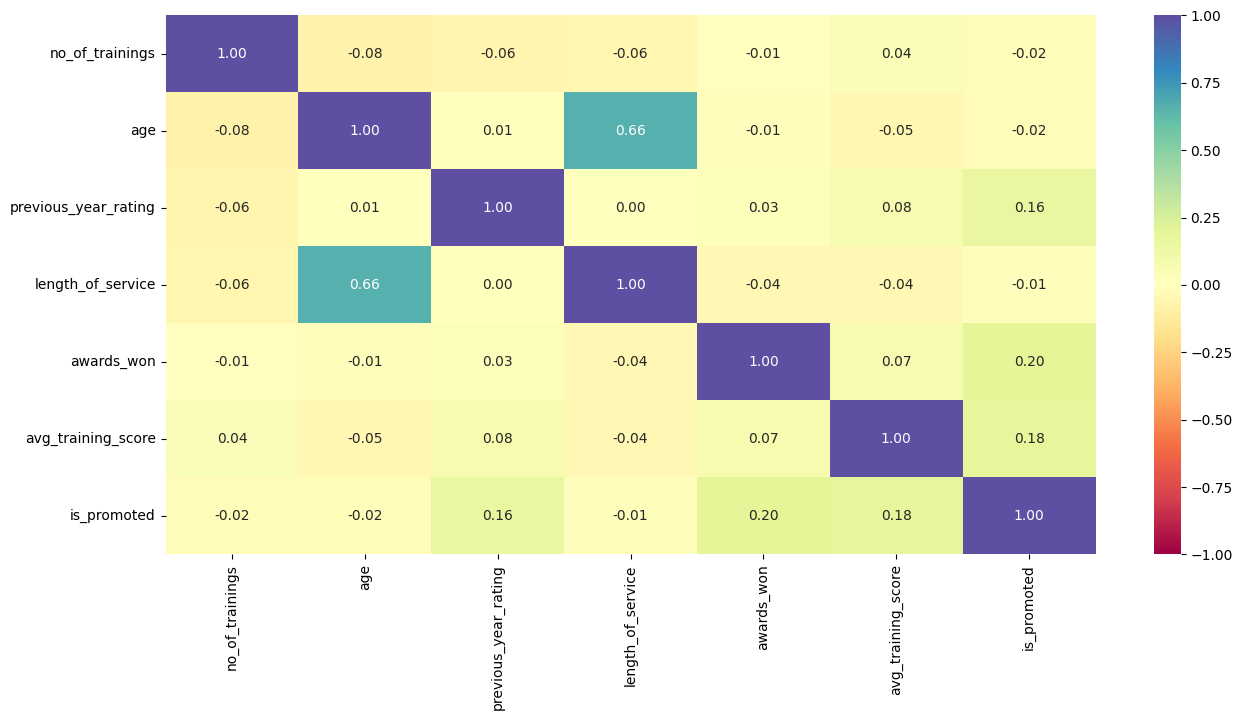
**Fig-22**

****

**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:

1. **length\_of\_service (-0.01)**:
   * Virtually it has no correlation, indicating that tenure alone doesn’t drive promotion.
2. **age (-0.02)**:
   * Slightly negative correlation; age is not a significant promotion factor.
3. **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.

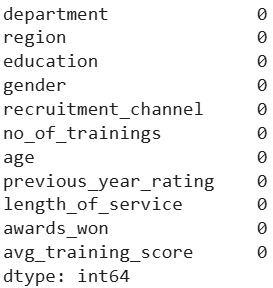
### **Actionable Insight**: Encourage award programs and performance-focused training to influence promotions.

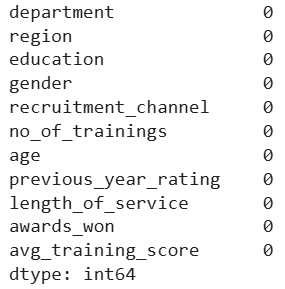
**4. DATA PREPROCESSING**

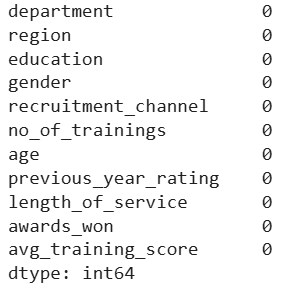
**4.1. Missing Value Treatment**:

* **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







**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.

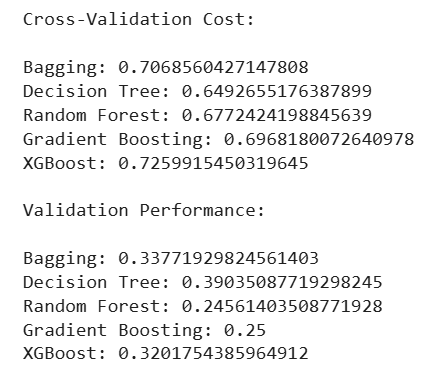
**How to reduce this loss i.e. need to reduce False Negatives as well as False Positives?**

* 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**

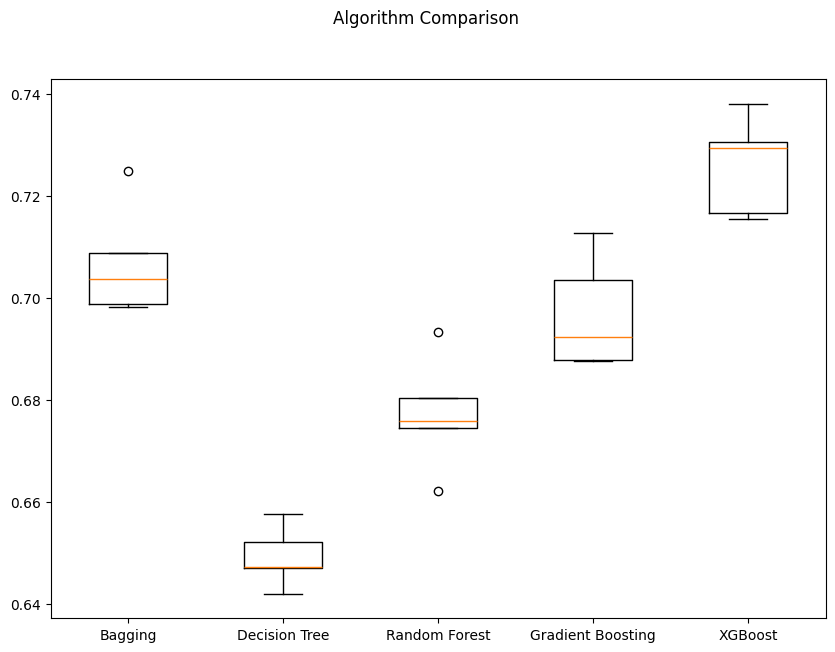


**Performance:**

* 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:**

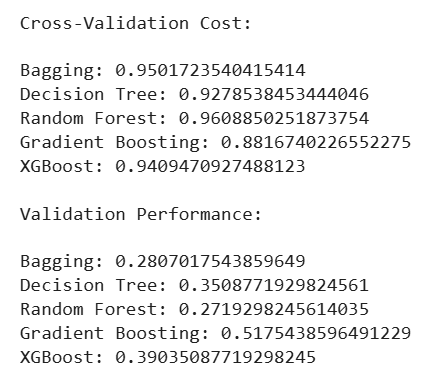
****

**Fig-25**

**5.3. Model Building-Oversampled data**

**Technique used:** SMOTE (Synthetic Minority Oversampling Technique)

**Models Used:** Same as above

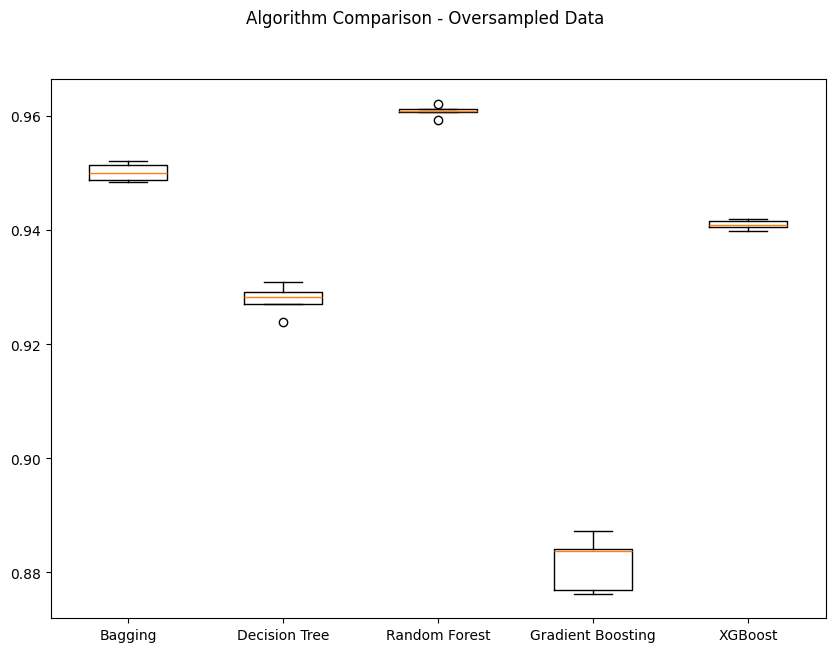


**Performance:**

* 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:**

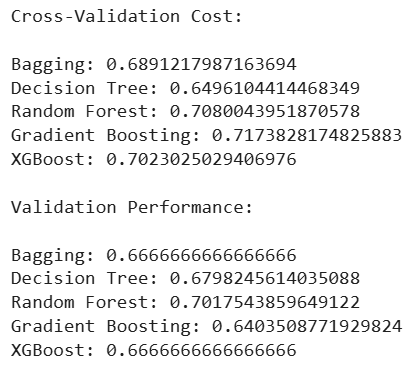


**Fig-26**

**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

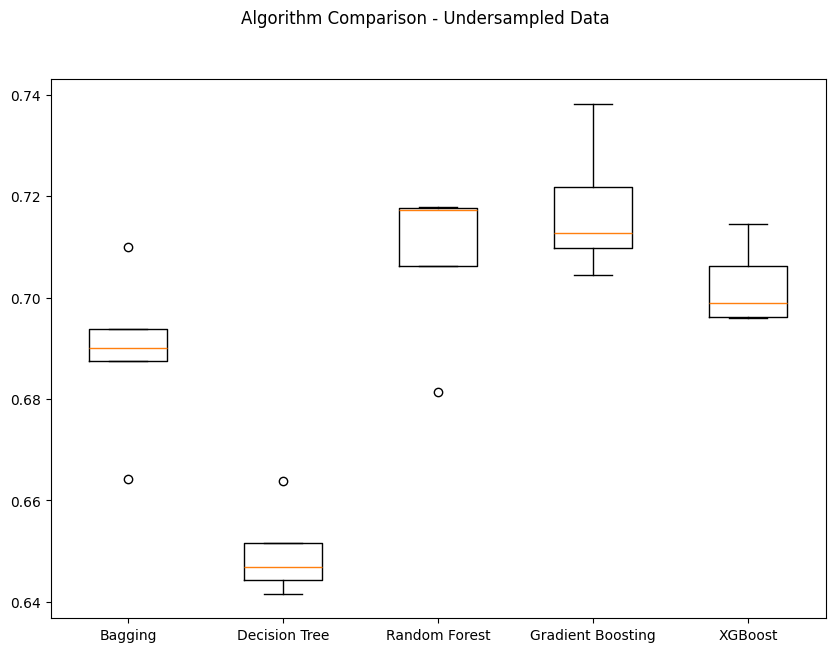


**Performance:**

* 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**

# ****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

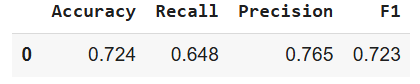
**Approach:**

* RandomizedSearchCV
* Metric: F1-score

**Result:**

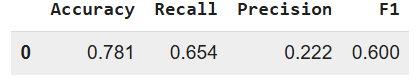
* Adaboost using under-sampled data

**Performance on train set:**

****

**Table-11**

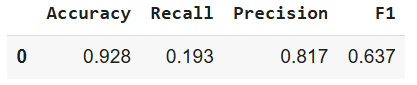
**Performance on validation set:**



**Table-12**

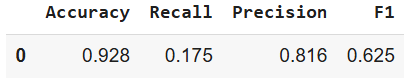
* Adaboost using Original data

**Performance on train set:**



**Table-13**

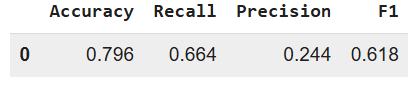
**Performance on validation set:**

****

**Table-14**

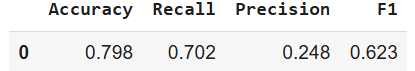
* Gradient boosting using under-sampled data

**Performance on train set:**



**Table-15**

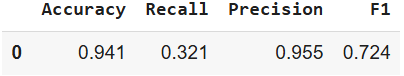
**Performance on validation set:**



**Table-16**

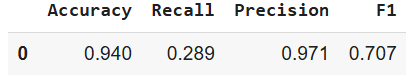
* Gradient boosting using Original data

**Performance on train set:**

****

**Table-17**

**Performance on validation set:**

****

**Table-18**

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

# ****MODEL PERFORMANCE COMPARISON & FINAL MODEL SELECTION****

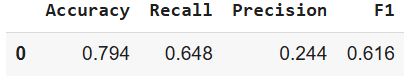
# ****Table-19****

**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:**

****

**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**

# download (32).pngFig-27

# ****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.