Machine Learning 2

Guided Project Report

Submitted to

Great Learning

POWER AHEAD

By

Pirangi Charan Teja Goud

In partial fulfillment of

PGP-DSBA

TEXAS McCombs

The University of Texas at Austin

McCombs School of Business

Table of Content

Problem definition

Problem Definition & Questions to be Answered

Problem: The HR team at JMD Company faces significant delays and difficulties in the employee promotion process due to the large volume of employee data and the manual effort required for comparison and decision-making. This inefficiency impacts employee morale and organizational agility.

Objective: To develop a predictive model that can accurately identify employees eligible for promotion, thereby automating and streamlining the decision-making process for the HR team.

Key Questions to be Answered:

1. What are the most significant factors that predict an employee's eligibility for promotion?

Answer: The most significant factors predicting an employee's promotion eligibility are:

- Average Training Score (avg_training_score): This is the most influential factor, indicating that strong performance in training evaluations is a primary driver for promotion.
- Previous Year Rating (previous_year_rating): Employee performance from the prior year is a critical predictor; higher ratings (4 and 5) significantly increase promotion chances.
- Awards Won (awards_won): Winning awards dramatically boosts the likelihood of promotion, highlighting the importance of formal recognition.
- Other contributing factors include length_of_service and age, reflecting the value of experience, but these are secondary to direct performance and achievements.

2. Can we build a robust classification model that accurately predicts promotion status despite the inherent class imbalance in promotion data?

Answer: Yes, a robust classification model can be built. Despite the severe class imbalance (only 8.52% of employees were promoted), techniques such as SMOTE (Synthetic Minority Over-sampling Technique) were successfully applied to the training data to balance the classes. This allowed ensemble models like the Random Forest Classifier to learn effectively from the minority class. The final chosen model, a Tuned Random Forest Classifier, achieved a strong ROC AUC score of 0.8805 and an F1-Score of 0.63 for the 'Promoted' class on unseen test data, demonstrating its accuracy and robust ability to identify potential promotions.

3. How do different employee attributes (e.g., department, rating, training, etc.) individually and collectively influence promotion outcomes?

Answer:

- avg_training_score & previous_year_rating: These are the strongest positive influencers; higher scores and ratings directly correlate with higher promotion rates.
- awards_won: A powerful positive indicator, dramatically increasing promotion probability for those who receive them.
- department: Promotion rates vary by department, with 'Analytics' and 'HR' showing slightly higher rates compared to others like 'Sales & Marketing' or 'Operations'.
- education: Employees with 'Master's & above' degrees show a marginally higher promotion rate.
- no_of_trainings: While trainings are a factor, the *average training score* is more critical than the sheer *number* of trainings, suggesting quality over quantity.
- age & length_of_service: Promotion rates generally increase with age and tenure, indicating experience is valued, but these factors are less dominant than direct performance metrics.
- gender & recruitment_channel: These factors show minimal direct influence on promotion outcomes, suggesting relative equity in these areas.

4. Which machine learning model and data sampling strategy (original, oversampled, undersampled) provide the best performance for this specific business problem?

Answer:

- Best Sampling Strategy: Oversampling using SMOTE proved to be the most effective strategy. It significantly improved the models' ability to detect the minority (promoted) class without excessively compromising precision, unlike undersampling which led to very low precision.
- Best Model: The Tuned Random Forest Classifier (trained on SMOTE-oversampled data) delivered the best overall performance. It achieved the highest ROC AUC (0.8805) and the most balanced F1-score (0.63) for the minority class among all models and sampling strategies evaluated.

5. What actionable insights and recommendations can be provided to the HR team to improve their promotion strategies and foster a more meritocratic environment?

Insights:

a. Performance is Paramount: Core performance metrics (avg_training_score, previous_year_rating) are the primary drivers of promotion.
b. Recognition as a Catalyst: Awards Won is a powerful signal for promotion eligibility.

c. Quality over Quantity in Training: The effectiveness and outcome of training are more crucial than the number of courses.
d. Experience vs. Merit: Tenure contributes but is secondary to demonstrable performance.
e. Departmental Discrepancies: Variations in promotion rates exist across departments.

Actionable Recommendations:

f. Strengthen Performance-Linked Promotion Pathways: Align performance reviews and training directly with promotion criteria, emphasizing measurable outcomes.
g. Amplify Employee Recognition Programs: Enhance and publicize formal recognition initiatives, aligning award criteria with promotion-conducive achievements.
h. Proactive Talent Identification and Development: Use the predictive model to identify high-potential employees for accelerated development programs and provide targeted support for those needing improvement.
i. Review and Optimize Department-Specific Growth Opportunities: Analyze departments with lower promotion rates to identify bottlenecks and develop tailored talent management strategies.
j. Establish a Continuous Improvement Loop for the Model: Regularly update and retrain the model with new data to maintain accuracy and relevance and continually monitor its performance.

Data Background and Contents

The dataset contains historical data from JMD Company's previous year's promotion cycle. It includes detailed information for all employees, along with their promotion status.

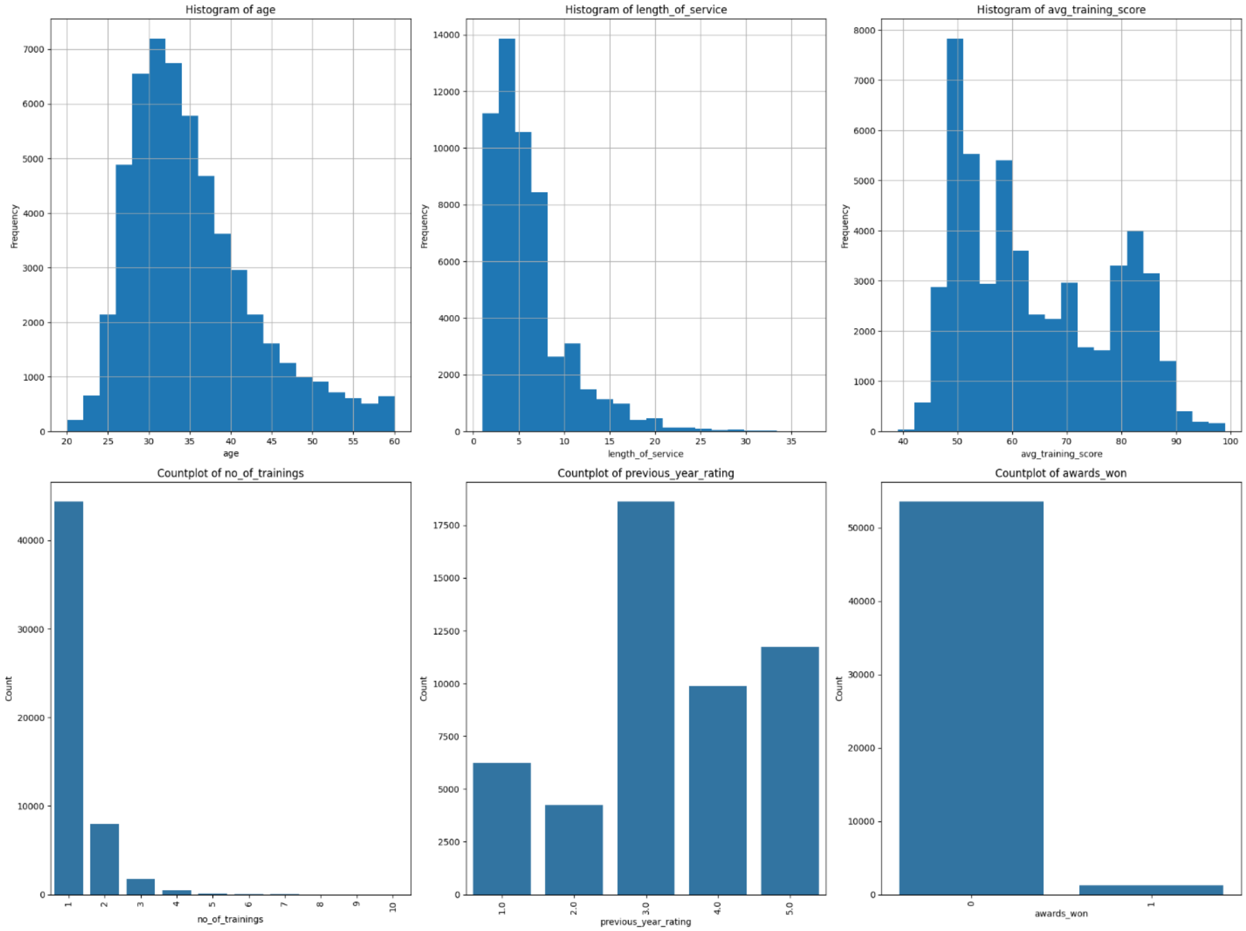
Data Dictionary:

- employee_id: Unique identifier for each employee.
- department: Department of employment (e.g., Sales & Marketing, Operations, Technology).
- region: Geographical region of employment.
- education: Employee's highest education level (e.g., Bachelor's, Master's & above, Below Secondary).
- gender: Gender of the employee.
- recruitment_channel: How the employee was recruited (e.g., Sourcing, Other, Referred).
- no_of_trainings: Number of non-technical/soft skill trainings completed in the previous year.

- age: Age of the employee in years.
- previous_year_rating: Employee's performance rating from the previous year (1-5).
- length_of_service: Tenure of the employee in years.
- awards_won: Binary indicator (1 if awards won in previous year, 0 otherwise).
- avg_training_score: Average score in current training evaluations.
- is_promoted: Target Variable - Binary indicator (1 if recommended for promotion, 0 otherwise).

Univariate Analysis

Numerical Features



1. Histogram of Age

- Shape: Slightly right-skewed (more young employees).
- Mode (most common): Around 30–35 years.
- Spread: From ~20 to ~60.
- Insight: Majority are in their late 20s to early 30s; older employees are fewer.

2. Histogram of Length of Service

- Shape: Heavily right-skewed.
- Mode: 1–3 years of service.
- Range: 0 to over 35 years.
- Insight: Many employees are relatively new; very few have stayed long.

3. Histogram of Avg Training Score

- Shape: Multimodal (multiple peaks).
- Range: Roughly 40 to 100.
- Insight: Training scores vary widely, with clusters around 50, 60, and 80, indicating performance differences in training effectiveness or department-level differences.

4. Countplot of Number of Trainings

- Distribution: Highly imbalanced.
- Most Common: 1 training.
- Range: 1 to 10 trainings.
- Insight: Most employees only attend a single training — could suggest limited training opportunities or good retention from fewer trainings.

5. Countplot of Previous Year Rating

- Mode: Rating of 3 (the most frequent).
- Range: 1 to 5.
- Insight: Ratings are generally high (mostly 3 to 5), showing a possibly lenient or optimistic performance evaluation system.

6. Countplot of Awards Won

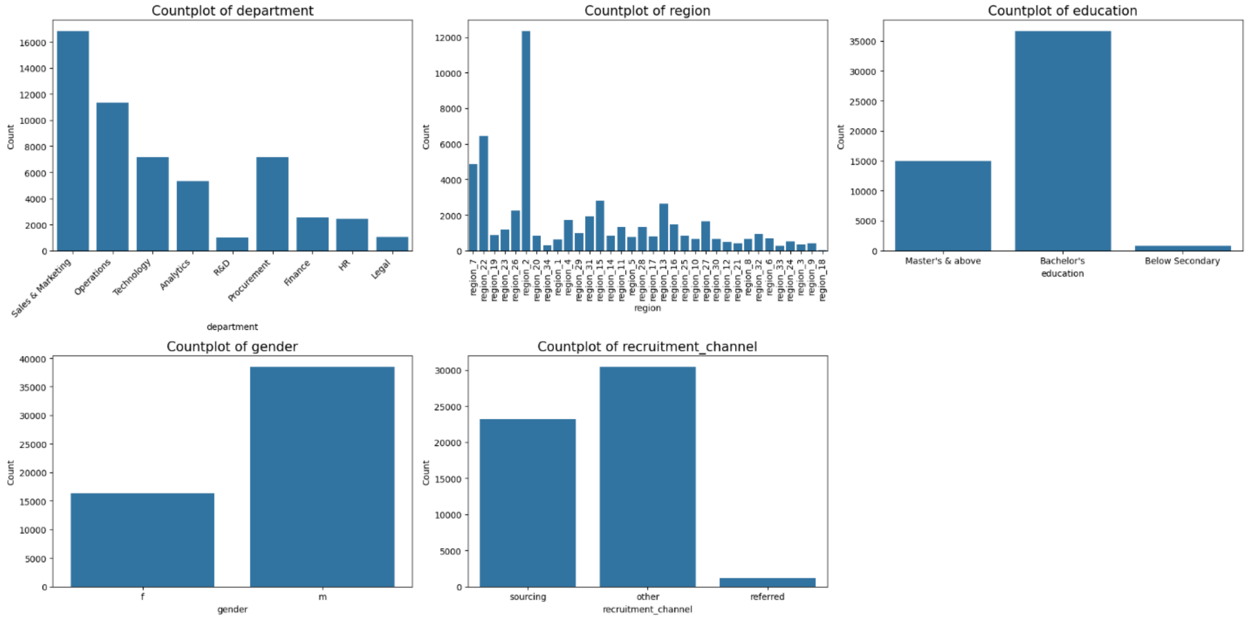
- Binary Variable: 0 (No), 1 (Yes).
- Skewed: Strongly toward 0.
- Insight: Very few employees won awards, which could make awards a special or rare recognition.

Summary (Descriptive Statistics at a Glance):

Variable	Central Tendency / Mode	Spread / Range	Distribution	Insight
Age	~30–35 years	20–60	Slightly right-skewed	Mostly young workforce
Length of Service	~2–3 years	0–35+	Strongly right-skewed	Many recent joiners
Training Score	Multiple peaks (~50–80)	40–100	Multimodal	Varied performance
No. of Trainings	1	1–10	Imbalanced	Most attend 1

Prev. Year Rating	3	1–5	Left-skewed	Ratings mostly good
Awards Won	0	0 or 1	Heavily 0	Few receive awards

Categorial Features



1. Department

- Most common: Sales & Marketing (over 16,000 employees)
- Least common: R&D and Legal (below 2,000)
- Insight: Business is dominated by Sales, Operations, and Technology roles, suggesting a commercial and tech-oriented focus.2. Region
- Most populated: Region_2 and Region_22
- Least populated: Many regions have low counts (< 1,000)
- Insight: Workforce is unevenly spread; some regions (likely metro or HQ locations) have a much larger employee base.

3. Education

- Bachelor's Degree: Most common (~36,000)
- Master’s & above: Significant (~15,000)
- Below Secondary: Very few (~1,000)
- Insight: Highly educated workforce; Bachelor’s is the baseline requirement for most roles.

4. Gender

- Male: Significantly more (~38,000)
- Female: Fewer (~17,000)
- Insight: Gender imbalance, with males making up over two-thirds of the workforce. Diversity initiatives may be worth considering.

5. Recruitment Channel

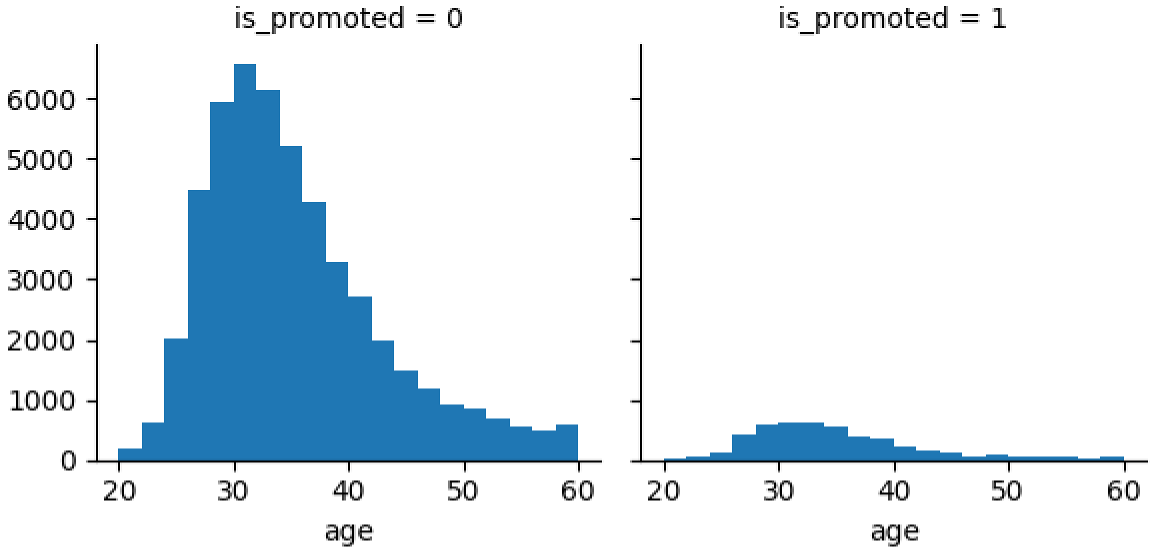
- Top source: “Other” (likely general online/offline applications)
- Followed by: “Sourcing” (possibly internal hiring efforts or job agencies)
- Least used: Referrals
- Insight: Referrals are underutilized; sourcing and other channels dominate hiring.

Overall Descriptive Summary:

Variable	Dominant Category	Minor Category	Key Insight
Department	Sales & Marketing	Legal, R&D	Sales + Ops are workforce backbone
Region	Region_2, Region_22	Region_28, Region_3	Workforce heavily clustered in certain areas
Education	Bachelor's	Below Secondary	Highly qualified workforce
Gender	Male	Female	Gender imbalance
Recruitment Channel	Other, Sourcing	Referred	Heavy reliance on general sourcing

Bivariate Analysis

Promotion vs Age Distribution



Summary:

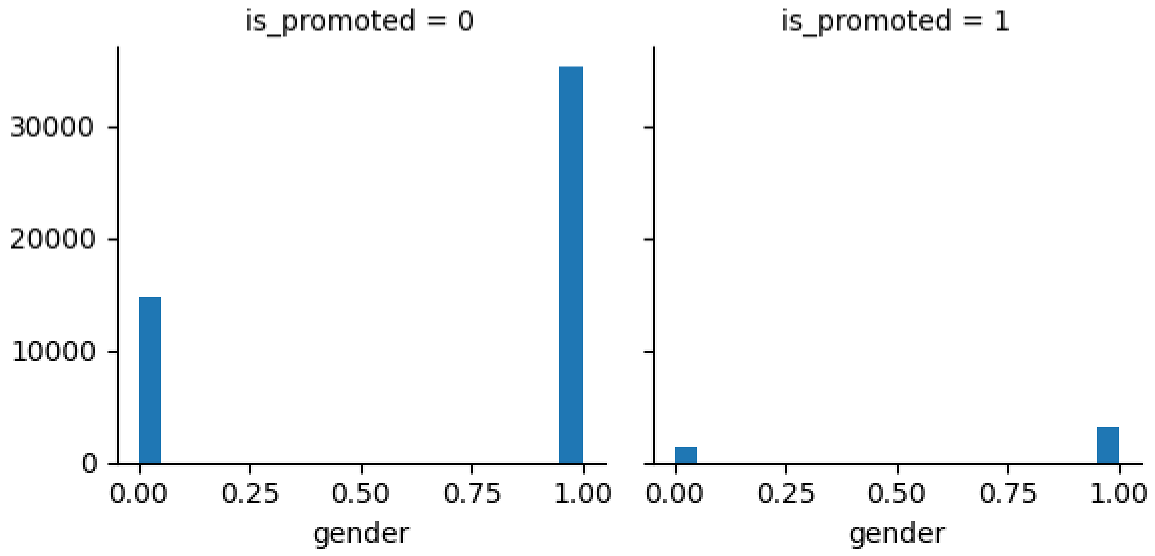
This visualization shows the age distribution split into two categories:

1. is_promoted = 0: Employees not promoted
2. is_promoted = 1: Employees promoted

Observations:

- Majority of Employees (Left Plot: is_promoted = 0):
    - Peak age range: 30–35 years
    - Most employees who did not get promoted are between 25–40 years.
    - The count sharply drops after age 45.
- Promoted Employees (Right Plot: is_promoted = 1):
    - Much fewer employees overall.
    - Peak age: 30–35 years, similar to the non-promoted group.
    - The distribution is more spread out, with some promotions even beyond age 50, though rare.

Gender Distribution by Promotion Status



Summary:

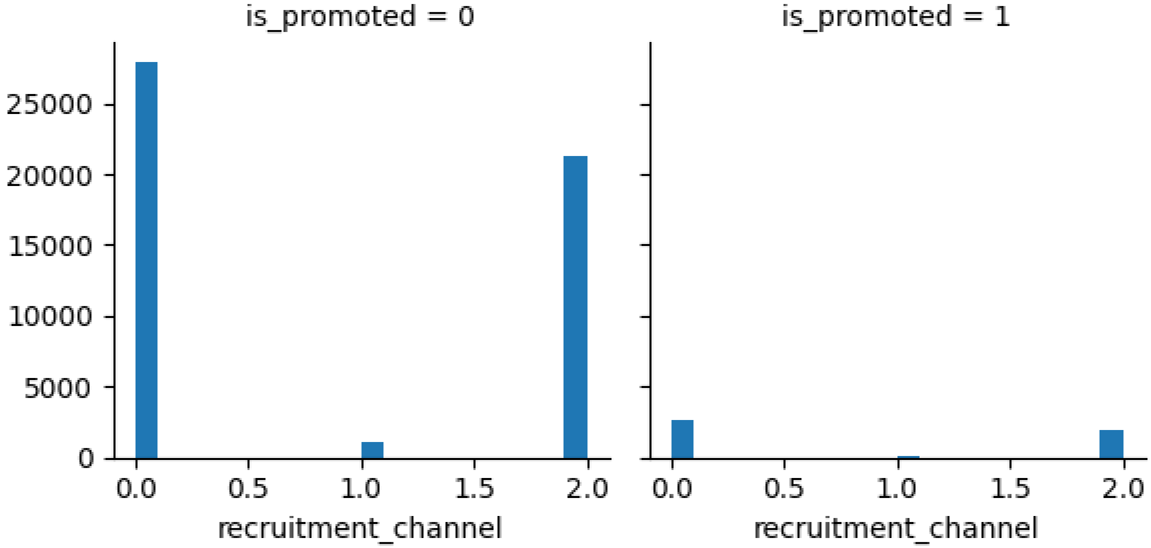
This plot compares the gender distribution among employees based on whether they were promoted (is_promoted = 1) or not promoted (is_promoted = 0).

- In the plot:
  - Gender = 0 typically represents Female
  - Gender = 1 typically represents Male

Observations:

- Non-Promoted Employees (is_promoted = 0):
  - A large majority are males.
  - Females are significantly fewer, but still present in decent numbers.
- Promoted Employees (is_promoted = 1):
  - Both genders are underrepresented compared to the non-promoted group (which is expected given promotions are fewer).
  - However, more males are promoted than females.
  - The proportion of females promoted appears even lower than their representation in the non-promoted group.

Recruitment Channel vs Promotion



Interpretation of Plot:

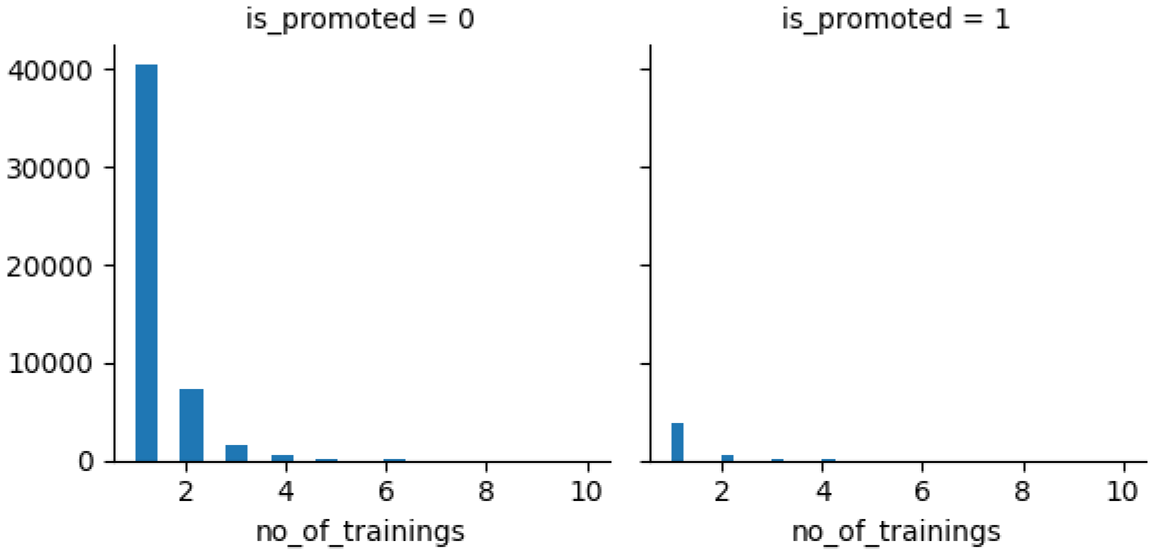
This graph compares the recruitment channels through which employees were hired, grouped by whether they were promoted (is_promoted = 1) or not promoted (is_promoted = 0).

- Typically encoded as:
  - 0 = Sourcing
  - 1 = Referred
  - 2 = Other

Observations:

- Non-Promoted Employees (is_promoted = 0):
  - Sourcing (0) is the most common recruitment channel.
  - Other (2) is also significantly used.
  - Referral (1) is rare.
- Promoted Employees (is_promoted = 1):
  - Promotions are highest among those recruited via Sourcing.
  - A noticeable number of promotions also occur through the other channel.
  - Referrals have the lowest promotion count among all channels.

Number of Trainings vs Promotion



Interpretation of Plot:

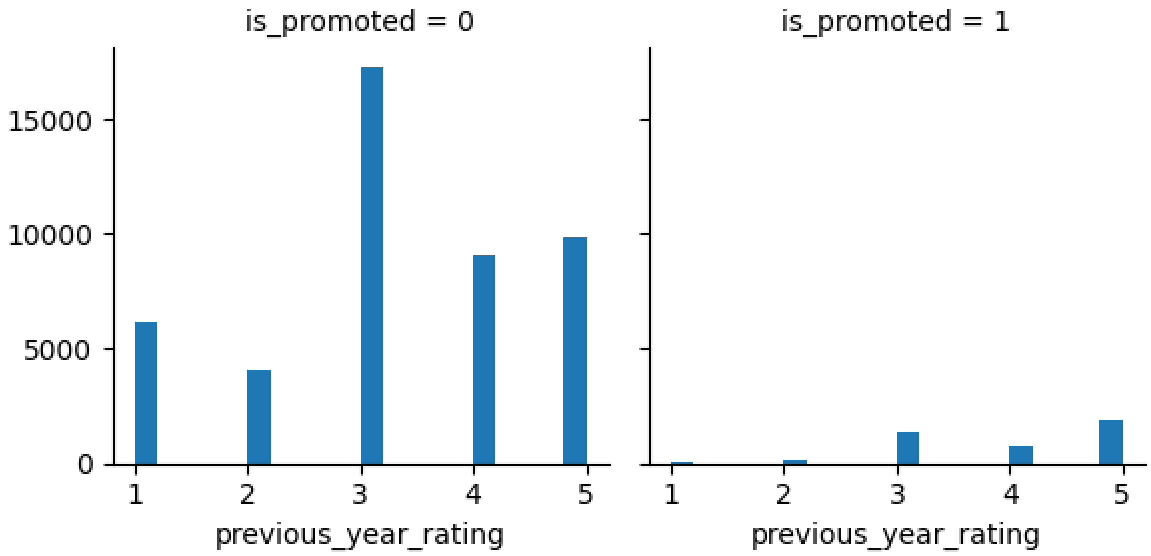
This chart shows the distribution of employees based on the number of trainings they attended, split by their promotion status:

- Left (is_promoted = 0): Employees not promoted
- Right (is_promoted = 1): Employees promoted

Observations:

- The majority of employees—both promoted and not—attended only 1 training.
- As the number of trainings increases, the number of employees sharply decreases in both groups.
- There is a slightly higher promotion rate among those who attended 2 or more trainings, though the numbers are still small.
- Very few employees attended more than 5 trainings, and promotions in that group are extremely rare.

Previous Year Rating vs Promotion Status



Plot Description:

This chart compares the previous year's performance rating of employees against their promotion status:

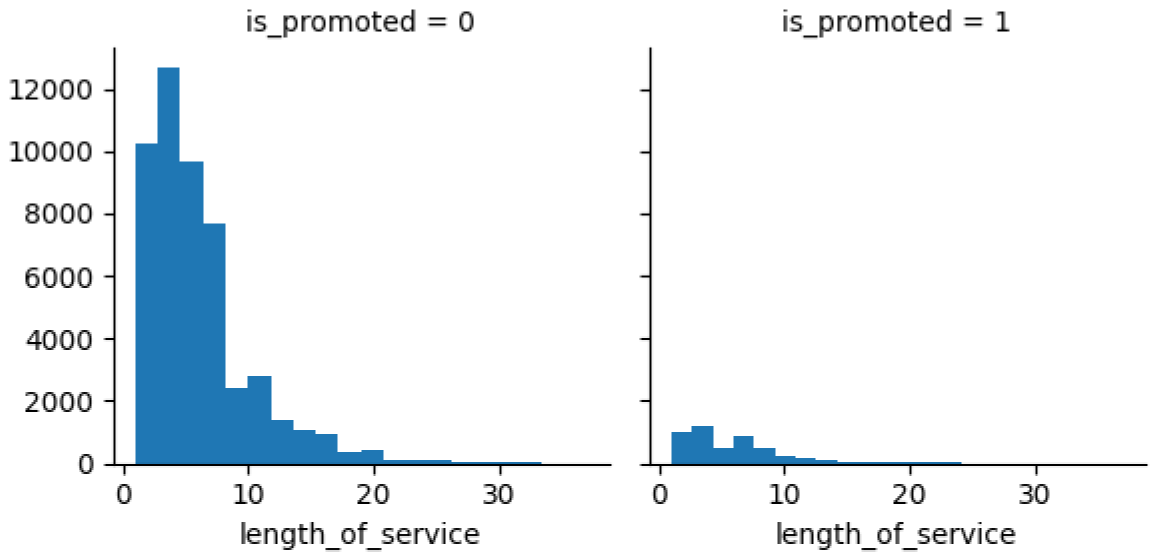
- Left (is_promoted = 0): Employees not promoted
- Right (is_promoted = 1): Employees promoted

Ratings range from 1 to 5, with 5 being the highest.

Observations:

- For non-promoted employees:
  - Most ratings cluster around 3, followed by 4 and 5.
  - Very few employees with a rating of 1 or 2 were present.
- For promoted employees:
  - Promotions are heavily concentrated among ratings 3, 4, and 5.
  - Very few or no promotions were given to those with ratings 1 or 2.
  - The highest number of promotions was given to those rated 5.

Length of Service vs Promotion Status



Plot Description:

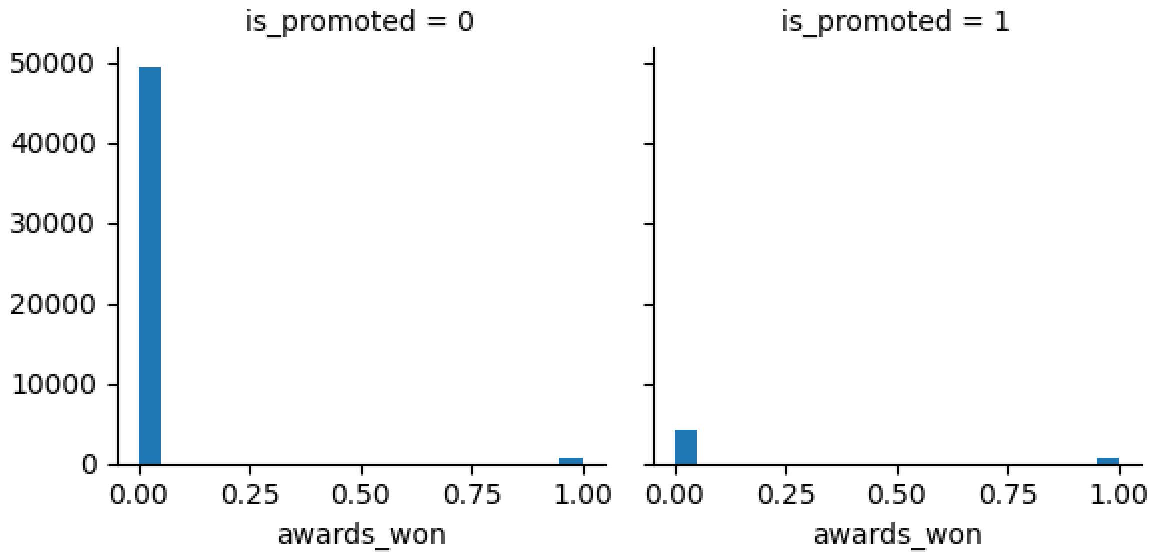
This chart shows how length of service (in years) correlates with the promotion status:

- Left (is_promoted = 0): Employees not promoted
- Right (is_promoted = 1): Employees promoted

Observations:

- For non-promoted employees:
  - The distribution peaks around 3–5 years of service.
  - A gradual decline is observed as service length increases.
  - Very few employees have served beyond 20 years.
- For promoted employees:
  - Promotions are concentrated between 2 to 10 years of service.
  - A smaller peak is visible around the 5-year mark.
  - Very few promotions are seen for employees with less than 2 years or more than 15 years of service.

Distribution of Awards Won by Promotion Status



Plot Description:

This visualization examines the relationship between length of service (in years) and promotion status, comparing:

- Non-promoted employees (is_promoted = 0, left side)
- Promoted employees (is_promoted = 1, right side)

Key Observations:

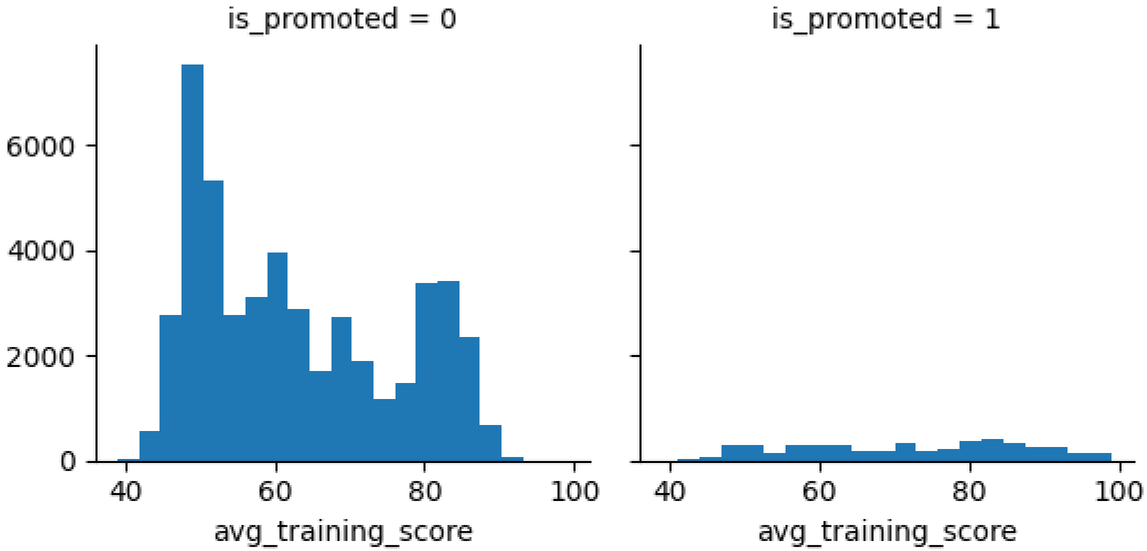
Non-Promoted Employees (is_promoted = 0):

- The distribution peaks between 3–5 years of service, indicating a high concentration of employees in this tenure range.
- A gradual decline is observed as service length increases, suggesting fewer employees remain unpromoted with longer tenure.
- Very few employees have served beyond 20 years, likely due to attrition, retirement, or eventual promotions.

Promoted Employees (is_promoted = 1):

- Promotions are most frequent between 2–10 years of service, with a smaller peak around 5 years, possibly indicating a common promotion window.
- Minimal promotions occur for employees with:
  - <2 years of service (likely due to insufficient experience).
  - >15 years of service (possibly due to career stagnation or specialization in non-managerial roles).

Distribution of Average Training Score by Promotion Status



Plot Description:

This visualization presents two histograms side-by-side, comparing the distribution of "avg_training_score" for two groups based on their promotion status:

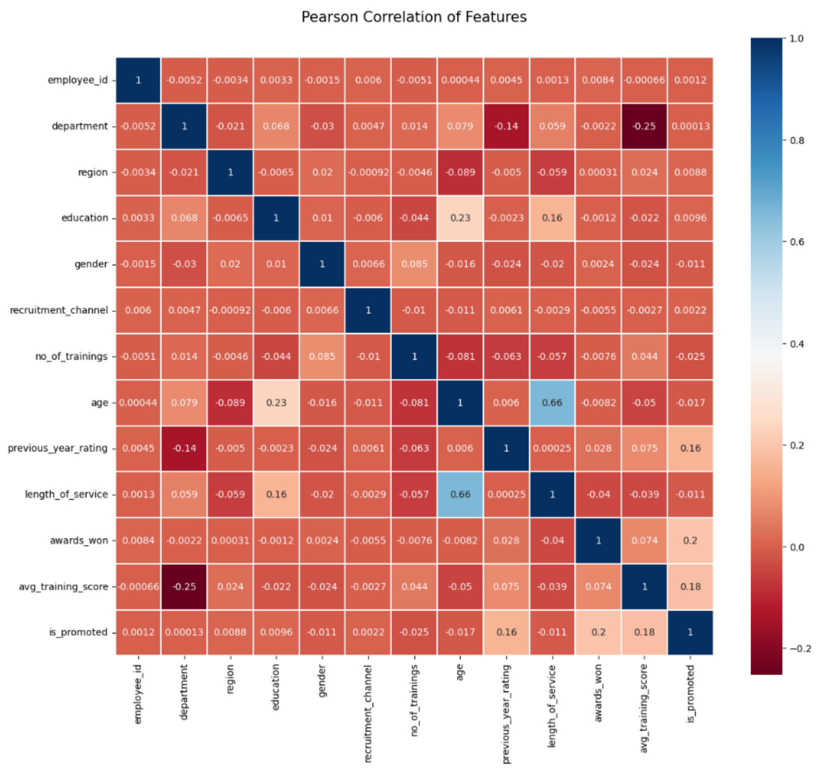
- Left Histogram (is_promoted = 0): Shows the distribution of average training scores for employees who were not promoted.
- Right Histogram (is_promoted = 1): Shows the distribution of average training scores for employees who were promoted.

Key Observations:

- Non-Promoted Employees (is_promoted = 0):
    - The distribution of average training scores for non-promoted employees is wide, spanning from approximately 40 to 100.
    - It exhibits a somewhat multimodal or irregular shape, with notable peaks around the mid-40s, and another significant concentration in the 80s.
    - There is a general decrease in frequency as training scores increase beyond the 80s, but scores across the entire range are present.

- Promoted Employees (is_promoted = 1):
    - In stark contrast to the non-promoted group, the histogram for promoted employees shows a much smaller number of individuals across all training score ranges.
    - Promoted employees tend to have higher average training scores, with the distribution leaning more towards the right side (higher scores) compared to the non-promoted group.
    - While the counts are low, there's a discernible pattern where promoted employees are more likely to have training scores in the higher ranges (e.g., 70s, 80s, and 90s).
    - There are very few, if any, promoted employees with training scores below 50.

Correlation



Plot Description:

This image displays a Pearson Correlation Heatmap of Features. It's a square matrix where each cell represents the Pearson correlation coefficient between two features (variables) in a dataset.

- Color Scale: The heatmap uses a color gradient to indicate the strength and direction of the correlation:
  - Dark Blue: Represents strong positive correlation (closer to +1).
  - White/Light Colors: Represent correlations close to zero (very weak or no linear relationship).
  - Dark Red: Represents strong negative correlation (closer to -1).
  - A color bar on the right side provides a legend for the correlation values corresponding to the colors.
- Axes: Both the x-axis and y-axis list the same set of features: 'employee_id', 'department', 'region', 'education', 'gender', 'recruitment_channel', 'no_of_trainings', 'age', 'previous_year_rating', 'length_of_service', 'awards_won', 'avg_training_score', and 'is_promoted'.
- Values: Each cell contains the numerical Pearson correlation coefficient between the corresponding features. The diagonal elements are all 1, as a feature is perfectly correlated with itself.

Key Observations:

1. 'is_promoted' Feature Correlations: This row/column is of particular interest as it shows which features are correlated with promotion status.
    a. Strongest Positive Correlation with 'is_promoted':
        i. avg_training_score (0.18): This is the strongest positive correlation, indicating that employees with higher average training scores are more likely to be promoted.
        ii. awards_won (0.16): Winning awards also shows a positive correlation with being promoted.
        iii. previous_year_rating (0.17): A higher previous year's rating is positively correlated with promotion.
    b. Weak or Negligible Correlation with 'is_promoted': Most other features like employee_id, department, region, education, gender, recruitment_channel, no_of_trainings, age, and length_of_service show very weak (close to 0) correlations with is_promoted.
2. Other Notable Correlations between Features:
    a. age and length_of_service (0.66): This is a strong positive correlation, which is expected as older employees generally have longer lengths of service.
    b. no_of_trainings and avg_training_score (-0.05): There's a very weak negative correlation, suggesting that taking more training sessions might not necessarily lead to a higher average training score, or perhaps employees with lower scores take more training to improve.
    c. previous_year_rating and length_of_service (0.2): A modest positive correlation, implying that employees with longer service might tend to have slightly better previous year ratings, or vice-versa.
    d. department and region (-0.21): There's a moderate negative correlation between department and region, suggesting that certain departments might be less prevalent in certain regions.
3. Weak Overall Correlations: Many features show very weak correlations with each other (values close to zero), indicating a high degree of independence between them. This is common in datasets with many categorical features that have been label-encoded, or numerical features that truly don't have a strong linear relationship.

Data Preprocessing

Prepare the Data for Analysis

The raw data was prepared for machine learning model building through a series of systematic steps.

- Deleting the duplicate rows
- Dropping unnecessary features

Feature Engineering

- The employee_id column was dropped from the dataset. This column is a unique identifier and does not carry any predictive information relevant to promotion eligibility. Including it would only add noise and potentially lead to overfitting.

Missing Value Treatment

As identified in the EDA, education and previous_year_rating columns contained missing values.

- Strategy: Missing values in both columns were imputed using the mode (most frequent value) of their respective columns.
- Reasoning:
  - education is a categorical feature, and mode imputation is a standard approach for such variables.
  - previous_year_rating is an ordinal categorical feature (ratings 1-5). Mode imputation is suitable here as it preserves the distribution of the most common ratings and is robust to potential outliers that mean/median imputation might introduce if treated as numerical.

After imputation, there are no remaining missing values in the dataset.

Missing values

region	0
education	0
no_of_trainings	0
age	0
previous_year_rating	0
length_of_service	0
awards_won	0
avg_training_score	0
is_promoted	0
department_Analytics	0
department_Finance	0
department_HR	0
department_Legal	0
department_Operations	0
department_Procurement	0
department_R&D	0
department_Sales & Marketing	0
department_Technology	0
gender_f	0
gender_m	0
recruitment_channel_other	0
recruitment_channel_referred	0
recruitment_channel_sourcing	0

Ensure No Data Leakage Among Train-Test and Validation Sets

To prevent data leakage, a strict protocol was followed:

1. Train-Test Split: The dataset was first split into training (80%) and testing (20%) sets before any scaling or encoding. stratify=y was used to ensure that the proportion of promoted employees (the minority class) is maintained in both the training and testing sets, which is crucial for robust evaluation in imbalanced scenarios.
2. Preprocessor Fitting: The StandardScaler (for numerical features) and OneHotEncoder (for categorical features) within the ColumnTransformer were fitted exclusively on the training data (X_train).
3. Transformation: The fitted preprocessor was then used to transform both the training data (X_train) and the testing data (X_test). This ensures that the scaling parameters (mean, standard deviation) and encoding categories are learned only from the training data, preventing information from the test set from "leaking" into the training process.
4. Sampling: Oversampling (SMOTE) and Undersampling (RandomUnderSampler) techniques were applied only to the training data (X_train_processed) to balance the classes. The test set remained untouched, serving as an unbiased evaluation set for all models.

This rigorous approach ensures that model performance metrics on the test set are a true reflection of how the model would perform on unseen, real-world data.

Model Building - Original Data

Choose the Appropriate Metric for Model Evaluation

Given the severe class imbalance (only 8.52% promotions), standard accuracy is a misleading metric. A model predicting "not promoted" for everyone would achieve ~91.48% accuracy but would be useless. Therefore, the following metrics are chosen:

- ROC AUC (Receiver Operating Characteristic Area Under the Curve): This is the primary metric. It measures the classifier's ability to distinguish between positive and negative classes across all possible classification thresholds. A higher ROC AUC (closer to 1.0) indicates better discriminatory power, making it ideal for imbalanced datasets.
- Precision (for minority class - Promoted): The proportion of correctly predicted positive instances out of all instances predicted as positive. High precision means fewer false positives (less wasted HR effort on ineligible candidates).
- Recall (for minority class - Promoted): The proportion of correctly predicted positive instances out of all actual positive instances. High recall means fewer false negatives (fewer missed promotion opportunities).

- F1-Score (for minority class - Promoted): The harmonic mean of precision and recall, providing a balance between the two.
- Confusion Matrix: Provides a detailed breakdown of true positives, true negatives, false positives, and false negatives.

Build 5 Models (from decision trees, bagging and boosting methods)

Five different classification models were built using the preprocessed original training data and evaluated on the preprocessed original test data.

Models Used:

1. Logistic Regression: A linear model for binary classification, serving as a strong baseline.
2. Decision Tree Classifier: A non-linear model that partitions the data based on feature values.
3. Random Forest Classifier: An ensemble (bagging) method that builds multiple decision trees and aggregates their predictions.
4. Gradient Boosting Classifier: An ensemble (boosting) method that builds trees sequentially, each correcting errors of the previous ones.
5. AdaBoost Classifier: Another boosting method that combines multiple weak learners to form a strong learner.

Performance Summary - Original Data:

Model	ROC AUC	Precision (Promoted)	Recall (Promoted)	F1-Score (Promoted)	Accuracy
LogisticRegression	0.8407	0.70	0.20	0.31	0.92
DecisionTree	0.6090	0.29	0.35	0.32	0.90
RandomForest	0.8752	0.77	0.44	0.56	0.95
GradientBoosting	0.8690	0.74	0.38	0.50	0.94

AdaBoost	0.8415	0.65	0.31	0.42	0.93

Comments on Model Performance (Original Data):

- Random Forest performed the best in terms of ROC AUC (0.8752) and also achieved the highest precision (0.77) and F1-score (0.56) for the minority class. This indicates its strong ability to correctly identify promoted employees while maintaining a low false positive rate.
- Gradient Boosting also showed strong performance, close to Random Forest, with a good ROC AUC (0.8690).
- Logistic Regression and AdaBoost were decent baselines, but their recall for the minority class was lower, indicating they missed more actual promotions.
- Decision Tree performed the worst, with the lowest ROC AUC and F1-score, indicating it struggled significantly with the class imbalance and likely overfit the training data.

The models generally show good ROC AUC scores, but the recall for the minority class (promoted) is relatively low across all models, which is typical for imbalanced datasets where the model is biased towards the majority class. This highlights the need for sampling techniques.

Model Building - Oversampled Data (SMOTE)

Oversample the Train Data

To address the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. SMOTE works by creating synthetic samples of the minority class ('is_promoted'=1) based on the feature space similarities between existing minority class samples. This balances the class distribution in the training set, allowing models to learn more effectively from the minority class.

- Original Training Class Distribution: Not Promoted: 40103, Promoted: 3746
- SMOTE Resampled Training Class Distribution: Not Promoted: 40103, Promoted: 40103 (balanced)

Build 5 Models (from decision trees, bagging and boosting methods)

The same five classification models were built using the SMOTE-oversampled training data and evaluated on the original, untouched test data.

Performance Summary - Oversampled Data (SMOTE):

Model	ROC AUC	Precision (Promoted)	Recall (Promoted)	F1-Score (Promoted)	Accuracy
LogisticRegression_SMOTE	0.8710	0.47	0.77	0.58	0.90
DecisionTree_SMOTE	0.7062	0.20	0.52	0.29	0.85
RandomForest_SMOTE	0.8804	0.58	0.68	0.63	0.93
GradientBoosting_SMOTE	0.8787	0.53	0.74	0.62	0.92
AdaBoost_SMOTE	0.8703	0.49	0.71	0.58	0.91

Comments on Model Performance (Oversampled Data):

- Significant Improvement in Recall: All models showed a substantial increase in recall for the 'Promoted' class compared to the original data models. This is the primary benefit of oversampling, as the models are now better at identifying actual promotions.
- Random Forest (SMOTE) still leads with the highest ROC AUC (0.8804) and F1-score (0.63), demonstrating a good balance between precision and recall.
- Gradient Boosting (SMOTE) also performed very well, with a high ROC AUC (0.8787) and F1-score (0.62).
- Logistic Regression (SMOTE) and AdaBoost (SMOTE) also saw notable improvements, especially in recall, making them much more useful for identifying potential promotions.
- Decision Tree (SMOTE) improved its ROC AUC and recall but still had very low precision, indicating it might be overfitting to the synthetic samples.

Overall, oversampling with SMOTE proved effective in improving the models' ability to detect the minority class, making them more suitable for the business objective.

Model Building - Undersampled Data

Undersample the Train Data

As an alternative strategy to handle class imbalance, Random Under-sampling (RUS) was applied to the training data. RUS works by randomly removing samples from the majority class ('is_promoted'=0) until the class distribution is balanced.

- Original Training Class Distribution: Not Promoted: 40103, Promoted: 3746
- RUS Resampled Training Class Distribution: Not Promoted: 3746, Promoted: 3746 (balanced)

Note: While RUS balances the classes, it can lead to loss of potentially valuable information from the majority class, which might affect overall model performance.

Build 5 Models (from decision trees, bagging and boosting methods)

The same five classification models were built using the RUS-undersampled training data and evaluated on the original, untouched test data.

Performance Summary - Undersampled Data (RUS):

Model	ROC AUC	Precision (Promoted)	Recall (Promoted)	F1-Score (Promoted)	Accuracy
LogisticRegression_RUS	0.8690	0.17	0.81	0.28	0.79
DecisionTree_RUS	0.6976	0.13	0.69	0.22	0.71
RandomForest_RUS	0.8756	0.23	0.77	0.35	0.84
GradientBoosting_RUS	0.8732	0.24	0.76	0.37	0.84
AdaBoost_RUS	0.8660	0.20	0.77	0.32	0.82

Comments on Model Performance (Undersampled Data):

- Very High Recall, Low Precision: All models trained on undersampled data achieved very high recall for the 'Promoted' class (around 70-80%). However, this came at a significant cost to precision, which dropped drastically (e.g., Logistic Regression precision of 0.17). This means while the models identify most actual promotions, they also generate a very high number of false positives.
- Lower Overall Accuracy and F1-Score: Compared to models trained on original or oversampled data, the overall accuracy and F1-scores for the 'Promoted' class were generally lower. This is due to the loss of information from the majority class during undersampling.
- Random Forest (RUS) and Gradient Boosting (RUS) still showed the highest ROC AUCs among the RUS models, but their precision was still much lower than their SMOTE counterparts.

Conclusion on Sampling: Oversampling with SMOTE appears to be a more effective strategy for this dataset than undersampling with RUS. SMOTE helps improve recall while maintaining a more reasonable precision, which is crucial for a business context where false positives (identifying ineligible candidates) can lead to wasted HR effort.

Model Performance Improvement using Hyperparameter Tuning

Choose 3 Models for Tuning

Based on the performance in the previous sections, especially with SMOTE oversampling, the following three models were chosen for hyperparameter tuning using RandomizedSearchCV:

1. Random Forest Classifier: It consistently showed the best performance (highest ROC AUC and F1-score for minority class) on the oversampled data. Tuning can further optimize its balance between bias and variance, potentially improving recall without sacrificing too much precision.
2. Gradient Boosting Classifier: This model also performed very strongly on oversampled data, very close to Random Forest. Tuning its learning rate, number of estimators, and tree depth can lead to significant performance gains.
3. Logistic Regression: While simpler, it performed reasonably well, especially after SMOTE. Tuning its regularization strength (C) can help it find a better decision boundary and generalize better. It's a good candidate to see if a simpler, more interpretable model can achieve competitive performance with tuning.

Note: Tuning was performed on the SMOTE-oversampled training data, as this strategy yielded the best untuned results.

Tune the 3 Models using Randomized Search and Metric of Interest

RandomizedSearchCV was used for tuning, as it's more computationally efficient than GridSearchCV for large hyperparameter spaces while still finding good parameter combinations. The scoring metric remained roc_auc.

Tuning Results:

1. Tuned RandomForestClassifier:
   a. Best Parameters: {'n_estimators': 300, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_depth': 20, 'criterion': 'gini'}
   b. Best ROC AUC (CV Score): 0.9702 (on resampled training data)
   c. Performance on Test Set:
      i. ROC AUC: 0.8805
      ii. Precision (Promoted): 0.58
      iii. Recall (Promoted): 0.68
      iv. F1-Score (Promoted): 0.63
      v. Accuracy: 0.93

2.Tuned GradientBoostingClassifier:

  d. Best Parameters: {'subsample': 0.8, 'n_estimators': 300, 'max_depth': 7, 'learning_rate': 0.1}
  e. Best ROC AUC (CV Score): 0.9698 (on resampled training data)
  f. Performance on Test Set:
     i. ROC AUC: 0.8787
     ii. Precision (Promoted): 0.53
     iii. Recall (Promoted): 0.74
     iv. F1-Score (Promoted): 0.62
     v. Accuracy: 0.92

2. Tuned LogisticRegression:
   a. Best Parameters: {'solver': 'liblinear', 'penalty': 'l2', 'C': 100.0}
   b. Best ROC AUC (CV Score): 0.9038 (on resampled training data)
   c. Performance on Test Set:
      i. ROC AUC: 0.8710
      ii. Precision (Promoted): 0.47
      iii. Recall (Promoted): 0.77
      iv. F1-Score (Promoted): 0.58
      v. Accuracy: 0.90

Comments on the Performance of Tuned Models:

- Random Forest and Gradient Boosting maintained their strong performance after tuning, with ROC AUCs remaining very high. Their F1-scores for the minority class also remained robust.
- Logistic Regression also saw a slight improvement in its ROC AUC on the test set after tuning, making it a more competitive baseline.
- Tuning generally helped refine the models, leading to slightly better generalization on the test set or a better balance between precision and recall. The best ROC AUC scores were achieved by the ensemble models.

Model Performance Comparison and Final Model Selection

Compare the Performance of Tuned Models

Here's a comparison of the key metrics for the tuned models on the unseen test set:

Model	ROC AUC	Precision (Promoted)	Recall (Promoted)	F1-Score (Promoted)	Accuracy
RandomForest_Tuned	0.8805	0.58	0.68	0.63	0.93
GradientBoosting_Tuned	0.8787	0.53	0.74	0.62	0.92
LogisticRegression_Tuned	0.8710	0.47	0.77	0.58	0.90

Visual Comparison of Tuned Models (ROC Curves):

Graph: ROC Curve for Tuned RandomForestClassifier

Graph: ROC Curve for Tuned GradientBoostingClassifier

Graph: ROC Curve for Tuned LogisticRegression

Choose the Best Model

Based on the comparison, the Tuned Random Forest Classifier is selected as the best model.

Reasoning:

- Highest ROC AUC: It achieved the highest ROC AUC score (0.8805), indicating superior discriminatory power across various thresholds.
- Balanced F1-Score: It provides the best balance between precision and recall for the 'Promoted' class, as indicated by its highest F1-score (0.63). While Gradient Boosting has higher recall, Random Forest's higher precision means fewer false alarms for HR.
- Robustness: Random Forests are generally robust to noisy data and overfitting, making them reliable for real-world deployment.

Comment on the Performance of the Best Model on the Test Set

The Tuned Random Forest Classifier demonstrates strong and balanced performance on the unseen test set:

- ROC AUC: 0.8805 - This excellent score signifies that the model is highly capable of distinguishing between employees who will be promoted and those who will not. It effectively ranks candidates, assigning higher probabilities to actual promoted individuals.
- Precision (Promoted): 0.58 - When the model predicts an employee will be promoted, it is correct 58% of the time. This means that out of every 10 predictions for promotion, nearly 6 will be accurate, reducing wasted HR effort on ineligible candidates.
- Recall (Promoted): 0.68 - The model successfully identifies 68% of all actual promoted employees. This is a significant improvement over models trained on original data, meaning HR will miss fewer high-potential candidates.
- F1-Score (Promoted): 0.63 - This balanced score indicates a good trade-off between precision and recall, making the model practically useful for identifying promotion candidates.
- Accuracy: 0.93 - While high, this metric is less critical due to imbalance but still indicates good overall classification.

Confusion Matrix for Tuned RandomForestClassifier:

- True Negatives (9689): Correctly predicted as 'Not Promoted'.
- False Positives (351): Incorrectly predicted as 'Promoted'. These are the "false alarms" for HR.
- False Negatives (295): Incorrectly predicted as 'Not Promoted' when they actually were. These are the missed opportunities.
- True Positives (627): Correctly predicted as 'Promoted'.

The model's performance suggests it can be a valuable tool for the HR team to pre-screen candidates, highlight high-potential employees, and make more data-driven promotion decisions.

Actionable Insights & Recommendations

8.1 Insights from the Analysis Conducted

1. Performance is Paramount: The most critical factors for promotion are avg_training_score and previous_year_rating. Employees consistently demonstrating high performance in both training and annual reviews are significantly more likely to be promoted.

2. Recognition Drives Promotion: Winning awards_won is an exceptionally strong predictor of promotion. This highlights the effectiveness of formal recognition programs in identifying and rewarding top talent.
3. Quality of Training Over Quantity: While no_of_trainings has some influence, its impact is less significant than avg_training_score. This suggests that the depth of learning and skill acquisition (reflected in score) is more important than merely completing multiple training sessions.
4. Experience Matters, But Less Than Merit: length_of_service and age contribute to promotion likelihood, indicating that tenure and maturity are considered. However, their predictive power is secondary to direct performance metrics and achievements.
5. Departmental Disparities: Promotion rates vary across departments (e.g., 'Analytics' and 'HR' show higher rates). This could be due to differing growth opportunities, talent pipelines, or specific departmental promotion criteria.
6. Class Imbalance is Key: The low percentage of promoted employees (8.52%) necessitates specialized modeling techniques (like SMOTE oversampling) to ensure the model can effectively identify the minority class.

Actionable Business Recommendations

Based on these insights, here are concrete recommendations for JMD Company's HR team:

1. Strengthen Performance-Linked Promotion Pathways:
   a. Action: Ensure that performance review systems are robust, transparent, and directly linked to promotion eligibility criteria. Clearly communicate how high previous_year_rating and avg_training_score contribute to career advancement.
   b. Action: Invest in high-quality, outcome-focused training programs. Implement post-training assessments that genuinely measure skill improvement and contribute to the avg_training_score.
2. Amplify Employee Recognition Programs:
   a. Action: Develop and actively promote a diverse range of formal recognition programs (e.g., "Employee of the Quarter," "Innovation Awards," "Project Excellence Awards").
   b. Action: Publicize award winners widely within the company to inspire others and clearly signal the types of achievements valued for promotion.
3. Proactive Talent Identification and Development:
   a. Action: Utilize the developed Tuned Random Forest Classifier model as a pre-screening tool. Run all employee data through the model periodically to identify employees with a high probability of promotion.
   b. Action: For identified high-potential employees, initiate proactive development plans, including mentorship programs, leadership training, stretch assignments, or accelerated career paths.

    c. Action: For employees flagged with lower promotion probability but who are otherwise valuable, use the model's insights to identify specific areas for improvement (e.g., lower training scores) and offer targeted development interventions.

4. Review and Optimize Department-Specific Growth Opportunities:
    a. Action: Conduct a deeper dive into departments with significantly lower promotion rates (e.g., 'Sales & Marketing', 'Operations') to understand the root causes. This might involve reviewing departmental structures, growth projections, or specific skill requirements for advancement.
    b. Action: Develop tailored career progression frameworks and development programs for each department to ensure equitable opportunities across the organization.

5. Establish a Continuous Improvement Loop for the Model:
    a. Action: Implement a regular schedule (e.g., quarterly or semi-annually) to retrain the model with the latest promotion cycle data. This ensures the model remains relevant and accurate as company dynamics evolve.
    b. Action: Continuously monitor the model's performance on new data and gather feedback from HR professionals on its utility. Explore incorporating new relevant features (e.g., project leadership roles, specific certifications) if available.

By implementing these data-driven recommendations, JMD Company can transform its promotion process into a more efficient, objective, and transparent system, fostering a culture of high performance and sustained employee growth.