IBM HR Analytics Employee Attrition & Performance

Introduction:

Attrition, in Human Resource terminology, refers to the phenomenon of the employees leaving the company. Attrition in a company is usually measured with a metric called attrition rate, which simply measures the no of employees moving out of the company (voluntary resigning or laid off by the company). Attrition Rate is also referred as churn rate or turnover.

High attrition is a cause of concern for a company as it presents a cost to the company. The company loses on the amount it spent to recruit and select these employees and to train them for their respective jobs. The company may also have to spend additional money to fill the vacancies left open by these employees. Hence it becomes critical for a company to keep a tab on the attrition rate which downsizes the employee base.

Now it is possible not only to predict unwanted attrition, but also to have action plans to help reduce it, based on the organization's unique attributes.

With Machine Learning, it is possible to gain more insight than ever before into employee engagement, and build action plans to reduce unwanted attrition and achieve the organization's business goals.

Target audience:

This document intends to provide data visualizations to communicate data insights to the data analysts and scientists.

Overview of data:

The data is taken from Kaggle, an online community of data scientists and machine learners, owned by Google, Inc. This is a fictional data set created by IBM data scientists. It contains categorical data and numeric data. It contains factors that affect attrition. Some of the factors have yes or no answers while some of them contain weightages on scale of 1 to 4. It has 1470 data rows.

This document is divided into following sections:

1. Exploratory data analysis
2. Implementing Machine Learning model – Random Forest classifier
3. Result and outcome of analysis

1. Exploratory data analysis:

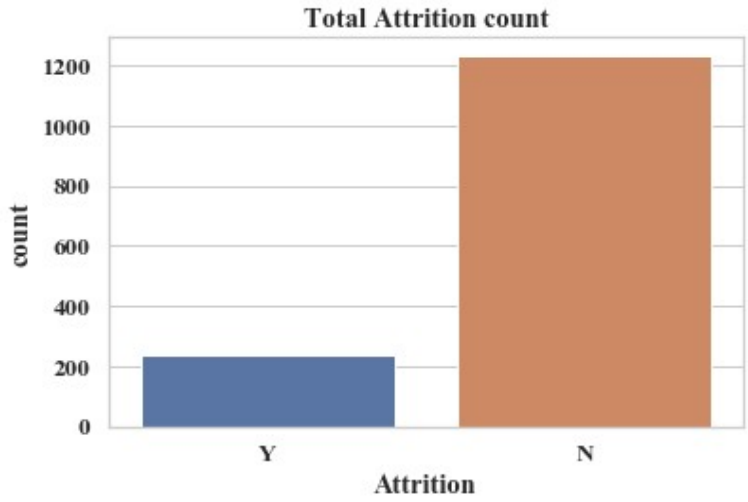
In this section, the dataset is explored by looking at the feature distributions, feature correlations, and creating some Seaborn visualisations.

List of visualisations:

- Attrition count plots
- Categorical plots between two features
- Feature correlation plot

Attrition count plots:

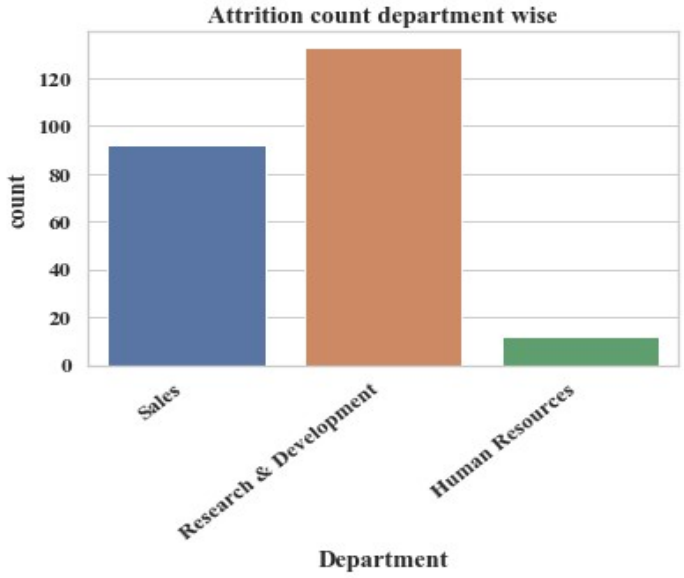
a. Yes vs No Attrition count:



This plot shows the total attrition count in the organisation. It can be seen that one class (N- NO attrition) is more than the other class (Y – Yes attrition). Hence there is a **class imbalance** in the data that needs to be dealt with in order to predict attrition.

Through Random Under Sampling technique (RUS) from Imblearn package, the class imbalance was treated.

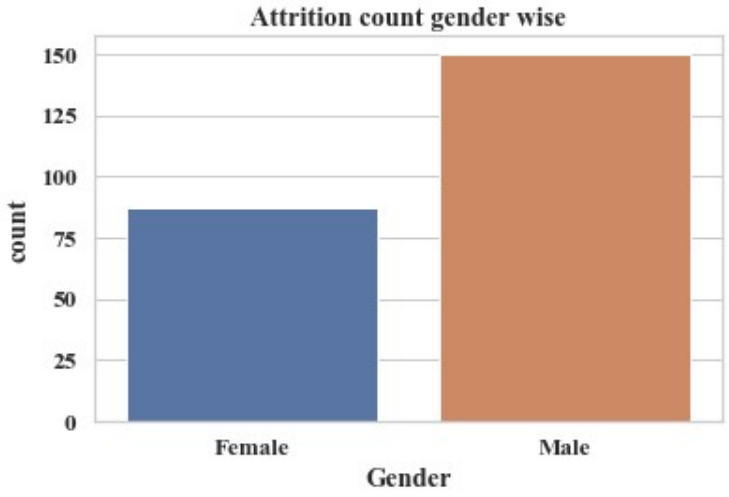
b. Attrition count department wise:



The Research & Development department sees a severe attrition followed by sales department. Hence employee concerns from these two departments must be immediately dealt with. Action plan must be taken to reduce the count.

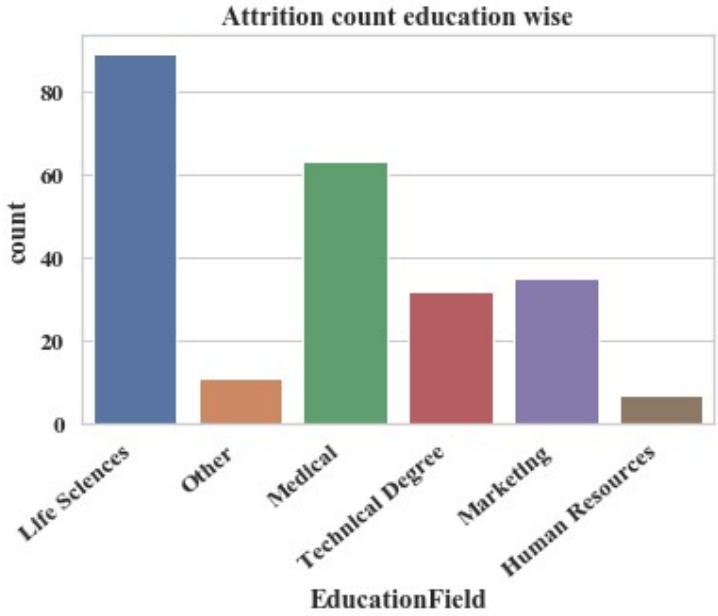
c. Attrition count gender wise:

Attrition is more common among men than women.



d. Attrition count education wise:

Employees with life sciences and medical background are more inclined towards leaving the organisation. The employees are those working in the R&D department where the highest attrition rate was noted, followed by marketing people in the sales department.



Categorical plots between two features:

a. Job satisfaction:

This plot shows the job satisfaction of the employees against number of years they have been working in the company. A scale of '1' means least satisfied and '4' is most satisfied. **Attrition is common among those with least job satisfaction. People who have worked for less than 5 years tend to leave company**, which means a good amount of money is spent in the process of losing one employee including the cost of initial training and appointing them. (Here we are not visualizing the cost factors.)

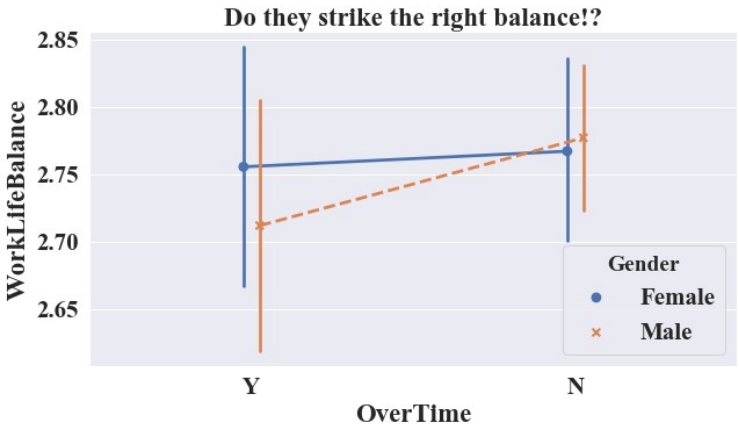
b. Employee manager relation:

Poor relationships with co-workers and management can cause many people to dread going to work each day. This friction between management and staff not only manifests itself in poor attitudes and morale, it can have a deleterious effect on a company's productivity and revenue.

The graph is interesting as it shows quite a dense plot for scale of 4 which means employee manager relationship is generally good in this organisation. But no employee has worked under a supervisor for more than 10 years.

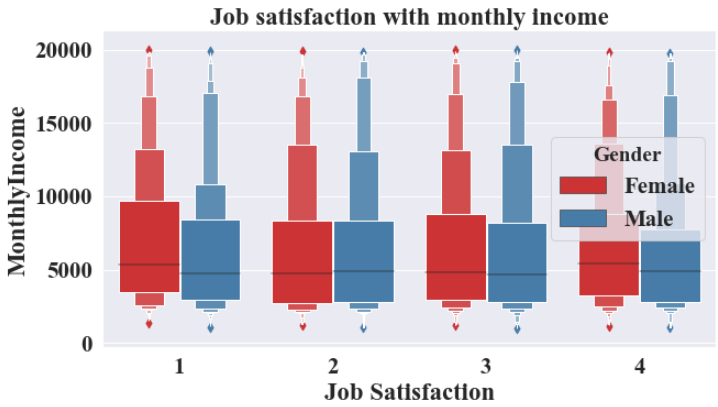
c. Overtime business:

Stress over the employees is one of the important factors of attrition. Overtime is, ironically, one of the least effective ways to meet deadlines. The men have poor work life balance score who does overtime than those who do not. This might be one of the cause of high attrition rate among them. Women have more or less the same central tendency irrespective of overtime.



d. Job satisfaction against monthly pay:

Salary satisfaction affects job involvement, work inspiration, employee performance and motivation. There is no significant disparities in job satisfaction among men and women although employees with higher pay have better job satisfaction than others.

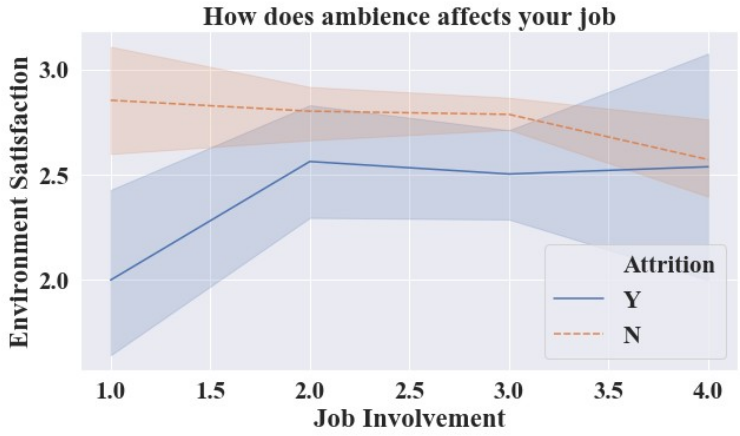


e. Promotion frequency:

The ‘No’ attrition red line is gradual and there are no peaks, which means these employees have been promoted within 2 or 3 years. The less promoted people are those in green line who are unmotivated and tend to leave the company. They have been in the current role for more than 10 years and they see no room for career development. Hence, it is essential for an organization to motivate their employees to work hard for achieving the organizational goals and objectives.

d. Impact of Working Environment on Job involvement:

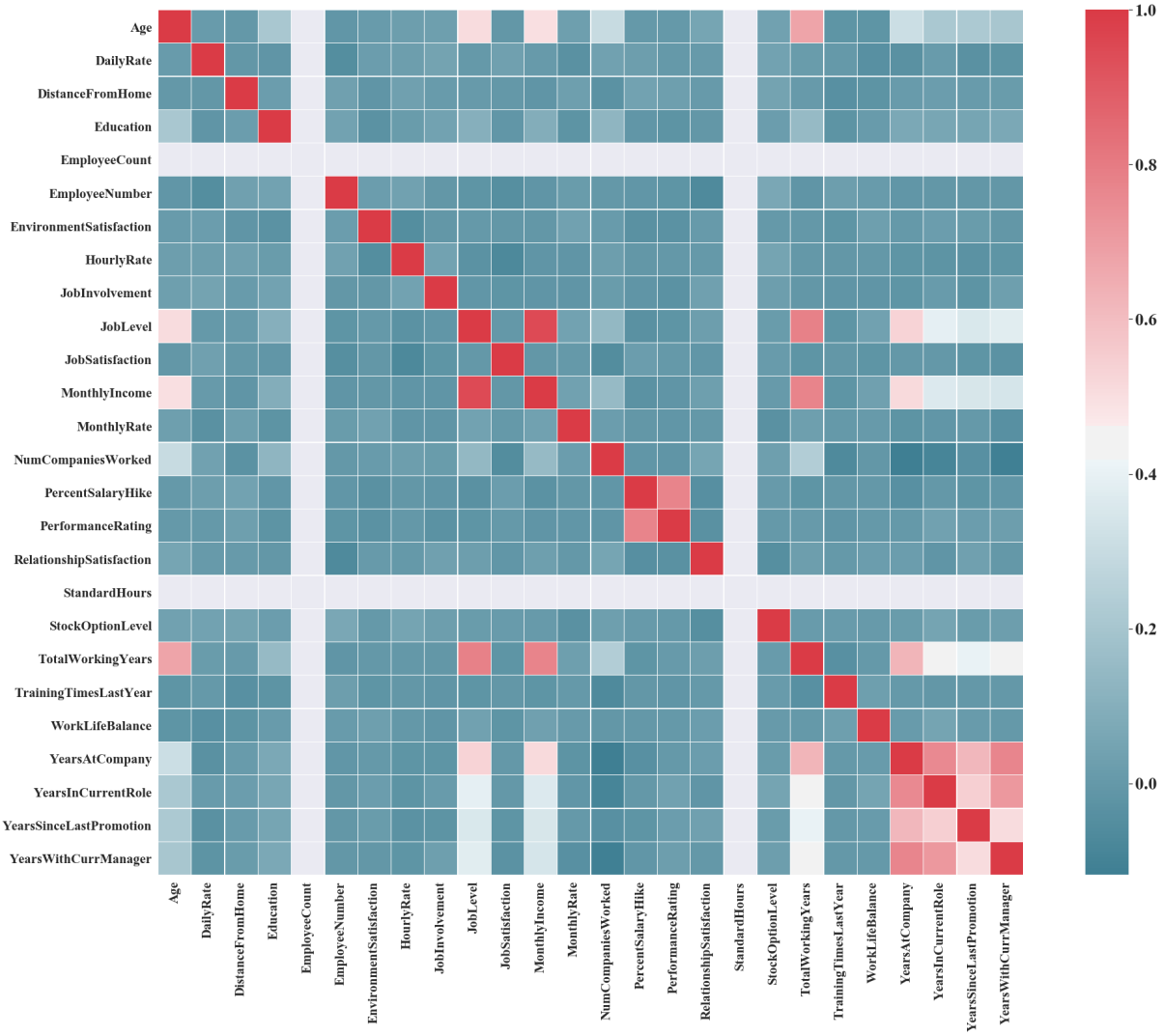
Employees who have given high scores for environment satisfaction perform job better. The organisation needs to realize the importance of good working environment for maximizing the level of job satisfaction.



Feature correlation plot

Correlation plots are awesome for exploratory analysis: it allows to quickly observe the relationship between every variable of the matrix. From the below plot, it is very clear that **features are correlated**. There is a strong correlation between job level and monthly income, percentage salary hike with performance rating, monthly income with respect to total service period of the employee and the like.

Hence, in order to deal with highly correlated features, the **Random Forest Classifier** is selected to predict attrition. Random Forest allows getting information on which **features are the most important** in the classification**. It has methods for balancing error in class population unbalanced data sets**. It computes proximities between pairs of cases that can be used in clustering, locating outliers or (by scaling) give interesting views of the data.

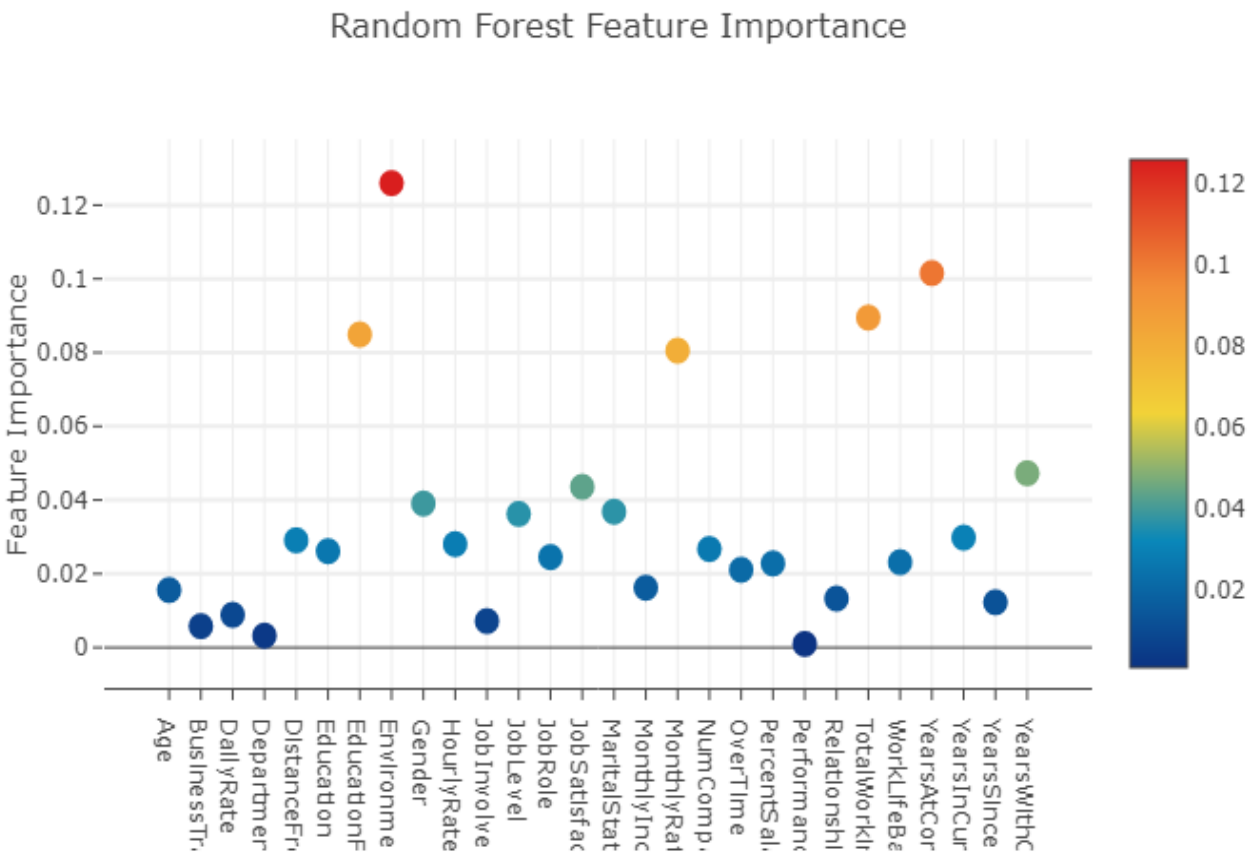


2. Implementing Machine learning model – Random Forest classifier

Feature importance plot gives estimates of what variables are important in the classification.

As can be seen from the below plot, **Environment satisfaction** plays the important role in determining attrition with an estimate of 0.125. Educational field, monthly rate, total working years and number of years at the company share equal importance between 0.1 and 0.08. The least factors contributing to attrition are age, business travel and department.

Since performance rating feature is correlated to percentage salary hike feature, we cannot completely rule out performance rating (0.0009) as percentage salary hike gives an estimate of 0.022. It is also slightly correlated to other features like job satisfaction, work life balance, years since last promotion and number of years in current role, which all have an importance estimate of 0.03 to 0.04 in determining attrition.



3. Result and outcome of analysis:

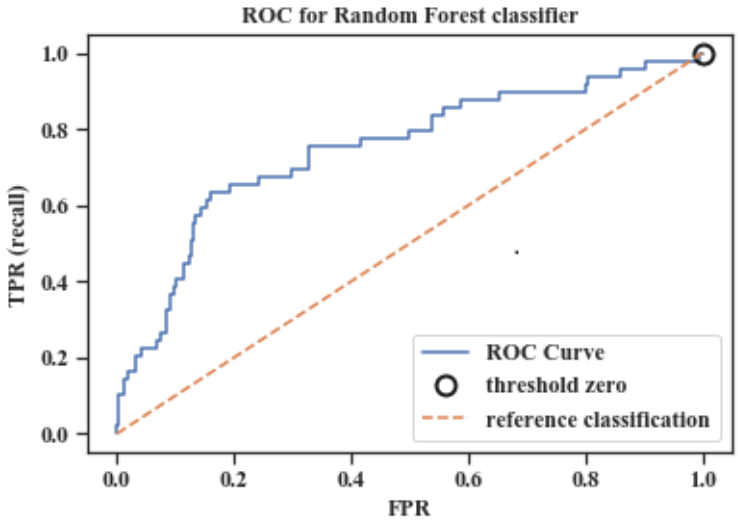
a. ROC Curve

```
Classification report:
                  precision    recall  f1-score   support

               0       0.91      0.86      0.88       245
               1       0.45      0.57      0.50        49

    micro avg          0.81      0.81      0.81       294
    macro avg          0.68      0.72      0.69       294
 weighted avg          0.83      0.81      0.82       294
```

In Machine Learning, performance measurement is an essential task. Therefore, when it comes to a classification problem, we can count on an AUC - ROC Curve. It is one of the most important evaluation metrics for checking any classification model's performance.

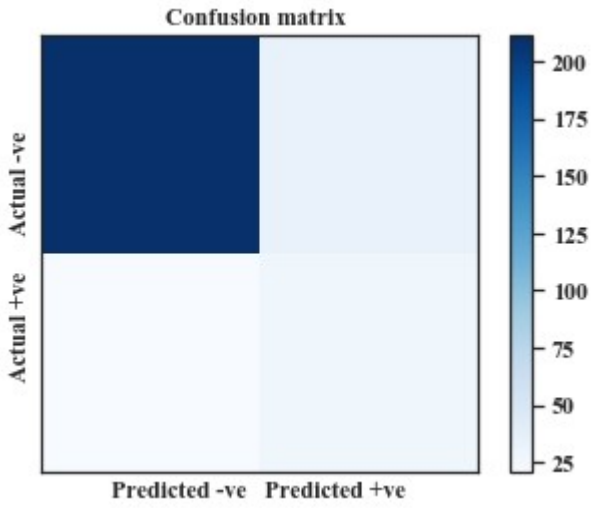


The recall factor for 0 and 1 are 0.86 and 0.57 respectively. Hence there is 57% chance of predicting correctly out of all positive cases(Yes- attrition) and 86% chance of predicting correctly out of all negaive cases(No- attrition). Precision for negative cases is better than positive cases. Since TPR increases as FPR increases, the developed model is stable and acceptable. The area under such curve is above 0.70 which means there is 70% chance that model will be able to distinguish between positive class and negative class.

b. Confusion Matrix:

```
Confusion matrix for above trained Random Forest Classifier
```

```
[[211  34]
 [ 21  28]]
```

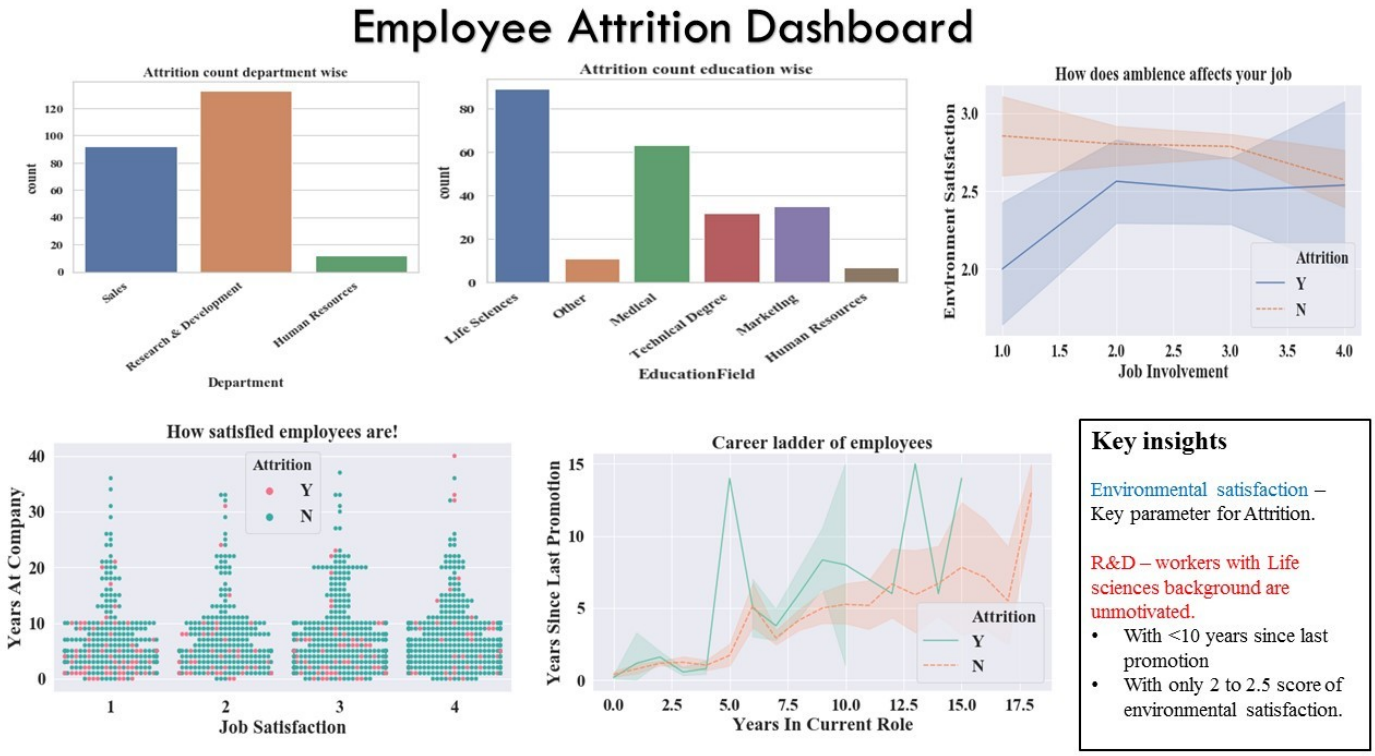


- ✓ Here the True negatives (actual and predicted No- attrition cases) are 211
- ✓ False positive (Actual No but predicted Yes attrition cases) are 34
- ✓ False negative (Actual Yes but predicted No attrition cases) are 21
- ✓ True positives (actual and predicted Yes- attrition cases) are 28

Conclusion 1:

From the employees perspective it can be concluded that employee turnover is not a natural phenomenon. There are reasons which lead to increase in attrition. It is observed that both the external and internal factors are responsible for employee turnover. Among the external factors opportunity for growth and promotion outside, location and work life space are the important. And among the internal factors compensation, work timing / shifts, working conditions, relations with supervisor / manager, opportunity to use skills, work load are important respectively.

From the perspective of managers, it can be concluded that the factors that lead to increase in employee turnover are majorly internal to the organisation. Although the external factors also influence, but as the management of the company does not have any control over the external factors it can focus on modifying the internal factors to enhance the retention of the employees in the organisation.



Key insights

Environmental satisfaction – Key parameter for Attrition.

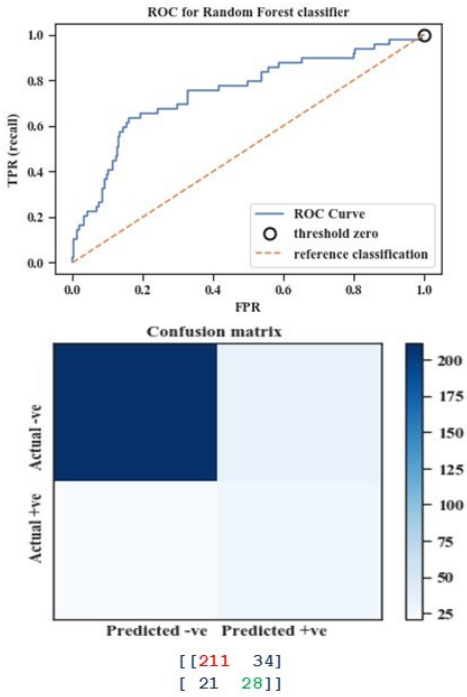
R&D – workers with Life sciences background are unmotivated.

- With <10 years since last promotion
- With only 2 to 2.5 score of environmental satisfaction.

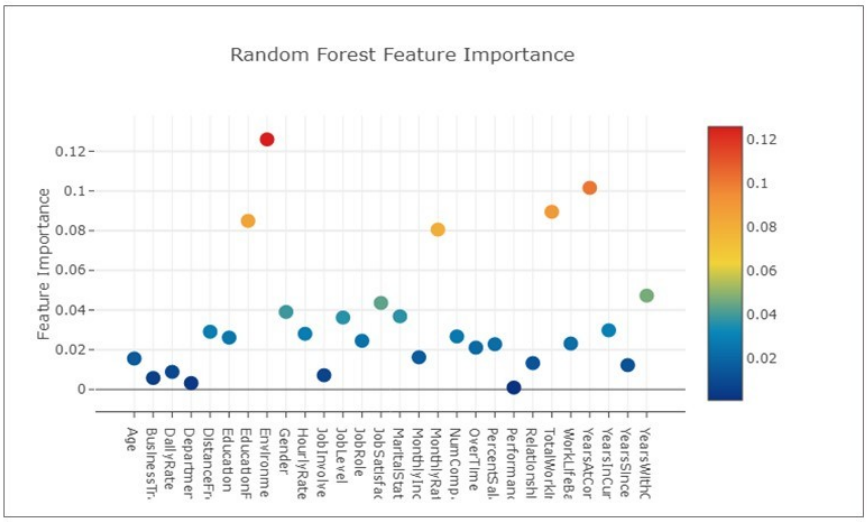
Conclusion 2:

Random Forest is a great algorithm to train early in the model development process, to see how it performs and it's hard to build a "bad" Random Forest, because of its simplicity. This algorithm is also a great choice, if we need to develop a model in a short period of time. On top of that, it provides a pretty good indicator of the importance it assigns to the features.

Random Forest performance report




```
[[211  34]
 [ 21  28]]
```




```
Classification report:
              precision    recall  f1-score   support

           0       0.91      0.86      0.88       245
           1       0.45      0.57      0.50        49
```