**Analysis and Prediction of Employee Promotions Using**

**Machine Learning**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Group 1**

1. **1. Anwar Sadath Erumpan**

**2. Akhiljith A S**

**3. Anjana Prasad**

**4. Nadiya K A**

**5. Niranjana V**

**6. Shreejith M Nair**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**July 2023**

**List of Figures**

* **Figure 1 . Bivariate Analysis Barplot 1**
* **Figure 2 . Bivariate Analysis Barplot**
* **Figure 3 . Bivariate Analysis Barplot 3**
* **Figure 4. Gender Distribution in Department**
* **Figure 5.Ratio of Promotion based on Education and Age**
* **Figure 6 .Distribution of Education by Age before Null Value handling**
* **Figure 7. Distribution of Education by Age after Null Value handling**
* **Figure 8. Outlier Detection**
* **Figure 9. Outlier Handling using Log Transformation**
* **Figure 10. Outlier Handling using Log Transformation**
* **Figure 11. Box Plot after outlier handling**

**List of Abbreviations**

* HR - Human resources
* ML - Machine learning
* MNC - Multinational Corporation
* KPI - Key Performance Indicator
* IQR - Inter quartile range
* KNN - K-nearest neighbor
* XGBOOST - Extreme Gradient Boosting
* SMOTE - Synthetic Minority Oversampling Technique
* R&D - Research and Development

**Table of Contents**

1. **Abstract**
2. **Problem definition**
3. **Introduction**
4. **Literature Survey**
5. **Exploratory Data Analysis**
6. **Data Pre- Processing 1**
7. **Data Pre- Processing 2**
8. **Modeling**
9. **Web Hosting**
10. **Result**
11. **Conclusion**
12. **Reference**

**Abstract**

This abstract introduces an HR dataset that can be used to predict employee promotions within an organization based on various factors/features such as employee ID, department, region, education, etc. The dataset contains (54809 observations and 14 Features) historical employee records and promotion outcomes, offering a valuable resource for understanding the factors that influence career advancement. The goal of this dataset analysis is to develop a predictive model using ML techniques to determine the likelihood of an employee being promoted. By leveraging the available features, the model aims to uncover the underlying patterns and relationships that affect promotion decisions.

The insights gained from this analysis can help organizations identify employees with high potential for advancement and enable targeted talent management strategies. Moreover, the model's predictions can assist HR departments in optimizing resource allocation, succession planning, and performance evaluation processes. Ultimately, using this dataset for promotion prediction can contribute to more effective HR decision-making, ensuring deserving employees receive appropriate recognition and opportunities for career growth.

**1. Problem Definition**

**1.1 Overview**

The aim of this project is to develop a data-driven solution for identifying and preparing eligible candidates for promotion within a large MNC. The MNC has nine broad verticals across the organization, and faces challenges in finding the right people for manager positions and below. The current promotion cycle is slow and inefficient, as the final promotions are only announced after the evaluation, which delays the transition to new roles. The company needs help in identifying the eligible candidates at a particular checkpoint, based on multiple attributes related to their past and current performance and demographics. This would enable the company to expedite the promotion cycle and improve employee satisfaction and retention.

**1.2 Problem Statement**

This is a company’s HR dataset. Employees are promoted within the organization on an annual basis. Therefore, using the dataset, we must determine whether an employee has been promoted or not

**2. Introduction**

There are many advantages that come with promotion in the workplace. It comes with an increase in prestige, respect and wages/salary. Promotion in a business environment also comes with increased freedom and agency. It is because of this that many employees at different levels in an organization try to climb the corporate ladder in order to enjoy the various advantages that come with it.Organizations also pay much attention to who is going to get promoted and how that will affect the business. This is because the activities of those in key positions in a business usually affects the welfare of the organization in general. Firms and businesses pay attention to many factors when making decisions on who to promote. For this reason resources are expended by businesses in the process of selection of employees to promote. This project is important for both employers looking to start training in arrears to promote employees. It is also important for employees to determine their strengths and weaknesses, opportunities and threats, as well as, key factors to work on in order to get promoted. Although this is a case study of a given firm, the findings of this study can be applied to various firms.Following are the features in the dataset :

* employee\_id: Unique ID for employee
* department: Department of employee
* region: Region of employment (unordered)
* education: Education Level
* gender: Gender of Employee
* recruitment\_channel: Channel of recruitment for employee
* no\_ of\_ trainings: no of other trainings completed in previous year on soft skills, technical skills etc.
* age: Age of Employee
* previous\_ year\_ rating: Employee Rating for the previous year
* length\_ of\_ service: Length of service in years
* awards\_ won?: if awards won during previous year then 1 else 0
* avg\_ training\_ score: Average score in current training evaluations
* is\_promoted: (Target) Recommended for promotion

**3. Literature Survey**

The prediction of employee promotions using machine learning algorithms is a multidisciplinary domain that integrates machine learning and human resource management. Recent studies and ongoing research have presented algorithms that have the potential to forecast employee promotion based on a range of performance indicators and other professional attributes.

Employee promotion is an important decision for businesses, as it can have a significant impact on employee morale, productivity, and turnover. In order to make informed promotion decisions, businesses need to be able to accurately predict which employees are most likely to be successful in a new role.

**3.1 Evolution of HR analytics**

HR analytics is a relatively new field, with its origins dating back to the early 2000s. In the early days, HR analytics was primarily focused on collecting and storing data on employee performance, turnover, and compensation. However, as technology has advanced, HR analytics has become more sophisticated and is now able to analyze a wider range of data, including employee surveys, social media data, and customer feedback.

Human resource analytics could offer insights related to organizational issues such as turnover of employees and seeking suitable candidates in recruitment. Thus, it is clear that human resource analytics could improve in decision making and it would be a waste of time and storage if the data are not being used for analytics. One of the approaches in human resource analytics is by using machine learning. One of branches in artificial intelligence is machine learning where it could be taught based on past history and no need interruption or help from outside (Dianah et al., 2021).

By using machine learning, human resource management now can apply the prediction model in their business case. For example, machine learning could help in predicting employee attrition, tracking a candidate journey throughout the interview process, predict employee performance and others.

**3.2 HR Analytics and Employee promotion**

HR analytics is the use of data and analytics to improve HR decision-making. HR analytics can be used to track employee performance, identify trends, and make predictions about future events. This information can be used to make better decisions about hiring, training, compensation, and other HR-related matters.

There is a growing body of literature on the use of HR analytics for employee promotion prediction. In a study published in the journal Personnel Psychology, researchers found that HR analytics can be used to predict employee promotion with a high degree of accuracy. The study found that the most important factors for promotion were performance ratings, education level, and experience.

**3.3 Future directions**

Recent advances in machine learning, such as deep learning, could be used to develop more accurate models for employee promotion prediction.

HR analytics can be used to analyze a wide range of data, including employee performance data, training data, and social media data. Using more data sources could lead to more accurate models for employee promotion prediction.It is important to ensure that the models used for employee promotion prediction are fair and equitable, and that they do not discriminate against any protected groups.

Future research could focus on developing models that are more fair and equitable.HR analytics can be used to promote diversity and inclusion in the workplace by identifying and removing barriers to promotion for underrepresented groups. Future research could focus on developing models that promote diversity and inclusion in the workplace.

**3.4 Significance of new search**

There is a lack of research on the fairness and equity of models used for employee promotion prediction. This is an important area of research, as it is essential to ensure that these models do not discriminate against any protected groups. The other area needed for research is the impact of HR analytics on employee motivation and performance. This is an important area of research, as it is essential to ensure that HR analytics is used in a way that is beneficial to employees.

**3.5 Conclusion**

Machine learning algorithms can be used to predict employee promotion with a high degree of accuracy. However, it is important to note that the accuracy of these models can vary depending on the dataset that is used. It is also important to ensure that the models are fair and equitable, and that they do not discriminate against any protected groups.

By using HR analytics, businesses can make more informed promotion decisions that are based on data rather than gut instinct. This can lead to a more effective and efficient workforce, as well as a more satisfied and productive workforce.

**4.Exploratory Data Analysis**

❖ **Load the Data**: The first step is to import the necessary libraries and load the data from a file or a database into a data frame object.

❖ **Data Preparation**: The next step is to check the quality and consistency of the data, such as data types, missing values, outliers, duplicates, etc.

❖ **Identifying Target Feature**: The target feature is the variable that we want to predict or explain using the data. It is important to identify the target feature and separate it from the rest of the features.

❖ **Removing Invalid and Duplicate Rows**: Some rows in the data may be invalid or duplicate, which can affect the accuracy and reliability of the analysis. We need to remove such rows from the data frame using appropriate methods.

❖ **Data Visualization**: Data visualization is a powerful tool to explore and communicate the patterns and relationships in the data.

❖ **Statistical Summary**: We can also use descriptive statistics to summarize the main characteristics of the data, such as mean, median, mode, standard deviation, variance, range, etc. These statistics can help us to compare and contrast different features and identify potential problems or opportunities in the data.

❖ **Understanding the Distribution of Dat**a: The distribution of data refers to how the values of a feature are spread across its range. It can reveal important information about the shape, symmetry, skewness, kurtosis, and modality of the data. We can use graphical methods like histograms or density plots to visualize the distribution of data or numerical methods like skewness or kurtosis coefficients to measure it.

**4.1 Observation of Univariate Analysis**

Key insights from the Employee Distribution Plots:

- The department with the largest number of employees is Sales and Marketing, followed by Operations and Technology.

- Only 8.5% of employees are promoted previous year(Target)

- Region 2 has the highest employee count, accounting for more than a quarter of the total workforce. Region 22 and Region 7 are also among the top regions in terms of employee distribution.

- The majority of the employees have a Bachelor's degree as their highest level of education, while only a small fraction have a Below Secondary.

- There is a gender imbalance in the employee population, with male employees outnumbering female employees.

- More than half of the employees joined the company through the other recruitment channel, indicating a strong external hiring strategy.

- Most of the employees have completed only one training session, suggesting a low investment in employee development and learning.

- The most common previous year rating that was given to the employees was 3.0, which implies a moderate level of performance and satisfaction,following 5.0.

- Only a minority of the employees achieved their KPI above 80%, and an even smaller percentage won an award in the previous year, indicating a low level of recognition and reward for outstanding performance

- The age range of the employees is between 20 and 60 years old, with the majority of them being between 25 and 40 years old.

- The length of service of the employees varies from 1 to 37 years, with the most common duration being between 1 and 5 years.

- The average training score of the employees ranges from 39 to 99.

- Education feature has 2409 null values and Previous year rating has 4124 null values

**4.2 Observation of Bivariate Analysis**

The promotion rate of employees in the last year was influenced by several factors, such as age, training score, and service length. The analysis of the data revealed that

One of the possible factors that influence the promotion criteria is the department of the employees. According to the data, technology had a higher promotion rate than sales and marketing, despite having fewer promotions in absolute numbers. This suggests that the promotion criteria may vary by department, and that technology may have more stringent or specific requirements than sales and marketing.

Based on the factor education level of the employees: The data indicates that employees with a master's degree or above had a higher promotion rate than those with below secondary or bachelor's degrees. This means that the promotion criteria may favor employees with higher education levels, and that having a master's degree or above may give an advantage over other employees.

Based on the gender of the employees: The data reveals that female employees had a higher promotion rate than male employees, despite having fewer promotions in absolute numbers. The female employees might have more opportunities or less competition than male employees.

One of the factors that may influence the promotion criteria is the recruitment channel of the employees. According to the data, employees who were hired from referrals had the highest promotion rate (18.2%), compared to those who were hired from sourcing (8.5%), other channels (8.4%). This suggests that referrals may have an advantage in terms of skills, qualifications, or connections that make them more likely to be promoted.

Another factor that may affect the promotion criteria is the training attendance of the employees. The data shows that employees who attended only one training session had the highest promotion rate (17.6%), while those who attended seven to ten sessions had no promotion. This implies that attending too many training sessions may indicate a lack of confidence, competence, or initiative, which may hinder the promotion prospects of the employees.

A factor that may play a role in the promotion criteria is the performance rating of the employees. The data reveals that employees who had a 5.0 rating had the highest promotion rate (16.4%), while those who had a 1.0 or 2.0 rating had the lowest promotion rate (4.3% and 1.4%, respectively). This indicates that performance rating is a strong predictor of promotion potential, as it reflects the quality and quantity of work done by the employees.

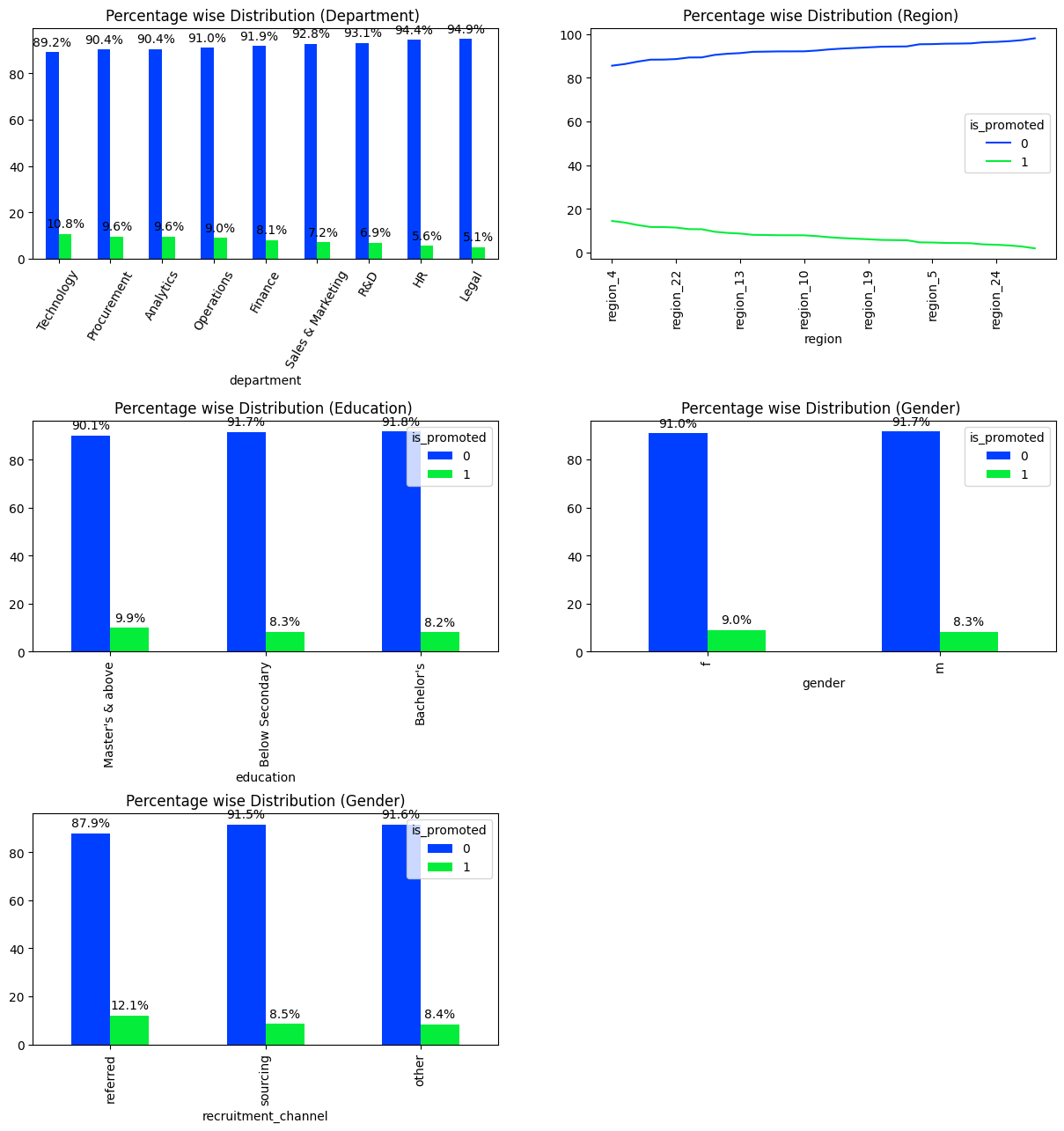
Based on factors that may contribute to the promotion criteria is the achievement of KPIs by the employees: The data demonstrates that employees who met their KPIs had the highest promotion rate (16.9%), while those who did not meet their KPIs had the lowest promotion rate (4.0%). This suggests that meeting KPIs is a key requirement for promotion, as it shows the alignment of the employees' goals with the organization's objectives.

Based on factors that may determine the promotion criteria is the award status of the employees: The data indicates that employees who won an award last year had the highest promotion rate (44.0%), while those who did not win an award had the lowest promotion rate (7.7%). This implies that winning an award is a significant achievement that can boost the promotion chances of the employees, as it recognizes their excellence and innovation.

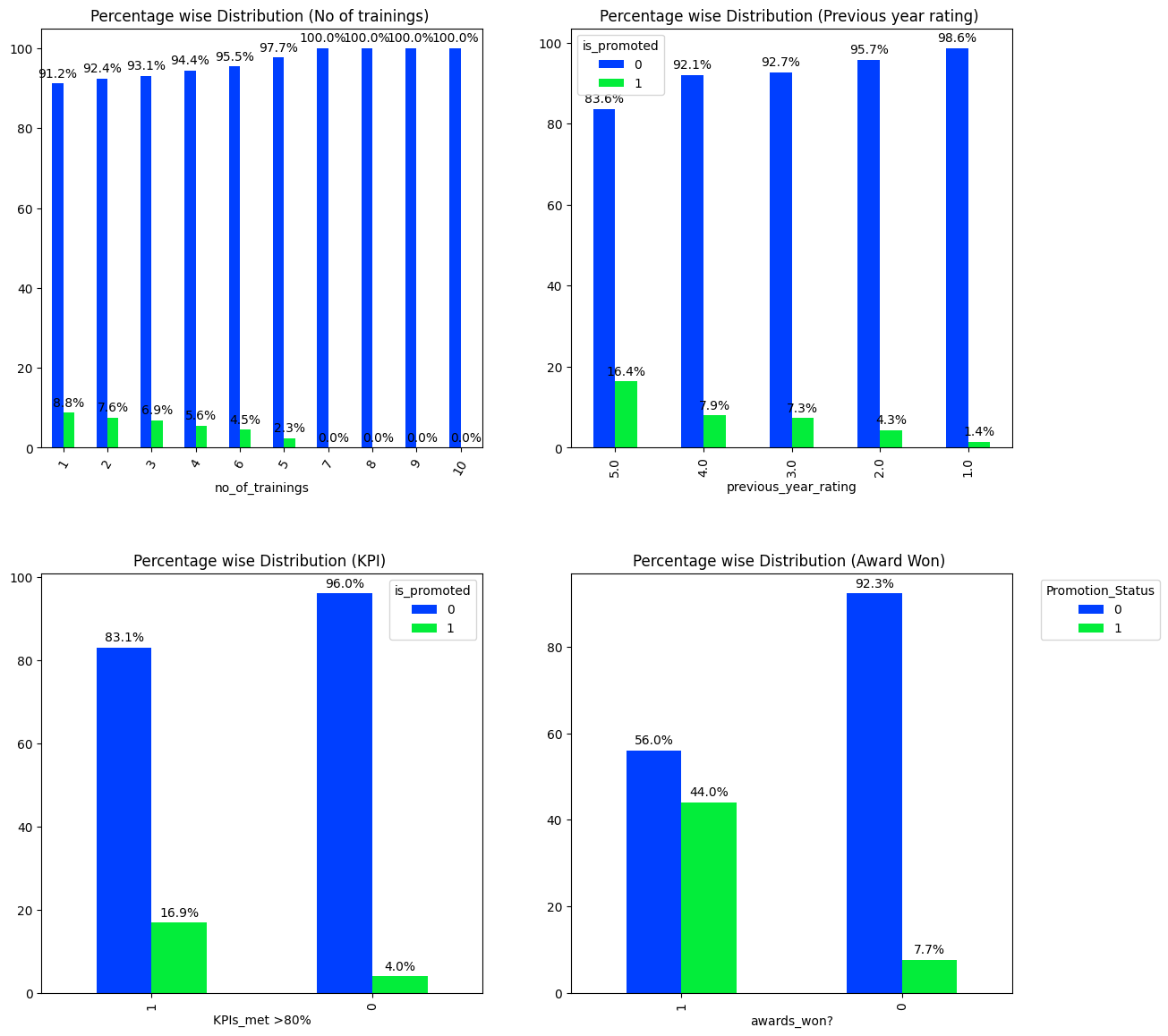
The age group of the employees was one of the factors that influenced their promotion rate. According to the data, the promotion rate was highest for employees who were between 30 and 39 years old, at 9.1%. The second highest promotion rate was for employees who were between 20 and 29 years old, at 8.6%. The promotion rate decreased for employees who were older than 40 years old, with 7.3% for those who were between 40 and 49 years old, and 6.7% for those who were between 50 and 60 years old. This indicates that employees who had a balance of experience and maturity were more likely to be promoted than those who were either too young or too old.

Another factor that affected the promotion rate of employees was their average training score. The data revealed that employees who had an average training score between 90 and 99 had a remarkable promotion rate of 76.8%, followed by those who had a score between 80 and 89 with 11.7%. This indicates that employees who performed well in their training sessions were more likely to be recognized and rewarded for their skills and knowledge.

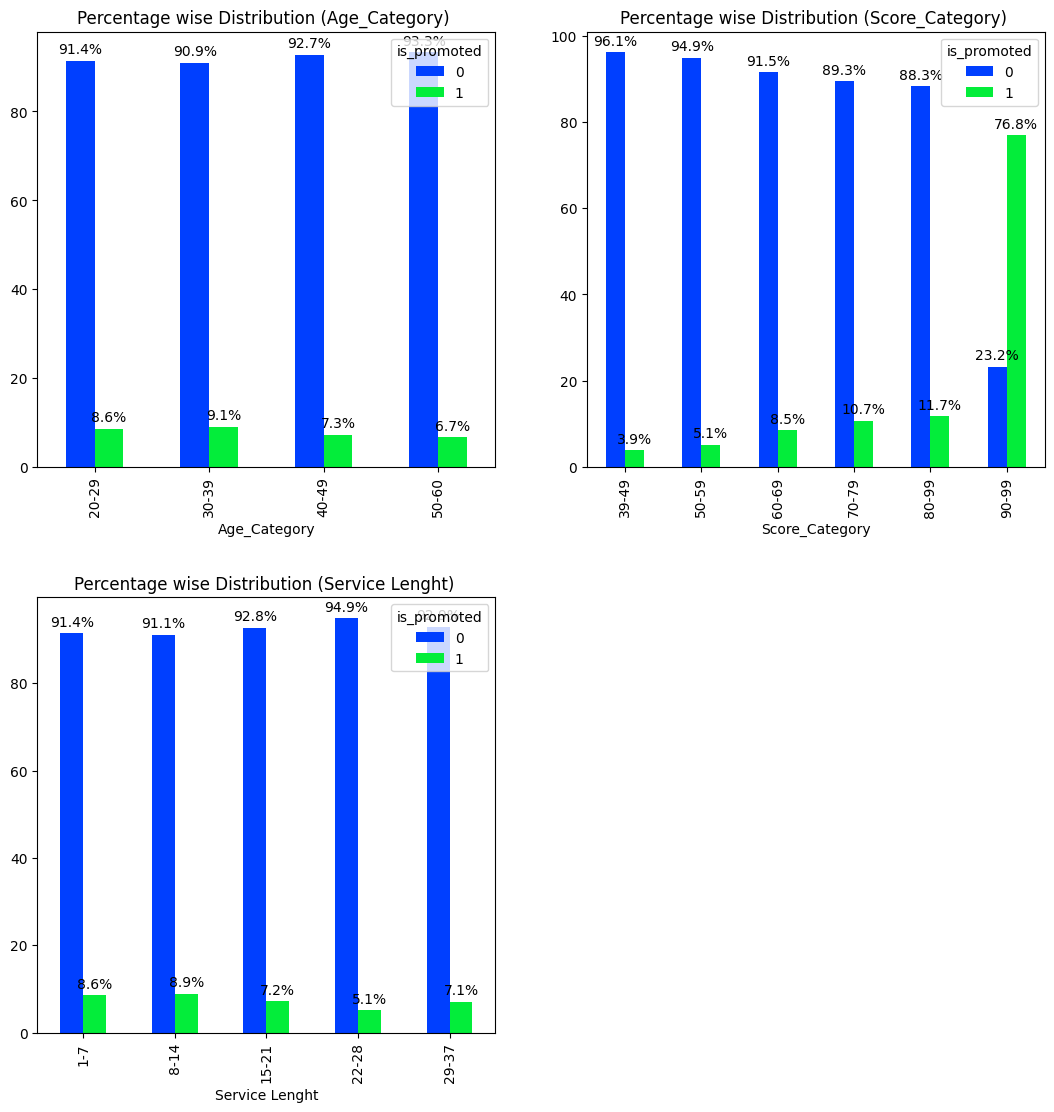
Based on factors that influenced the promotion rate of employees was their service length. The data showed that employees who had a service length of 8-14 years had a slightly higher promotion rate of 8.9%, compared to those who had a service length of 1-7 years with 8.6%. This suggests that employees who had a moderate amount of service time were more likely to be promoted than those who were new or veteran. Employees who did not win any awards had a low promotion rate of 7.7%. This indicates that employees who did not receive any recognition for their achievements were less likely to be promoted than those who did.

****

**Figure 1 Bivariate Analysis Barplot 1**

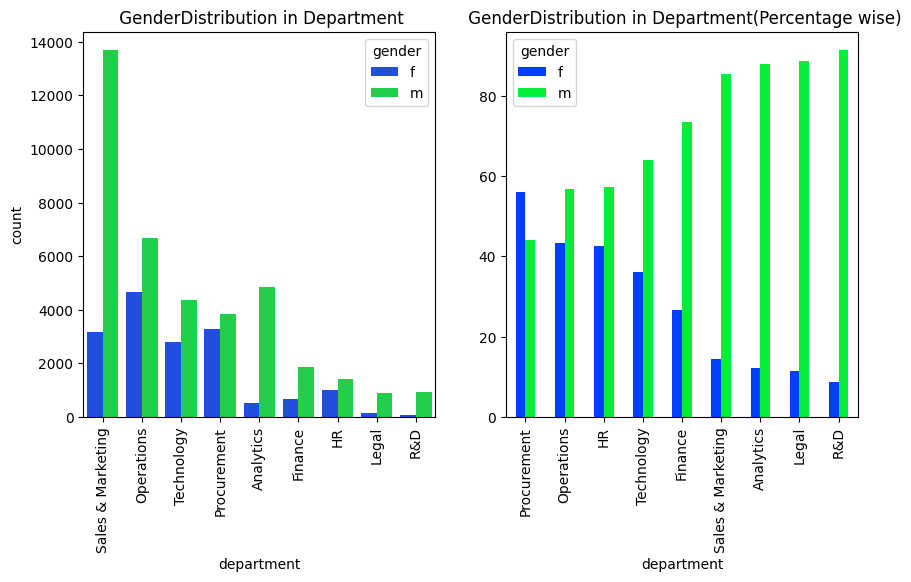


**Figure 2 Bivariate Analysis Barplot 2**



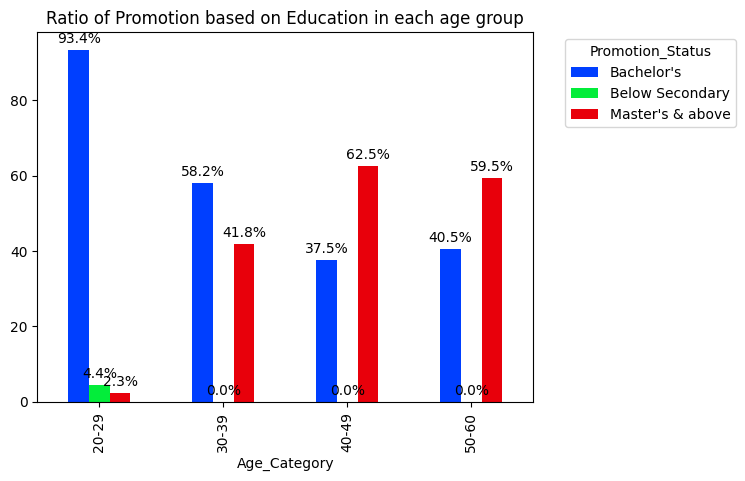
**Figure 3 Bivariate Analysis Barplot 3**

The promotion rate of female employees varies across different departments. According to the analysis, the Procurement Department has the highest promotion rate for females (56%), followed by Operation and HR. On the other hand, the R&D Department has the lowest promotion rate for females, which may be due to the low number of female employees in that department.



**Figure 4 Gender Distribution in Department**

The analysis also reveals some interesting patterns among the promoted employees based on their age and education level. For instance, most of the promoted employees in the age group of 20-29 have a bachelor's degree (93.4%), while only a few have a master's degree or above (2.3%) or below secondary education (4.4%). However, as the age group increases, the percentage of promoted employees with a master's degree or above also increases. For example, in the age group of 30-39, 41.8% of the promoted employees have a master's degree or above, and in the age group of 40-49, this percentage rises to 62.5%. Similarly, in the age group of 50-60, 59.5% of the promoted employees have a master's degree or above, while 40.5% have a bachelor's degree.



**Figure 5 Ratio of Promotion based on Education and Age**

**5.Data Pre-Processing 1**

Data pre-processing is an important step in data analysis and machine learning. It involves transforming raw data into a suitable format for further processing and modeling.

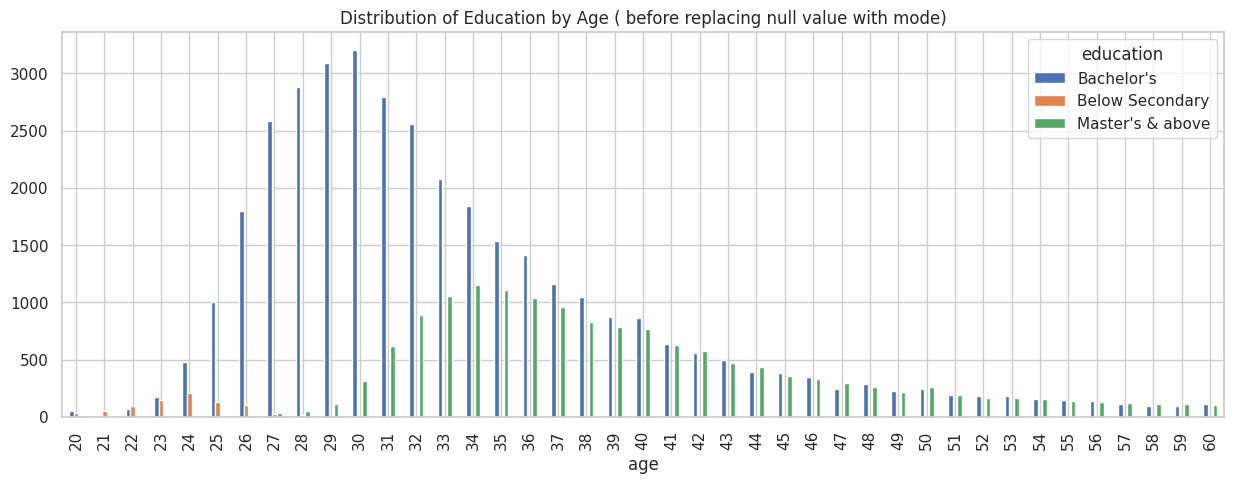
Here we have taken the steps:

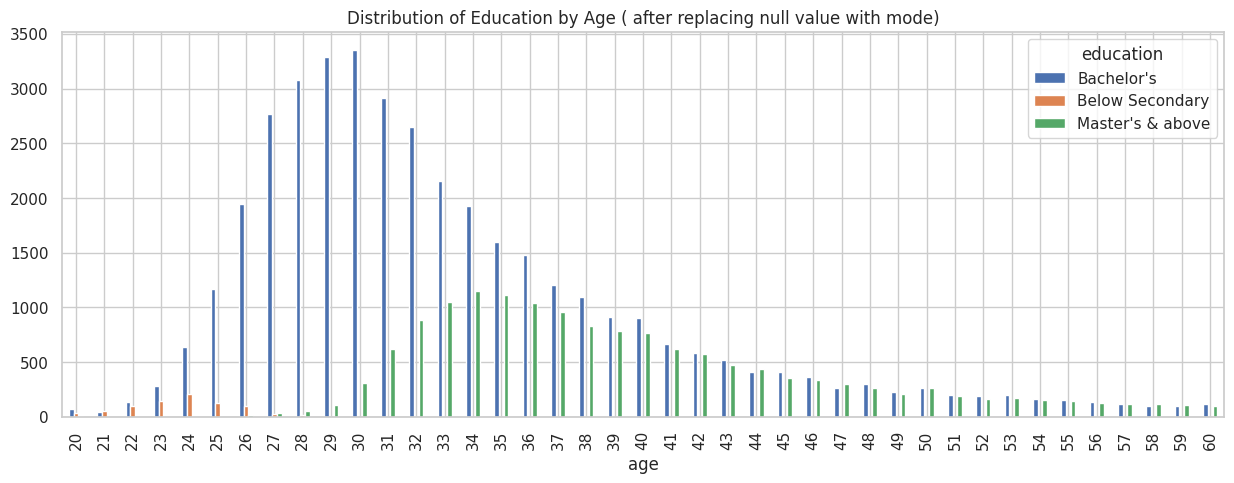
* Null value handling: This refers to dealing with missing or incomplete values in the data. There are different methods to handle null values. The choice of method depends on the nature and amount of null values, as well as the goal of the analysis.
* Outlier identification: This refers to detecting and removing values that are significantly different from the rest of the data. Outliers can affect the accuracy and performance of the models, as well as the interpretation of the results.
* Encoding: This refers to converting categorical or textual data into numerical or binary data. Encoding is necessary for some models that can only work with numerical data.

**5.1 Null Value Handling**

The dataset contains two columns with missing values: education and previous\_year\_rating. The percentage of missing values for education is 4.3% and for previous\_year\_rating is 7.5%. There is no correlation between education and age or gender variables. The following steps were taken to handle the missing values:

* Previous\_year\_rating: The missing values were replaced with 0, as they correspond to employees with a length of service of 1 year, meaning they are new hires who have not been rated yet.
* Education:There is no significant relationship between education with any other features. Mode imputation was applied to fill the missing values with the most frequent category, which is "Bachelor's". This category accounts for 71% of the data. The distribution of the data did not change significantly after the imputation.



**Figure 6 Distribution of Education by Age before Null Value handling**

**Figure 7 Distribution of Education by Age After Null Value handling**

**5.2 Outlier Detection**

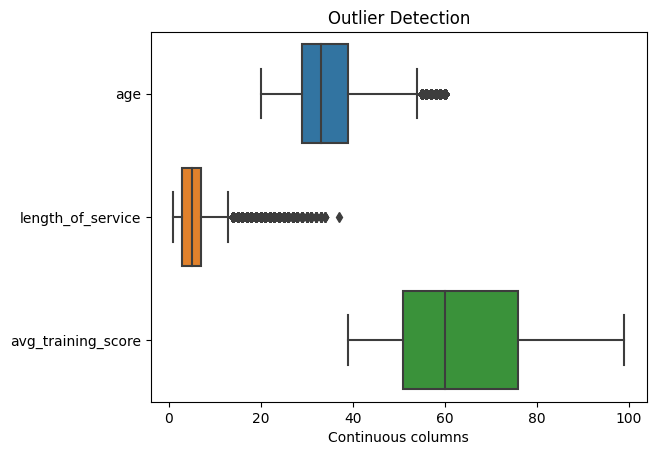
Using the IQR method for outlier detection. The IQR method uses the Interquartile Range (IQR) to identify outliers. Points that fall below Q1–1.5 \* IQR or above Q3 + 1.5 \* IQR are considered outliers, where Q1 and Q3 represent the 25th and 75th percentiles, respectively².

There are no outliers in average training score but outliers are present in length of service and age (Right skewed data). Based on the data we have calculated the Upper and lower values of columns using formula Upper limit = Q3 + 1.5\*IQR Lower limit = Q1 – 1.5\*IQR. The Upper limit for Age is 54 and the Lower Limit for Age is 14. The Upper limit for Length of Service is 13 and the Lower Limit for Length of Service is -3.

Saved outlier data in Data Frame age\_out, sl\_out respectively.

Percent of Age outlier in data is 2.6% and Percent of Length of Service outlier in data is 6.4%¹.

Visualizing Outliers using boxplot

**Figure 8 Outlier Detection**

**5.3 Encoding**

Categorical columns are variables that have a finite number of possible values, such as department, region, education, gender, or recruitment channel. Depending on the type and number of values, different encoding techniques can be applied to transform them into numerical values that can be used for analysis or modeling.

One type of categorical column is ordinal, which means that the values have a natural order or ranking. , education can be ordinal if it has values like masters, bachelor's, and secondary. To encode ordinal columns, we can use **Ordinal Encoding**, which assigns a numerical rank to each value based on the order. For example, Masters and above can be encoded as 2, bachelor's as 1, and secondary as 0.

Another type of categorical column is nominal, which means that the values have no inherent order or ranking. For example, department, gender, and recruitment channel are nominal columns. To encode nominal columns, we can use **One Hot Encoding**, which creates a new binary column for each unique value and sets it to 1 if the row has that value and 0 otherwise. To avoid the dummy variable trap, which is a problem of multicollinearity that can affect some models, we can drop one of the new columns as it can be inferred from the others.

A special case of nominal columns is when they have a large number of unique values, such as region. In this case, one hot encoding can create too many new columns and increase the dimensionality of the dataset. To reduce the number of columns, we have simplify the values by extracting some common features or patterns. Wwe have region values like region\_1, region\_2,...region\_34, we can separate the region name from the region number by splitting the string and replacing the number with an integer. For example, region\_7 can be replaced with 7. Then we can convert the column from object to integer type.

**6. Data Pre-Processing II**

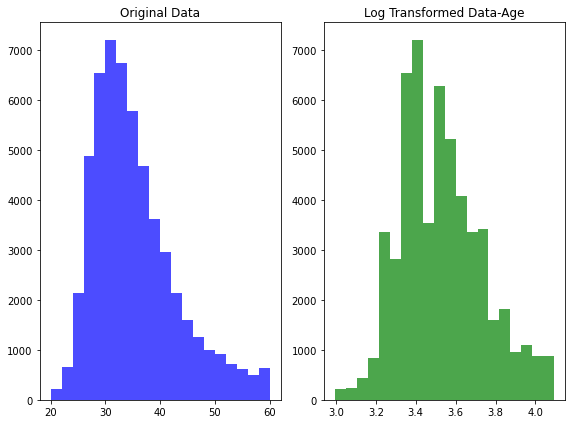
In the subsequent preprocessing stage, we focus on refining the dataset for optimal model performance by employing several key techniques. These steps, namely outlier handling, feature engineering, feature scaling, and feature reduction, play a pivotal role in shaping the data and preparing it for effective machine learning model training and evaluation.

* Outlier Handling: Outliers are data points that significantly deviate from the majority of the dataset. These can skew the model's performance and lead to incorrect predictions. Outlier handling involves identifying and either correcting or removing these outliers.
* Feature Engineering: Feature engineering entails the creation of new features or transforming existing ones to enhance the predictive power of the model. This step involves a deep understanding of the domain and the problem .
* Feature Scaling: Feature scaling ensures that all features are on the same scale, which prevents certain algorithms from being biased towards features with larger magnitudes.
* Feature Reduction: As datasets often contain a high number of features, some of them may be redundant or noisy. Feature reduction methods aim to select the most informative features or reduce the dimensionality of the dataset.

**6.1 Outlier Handling**

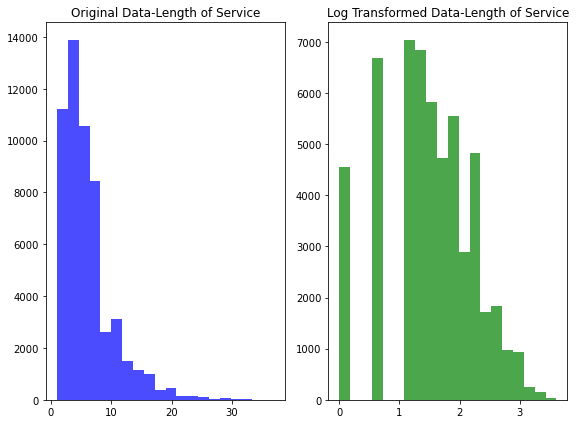
The outliers were detected in Age and Length of Servic column. Since both are right skewed distributions, log transformation is used for outlier handling.Log transformation also de-emphasizes outliers and allows us to potentially obtain a bell-shaped distribution. The idea is that taking the log of the data can restore symmetry to the data.

Initially, the 'age' column exhibited a skewness of approximately 1.0074, indicating a non-symmetric distribution. By applying a log transformation, the distribution's symmetry improved significantly, with the skewness reduced to about 0.4823. This transformation not only mitigated the skewness but also made the distribution more balanced and suitable for modeling.

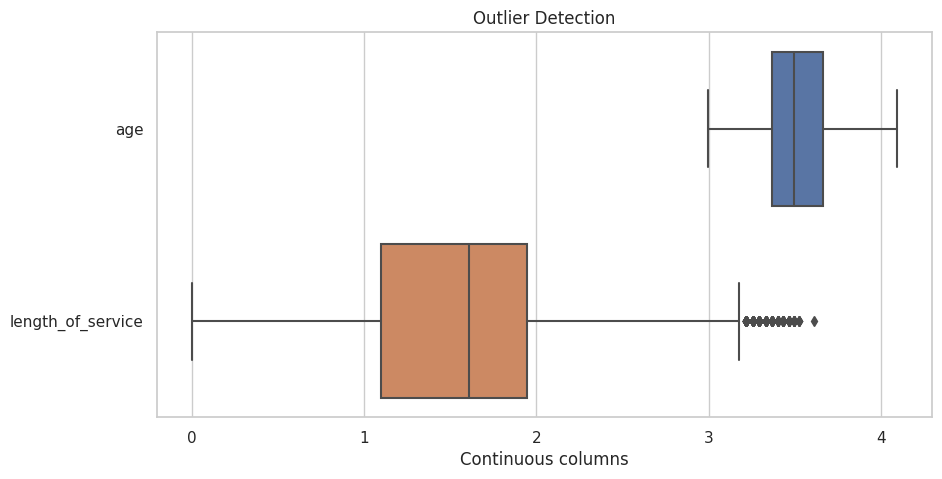


**Figure 9 Outlier Handling using Log Transformation**

Similarly, the 'length\_of\_service' column displayed a high skewness of approximately 1.7381, suggesting a skewed distribution. Employing the same log transformation technique, the skewness was effectively addressed, resulting in a log-transformed skewness of around -0.2899. Remarkably, this transformation substantially reduced the skewness and rendered the distribution more symmetric, enhancing its suitability for subsequent analysis.



**Figure 10 Outlier Handling using Log Transformation**



**Figure 11 Box Plot after outlier handling**

**6.2 Feature Engineering**

Feature engineering is an important stage in the process of preparing and transforming raw data into a format appropriate for machine learning algorithms. It entails adding new features or adjusting existing ones in order to improve a model's predictive ability and performance

A new novel feature called **'acheivement\_metric'** was created by combining 'previous\_year\_rating' and 'awards\_won?' to potentially capture an employee's overall achievement level. Additionally, calculated a **'total\_score**' by multiplying the 'avg\_training\_score' with 'no\_of\_trainings.' These newly engineered features laid the foundation for exploration. A greater 'total\_score' was associated with a higher chance of promotion.

Overall, the analysis provided valuable insights into the factors affecting employee promotions. It suggested that certain achievements and a higher 'total\_score' might contribute to an increased likelihood of promotion.

**6.3 Feature Scaling**

The utilization of standard scaling for the columns 'avg\_training\_score' and 'total\_score' in our dataset is a critical preprocessing step that profoundly impacts the effectiveness of our subsequent analysis and modeling.

These two columns contain numerical values that represent different aspects of employee performance and training. However, they might have distinct scales, which can lead to skewed model performance. Standard scaling addresses this issue by transforming the data in a way that each value becomes comparable, improving the overall reliability of our analysis.

Specifically, standard scaling, implemented through the StandardScaler module, standardizes these two columns by subtracting the mean value and dividing by the standard deviation for each data point. This normalization process centers the data around zero and scales it to have a unit variance.

The primary goal of this normalization is to ensure that the features with larger scales do not unduly influence the modeling process compared to those with smaller scales. In our context, 'avg\_training\_score' and 'total\_score' could potentially have significantly different ranges. For instance, 'avg\_training\_score' could range from 0 to 100, while 'total\_score' might span from 0 to 1000. If left unscaled, the latter could dominate the modeling process, leading to suboptimal results.

By applying standard scaling, we bring both columns to a common scale, facilitating fair and unbiased comparisons between them. This treatment enhances the stability and accuracy of machine learning models, as they are now less sensitive to varying scales and can provide more meaningful insights.

**6.4 Feature Reduction**

In the process of feature reduction, a crucial step is to identify and eliminate irrelevant or redundant features that do not significantly contribute to the predictive power of the model. In this context, the removal of the 'employee\_id' column is a strategic decision that aligns with this principle.

The 'employee\_id' column contains unique identifiers for each employee, essentially serving as a nominal label without any inherent predictive value regarding the likelihood of employee promotion. Since every employee possesses a distinct ID, there is no meaningful correlation between 'employee\_id' and the target variable, which is the 'is\_promoted' label indicating promotion status.

By dropping the 'employee\_id' column from the dataset, we eliminate a feature that does not offer any meaningful insights or predictive information for our analysis. This action contributes to feature reduction by simplifying the dataset and removing noise, resulting in a more streamlined and effective modeling process.

By drop the columns for achievement metric and total score from the dataset, as they do not have a significant impact on the target variable. This will reduce the dimensionality and complexity of the data, and make it easier to identify the most relevant features for the prediction task.

**7. Modeling**

**7.1 Machine Learning Models**

Before applying the classification model we split our data into train and test datasets.

(67% for training and 33% for testing respectively).

In the first step we tried some models without sampling, to check how the model works with our imbalanced dataset. As part of this we tried the basic models like Logistic Regression, KNN, Decision tree, Naive bayes, and some advanced models like Random forest and XG boost.

**Random Forest:**

The supervised machine learning approach known as "Random Forest Classification" combines the predictions from multiple decision trees to produce a more reliable and accurate final classification model.

**Logistic Regression:**

Logistic Regression models the relationship between features and class probabilities using the sigmoid function and optimizes parameters to make accurate predictions..

**KNN:**

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning method that makes predictions based on the similarity of data points.

**Decision Tree:**

Decision Tree Classification is a fundamental supervised machine learning algorithm which employs a tree-like structure to make decisions based on the values of input features and eventually assigns a class label to each data point.

**Bernoulli Naive Bayes:**

Bernoulli Naive Bayes is a specialized Naive Bayes algorithm designed for binary features. It leverages probabilistic calculations to predict class labels based on the presence or absence of specific features.

**Gaussian Naive Bayes:**

Gaussian Naive Bayes is a probabilistic algorithm that leverages the Gaussian distribution of continuous features within each class to make classification predictions.

**XGB Classifier:**

XGBoost is an ensemble learning algorithm that uses decision trees to create a predictive model. Iteratively, it incorporates new trees to correct previous errors, using gradient boosting to minimize residuals . Unlike Random forest classification, this algorithm works in the sequential manner thus refining the model's performance.

**7.2 Balancing dataset:**

Class imbalance can lead to biased models, affecting their overall performance. This project focuses on mitigating class imbalance challenges using the **SMOTE-TOMEK** hybrid approach.

The combined technique involves oversampling the minority class using SMOTE(Synthetic Minority Oversampling Technique) and then undersampling using Tomek Links to remove potentially noisy samples from both classes.

**7.3 Evaluation Metrics and Model Selection**

Evaluation metrics are used to measure the performance of machine learning models. They provide a quantitative measure of how well a model is performing in terms of its ability to predict outcomes accurately.

**Accuracy Score** :

Accuracy Score measures a classification model's effectiveness by quantifying the proportion of correctly classified instances in a dataset. A higher score indicates better performance, while a 0 score indicates incorrect predictions. An accuracy score of 1 (or 100%) means that all predictions were correct, while a score of 0 (or 0%) indicates that all predictions were incorrect.

**Precision Score:**

Precision Score is a performance metric used in classification tasks to measure the accuracy of positive predictions made by a machine learning model. It focuses on evaluating the proportion of correctly predicted positive instances out of all instances that the model predicted as positive.

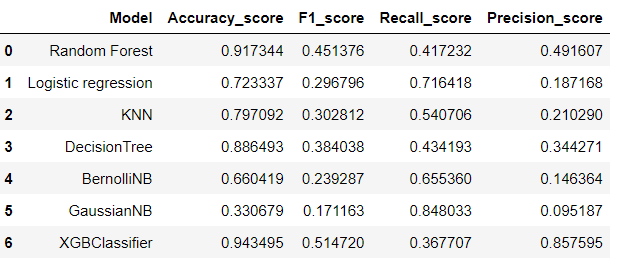
**Recall Score** ;

Recall is a performance metric used in classification tasks to measure a machine learning model's ability to correctly identify positive instances out of all actual positive instances in the dataset.

**F1 Score:**

The F1 Score is a commonly used performance metric in classification tasks that combines both precision and recall into a single value. It provides a balanced assessment of a machine learning model's ability to make accurate positive predictions while correctly identifying actual positive instances.

Since this is a balanced data, we used **Accuracy score** as the evaluation metric for the selection of machine learning model.



Based on the accuracy scores of the different models, we selected **Extreme Gradient Boosting** to predict the promoted employees .

**7.3.1. Cross Validation**:

Stratified K-Fold cross-validation strategy is used to evaluate the model's performance, ensuring robust assessment and validation. This technique preserves class distribution within each fold, preserving a representative dataset across iterations, and safeguards against overfitting. This approach ensures accurate performance estimates and generalizes well to unseen data.

After cross-validation and hyperparameter adjustment, the model's accuracy score was 0.93, which was an ideal outcome for model prediction.

**7.3.2. Hyperparameter Tuning**

Hyperparameter tuning plays a crucial role in optimizing the performance and robustness of machine learning models. It involves the systematic exploration and selection of hyperparameters, which are parameters that govern the behavior of a model during training. Randomized Search and Grid Search are techniques used here.

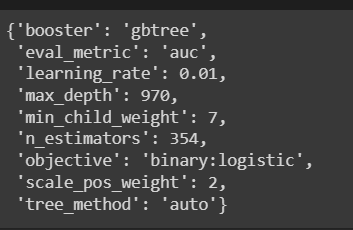
**Randomized Search**

The hyperparameter tuning process utilizes a randomized search to explore various hyperparameters, identifying high-performing regions within the hyperparameter space. This technique uses random sampling to discover promising hyperparameter combinations for improved model performance.

The parameters included are learning rate, number of estimators, maximum depth, minimum child weight, scale positive weight, objective, booster, tree method, and evaluation metric.

**Grid Search**

The grid search narrows down the hyperparameter space based on the findings from the randomized search. This iterative refinement process allows us to systematically evaluate specific combinations of hyperparameters that show promise. The grid search serves as a targeted optimization step, fine-tuning the model for enhanced accuracy and robustness. Through this iterative process, we aim to pinpoint the optimal hyperparameters that result in superior classification performance.

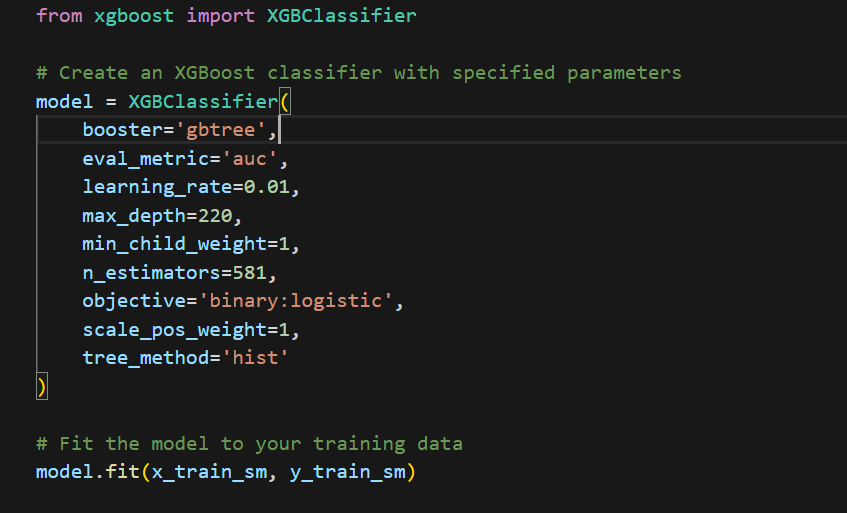
After fine tuning, below are the hyperparameters used for the model training ,

**7.3.3. Training Selected Model and Pickel file creation** :

We are utilizing the XGBoost algorithm, a powerful gradient boosting technique, to build a binary classification model. XGBoost stands for "Extreme Gradient Boosting" and is well-known for its exceptional performance in various machine learning

Model Configuration

1. The classifier is initialized with the following hyperparameters:
2. booster=‘gbtree’: The algorithm is based on decision tree boosting.
3. eval\_metric=‘auc’: The evaluation metric used during training is the Area Under the ROC Curve (AUC), which is a common choice for binary classification problems.
4. learning\_rate=0.01: A small learning rate helps prevent overfitting and improves the model’s convergence.
5. max\_depth=220: The maximum depth of each individual decision tree in the boosting process. This parameter controls the complexity of the trees.
6. min\_child\_weight=1: Specifies the minimum sum of instance weight (hessian) needed in a child. It can be used to control over-fitting.
7. n\_estimators=581: The number of boosting rounds or trees in the ensemble. More trees can improve model performance, but can also increase computation time.
8. objective=‘binary:logistic’: The objective function for binary classification using logistic regression.
9. scale\_pos\_weight=1: Controls the balance of positive and negative weights, useful for imbalanced datasets.
10. tree\_method=‘hist’: A histogram-based algorithm for parallel and faster tree construction.



**Pickel file creation**

Pickle is a Python module that allows us to serialize and deserialize Python objects, which means we can convert them into a binary format and store them in a file. This way, we can reuse them later without having to retrain or preprocess them. To save our objects, we will use the pickle.dump function, which takes two arguments: the object to be saved and the file object to write

* pickle.dump(scalar,open('scale.pkl','wb') )
* pickle.dump(ohe,open('ohe.pkl','wb') )

To create two more files named scale.pkl and ohe.pkl, which contain the binary representations of our scaler and one-hot encoder objects. To load these objects later, we will use the pickle.load function, which takes a file object as an argument and returns the deserialized object

This will create a file named xgmodel.pkl in our current directory, which contains the binary representation of our model

* model = pickle.load(open('xgmodel.pkl','rb') )

These will assign the scaler and one-hot encoder objects to the variables scalar and ohe, which we can use for preprocessing new data.

**8. Web Hosting**

Flask is a micro web framework that empowers us to construct web applications using Python. Complementing this, HTML and CSS are markup languages responsible for defining the structure and style of web pages. Our project entails learning how to leverage Flask for tasks like request handling, template rendering, session and cookie management, and eventually deploying our web application on a server.

Furthermore, we will delve into the application of HTML and CSS to craft responsive and visually appealing web pages that can dynamically exhibit content from Flask. As we conclude this endeavor, we will have successfully developed a fully functional web hosting service capable of hosting multiple websites featuring diverse domains and functionalities.

**8.1 HTML and CSS**

Created a folder called "templates" inside your project directory, and store all your HTML files there and folder called "static" inside project directory, and store all your static files, such as images, and CSS there. use the url() function in CSS to specify the path to the image file and stylesheet.

**Index Page**

CSS Styling Links:

The head section contains two link elements that reference external CSS stylesheets. One is linked from the "../static/css/style.css" path, and the other is commented out (likely intended for an alternative stylesheet).

Page Title:

The title of the web page is set to "Employee Promotion Prediction."

Header Section:

Inside the body tag, there's a header section with the class "speaker-form-header." It includes an h1 tag displaying the title "Employee Promotion Prediction" and an em tag containing informative text. The text encourages users to input employee details for predicting promotions and emphasizes that all fields are mandatory.

Input Form:

The main content of the page is the input form contained within a div element with the class "input form." The form tag specifies the action as "/predict," indicating where the form data will be submitted using the POST method. The form also has an id attribute set to "input head."

Form Fields:

The form contains various legend and input field pairs, each representing a specific employee attribute. These include:

Employee ID (numeric input)

Age (numeric input with defined range)

Education (select dropdown with educational levels)

Gender (select dropdown for gender)

Department (select dropdown for department)

Region (numeric input with defined range)

Recruitment Channel (select dropdown for recruitment sources)

Num of trainings (numeric input with defined range)

Previous year ratings (numeric input with defined range)

Length of service (numeric input with defined range)

Avg train score (numeric input with defined range)

KPI above 80% (radio buttons for "Yes" or "No" options)

Awards Won (radio buttons for "Yes" or "No" options)

Submission Button:

At the end of the form, there's a "Submit" button (with the id "submit") that users can click to submit their input.

Background Image:

The body tag has a background attribute that sets the background image to "/static/image/background.jpeg."

This HTML code creates a user-friendly web page for predicting employee promotions. Users are guided through a form where they can input various attributes of an employee, and the form data is then submitted to the "/predict" endpoint.

**Result Page**

Heading and Dynamic Content:

An h2 heading is present on the page. It contains an id attribute with the value "Iris is." The content inside the heading is enclosed within double curly braces {{ }}. This indicates that the content is intended to be dynamically populated from a template context, likely through a web framework like Flask.

Specifically, the {{prediction\_text}} syntax suggests that the content of this heading will be replaced with the value of the "prediction\_text" variable from the context. This is a common approach used in web frameworks to inject dynamic data into HTML templates.

This HTML code is meant to be part of a web page that displays the result of a promotion status prediction. The page features a background image, imports a stylesheet for styling, and includes a dynamic heading element that will display the prediction result. The actual prediction result will be provided by the backend logic of the web application, which will replace the {{prediction\_text}} placeholder with the appropriate value.

**CSS**

CSS code is used to style a web page containing a form for predicting employee promotions. The styles enhance the visual appeal, responsiveness, and user experience of the page. Elements are styled to have consistent padding, borders, and fonts, with interactive elements like buttons changing appearance on hover. The responsive design ensures that the form looks well-presented on different screen sizes.

To use a background image in CSS, save the image file in a folder that is accessible by the web server. create a folder called "static" inside project directory, and store all your static files, such as images, and CSS there. use the url() function in CSS to specify the path to the image file and stylesheet.

**8.2. Flask**

Flask web application code that uses a trained machine learning model to predict whether an employee should be promoted or not based on various input features

The / route corresponds to the home page, which is rendered using the index.html template.

The /predict route is where the actual prediction happens. This route expects a POST request with input data from the user interface.

**Loading Pretrained Models:**

You load three pretrained artifacts: the machine learning model (xgmodel.pkl), the OneHotEncoder (ohe.pkl), and the scaler (scale.pkl).

**User Input Handling:**

In the predict route, you retrieve user input from the HTML form. Make sure the form fields in the HTML template correspond correctly to these inputs.

**Feature Preparation:**

The categorical features like gender, recruitment, and department are transformed using the loaded OneHotEncoder.

The score is scaled using the loaded scaler.

The remaining numerical features are concatenated with the transformed categorical features to form the complete feature vector.

**Prediction and Result:**

The model's predict method is called with the feature vector, and the output is used to determine whether the employee should be promoted or not.The prediction result is then displayed in the rendered result.html template.

Run Flask application script,the development server starts and listens on the specified port ( port 9000).When you access **http://127.0.0.1:9000/** in web browser, the Flask server responds to the request and serves the content defined in your application routes, templates, and static files.

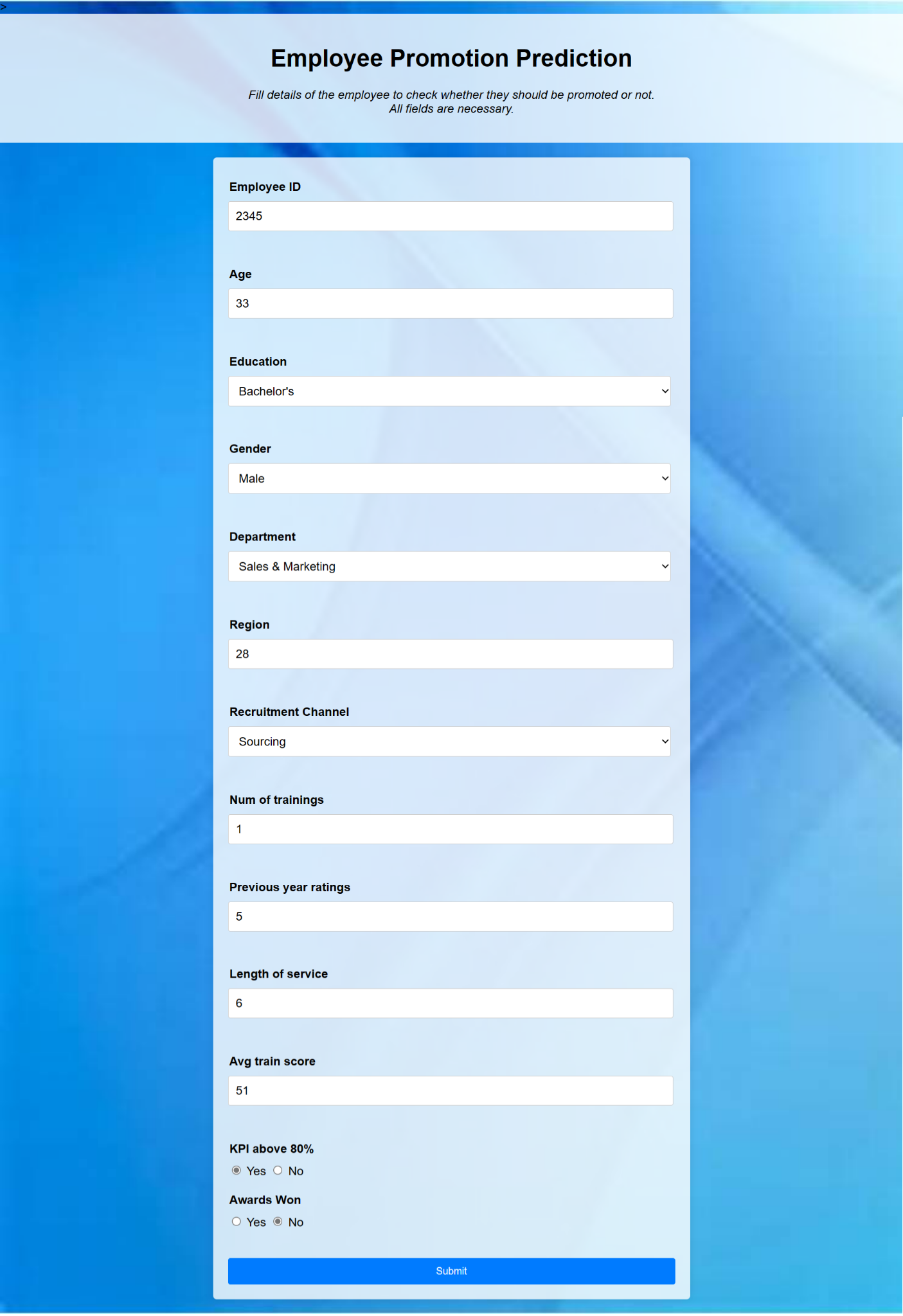
**9. Result**

There are many web hosting platform. We are using PythonAnywhere which is a popular web hosting platform that allows you to deploy and run web applications built with Python. It provides a simple and convenient way to host your Python web applications without the need to manage complex server configurations.

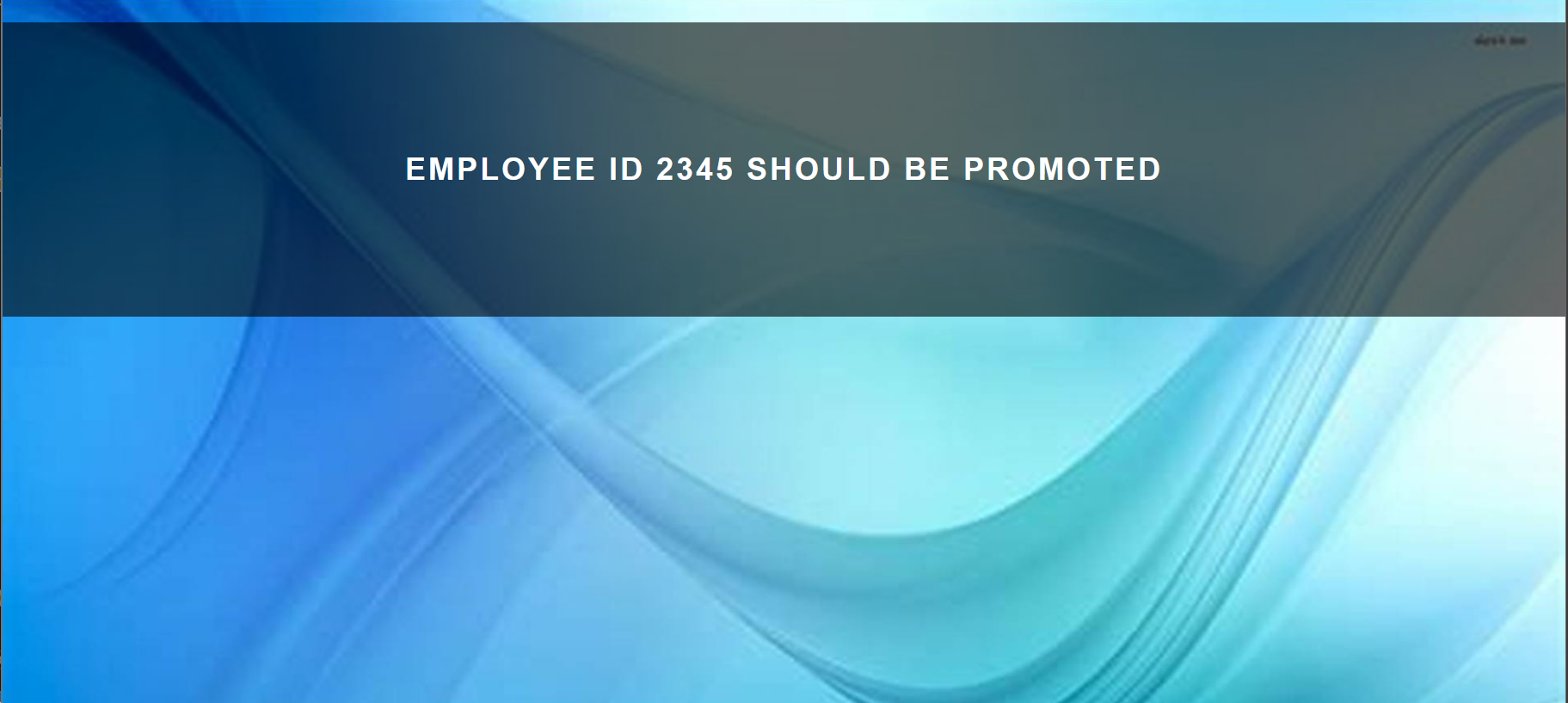
As file could not be uploaded in Python anywhere output is from local hosting

[**http://127.0.0.1:9000/**](http://127.0.0.1:9000/)

After running the hosted webpage, the result is



**Prediction Result:**

****

**8. Conclusion**

In conclusion, the project focused on predicting the promotion status of employees using the powerful Boost model. The dataset contained a variety of attributes, each providing unique insights into an employee's profile and performance. These attributes encompassed the user's age, education, gender, department, region, recruitment channel, number of training sessions, rating, length of service, average score, KPI achievement above 80%, and award achievements. The successful completion of this project yields a valuable tool for organizations to make informed decisions about promoting employees. By harnessing the capabilities of the XG Boost model, businesses can better allocate resources, identify high-potential individuals, and create an environment that fosters growth and advancement.

**References**

● https://www.w3schools.com/python

● https://www.kaggle.com

● https://towardsdatascience.com/

● https://www.geeksforgeeks.org/

● https://scikit-learn.org/

● https://www.medium.com