Project Proposal

Managing promotions effectively is one of the most powerful ways leaders can drive their company's success. Promotion's process allows leaders to evaluate each employee and their potential to be promoted. Some analysis needs to be conducted to estimate the probability of getting promotion based on some characteristics of the employee. Therefore, this project aims to perform predictive analysis to identify the employees most likely to get promoted.

Dataset

The data size is over 54K observations and obtains from Kaggle. The variables which are available in this dataset are:

- **Employee ID:** Unique ID for employee
- **Department:** Department of employee
- **Region:** Region of employment
- **Education:** Education Level
- **Gender:** Gender of Employee
- **Recruitment channel:** Channel of recruitment for employee
- **Number 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 variable):** Recommended for promotion

Tools

In order to achieve the goal of this project the following tools will be used:

Manipulation tools: NumPy and Panda will used for scientific computing. It provides support for large multi-dimensional arrays and matrices.

Visualization tool: Matplotlib and Seaborn will be used to visualize the features that are available in the dataset. This will help us to understand the nature of the features and how they are distributed. Moreover, Plotly will be used to create beautiful interactive web-based visualizations that can be displayed in Jupyter notebook.

Modelling tool: sklearn packages will be used to perform the analysis.

Work frame

This project is structured as following:

- **Exploratory data analysis**: In this section we aim to perform initial investigations on data. We will summarize their main characteristics and identify significant patterns. Each of graphical and non-graphical methods will be used in this section.

- **Formal data analysis:** In this section we aim to reach our goal of the project by perform Machine Learning models on the data. We will use two approaches to classify our target. The first one is the linear classification such as Logistic Regression. The second one is nonlinear classification such as Random Forest. Both models will be evaluated based on prediction accuracy and select the one with high accuracy.