Project Title: Reducing Employee Attrition: A Data-Driven Approach

1. Introduction

Employees play a critical role in an organization's success. The quality of the workforce significantly impacts overall performance. However, organizations often face challenges related to employee attrition. Let's explore some of these challenges:

1. Costly Training: Hiring and training new employees can be expensive in terms of both time and money.

2. Loss of Experience: When experienced employees leave, the organization loses valuable knowledge and expertise.

3. Productivity Impact: High attrition rates can negatively affect productivity.

4. Financial Implications: Attrition can impact the company's profitability.

Before getting our hands dirty with the data, first step is to frame the business question. Having clarity on below questions is very crucial because the solution that is being developed will make sense only if we have well stated problem.

Framing Business Questions

Before diving into data analysis, it's essential to define clear business questions. Having well-stated problems ensures that the solutions developed are meaningful. Let's consider the following questions:

1. Factors Contributing to Attrition: What specific factors contribute to employee attrition?

2. Retention Strategies: What measures can the company implement to retain employees?

3. Business Value of Solutions: How will the proposed model benefit the organization?

4. Cost Savings: Can the model lead to significant cost savings?

5. Identifying Problem Areas: Which business units face the most significant attrition challenges?

2. Dataset Analysis

There are no empty values or duplicate rows in the dataset.

We observe that **'EmployeeCount'**, **'Over18'**, and **'StandardHours'** each have only one unique value, while **'EmployeeNumber'** has 1470 unique values. These features hold no utility for our analysis, thus we intend to remove these columns."

There are 8 column with categorical values apart from attrition column and 16 colums with numerical values. Total colums are 31.

The workers with low `JobLevel`, `MonthlyIncome`, `YearAtCompany`, and `TotalWorkingYears` are more likely to quit there jobs.

- `BusinessTravel` : The workers who travel alot are more likely to quit then other employees.

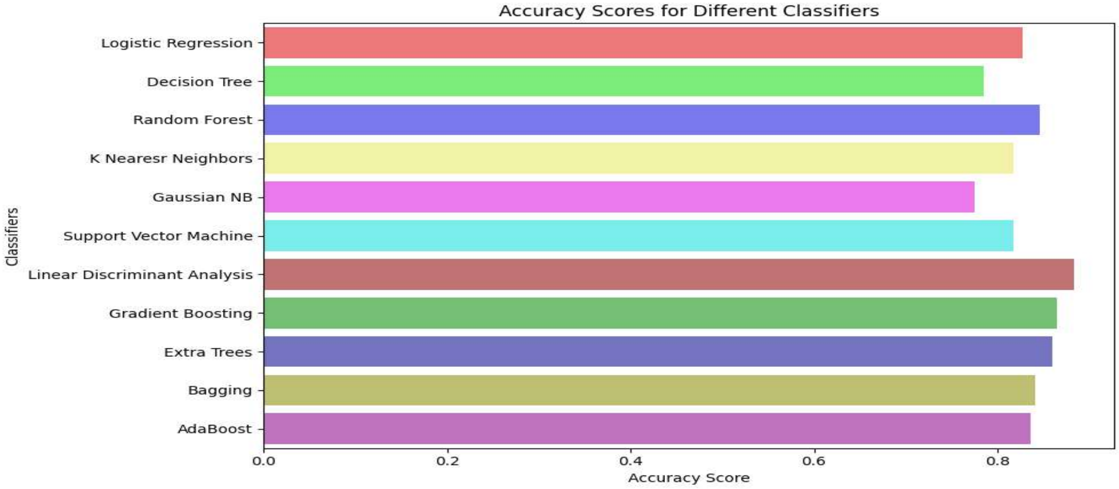
- `Department` : The worker in `Research & Development` are more likely to stay then the workers on other departement.
- `EducationField` : The workers with `Human Resources` and `Technical Degree` are more likely to quit then employees from other fields of educations.
- `Gender` : The `Male` are more likely to quit.
- `JobRole` : The workers in `Laboratory Technician`, `Sales Representative`, and `Human Resources` are more likely to quit the workers in other positions.
- `MaritalStatus` : The workers who have `Single` marital status are more likely to quit the `Married`, and `Divorced`.
- `OverTime` : The workers who work more hours are likely to quit then others.
- Monthly income is highly correlated with Job level.
- Job level is highly correlated with total working hours.
- Monthly income is highly correlated with total working hours.
- Age is also positively correlated with the Total working hours.
- Marital status and stock option level are negatively correlated

'MonthlyIncome', 'TotalWorkingYears', 'YearsAtCompany', ‘YearsInCurrentRole’ and 'YearsSinceLastPromotion' has outliers, we removed it using z-score method.

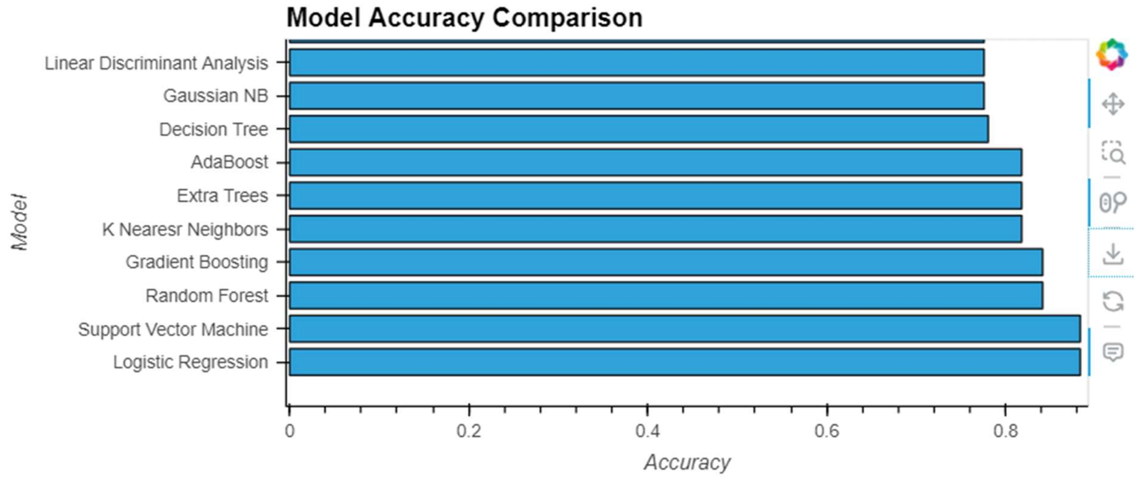
3. Outcome:

Accuracy of different ml models before hyperparameter tuning and after hyperparameter tuning is given below:

{'Linear Discriminant Analysis': 0.883177570093458, 'Gradient Boosting': 0.8644859813084113, 'Extra Trees': 0.8598130841121495, 'Random Forest': 0.8457943925233645, 'Bagging': 0.8411214953271028, 'AdaBoost': 0.8364485981308412, 'Logistic Regression': 0.8271028037383178, 'K Nearesr Neighbors': 0.8177570093457944, 'Support Vector Machine': 0.8177570093457944, 'Decision Tree': 0.7850467289719626, 'Gaussian NB': 0.7757009345794392}



```
{'Logistic Regression': 0.883177570093458, 'Support Vector Machine': 0.883177570093458, 'Random
Forest': 0.8411214953271028, 'Gradient Boosting': 0.8411214953271028, 'K Nearesr Neighbors':
0.8177570093457944, 'Extra Trees': 0.8177570093457944, 'AdaBoost': 0.8177570093457944,
'Decision Tree': 0.780373831775701, 'Gaussian NB': 0.7757009345794392, 'Linear Discriminant
Analysis': 0.7757009345794392, 'Bagging': 0.7757009345794392}
```



```
Logistic regression is giving the best accuracy(88%) over other ml models.
```