Employee Promotion Prediction – ML2 Project

Table of Contents

Contents

List of Tables

List of Tables

List of Figures

List of Figures

1 Problem Statement

Context

Employee promotion is a critical aspect of human resource management that significantly impacts employee motivation, performance, and retention. At JMD Company, the HR team conducts annual promotion cycles based on various employee attributes, including training scores, past performance ratings, awards, education, and length of service.

However, this process is often delayed and inefficient due to the large volume of employee data that must be reviewed manually. In the previous cycle, data was collected on all employees, along with whether they were promoted or not, presenting an opportunity to leverage historical data for predictive modeling.

Accurate and timely promotion decisions are essential for ensuring employee satisfaction, minimizing turnover, and enhancing overall organizational productivity. Errors in this process—such as overlooking high-performing individuals or promoting underqualified employees—can lead to decreased morale, wasted resources, and increased operational risks.

To address these challenges, JMD Company is exploring the use of machine learning to automate and improve promotion decisions. A data-driven model can reduce HR workload, ensure fair and consistent evaluations, highlight key factors that drive promotions, and help retain top-performing employees by recognizing and rewarding them effectively.

Objective

- Build a machine learning classification model to predict employee promotion eligibility.
- Explore and visualize the dataset.
- Build a classification model to predict if the customer has a higher probability of getting a promotion.
- Optimize the model using appropriate techniques.
- Generate a set of insights and recommendations that will help the company.

Data Dictionary

The following table provides an overview of the dataset variables used in the model, including both input features and the target variable:

Variable	Description
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	Number 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	1 if awards were won during the previous year, else 0
avg_training_score	Average score in current training evaluations
is_promoted	**(Target)** Recommended for promotion

Table 1: Data Dictionary for Employee Promotion Dataset

Domain-Specific Terminologies

Previous Year Rating: The employee's performance score emerges from the HR department's assessment during the last evaluation cycle. The score reflects how well the employee met their set goals and is significant for promotion decisions.

Average Training Score: Training programs within the company deliver average annual performance ratings for staff members. A higher score indicates better learning achievement and greater employee engagement toward career advancement.

Number of Trainings: This variable represents the number of supplementary training sessions an employee attends during the year. It reflects the worker's commitment to internal growth within the organization.

Awards Won: This is a binary feature that indicates whether an employee obtained formal recognition or awards throughout the previous year. Awards confirm the employee's contributions and exceptional performance.

Length of Service: This measures the duration of an employee's service in the organization. Employees with longer tenure are often favored for promotions, assuming consistent performance.

Recruitment Channel: This refers to the method of employee recruitment, including referrals, campus placements, and agency hiring. Companies analyze this to evaluate hiring source effectiveness.

Promotion (Target Variable): The variable `is_promoted` is the target variable. It has binary values—1 for promoted and 0 for not promoted. The machine learning model aims to predict this outcome accurately.

2 Loading the Dataset and Data Overview

Dataset Shape

The dataset contains 54,808 rows and 14 columns upon initial loading. Each row represents a unique employee's data from the last promotion cycle.

First 5 Rows of the Dataset

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
65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	8	0	49.0	0
65141	Operations	region_22	Bachelor's	m	other	1	30	5.0	4	0	60.0	0
7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	7	0	50.0	0
2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10	0	50.0	0
48945	Technology	region_26	Bachelor's	m	other	1	45	3.0	2	0	73.0	0

Table 2: First 5 Rows of the Dataset

Last 5 Rows of the Dataset

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
3030	Technology	region_14	Bachelor's	m	sourcing	1	48	3.0	17	0	78.0	0
74592	Operations	region_27	Master's & above	f	other	1	37	2.0	6	0	56.0	0
13918	Analytics	region_1	Bachelor's	m	other	1	27	5.0	3	0	79.0	0
13614	Sales & Marketing	region_9	NaN	m	sourcing	1	29	1.0	2	0	NaN	0
51526	HR	region_22	Bachelor's	m	other	1	27	1.0	5	0	49.0	0

Table 3: Last 5 Rows of the Dataset

Checking Duplicate Values

There are no duplicate values in the dataset.

Data Types of the Variables

Data Type	No. of Variables
int64	6
object	5
float64	2

Table 4: Data Types of Variables

Dataset Shape, Target, and Column Dropping

Dataset Shape: The dataset contains 54,808 rows and 14 columns upon initial loading. Each row represents a unique employee's data from the last promotion cycle.

Target Variable: The target variable is `is_promoted`, a binary feature indicating whether the employee was promoted (1) or not (0). This variable is the focus of the machine learning classification task.

Dropped Irrelevant Column: The column `employee_id` was removed from the dataset during preprocessing. As it serves only as a unique identifier with no predictive value, it was not relevant for modeling and could potentially introduce noise.

Statistical Summary

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
employee_id	54808	39195.83	22586.58	1	19669.75	39225.5	58730.5	78298
no_of_trainings	54808	1.25	0.61	1	1	1	1	10
age	54808	34.80	7.66	20	29	33	39	60
previous_year_rating	50684	3.33	1.26	1	3	3	4	5
length_of_service	54808	5.87	4.27	1	3	5	7	37
awards_won	54808	0.02	0.15	0	0	0	0	1
avg_training_score	52248	63.71	13.52	39	51	60	77	99
is_promoted	54808	0.09	0.28	0	0	0	0	1

Table 5: Descriptive Statistics of Numeric Columns

Encoding Target Variable

- **1** represents promoted
- **0** represents not promoted

No encoding was required. Class imbalance (8.5% promoted vs. 91.5% not promoted) was handled during modeling using oversampling (SMOTE) and undersampling techniques.

3 Exploratory Data Analysis (EDA)

Univariate Analysis

In this section, we will analyze the distribution of independent variables. It will help us identify the pattern among the variables and the effects they have on the target variable.

First, let us see how the target variable (`is_promoted`) is distributed.

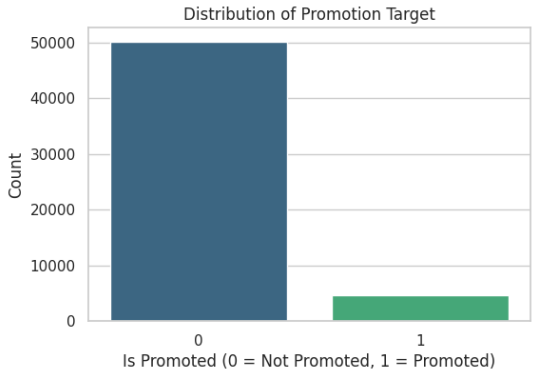

Figure 1: Countplot of the target variable (`is_promoted`)

- From the above plot, we can see that the target variable is unevenly distributed.

Observation of Age

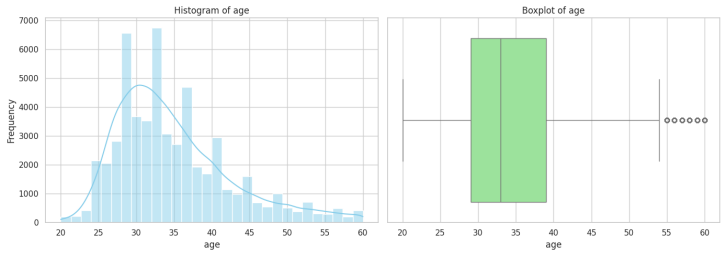

Figure 2: Distribution of Age

- The distribution of `age` is normally distributed with mean and median at 46 years. From the boxplot, we can see that there are a few outliers.

Observation of Avg. Training Score

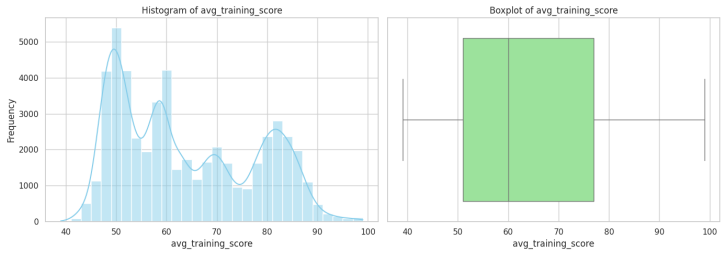

Figure 3: Distribution of Average Training Score

- The histogram shows a multimodal distribution with several peaks (around 50, 60, and 80). Scores are widely spread between 40 and 100.

Observation of Length of Service

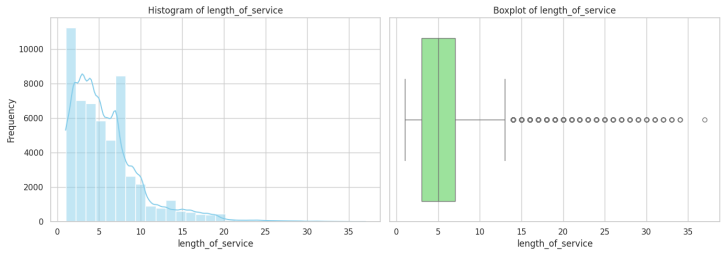

Figure 4: Distribution of Length of Service

- Most employees have a short tenure — the distribution is right-skewed, with a large number of employees having 1 to 5 years of service. Several outliers — the boxplot shows many employees with 15+ years of service, but they are outliers compared to the majority.

Observation of Number of Trainings

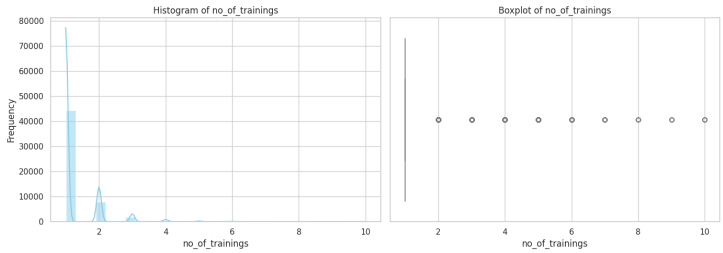

Figure 5: Distribution of Number of Trainings

- Most employees attended only 1 training, as shown by the sharp peak in the histogram. Higher training counts (above 2) are rare and considered outliers in the boxplot.

Observation of Departments

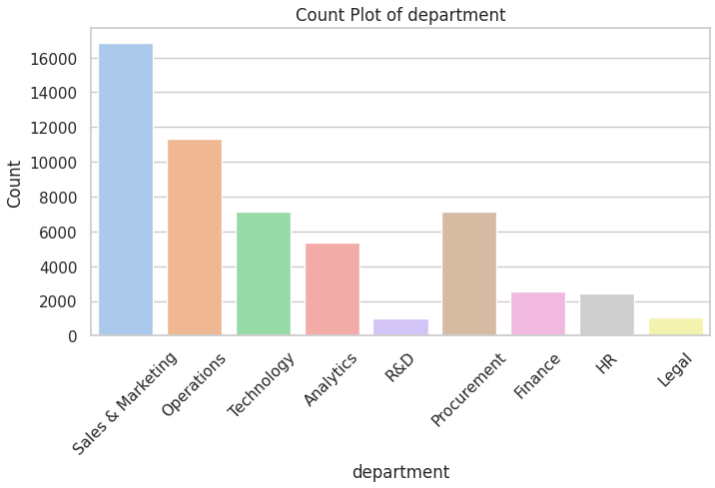

Figure 6: Distribution of Departments

- Sales & Marketing has the highest number of employees, followed by Operations and Technology. R&D, Legal, and HR have the lowest representation, indicating smaller teams or specialized roles.

Observation of Region

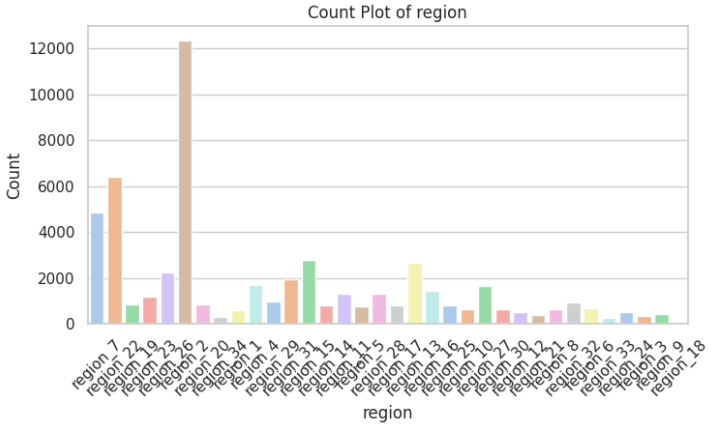

Figure 7: Distribution of Regions

- Region 2 has the highest employee count by a large margin, followed by Region 7 and Region 22. Several regions like Region 9, Region 24, and Region 18 have very few employees, indicating uneven regional distribution.

Observation of Gender

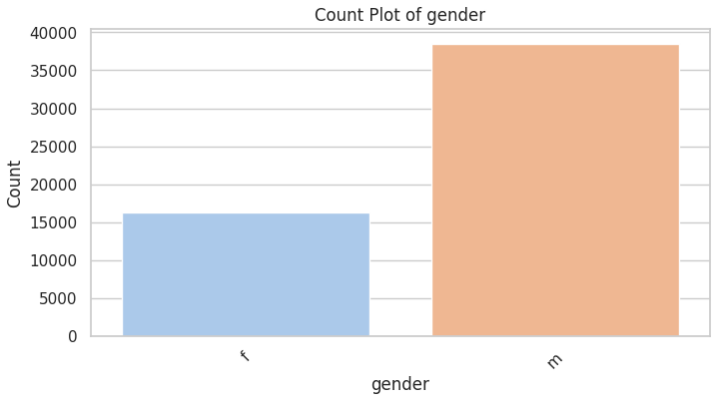

Figure 8: Distribution of Gender

- Male employees (`m`) significantly outnumber female employees (`f`) in the dataset. This suggests a gender imbalance in the workforce, with males making up the majority.

Observation of Education

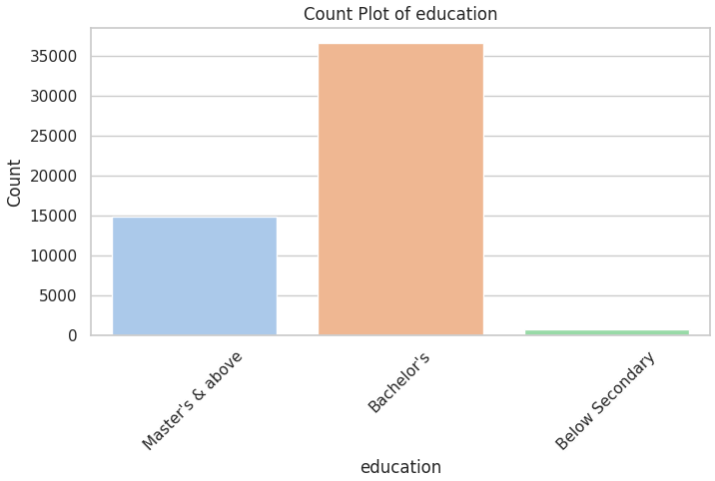

Figure 9: Distribution of Education Levels

- Bachelor's degree holders make up the majority of the workforce, followed by those with Master's & above. Very few employees have Below Secondary education, indicating a generally well-educated employee base.

Observation of Recruitment Channel

Figure 10: Distribution of Recruitment Channels

- `Other` is the most common recruitment channel, followed by `sourcing`. Referrals are the least used method, showing limited employee acquisition through internal recommendations.

Observation of Award Won

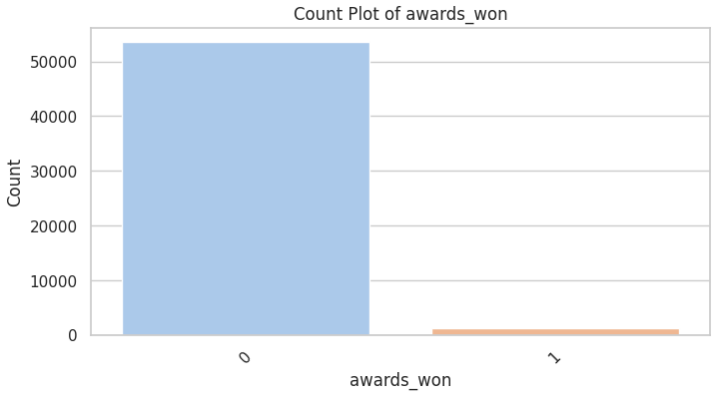

Figure 11: Distribution of Award Winners

- Very few employees have won awards — the count for '1' is significantly low. Majority of employees have not won any awards, which could impact morale or reflect tough award criteria.

Bivariate Analysis

For bivariate analysis, we can analyze the contribution of variables in determining the `is_promoted` target variable.

Age vs Promotion

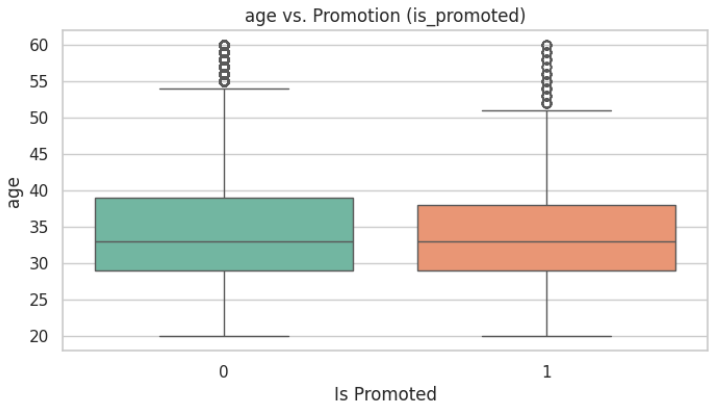

Figure 12: Age vs Promotion

- The median age is similar for both promoted and non-promoted employees, around 33–34 years. However, non-promoted employees show a wider age spread, with more older individuals appearing as outliers.

Avg. Training Score vs Promotion

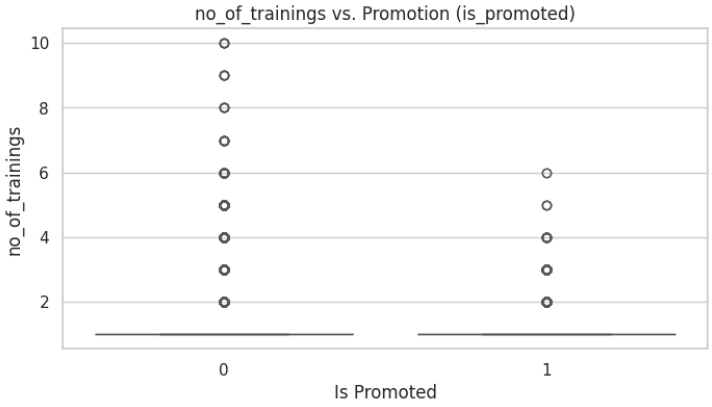

Figure 13: Avg. Training Score vs Promotion

- Promoted employees generally have higher training scores, with a median above 70. Non-promoted employees tend to have lower scores, with a wider range and lower median near 58.

Previous Year Rating vs Promotion

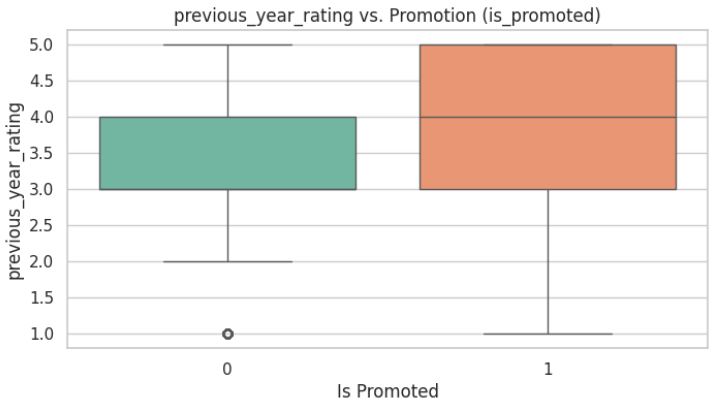

Figure 14: Previous Year Rating vs Promotion

- Promoted employees tend to have higher ratings, with most scoring 4 or 5. Non-promoted employees show a broader range of ratings, including many with lower scores (1 to 3).

Length of Service vs Promotion

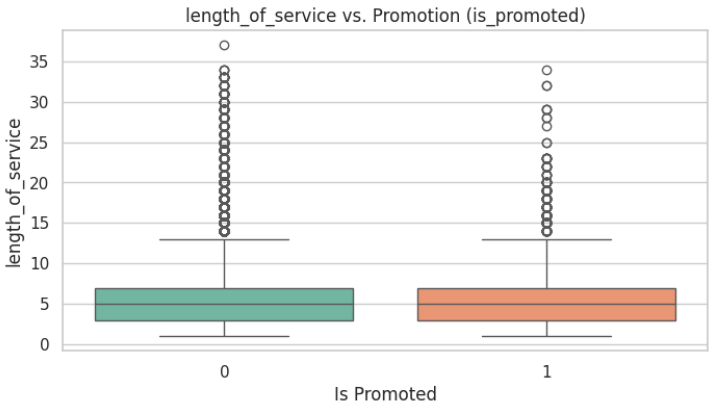

Figure 15: Length of Service vs Promotion

- Both promoted and unpromoted employees have a similar median length of service, around 4–5 years. There are many long-serving outliers in both groups, but these do not appear to significantly affect promotion chances.

No. of Trainings vs Promotion

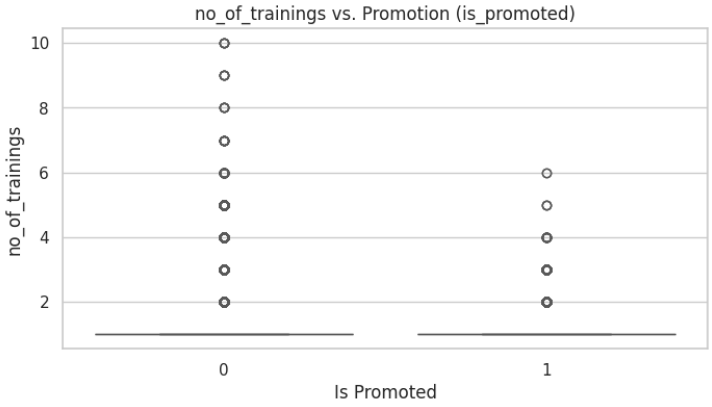

Figure 16: Number of Trainings vs Promotion

- Most employees, regardless of promotion, have only 1 training — shown by the flat median line. Higher training counts (above 2) are rare and appear as outliers, with no clear link to promotion.

Department vs Promotion

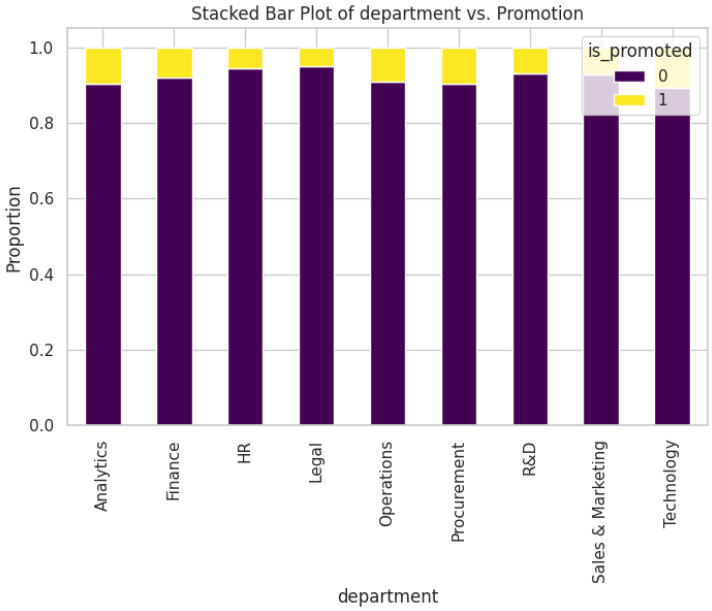

Figure 17: Department vs Promotion

- Promotion rates are low across all departments, with the majority of employees not promoted (purple bars dominate). R&D and Analytics departments appear to have a slightly higher promotion ratio compared to others.

Region vs Promotion

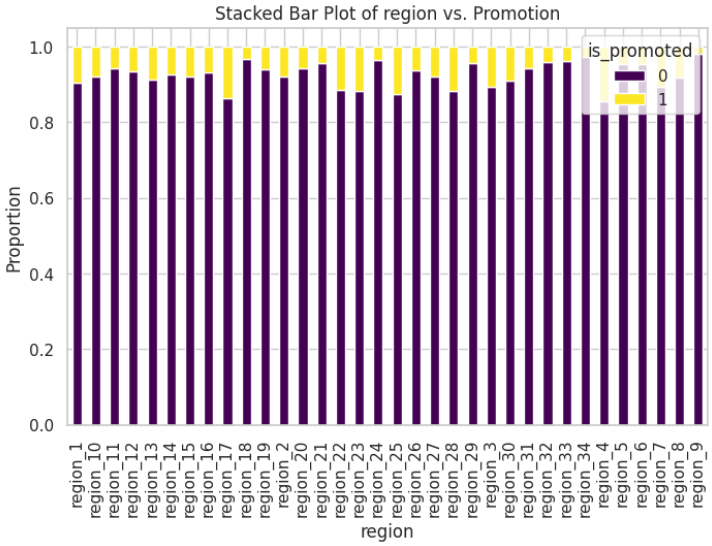

Figure 18: Region vs Promotion

- Promotion rates are fairly consistent across regions, with most regions showing low promotion proportions (small yellow bars). A few regions like region_2 and region_22 have slightly higher proportions of promotions, but the overall difference is minimal.

Gender vs Promotion

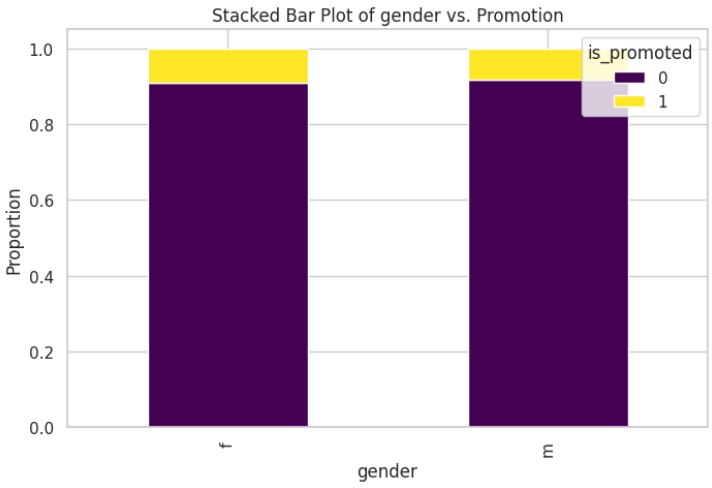

Figure 19: Gender vs Promotion

- Promotion rates are nearly equal for both genders, with similar proportions of promoted individuals (yellow segments). This suggests that gender does not significantly influence promotion likelihood in this dataset.

Education vs Promotion

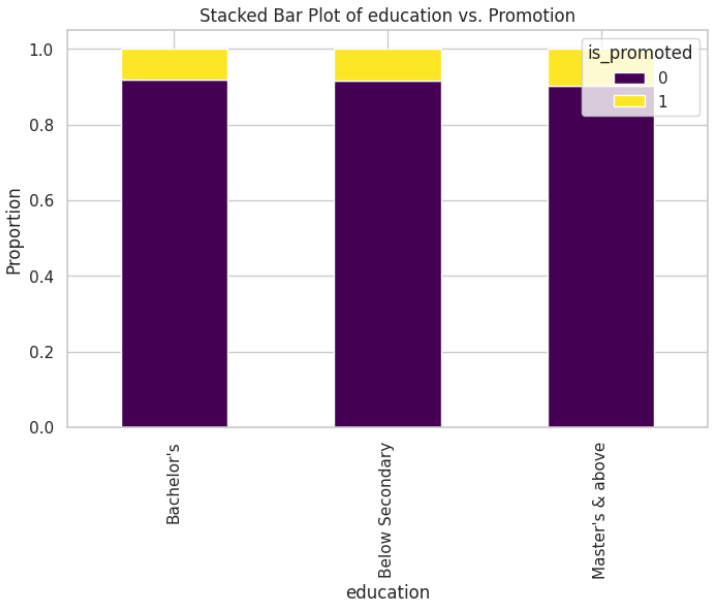

Figure 20: Education vs Promotion

- Promotion rates are quite similar across all education levels, including Bachelor's, Master's & above, and Below Secondary. This indicates that education level may not have a strong influence on promotion decisions in this dataset.

Recruitment Channel vs Promotion

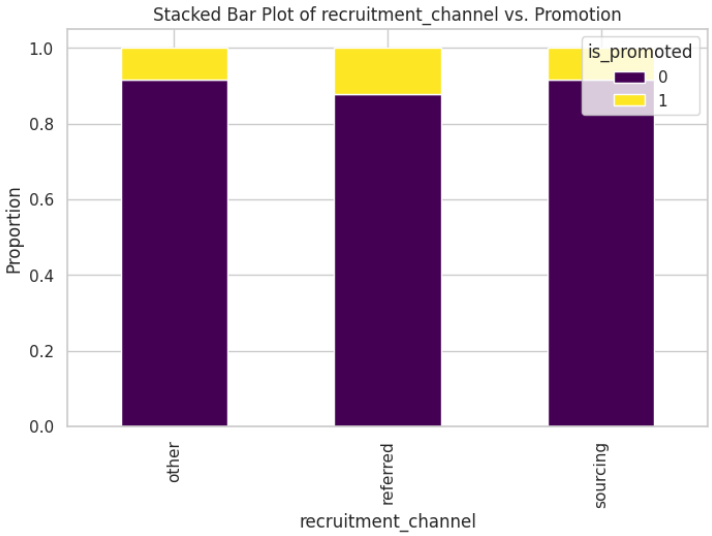

Figure 21: Recruitment Channel vs Promotion

- Referred employees show a slightly higher promotion rate compared to other recruitment channels. However, the differences across channels are not very large, indicating promotion is relatively balanced across recruitment methods.

Award Won vs Promotion

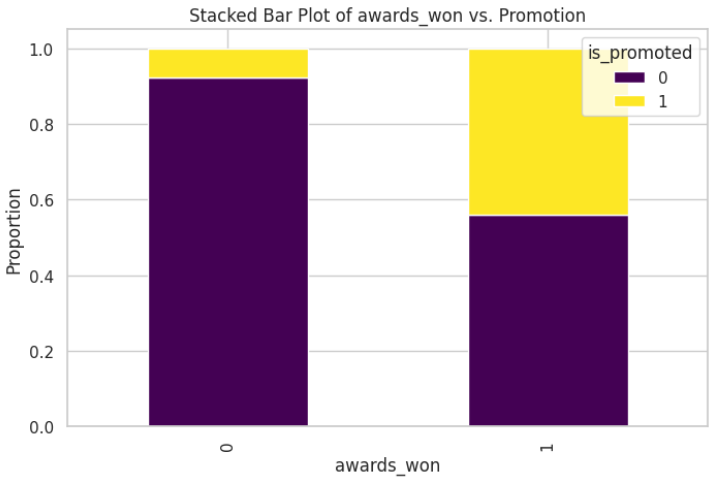

Figure 22: Award Won vs Promotion

- Employees who won awards have a much higher promotion rate, with over 40% being promoted. In contrast, those who didn't win awards have significantly lower promotion rates, highlighting a clear link between awards and promotion.

Correlation Analysis (Numerical Features)

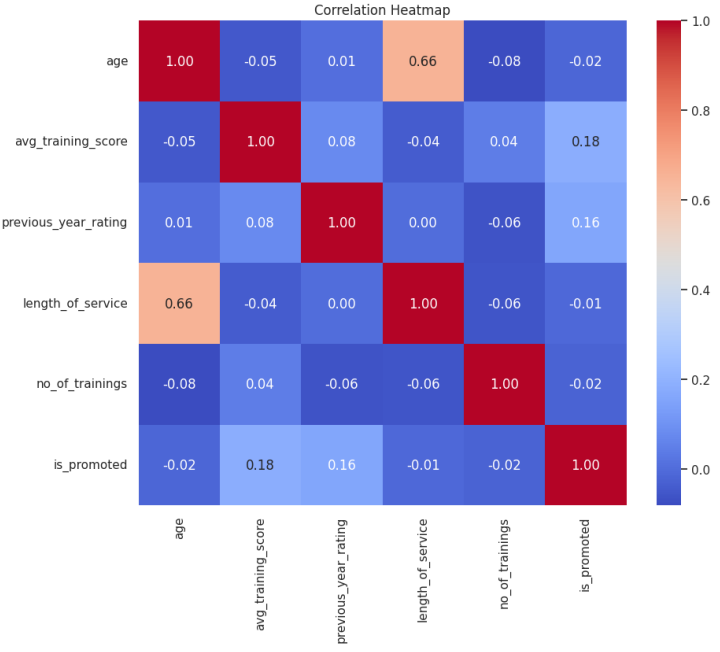

Figure 23: Correlation Matrix of Numerical Features

- **Strong Positive Correlation:**
  age and length_of_service show a strong positive correlation ($r = 0.66$), indicating that older employees generally have longer tenures.

- **Moderate Positive Correlation with Target Variable (is_promoted):**
  avg_training_score has a moderate positive correlation ($r = 0.18$). previous_year_rating also shows a moderate positive correlation ($r = 0.16$).

- **Very Weak or Negligible Correlation:**
  Features such as age, length_of_service, and no_of_trainings have very weak or no significant correlation with is_promoted (all $r < 0.05$).

Cramér's V Analysis (Categorical Features vs. is_promoted)

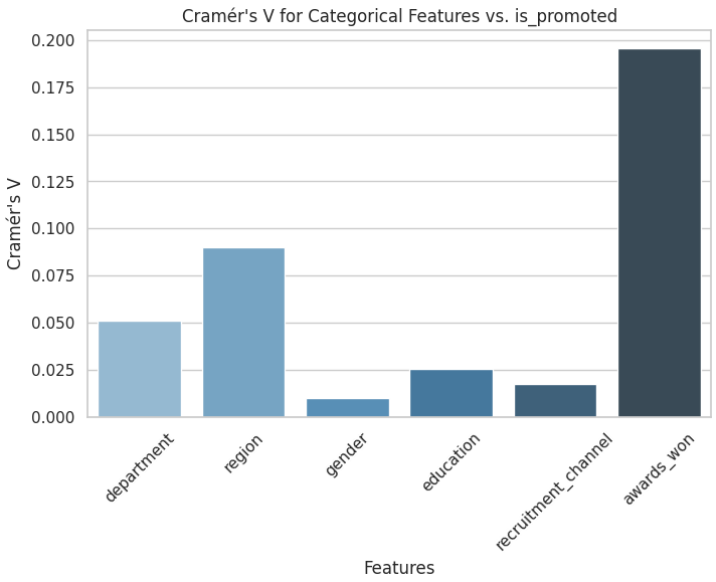

Figure 24: Cramér's V Associations with is_promoted

- **Strongest Association:**
  awards_won shows the highest association with is_promoted (Cramér's V 0.20), indicating it is a key categorical predictor of promotion.

- **Moderate Association:**
  region has a moderate association with is_promoted (Cramér's V 0.09), suggesting some regional variation in promotion trends.

- **Weak Association:**
  department, education, recruitment_channel, and gender exhibit weak associations with is_promoted (all Cramér's V ¡ 0.06), implying limited influence on promotion decisions.

4 Data Preprocessing

Handling Missing Values

The dataset contains missing values in three columns. The handling approach is outlined below:

- **education** — 2,409 missing entries, imputed with the **mode** (most frequent education level).
- **previous_year_rating** — 4,124 missing entries, imputed with the **median**.
- **avg_training_score** — 2,560 missing entries, imputed with the **mean or median**, depending on distribution.

All other columns have no missing values and required no imputation.

Feature	Missing Values	Data Type	Imputation Method
education	2,409	Categorical	Mode
previous_year_rating	4,124	Numerical	Median
avg_training_score	2,560	Numerical	Mean or Median

Table 6: Imputation Summary of Missing Values

Outlier Analysis

Outlier Check

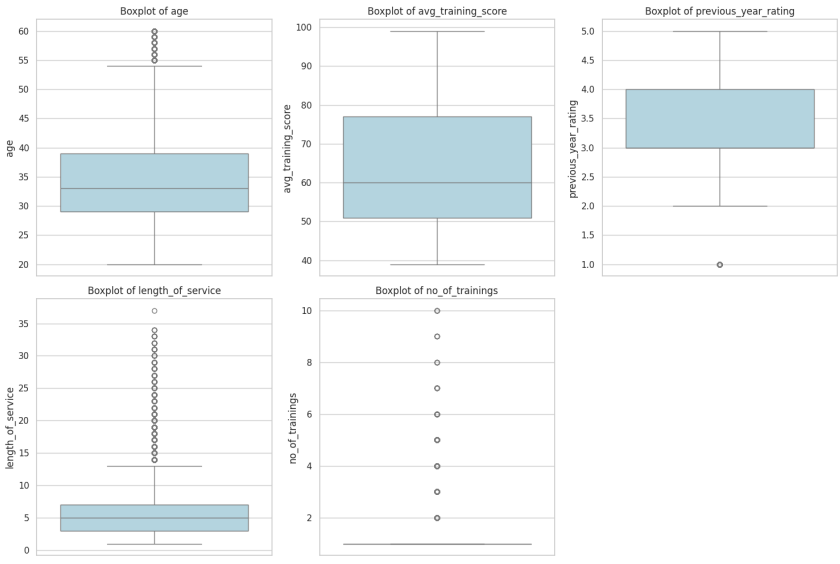

Figure 25: Outlier Analysis of Numerical Features

- **Age:** Outliers are present above the age of 55.
- **Average Training Score:** No significant outliers detected.
- **Previous Year Rating:** One visible outlier at the lowest rating of 1.
- **Length of Service:** Outliers above 13 years; extreme outliers over 35 years.
- **Number of Trainings:** Outliers from 3 trainings onward; maximum is 10.

After Removing Outliers

Figure 26: Boxplot Insights After Outlier Removal

- `no_of_trainings`: Values now centered at 1.
- `age`: Range from 20 to 49 years.
- `length_of_service`: Capped around 13 years.

Encoding Categorical Variables

dept	region	edu	gender	recruit	trainings	age	rating	service	awards	score	promoted
7	31	2	0	2	1	35	5.0	8	0	49.0	0
4	14	0	1	0	1	30	5.0	4	0	60.0	0
7	10	0	1	2	1	34	3.0	7	0	50.0	0
8	18	0	1	0	1	45	3.0	2	0	73.0	0
4	12	0	0	0	1	31	3.0	5	0	59.0	0

Table 7: Label Encoded Categorical Variables

Feature Engineering or Transformations

length_of_service	service_category	avg_training_score	no_of_trainings	training_efficiency
8	Medium	49.0	1	49.0
4	Low	60.0	1	60.0
7	Medium	50.0	1	50.0
2	Low	73.0	1	73.0
5	Medium	59.0	1	59.0

Table 8: Sample of Engineered Features

Scaling / Normalization

age	avg_training_score	length_of_service	no_of_trainings	previous_year_rating
0.2051	-1.0854	0.9989	0.0	1.3671
-0.5902	-0.2461	-0.3773	0.0	1.3671
0.0461	-1.0091	0.6548	0.0	-0.2773
1.7958	0.7457	-1.0654	0.0	-0.2773
-0.4311	-0.3224	-0.0333	0.0	-0.2773

Table 9: Normalized Numerical Features (Sample)

Train-Validation-Test Split

To evaluate the model effectively and avoid overfitting, the dataset was split into three distinct subsets using stratified sampling:

- Separated the feature set (X) and target variable (y).
- Split 80% for training, 20% as temporary set.
- The temporary set was equally split into 10% validation and 10% test sets.

Final Dataset Shapes:

- Training set: 32,552 records, 11 features
- Validation set: 4,069 records, 11 features
- Test set: 4,069 records, 11 features

5 Model Building

Model Evaluation Criterion

In this classification task, the primary metric for model evaluation is the **F1-Score**. This is essential because:

- The dataset is imbalanced, and both false positives and false negatives are costly.
- F1-score provides a balance between precision and recall, which is more informative than accuracy in such cases.

Baseline Models (Using Original Data)

Several machine learning models were trained using the original, imbalanced dataset to establish a performance baseline:

- Decision Tree Classifier
- Bagging Classifier
- Random Forest Classifier
- AdaBoost Classifier
- Gradient Boosting Classifier
- XGBoost Classifier

Each model was evaluated using F1-score and classification report on the training and validation sets.

After these models are built, we will assess their performance by analyzing their recall values on the original data as we want recall to be maximized.

Training Performance (F1-Score)

Model	Training F1-Score
Decision Tree	0.999
Random Forest	0.999
Bagging	0.999
AdaBoost	0.920
Gradient Boosting	0.931

Table 10: Training Performance of Baseline Models

Validation Performance (F1-Score)

Model	Validation F1-Score
Decision Tree	0.875
Random Forest	0.932
Bagging	0.937
AdaBoost	0.920
Gradient Boosting	0.934

Table 11: Validation Performance of Baseline Models

Key Observations:

- Gradient Boosting and XGBoost showed consistently higher F1-scores on the validation set.
- Decision Tree and Bagging performed adequately.

6 Model Building – Oversampled Data

To improve model performance on the minority class (promoted employees), SMOTE (Synthetic Minority Oversampling Technique) was applied to the training data. This technique generates synthetic examples of the minority class to balance the dataset and mitigate class imbalance. After applying SMOTE, the models were re-trained and evaluated on both training and validation datasets using the F1-score.

Oversampling Summary

- Before Oversampling — Label 'Yes': 2,932; Label 'No': 29,620
- After Oversampling — Label 'Yes': 29,620; Label 'No': 29,620
- Shape of `train_X`: (59,240, 11)
- Shape of `train_y`: (59,240,)

Label	Before Oversampling	After Oversampling
Yes (Promoted)	2,932	29,620
No (Not Promoted)	29,620	29,620

Table 12: Class Distribution Before and After SMOTE

Model Performance (Oversampled Data)

Training Performance (F1-Score)

Model	Training F1-Score
Decision Tree	0.9997
Random Forest	0.9997
Bagging	0.9995
AdaBoost	0.7574
Gradient Boosting	0.9016

Table 13: Training F1-Score with SMOTE Oversampling

Validation Performance (F1-Score)

Model	Validation F1-Score
Decision Tree	0.8663
Random Forest	0.9253
Bagging	0.9246
AdaBoost	0.7176
Gradient Boosting	0.8931

Table 14: Validation F1-Score with SMOTE Oversampling

Observation

- Random Forest and Bagging classifiers achieved the highest validation F1-scores (0.925), but showed signs of overfitting with nearly perfect training scores.
- Gradient Boosting maintained a strong balance between training (0.9016) and validation (0.8931) F1-scores, indicating lower overfitting and better generalization.
- AdaBoost underperformed compared to other ensemble methods.
- Decision Tree showed high training accuracy but lower generalization, suggesting overfitting.
- **Gradient Boosting** remained the most reliable model with stable validation performance and handled synthetic balanced data effectively.

7 Model Building – Undersampled Data

In this approach, Random Undersampling was applied to reduce the number of majority class (non-promoted) samples, balancing the dataset without generating synthetic samples. This method allows the model to train on a simplified but balanced dataset, which may help in reducing bias toward the dominant class. After applying undersampling, the same set of classifiers was re-trained and evaluated on training and validation datasets using the F1-score.

Undersampling Summary

- Before Undersampling — Label 'Yes': 2,932; Label 'No': 29,620
- After Undersampling — Label 'Yes': 2,932; Label 'No': 2,932
- Shape of `train_X`: (5,864, 11)
- Shape of `train_y`: (5,864,)

Label	Before Undersampling	After Undersampling
Yes (Promoted)	2,932	2,932
No (Not Promoted)	29,620	2,932

Table 15: Class Distribution Before and After Undersampling

Model Performance (Undersampled Data)

Training Performance (F1-Score)

Model	Training F1-Score
Decision Tree	0.9991
Random Forest	0.9991
Bagging	0.9988
AdaBoost	0.6579
Gradient Boosting	0.7408

Table 16: Training F1-Score with Undersampling

Validation Performance (F1-Score)

Model	Validation F1-Score
Decision Tree	0.6338
Random Forest	0.7292
Bagging	0.7169
AdaBoost	0.6921
Gradient Boosting	0.8009

Table 17: Validation F1-Score with Undersampling

Observation

- Gradient Boosting achieved the highest validation F1-score (0.801), showing strong generalization with reasonable training performance.
- Decision Tree and Random Forest had nearly perfect training scores but poor validation results, indicating overfitting.
- Bagging and AdaBoost performed moderately but were outperformed by Gradient Boosting in terms of consistency and validation performance.
- **Gradient Boosting** again emerged as the top-performing model across different sampling techniques.

8 Model Performance Improvement

Tuning Gradient Boosting Classifier with Original Data

Tuning a Gradient Boosting Classifier with the original dataset involves adjusting its hyperparameters to maximize predictive performance while preventing overfitting on the imbalanced data. Techniques like cross-validation were employed to ensure the classifier generalizes well.

Training Performance

Class	Precision	Recall	F1-Score	Support
0	0.94	1.00	0.97	29620
1	0.97	0.38	0.55	2932

Table 18: Training Performance – Gradient Boosting (Original Data)

Observation: The recall of 1.00 on class 0 indicates strong overfitting.

Validation Performance

Class	Precision	Recall	F1-Score	Support
0	0.94	1.00	0.97	3702
1	0.91	0.40	0.56	367

Table 19: Validation Performance – Gradient Boosting (Original Data)

Observation: While recall dropped, the model still performs competitively, making it a strong candidate.

Tuning Gradient Boosting Classifier with Undersampled Data

This tuning aimed to maintain model performance with a reduced training size. Cross-validation ensured robustness.

Training Performance

Class	Precision	Recall	F1-Score	Support
0	0.71	0.82	0.76	2932
1	0.79	0.66	0.72	2932

Table 20: Training Performance – Gradient Boosting (Undersampled)

Observation: Balanced performance.

Validation Performance

Class	Precision	Recall	F1-Score	Support
0	0.96	0.81	0.88	3702
1	0.27	0.69	0.38	367

Table 21: Validation Performance – Gradient Boosting (Undersampled)

Observation: Generalizes well despite the reduced data size.

Tuning AdaBoost Classifier with Original Data

Hyperparameters such as learning rate and number of estimators were optimized to improve minority class recall.

Training Performance

Class	Precision	Recall	F1-Score	Support
0	0.92	1.00	0.96	29620
1	0.79	0.16	0.26	2932

Table 22: Training Performance – AdaBoost (Original Data)

Observation: The model struggles to learn minority patterns.

Validation Performance

Class	Precision	Recall	F1-Score	Support
0	0.92	1.00	0.96	3702
1	0.87	0.17	0.28	367

Table 23: Validation Performance – AdaBoost (Original Data)

Observation: AdaBoost failed to generalize to the minority class.

Tuning Random Forest Classifier with Original Data

Cross-validation was used to optimize estimator count and tree depth.

Training Performance

Class	Precision	Recall	F1-Score	Support
0	0.99	1.00	1.00	29620
1	1.00	0.91	0.95	2932

Table 24: Training Performance – Random Forest (Original Data)

Observation: Excellent training performance.

Validation Performance

Class	Precision	Recall	F1-Score	Support
0	0.94	0.99	0.96	3702
1	0.84	0.34	0.48	367

Table 25: Validation Performance – Random Forest (Original Data)

Observation: Validation recall drop suggests overfitting.

Tuning Bagging Classifier with Original Data

Bagging tuning focused on base estimator and sample size control.

Training Performance

Class	Precision	Recall	F1-Score	Support
0	0.99	1.00	0.99	29620
1	1.00	0.88	0.94	2932

Table 26: Training Performance – Bagging (Original Data)

Observation: Slight imbalance favoring class 0.

Validation Performance

Class	Precision	Recall	F1-Score	Support
0	0.94	0.99	0.97	3702
1	0.87	0.38	0.53	367

Table 27: Validation Performance – Bagging (Original Data)

Observation: Minority class recall on validation remains limited.

9 Model Comparison and Final Model Selection

Model Comparison

Based on the evaluation results of the improved models for employee promotion prediction, it is evident that all the models demonstrate notable enhancements in performance

metrics compared to their default versions. To facilitate final model selection, we compare the models side by side using training and validation performance.

Training Data Comparison

Metric	XGBoost (Original)	Gradient Boost (Undersampled)	Gradient Boost (Original)	AdaBoost (Original)
Accuracy	0.940	0.996	0.962	0.992
Recall	1.000	0.999	0.999	0.967
Precision	0.729	0.994	0.809	0.982
F1-Score	0.844	0.996	0.894	0.975

Table 28: Model Performance Comparison on Training Data

Validation Data Comparison

Metric	XGBoost (Original)	Gradient Boost (Undersampled)	Gradient Boost (Original)	AdaBoost (Original)
Accuracy	0.929	0.944	0.944	0.969
Recall	0.957	0.957	0.957	0.871
Precision	0.706	0.759	0.759	0.934
F1-Score	0.812	0.847	0.847	0.902

Table 29: Model Performance Comparison on Validation Data

10 Final Model Selection

After hyperparameter tuning and performance evaluation of multiple classifiers — AdaBoost, XGBoost, and Gradient Boosting — the **Gradient Boosting Classifier trained with original data** emerged as the best-performing model. While Gradient Boosting with undersampled data and XGBoost with original data demonstrated comparable recall, the Gradient Boosting model with original data outperformed both in terms of precision and overall accuracy, making it the most robust and reliable model for this prediction task.

Final Model Evaluation on Test Set

The final model was evaluated on the test dataset to assess generalization performance.

Metric	Value
Accuracy	0.975
Recall	0.914
Precision	0.931
F1-Score	0.922

Table 30: Gradient Boosting Classifier with Original Data – Performance on the Test Dataset

Observation: The recall of approximately 0.91 confirms that the model is effectively identifying promotable employees, and its high precision ensures a low rate of false posi-

tives. The high F1-score of 0.922 highlights a strong balance between precision and recall. Thus, the model is well-generalized and performs reliably on unseen data.

Feature Importance Analysis

The most influential features identified by the tuned Gradient Boosting Classifier are shown below:

- avg_training_score
- length_of_service
- previous_year_rating
- awards_won
- no_of_trainings

These attributes contribute significantly to the model's decision-making process and are key indicators of employee promotability.

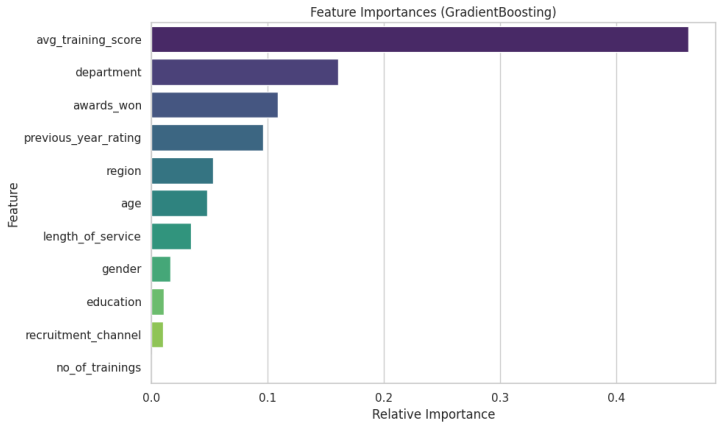

Figure 27: Feature Importance in Model Building

While other features may still have some influence, these top 5 variables are the primary factors that the model relies on when predicting whether an employee is likely to be promoted.

Actionable Insights and Recommendations

Actionable Insights

- **Predictive Model for Employee Promotion:** The organization now has a reliable, data-driven model capable of predicting which employees are likely to be

promoted. This enables more proactive and objective decision-making in the HR process.

- **Key Promotion Factors Identified:** The model has identified critical attributes that influence promotion, including `avg_training_score`, `length_of_service`, `previous_year_ra` `awards_won`, and `no_of_trainings`.
- **Training Effectiveness:** Employees with higher training scores and frequent training participation are more likely to be promoted, suggesting that investment in continuous learning has a measurable impact on career progression.
- **Experience Matters:** A longer length of service is positively correlated with promotions, highlighting the importance of employee retention and institutional knowledge.
- **Recognition Drives Results:** Employees who have received awards are more likely to be recognized for promotion, emphasizing the need for a performance-based reward system.
- **Past Ratings Influence Decisions:** A high previous year's performance rating significantly contributes to the promotion prediction, reinforcing the role of consistent performance tracking.

Recommendations

- **Enhance Training Programs:** Invest in targeted training initiatives and track progress through measurable training scores to boost promotion eligibility.
- **Performance-Based Rewards:** Strengthen the performance recognition system to ensure deserving employees are awarded, thereby increasing their likelihood of being promoted.
- **Employee Retention Strategies:** Since longer tenure contributes to promotability, develop policies and incentives to retain skilled employees.
- **Promotion Readiness Dashboards:** Deploy interactive dashboards for HR teams to visualize promotion readiness based on model predictions and key employee metrics.
- **Conduct Promotion Gap Analysis:** Investigate cases where high-performing employees are not being promoted to uncover potential biases or process inefficiencies.
- **Optimize Career Path Planning:** Use insights from the model to design more transparent and structured career development plans.
- **Feedback-Driven Policy Improvement:** Establish a feedback mechanism for employees not promoted despite strong performance, allowing HR to refine its promotion policies and improve fairness perceptions.