Employee Promotion Prediction

Understanding the Problem Statement

- HR analytics is revolutionising the way human resources departments operate, leading to higher efficiency and better results overall.
- The collection, processing and analysis of data is manual, and nature of HR dynamics and KPIs has been constraining HR Dept.
- Try this predictive analytics and identify the employees who are most likely to get promoted in an Organization.

Who is the Client?

It is a Large MNC and they have 9 broad verticals across the organisation. One of the problems they face is around identifying the right people for promotion and preparing them in time.

Currently there Process involves:

- Identify employees based on Past performance and Recommendation.
- Selected Employees go through Training and Evaluation based on Skills Required.
- At the end, The employee gets Promotion based on Training Score, and KPIs.

Descriptive Statistics

It gives us Summary about all the continuous and Categorical Variables present in the dataset.

Continuous Variables

- Count,
- Average
- Standard Deviation
- Min and Max
- 25, 50, and 75 Percentiles.

Categorical Variables

- Count,
- Unique
- Top category
- No. of records in the Top Category.

Missing Values Treatment

Reasons of having Missing Values in Data set,

- Unavailability of Data
- Loss of Data
- Data Entry Error
- Incomplete Forms etc.

Treatment of Missing Values is very Important as Machine Learning Predictive Models cannot work with Missing Values.

Outliers Treatment

- The presence of outliers in a classification or regression dataset can result in a poor fit and lower predictive modeling performance.
- Outliers Detection and Treatment becomes necessary in some cases.
- Outliers can be Found in Numeric Columns of the Data.

- Columns in Dataset which might have Outliers
  - **Average Training Score**
  - **Length of Service.**

- Box plots can be used for identifying Outliers in Data.

Univariate Analysis

- Univariate analysis is the simplest form of statistical analysis.
- The key fact is that only one variable is involved in Univariate Analysis.
- Univariate analysis can yield misleading results in cases in where multivariate analysis is more appropriate.
- This is an Essential step to understand the variables present in the dataset step by step.
- We can use Distribution plots to check the distribution of the Numerical Columns.
- We can check distribution of Categorical Columns using Pie charts, Count plots etc.

Bivariate Analysis

Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them.

There are three Types of Bivariate Analysis which can be used to understand the association between two variables in a dataset.

- **Categorical vs Categorical**
- **Categorical vs Numerical**
- **Numerical vs Numerical**

Multivariate Analysis

Multivariate analysis is based on the principles of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time.

- We will use Heatmaps for finding relation between all the variables in the dataset.
- A heatmap is a graphical representation of data that uses a system of color-coding to represent different values.

Before Using Heatmaps, Let’s understand how to analyze Correlation.

- If the Correlation Value is +1, it means that the two columns highly similar
- If the Correlation Value is 0, It means that the two columns are having no relation, and
- If the Correlation Value is -1, It means that the two columns are completely Opposite to each other.

Feature Engineering

Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms.

Let's discuss the ways how we can perform feature engineering

- We can perform Feature Engineering by Removing Unnecessary Columns
- We can do it by Extracting Features from the Date and Time Features.
- We can do it by Extracting Features from the Categorical Features.
- We can do it by Binning the Numerical and Categorical Features.
- We can do it by Aggregating Multiple Features together.

Categorical Encoding

- We already know that Machine Learning algorithms working only with Numeric Data.
- So, we have to encode our Object Data and Convert that into Numeric. so that Machine Learning Model can accept our Data.
- There are three Columns in our Dataset which needs encoding
  - Department
  - Gender
  - Education

Data Processing

- In this case we are going to perform the following Steps
  - Target Column Splitting.
  - Validation set Splitting from Training Data.
  - Statistical Sampling to make the data balanced.
- It is very Important to Process the data before feeding it to the Machine Learning Predictive Model to avoid any error.

Feature Scaling

- Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

- Feature Scaling basically helps to normalise the data within a particular range. Sometimes, It also helps in speeding up the calculations in an algorithm.

Predictive Modelling

- Predictive modeling uses statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred

- In this Case, we are going to use a Decision Tree, so let’s also know a little about Decision Trees also.

- A Decision Tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Why Decision Trees?

Let’s understand why we are preferring Decision Trees over other Machine Learning Algorithms.

1. It is easy to implement and Explain.
2. It works well with Datasets with categorical values, and we know that many columns present in the data are categorical in nature.
3. The Size of the Data is not too huge or not too small.
4. The Tree based Splitting will be good for separating the Promoted and unpromoted employees easily.

Performance Metrics

- We are going to use Confusion Matrix to analyze the Performance of the Model.
- Confusion Matrix is a table with 4 different combinations of predicted and actual values.
- It is extremely useful for measuring Recall, Precision, Specificity, Accuracy and most importantly AUC-ROC Curve.

Confusion Matrix

True Positive:

Interpretation: You predicted positive and it’s true.

You predicted that an employee is promoted and he/she actually is.

True Negative:

Interpretation: You predicted negative and it’s true.

You predicted that an employee is not promoted and he/she actually is not.

False Positive: (Type 1 Error)

Interpretation: You predicted positive and it’s false.

You predicted that an employee is promoted but he/she actually is not.

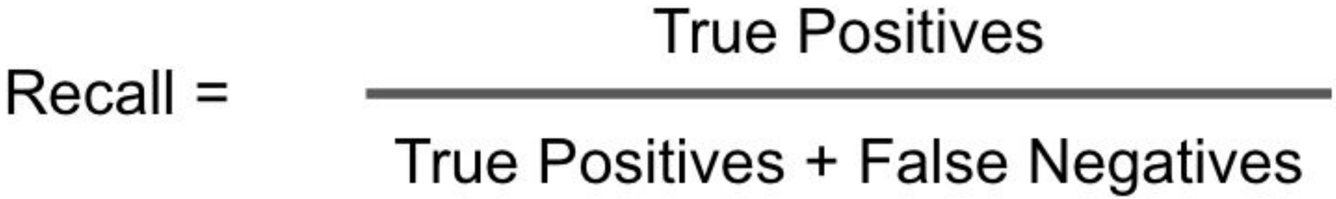
False Negative: (Type 2 Error)

Interpretation: You predicted negative and it’s false.

You predicted that an employee is not promoted but he/she actually is.

Recall

Out of all the positive classes, how much we predicted correctly. It should be as high as possible.



$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Precision

Out of all the positive classes we have predicted correctly, how many are actually positive.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

F1-Score

It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Possible Improvements

It is very Important to analyze the Improvements that we can perform to improve our Model Performance.

1. Creating some extra features, using the existing features can help us in achieving better results.
2. Instead of removing the region column, We can categorize 32 columns into two parts i.e., regions where employees have higher chances of promotion, and regions where employees have lower chances of promotion.
3. We can remove the gender column also, as we can see that there is a minute difference between the probability of a male and a female getting promotion.
4. As we are using Decision Trees, Feature Scaling is not an Important step, so we might ignore that.
5. We can tune our Decision Trees using Grid Search,for better accuracy and results.

Major Takeaways

- You came to know how to deal with Outliers in real life scenarios.
- You came to understand the Importance of Data Analysis and Visualization for understanding patterns, relation, association and most importantly Model Building.
- You also understood why it is important to perform feature engineering and remove unnecessary columns from the dataset.
- And the last, You understood, How to Handle Imbalanced Datasets for Predictive Analysis.