Problem Statement: McCurr Consultancy – Attrition Analysis

Background:

McCurr Consultancy is an MNC that has thousands of employees spread across the globe. The company believes in hiring the best talent available and retaining them for as long as possible. A huge amount of resources is spent on retaining existing employees through various initiatives. The Head of People Operations wants to bring down the cost of retaining employees. For this, he proposes limiting the incentives to only those employees who are at risk of attrition. As a recently hired Data Scientist in the People Operations Department, you have been asked to identify patterns in characteristics of employees who leave the organization. Also, you have to use this information to predict if an employee is at risk of attrition. This information will be used to target them with incentives and in turn curb attrition.

Objective:

To predict if an employee is going to attrite or not. Apply feature engineering and model tuning to obtain 85% to 95% accuracy.

Steps and Tasks:

- ***Exploratory data quality report reflecting the following:***
  1. Univariate analysis – data types and description of the independent attributes which should include (name, meaning, range of values observed, central values (mean and median), standard deviation and quartiles, analysis of the body of distributions / tails, missing values, outliers **(10 Marks)**

  2. Bi-variate analysis between the predictor variables and between the predictor variables and target column. Comment on your findings in terms of their relationship and degree of relation if any. Visualize the analysis using boxplots and pair plots, histograms or density curves. **(10 marks)**

  3. Feature Engineering techniques **(10 marks)**
     a) Identify opportunities (if any) to create a composite feature, drop a feature (if required)
     b) Get data model ready and do a train test split.
     c) Decide on complexity of the model, should it be simple linear model in terms of parameters or would a quadratic or higher degree.

- ***Creating the model and tuning it***
  1. Algorithms that you think will be suitable for this project (one tree based and one bagging algorithm and one boosting algorithm). Use Kfold and Cross Validation to evaluate model performance. Use appropriate metrics and make a DataFrame to compare models w.r.t their metrics. **(15 marks)**

  2. Techniques employed to squeeze that extra performance out of the model without making it over fit or under fit. Use Grid Search or Random Search on any of the two models used above. Make a DataFrame to compare models after hyperparameter tuning and their metrics as above. **(15 marks)**

Data and Data Dictionary:

Data and data dictionary are attached separately with this assignment.