Contents

[1. HR Analytics 4](#_Toc155045856)

[1.1 Problem Understanding 4](#_Toc155045857)

[1.2 Exploratory Data Analysis 4](#_Toc155045858)

[1.3 Missing Value Imputation 17](#_Toc155045877)

[1.4 Categorical Data Encoding 18](#_Toc155045880)

[1.5 Correlation Analysis 19](#_Toc155045881)

[1.6 Feature Selection 20](#_Toc155045883)

[1.7 Feature Scaling 23](#_Toc155045886)

[1.8 Stratified Sampling 23](#_Toc155045887)

# 1. HR Analytics

## 1.1 Problem Understanding

With this dataset, the end goal is to use the features to predict if the employee will get promoted or not. However, before that analysis of the data is crucial as it will allow me to extract hidden insights which can be helpful. This will be the main aim of the project which is to analyse the data and prepare the data to pass it to the machine learning model.

The features that were given in the dataset includes a wide range of employee-related information, including demographics, performance history, training, and promotion status.

## 1.2 Exploratory Data Analysis

Basic Analysis

At first, I explored the dataset by using basic pandas dataframe functions such as .head() and .shape() etc. This exploration allowed me to get used to the dataset and understand it from the surface level.



Shape of Dataset

## A screenshot of a computer program Description automatically generated

## 

Dataset’s column Information

A screenshot of a computer

Description automatically generatedThere are 5 categorical columns while 9 columns are numerical.

Dataset’s information on numerical columns

## **Here are some observations of the numerical columns:**

* **Count:** There are **54,808** observations in the dataset with **14** columns.
* **no\_of\_trainings:** On average, employees have undergone around **1.25** training sessions, with a minimum of **1** and a maximum of **10**.
* **age:** The average age of employees is approximately **34.80 years**, with a minimum age of **20** and a maximum age of **60**.
* **previous\_year\_rating:** The average rating from the previous year is **3.33**, with a minimum rating of **1** and a maximum rating of **5**.
* **length\_of\_service:** The average length of service is approximately **5.87 years**, with a minimum of **1** year and a maximum of **37 years**.
* **KPIs\_met >80%:** A moderately large percentage of employees **35.19%** have met 80% of KPIs (values are binary).
* **awards\_won?:** A small percentage of employees **2.32%** have won awards (values are binary).
* **avg\_training\_score:** The average training score is around **63.39**, with a minimum score of **39** and a maximum score of **99**.
* **is\_promoted:** Approximately **8.52%** of employees have been promoted (values are binary).

## A screenshot of a computer Description automatically generated

## 

Null values in Dataset

Reasons for missing values in the columns:

For Education:

* Non-Reporting: Some employees might not have provided their educational details during data collection or entry.
* Educational Background: In some cases, certain roles or positions might not require specific educational qualifications, leading to non-disclosure.

For Previous\_year\_rating:

* New Employees: Missing ratings might be due to new employees who haven't completed a year to receive a previous year's rating.

In the section of Missing Value Imputation, I will discuss on how I imputated the null values.

Categorical Data Analysis

After carrying out some basic analysis, I went on to create some charts to further understand the data of categorical variables and how they affect is\_promoted which is the target column.

## A pie chart with text Description automatically generated

There is a **huge disparity** in the target column, is\_promoted, where there are close to **10 times** more of not promoted employees then promoted employees and this can be quite problematic when creating a logistic regression model as this may lead to **biasness** in the model.

## A screenshot of a graph Description automatically generated

* Sales & Marketing and Operations have the most substantial number of employees, indicating these departments might be pivotal in the company's operations.
* Analytics, Finance, HR, Legal, and R&D have smaller employee counts compared to Sales & Marketing and Operations, signifying potentially specialized or smaller-focused departments within the organization.
* The departments that have the highest rate of promotions are Technology, Procurement and Analytics while HR and Legal departments have a low promotion rate.

## A screenshot of a graph Description automatically generated

* A substantial majority of employees hold a Bachelor's degree, indicating that it is the most prevalent educational qualification among the workforce.
* Employees with education levels below secondary school are notably fewer in count, indicating that this category represents a minority within the organization.
* The null category encompasses a considerable count, signifying either a lack of available data regarding education or potentially employees who haven't provided this information.
* Among all the education, those who have qualifications of Master's & above have a higher chance of getting promoted. This makes sense as usually those who took Master's & above should have more skills and knowledge compared to others and therefore are able to perform better in their tasks.

## A screenshot of a graph Description automatically generated

* There is a noticeable gender gap within the organization, with a significantly higher count of male employees compared to female employees.
* Female employees, while fewer in count, has a higher chance of getting promoted then male employees by 0.7%.

## A screenshot of a graph Description automatically generated

* The organization seems to heavily rely on the 'Other' and 'Sourcing' channels for recruitment.
* Even though those employees who are referred are very few, this group has a larger chance of getting promoted then the others. Reason is because:
  + **Quality of Referrals:** Referred candidates often come with a recommendation from a current employee. This implies that the referring employee sees potential in the candidate, knows their skills and work ethic, and believes they would be a good fit for the company culture. Consequently, referred candidates might be of higher quality, making them more likely to perform well and be considered for promotions.
  + **Cultural Fit:** Referrals tend to have a better understanding of the company's culture and values since they have been recommended by an existing employee. This alignment with the company culture could result in quicker adaptation, better teamwork, and a stronger commitment, which are factors highly valued for promotion considerations.

## A screenshot of a graph Description automatically generated

* Ratings 3.0, 4.0, and 5.0 indicate a significant portion of the employee population, suggesting a more prevalent performance distribution within this range.
* For the promotion, as expected the employees with higher rating like 5.0 and 4.0 has a higher chance of getting promoted as usually those who have a high rating have better skills and qualities that help the company grow better.

## A screenshot of a graph Description automatically generated

* There is a great number of employees who has only done 1 training and as the number of trainings increase the number of employees going for the trainings decrease.
* The employees have a high chance of getting promoted when the number of trainings is low.
* When an employee undergoes more than 6 trainings, they are 100% not going to be able to get promoted.

A screenshot of a graph

Description automatically generated

* There are more employees who are not able to hit 80% of KPIs then those who can hit 80% of KPIs.
* Those who meet 80% of KPIs are more likely to get promoted than those who did not meet 80% of KPIs by close to 12.9%.
* Those who also manage to hit 80% of KPIs are more likely to get promoted as they were able to help the company reach its goals faster and also this group of employees tend to have more skills than the other employees.

A screenshot of a graph

Description automatically generated

* There are close to 42x more employees who have not won awards than those who won awards.
* Those who won awards are more likely to get promoted than those who did not win any awards by close to 35%.

## A screenshot of a graph Description automatically generated

* For departments like Technology, Analytics and R&D, they usually have high average training scores where the minimum is around 70 already. This shows that these departments are very competitve and in a way more prestigious than the other departments. Even those who are not promoted in these departments usually score highly as well.
* For other departments, even if the employee did not score very well or even fail (get below 50) they still have a chance to get promoted. For example, for the Sales & Marketing department, there are employees who scored a 41 and failed the training but still got promoted as well.
* However, there can also be a case where the training for some departments are simpler and lenient compared to other departments where the training can be very tough causing some people to not do well in the training.

## A graph of a graph Description automatically generated with medium confidence

* As the average training score increases, the chance of getting promoted also increases as well.
* The chance of getting promoted is much higher once the employee scores above 90.
* As the employee scores above 95, he/she will have a close to 100% chance to get promoted.

## A graph of a bar chart Description automatically generated with medium confidence

* As the previous year rating increases, the chance to get promoted also increases.
* Similarly, as the previous year rating increases, the percentage of those who did not get promoted slowly started to decrease.

## A graph of blue and red bars Description automatically generated

* As the no. of trainings taken by the employee increases, the chance of the employee getting promoted decreases.
* When a employee undergoes more than 6 trainings, they are 100% not going to be able to get promoted.

Numerical Data Analysis

## A graph with a line and a red line Description automatically generated

* age is right skewed.
* Some outliers are present.

A blue line and red line

Description automatically generated

* length\_of\_service is right skewed.
* Many outliers are present.

## A blue line with red line Description automatically generated

* avg\_training\_score is right skewed.
* No outliers are present.

## 1.3 Missing Value Imputation

There are 2 columns in the dataset that has missing values, and I will show how I have imputated the missing values.

## A screenshot of a computer program Description automatically generated

In the basic data exploration, I had shared how I assumed that the missing values had occurred. In fact, for the previous\_year\_rating column, I had found evidence to support my theory.

## A screenshot of a computer Description automatically generated

Based on the figure above, this can be interpreted to mean that workers with only 1 year of service to the firm have missing values for previous\_year\_rating because they were not part of the firm in the previous year. However, for the education column, I was not able to find evidence to support why there were missing values. I shall now discuss how I carried out missing value imputation.

**education**

Due to the significant number of null values for education, I used imputation methods such that they are able to capture the value "Missing". This would not only not distort the natural distribution too much due to number of null values that exist, but allow our machine learning or prediction models to have a distinction between those null values compared to using a imputation method such as those based on frequency which may distort it and worsen the models ability to identify effective relationships between those values, allowing the machine learning algorithms to establish a relationship between "regular" values and null values.

**previous\_year\_rating**

Since previous\_year\_rating is null only when the length\_of\_service of the employee is 1, this means that the employee just joined the firm and he could not have a rating of last year as he/she was not present in the company last year. Therefore, I replaced the null values to 0 to indicate that those who joined the company newly will start with a rating of 0.

## 1.4 Categorical Data Encoding

Since machine learning models do not understand object datatypes, all the features will have to be converted to numerical datatype. Therefore, I had to encode the following features: region, recruitment\_channel, gender, education, department. Now, I will share the encoding steps I had done and also the rational behind it.

**region:** Ordered Ordinal Encoding

There seem to be a natural order to the region. Due to this nature forming a ordinal relationship, I decided to use ordinal encoding for this feature. Moreover, doing one hot encoding could lead to problems as it will result in curse of dimensionality as this column has a high cardinality.

**recruitment\_channel:** Dummy Encoding

The values that make up the cardinality of this feature is low enough that One Hot Encoding is less likely to lead to the curse of dimensionality.

**gender:** Dummy Encoding

The cardinality of this feature is only 2. This seems perfect for a binary representation of the values such as 0 meaning the gender is male else if it is 1, it means female.

**education:** Dummy Encoding

The values that make up the cardinality of this feature seems to be very distinct and low enough that One Hot Encoding is less likely to lead to the curse of dimensionality.

**department:** Dummy Encoding

I want to preserve the distinction between different departments as this column is important and therefore, I did one hot encoding as it removes bias by creating separate columns, ensuring equal importance for each category. Moreover, the cardinality of this column is 9 and this is not too much that it will cause curse of dimensionality.

## 1.5 Correlation Analysis

Correlation refers to a statistical measure that describes the strength and direction of a relationship between two variables. It assesses how changes in one variable are associated with changes in another variable. In data analysis and modeling, correlation helps identify which variables are strongly related, aiding in feature selection for predictive models.

How it Works:

* **Positive Correlation:** When one variable increases, the other tends to increase as well (correlation coefficient close to +1).
* **Negative Correlation:** When one variable increases, the other tends to decrease (correlation coefficient close to -1).
* **No Correlation:** When there's no discernible relationship between the variables (correlation coefficient close to 0).

Here, I will check correlation between the variables for redundency.

## A screenshot of a graph Description automatically generated

Some significant correlation between the features and the target variable:

* + region
  + previous\_year\_rating
  + KPIs\_met >80%
  + awards\_won
  + avg\_training\_score

## 1.6 Feature Selection

Feature Selection is important as it helps determine which features will be useful for the model to predict the target variable. I have taken several steps to carry out feature selection.

## A table of numbers and words Description automatically generated

* From the above model summary, we can see that region, no\_of\_trainings, age, previous\_year\_rating, length\_of\_service, KPIs\_met >80%, awards\_won and avg\_training\_score have a 0 p-value. It means that there is a very high confidence on the coefficients of these features, or say these features are the strongest indicators to predict is\_promoted.
* On the other hand, the rest of the columns have either a **p-value** of 1 or nan
* Therefore, utilizing this method to find important features is not helpful and therfore I have to find other ways to carry out feature selection.

Now I used Chi-Values to determine the importance of the features.

A graph with blue bars and white text

Description automatically generated

* The higher the chi value, the higher the importance of the feature
* avg\_training\_score, awards\_won, KPIs\_met >80%, region, previous\_year\_rating seem to have a high importance
* recruitment\_channel\_sourcing, education\_BelowSecondary, department\_Finance, recruitment\_channel\_other, gender\_m are features that have low importance.

And lastly, I have used F-scores to identify important features and not important features:A graph of a number of people

Description automatically generated with medium confidence

## A graph with blue bars Description automatically generated

After all the tests done, I removed 3 columns, 'education\_Below Secondary', 'department\_Finance', 'recruitment\_channel\_sourcing' as these columns will not help the target variable based on the tests I did above.

## 1.7 Feature Scaling

Even though logistic regression is less sensitive to feature scaling compared to some other algorithms, scaling features can still be advantageous.

* Feature scaling can help the optimization algorithm converge more quickly. Gradient-based optimization methods, like those used in logistic regression, can converge faster when features are on a similar scale.
* A comparison of a graph

  Description automatically generated with medium confidenceWhen features are on different scales, the coefficients become challenging to interpret since a unit change in one feature might have a different effect compared to a unit change in another feature. Feature scaling helps in making these coefficients comparable and interpretable.

## 1.8 Stratified Sampling

At the start of the report I mentioned that the target variable is imbalanced and this is not suitable when creating a logistic regression model. Therefore, I have downsampled the data and reduced the columns from 54808 to 9336. Now the data in the target variable is balanced and it is now more suitable to create the model.

## 