IBM HR Analytics Employee Attrition & Performance

By Rahul M Ramchandani

Exploring the factors affecting Attrition

Problem Statement:

Human Resources are critical resources of any organization. Organizations spend huge amount of time and money to hire and nurture their employees. It is a huge loss for companies if employees leave, especially the key resources. Reasons for attrition can be plenty and range from dissatisfaction due to low salaries, less or no career growth opportunities, inferior employee supervision, eagerness to get into companies with global presence, lack of recognition, lack of freedom of expression in the organization and underutilization of talents and skills of the individuals. Thus in a situation when more and more employees are quitting the organization, the attrition rate is on a rise.

So if HR can predict weather employees are at risk for leaving the company, it will allow them to identify the attrition risks and help understand and provide necessary support to retain those employees or do preventive hiring to minimize the impact to the organization.

Objective:

The objective of the present report is to study factors like salary, satisfactory level, growth opportunities, facilities, policies and procedures, recognition, appreciation, suggestions of the employee's by which it helps to know the Attrition level in the organizations and factors relating to retain them. This study also helps to find out where the organizations are lagging in retaining.

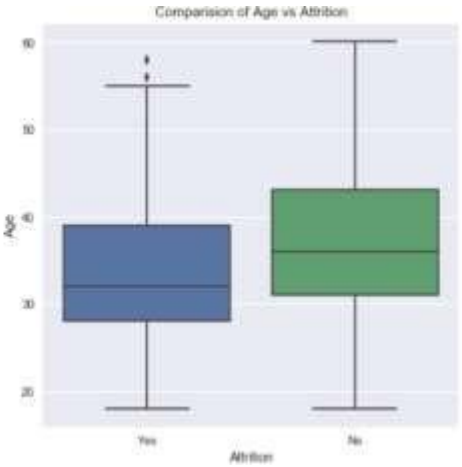
Hypothesis:

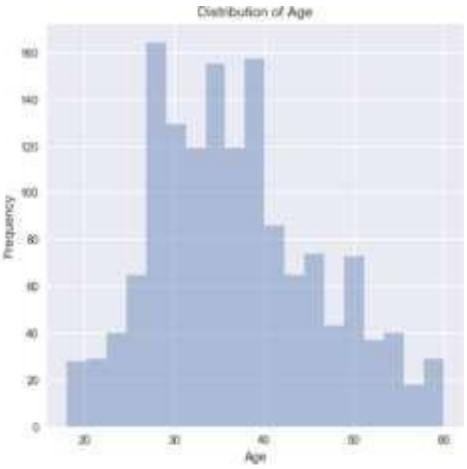
1. Employee attrition increases costs of recruitment, hiring and training replacement in the industries.
2. Employee attrition reduces production, and profit in the industries.

Exploratory Data Analysis:

Attrition Vs Age:

Hypothesis: Age has impact on attrition.



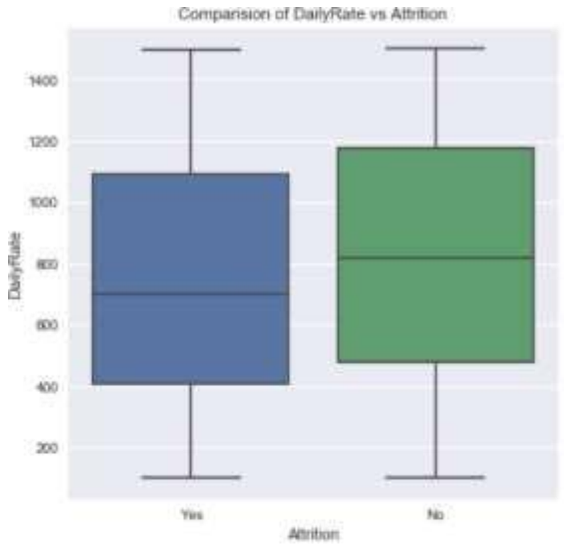


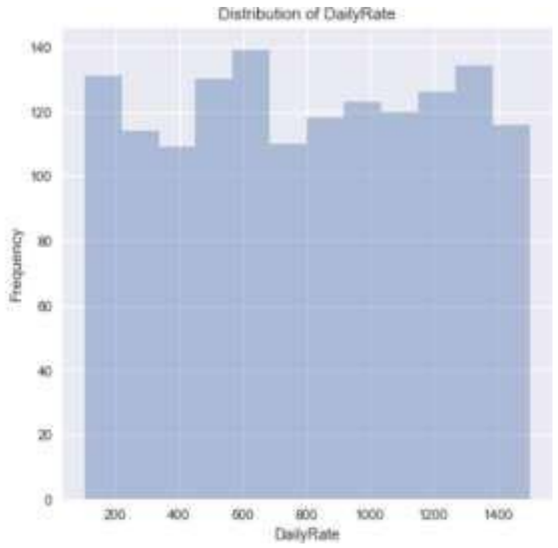
1. We found that median age of employee's in the company is 30 - 40 Yrs. Minimum age is 18 Yrs. and Maximum age is 60 Yrs.
2. From the Age Comparison boxplot, majority of people who left the company are below 40 Yrs. and among the people who didn't left the company are of age 32 to 40 years.

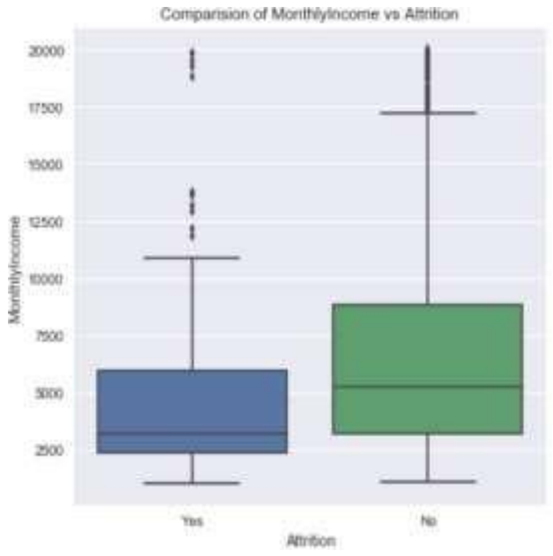
   Age has an effect on attrition. So it is considered as influential variable for attrition.

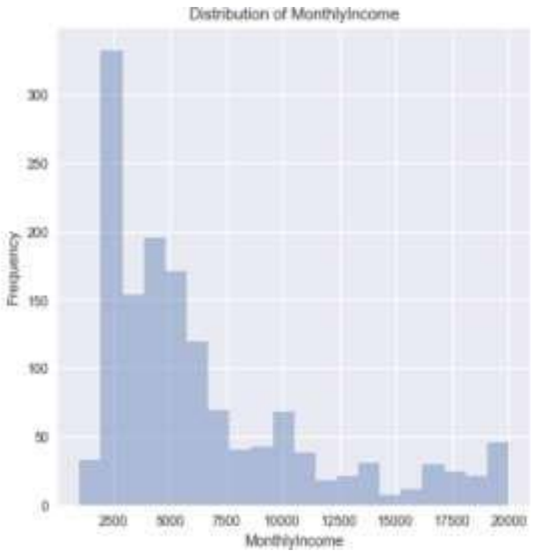
Attrition Vs Daily Rate, Monthly Rate, Hourly Rate:

Hypothesis: Hourly rate is influential variable of attrition

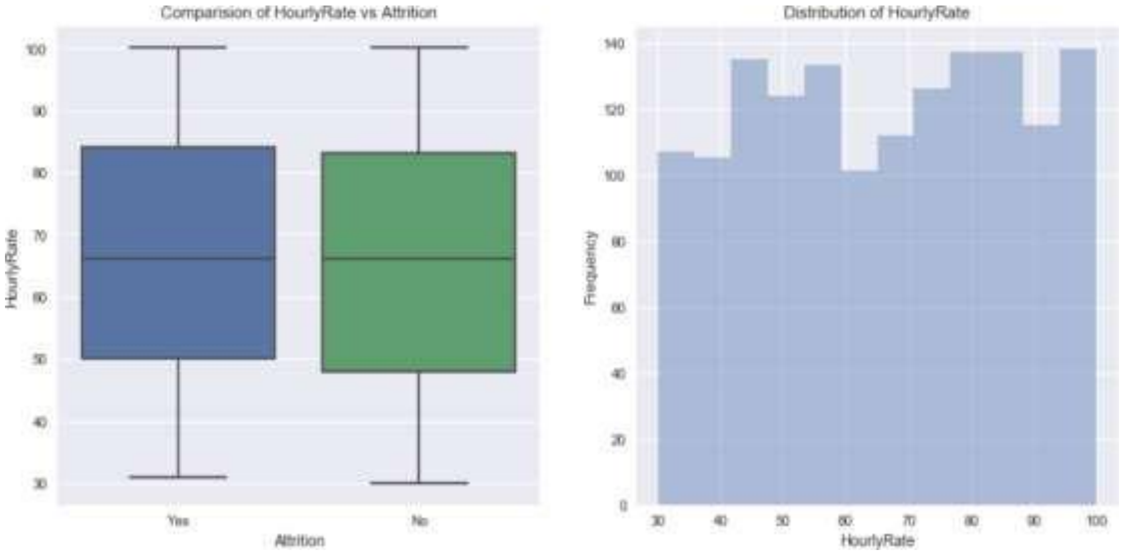








1. Employee's working with lower daily rates are more prone to leave the company than compared to the employee's working with higher rates. The same trend is resonated with monthly income too.

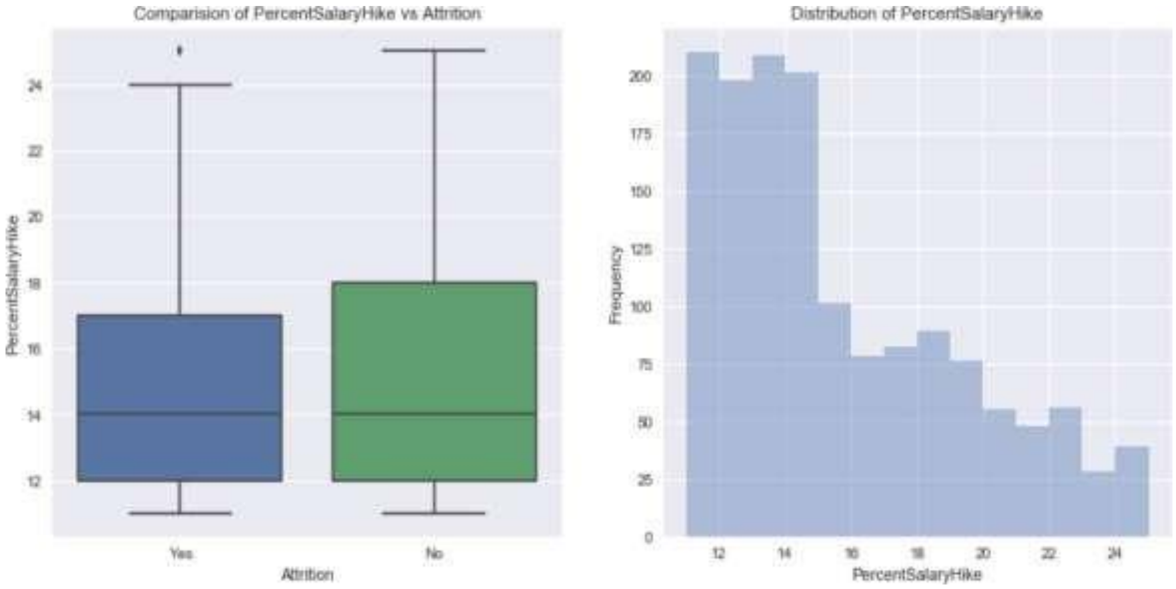


1. From plot we have seen that there is no significant difference in the hourly rate and attrition. Therefore hourly rate is considered as not significant to attrition

So it has no significant effect on attrition, hence it is not considered as important variable.

Attrition Vs Percent Salary Hike:

Hypothesis: Percent salary has significant importance on attrition.

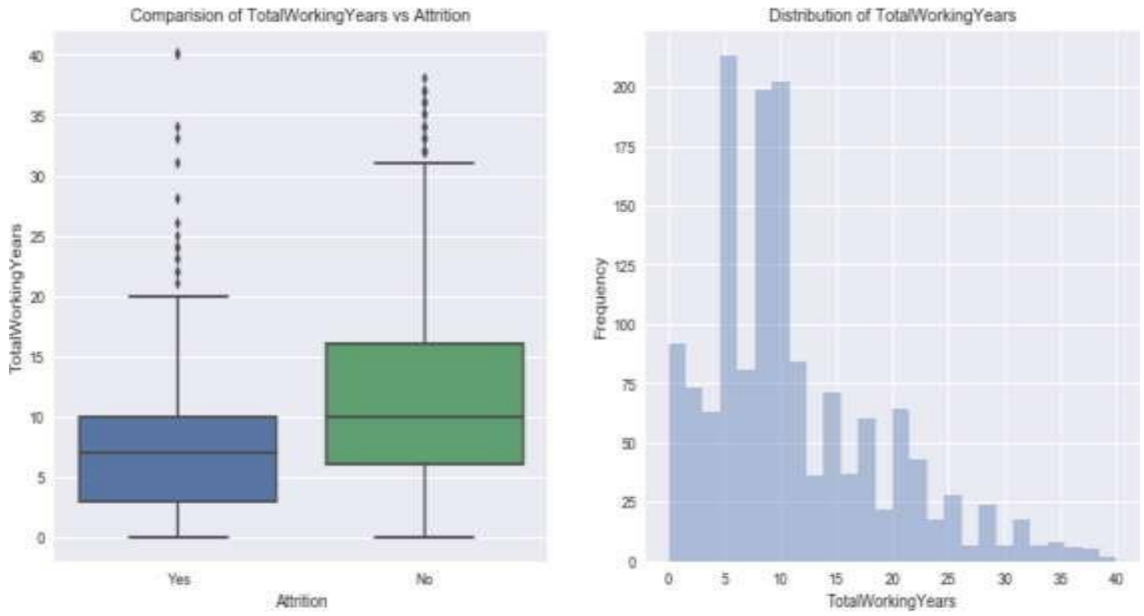


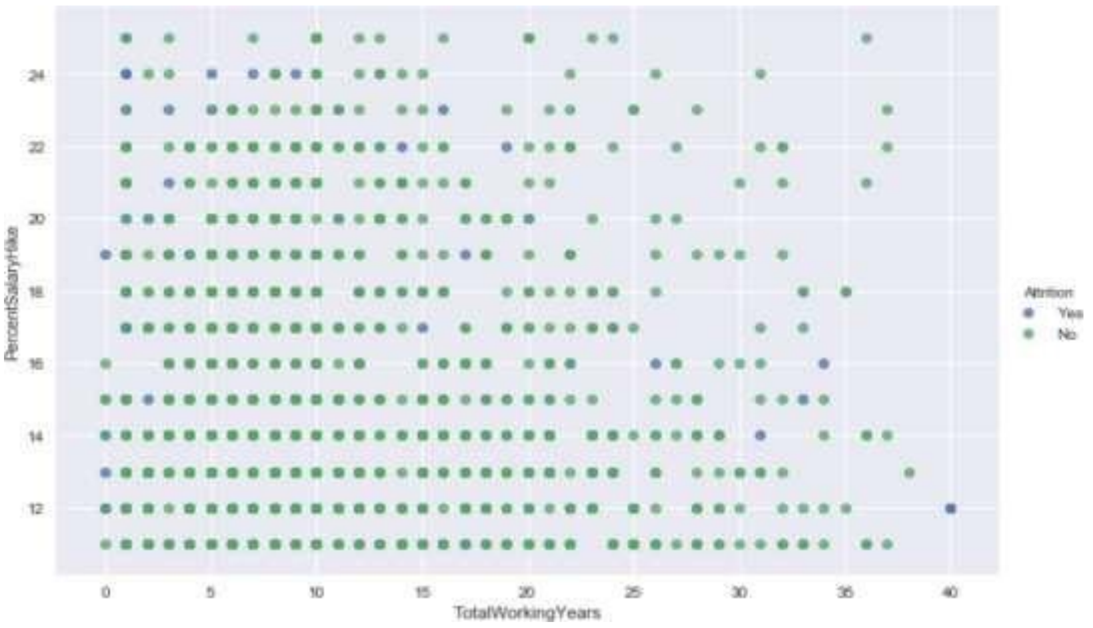
1. Majority (60% of total strength) of employee's receive 16% salary hike in the company, employees who received less salary hike have left the company.

So percent salary is considered has significanteffect on attrition and is considered as important variable

Attrition Vs Total Working Hours:

Hypothesis: Total working Years and percent salary hike together has an effect on attrition.



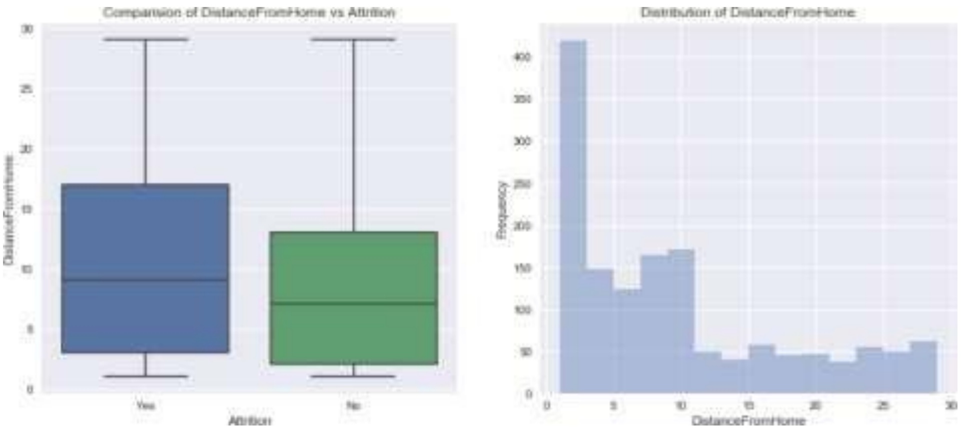


1. Employee's with less working years have received 25% Salary hike when they switch to another company, but there is no linear relationship between working years and salary hike.
2. Attrition is not seen among the employee's having more than 20 years of experience if their salary hike is more than 20%, even if the salary hike is below 20% attrition rate among the employee's is very low.
3. Employee's with lesser years of experience are prone to leave the company in search of better pay, irrespective of salary hike.

So percent salary hike and total working years together has and effect of attrition and is considered as important variable.

Attrition Vs Distance From Home:

Hypothesis: Distance from home plays an important role on attrition.

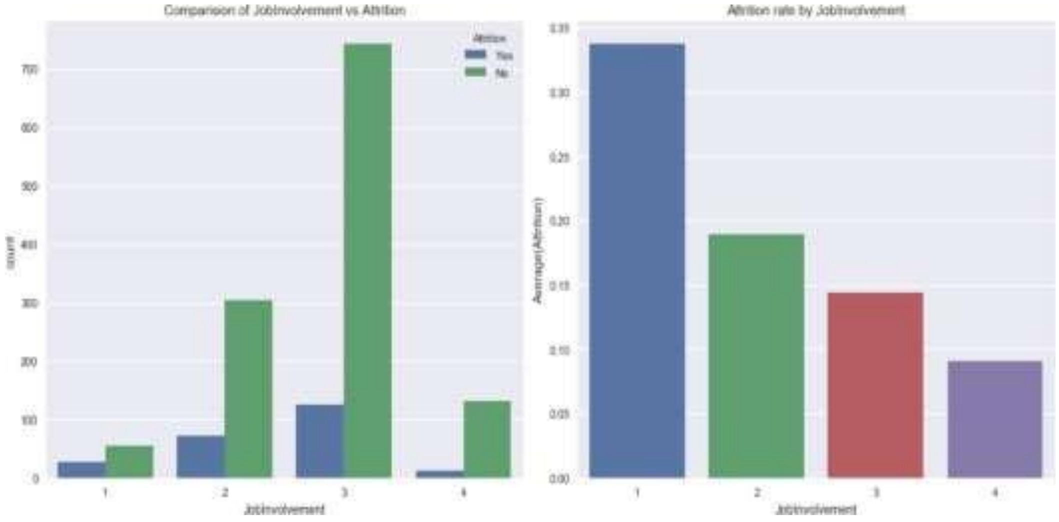


1. There is a higher number of people who reside near to offices and hence the attrition levels are lower for distance less than 10. With increase in distance from home, attrition rate also increases.

So it is observed that distance from home has significant effect on attrition.

Attrition Vs Job Involvement:

Hypothesis: Job involvement has significant effect on attrition

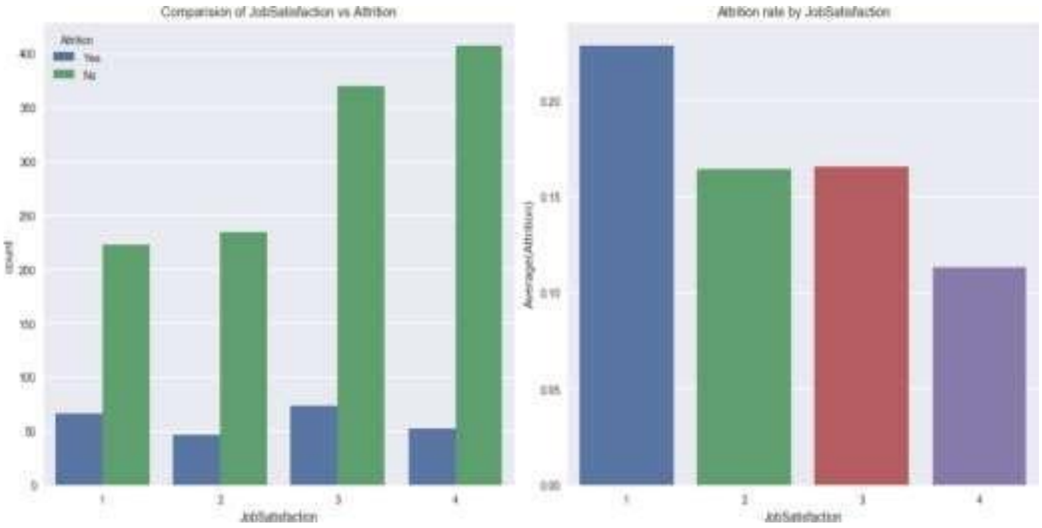


1. In the total data set, 59% have high job involvement whereas 25% have medium involvement rate
2. From above plot we can observe that round 50% of people in low job involvement (level 1 & 2) have left the company.
3. Even the people who have high job involvement have higher attrition rate around 15% in that category have left company.

The job involvement has important effect on attrition so it is considered as important variable

Attrition Vs Job Satisfaction:

Hypothesis: Job satisfaction plays an important role in attrition rate.

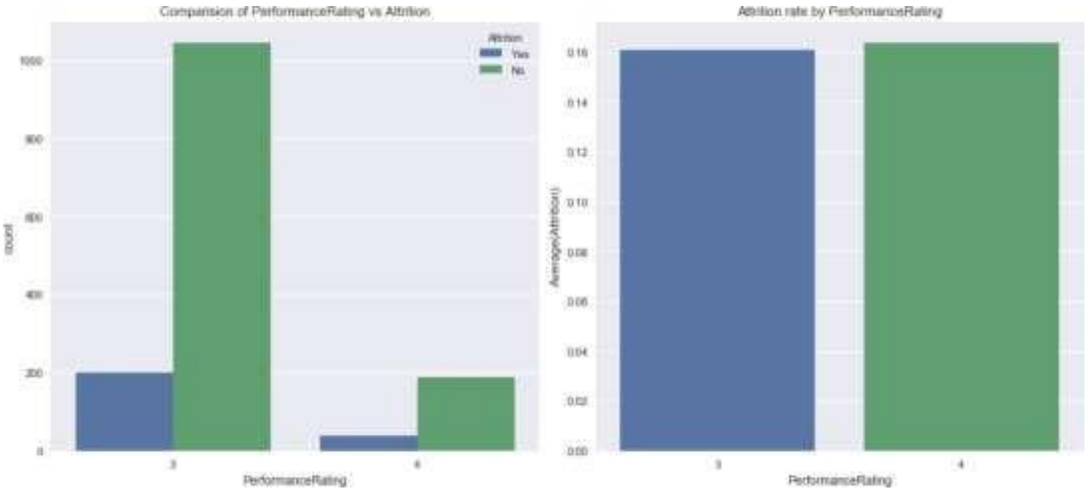


1. As expected, people with low satisfaction have left the company around 23% in that category. what surprising is out of the people who rated medium and high job satisfaction around 32% has left the company. There should be some other factor which triggers their exit from the company

Job satisfaction has significant effect on attrition so it is taken under consideration.

Attrition Vs Performance Rating:

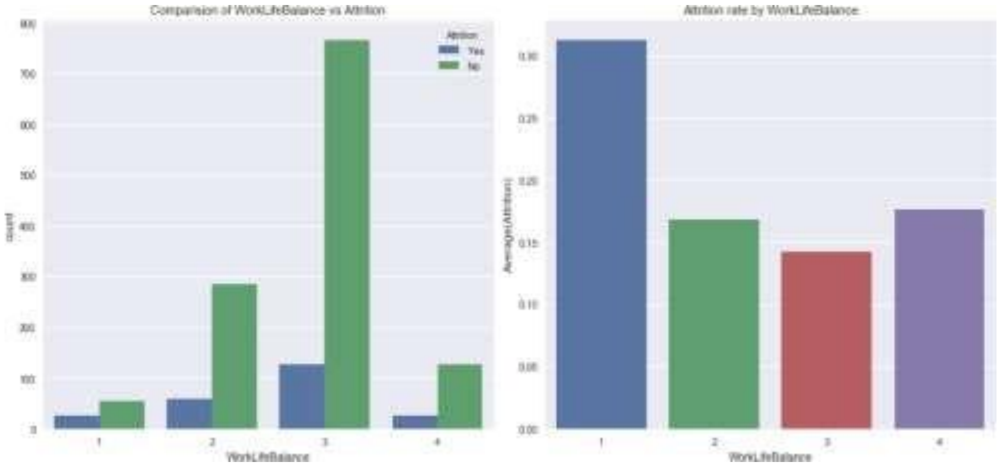
Hypothesis: Performance rating has significant effect on attrition



1. Contrary to normal belief that employee's having higher rating will not leave the company. It may be seen that there is no significant difference between the performance rating and Attrition Rate.

Attrition Vs Work Life Balance:

Hypothesis: Work life balance has direct impact on attrition.

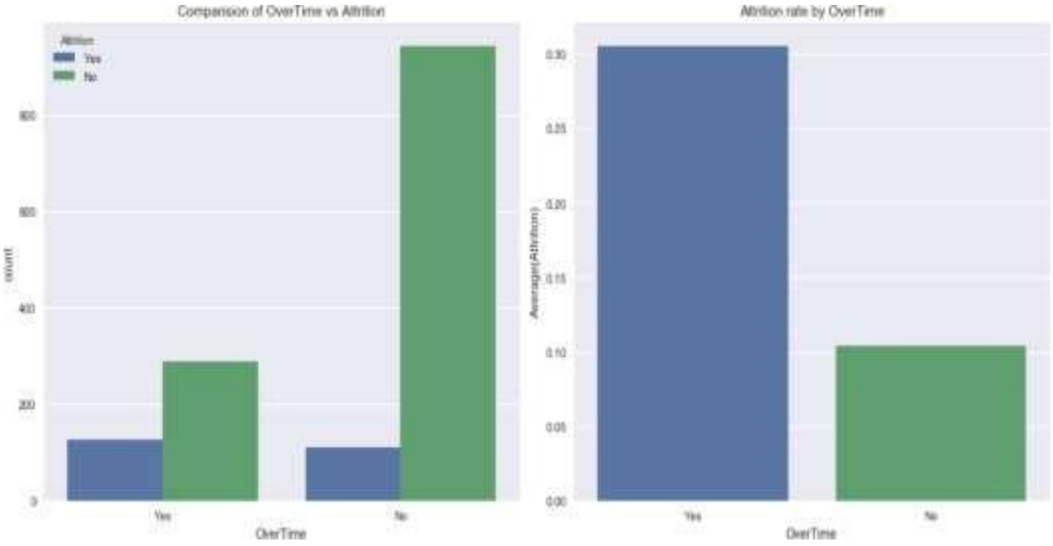


1. As expected more than 30% of the people who rated as Bad Work Life Balance have left the company and around 15% of the people who rated for Best Work Life Balance also left the company

Work life balance has an impact on attrition so it is the important variable to be considered.

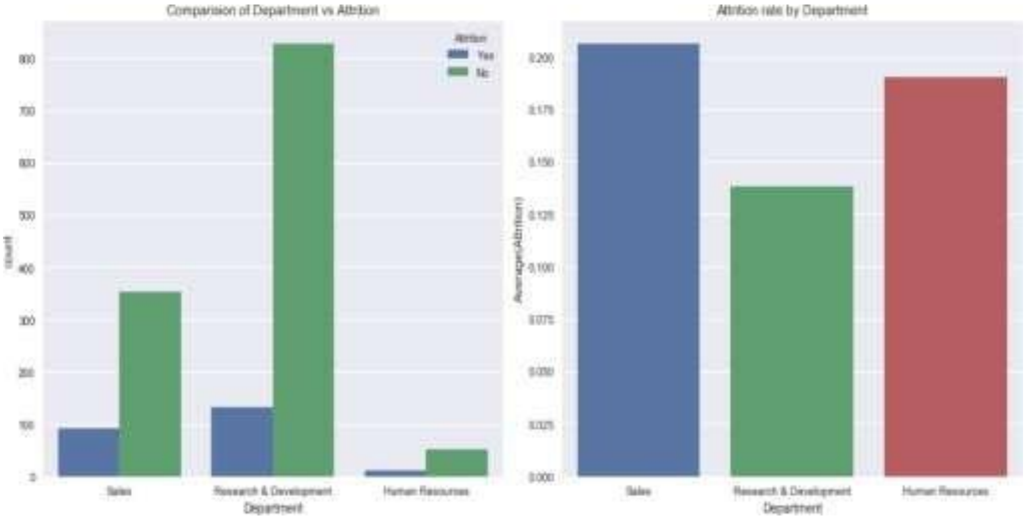
Attrition Vs Overtime:

Hypothesis: Overtime of employees may show impact on attrition rate.



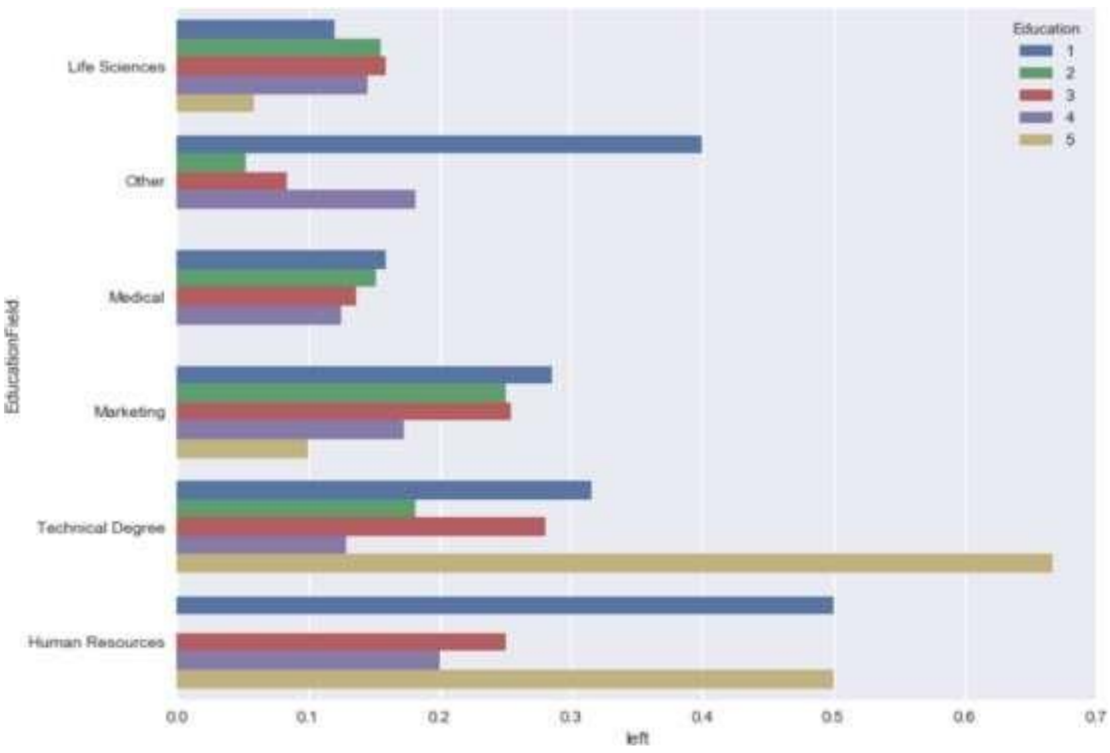
1. More than 30% of employee's who worked overtime has left the company, where as 90% of employee's who have not experienced overtime has not left the company. Therefore overtime is a strong indicator of attrition

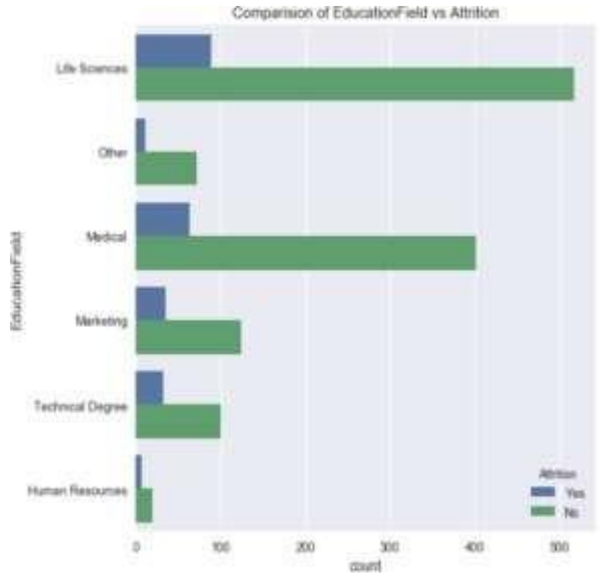
Attrition Vs Department:

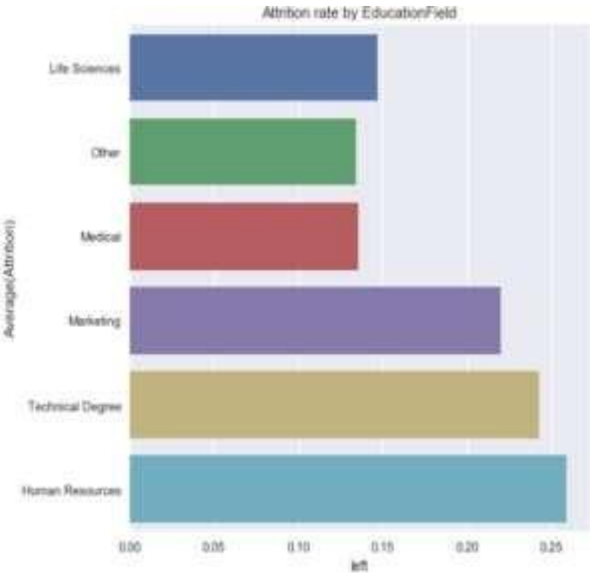


1. On comparing department wise, we can conclude that HR has seen only a marginal high in turnover rates whereas the numbers are significant in sales department with turnover rates of 39 %. The attrition levels are not appreciable in R & D where 67 % have recorded no attrition.
2. Sales has seen higher attrition levels about 20.6% followed by HR around 18%.

Attrition Vs Education Field:

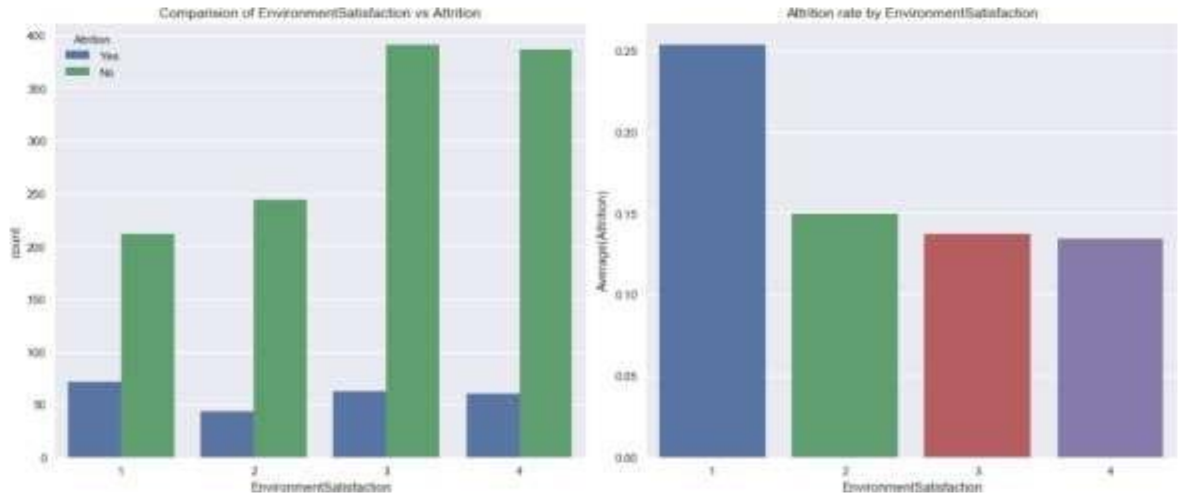






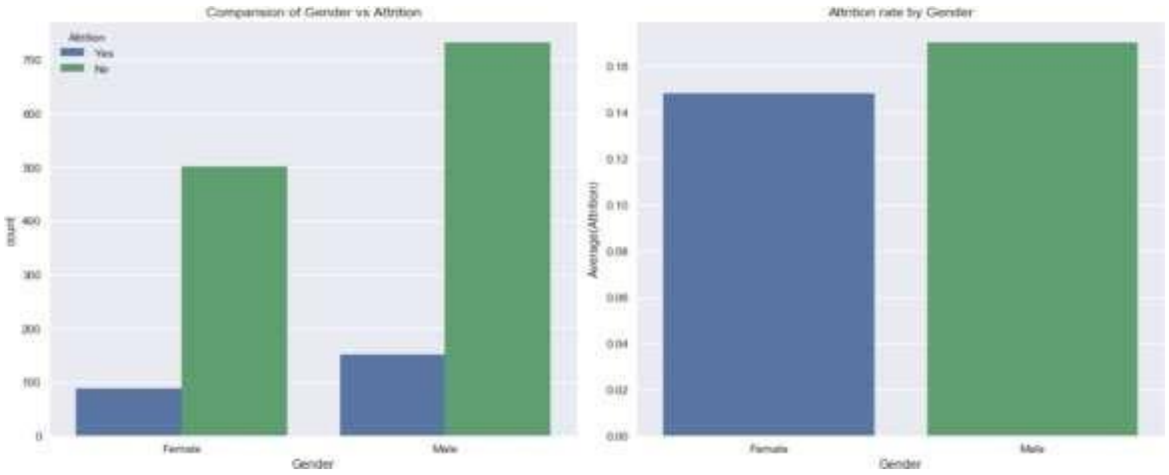
1. There are more people with a Life sciences followed by medical and marketing.
2. Employee's in the Education Field of Human Resources and Technical Degree have highest attrition levels around 26% and 23% respectively.
3. When compared with Education level, we have observed that employees in the highest level of education in their field of study have left the company. We can conclude that Education Field is a strong indicator of attrition.

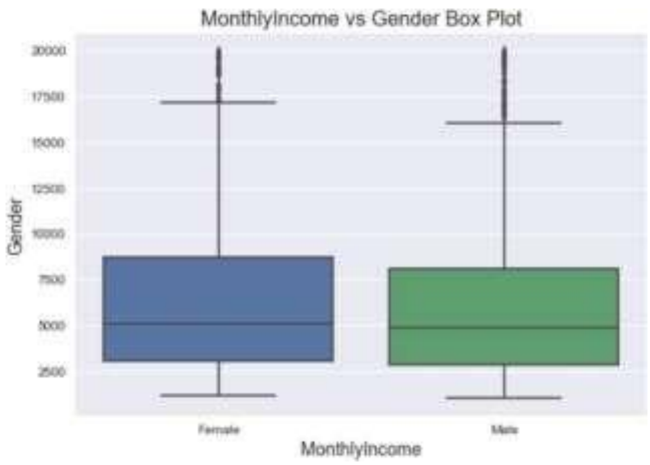
Attrition Vs Environmental Satisfaction:



1. We can see that people having low environment satisfaction 25% leave the company.

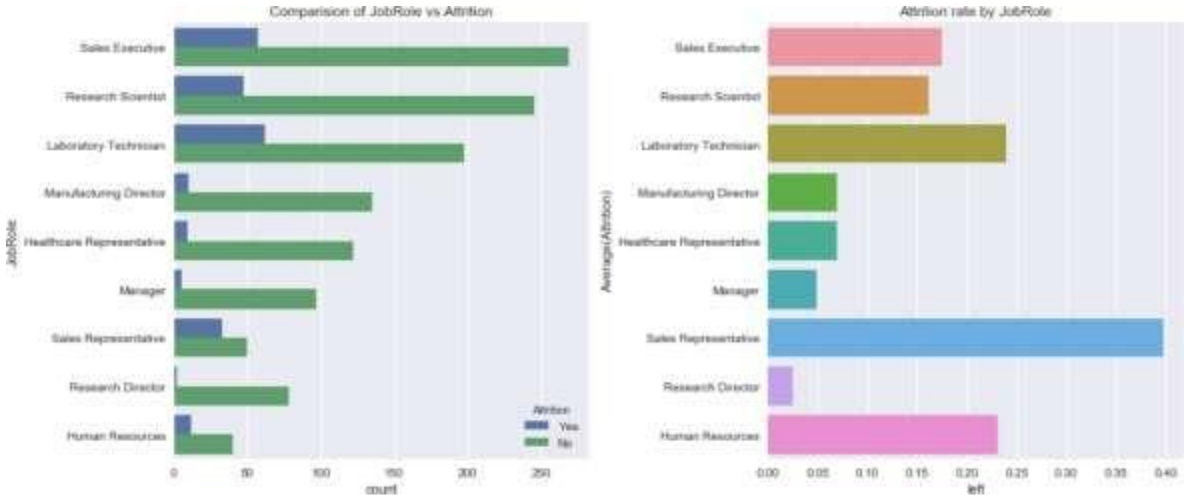
Attrition Vs Monthly Income & Gender:





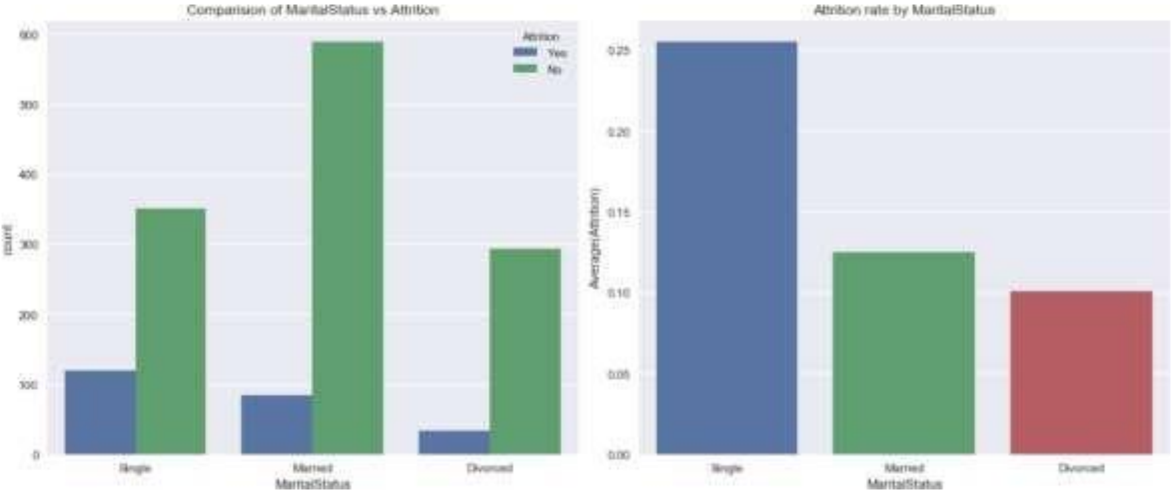
1. Monthly Income distribution for Male and Female is almost similar, so the attrition rate of Male and Female is almost the same around 15%. Gender is not a strong indicator of attrition.

Attrition Vs Job Role:



1. Jobs held by the employee is maximum in Sales Executive, then R&D , then Laboratory Technician
2. People working in Sales department is most likely quit the company followed by Laboratory Technician and Human Resources there attrition rates are 40%, 24% and 22% respectively.

Attrition Vs Marital status:



1. From the plot, it is understood that irrespective of the marital status, there are large people who stay with the company and do not leave. Therefore, marital status is a weak predictor of attrition

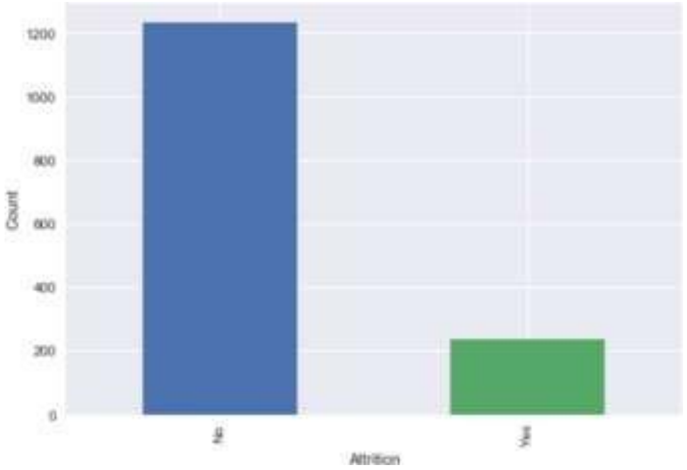
Decision Tree Modelling:

I have used Decision tree to create model. Decision Tree is a greedy algorithm it searches the entire space for possible decision trees. so we need to find an optimum parameter(s) or criteria for stopping the decision tree at some point. We use the hyperparameters to prune the decision tree.

By using grid search best parameters was found to be –

```
{'decisiontreeclassifier__max_depth': 3,
 'decisiontreeclassifier__max_features': 4,
 'decisiontreeclassifier__min_samples_leaf': 1,
 'decisiontreeclassifier__min_samples_split': 2}
```

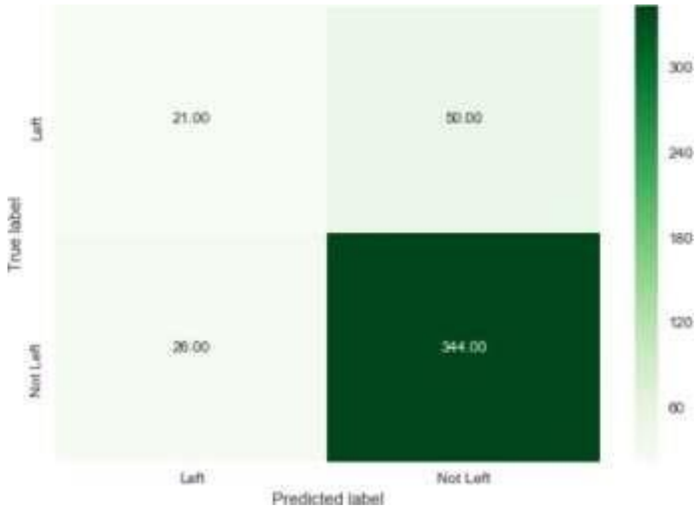
The major hurdle was we had 1223 unlabelled (No in Attrition) and 237 labelled (Yes in Attrition), which is a highly imbalanced data. So I used stratified samplingbased on proportionof attrition in overall data.



Model Evaluation:

I have used the k fold cross validation technique for assessing how the results of a model will generalize to an independent test data set. I used k =5 i.e.. 5 fold cross validation. Model was fit on the stratified sample data and tested on unmarked dataset.

Confusion Matrix: The confusion matrix is a way of tabulating the number of misclassifications, i.e., the number of predicted classes which ended up in a wrong classification bin based on the true classes.

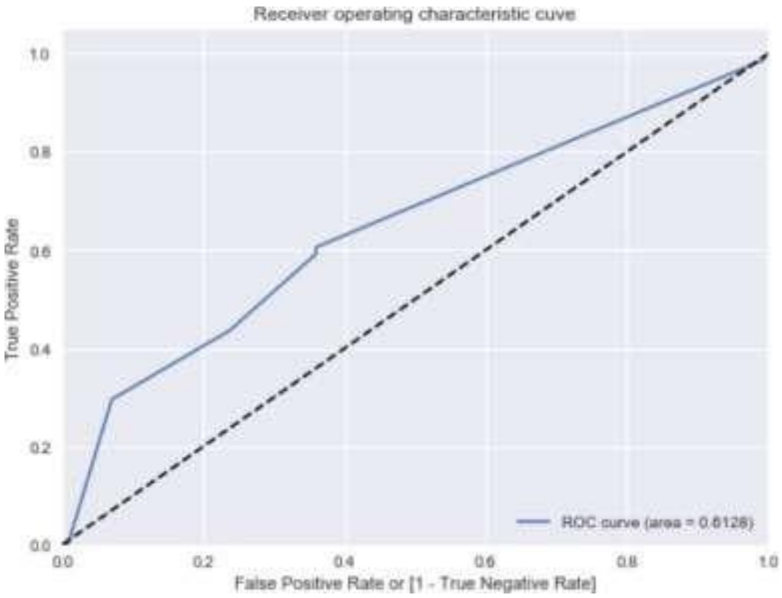


From the confusion matrix we have seen there are lot of misclassifications i.e.. 50 out of 71 were misclassified, this is due the fact that the data is highly imbalanced. Decision Tree algorithm is bias towards classes which have number of instances, here in this dataset we have more number of employees didn't leave the company. So the decision tree algorithm tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class.

In the cases like this accuracy measures tell the story that you might have excellent accuracy but the accuracy is only reflecting the overall accuracy of the data distribution. Instead of using accuracy as a performance metric we will use 'Precision', 'Recall' and ROC Curve.

ROC Curve:

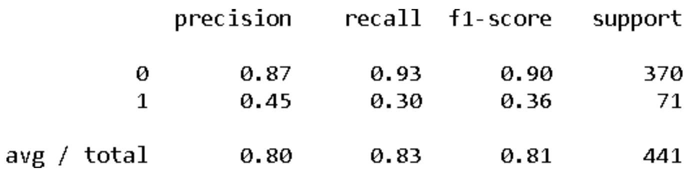
It is a curve between True Positive rate (Recall) and False Positive rate (1 – True Negative rate).



The ROC curve is a simple plot that shows the trade-off between the true positive rate and the false positive rate of a classifier for various choices of the probability threshold.

From the ROC Curve, we have a choice to make depending on the value we place on true positive and tolerance for false positive rate. If we wish to find the more people who are leaving, we could increase the true positive rate by adjusting the probability cut-off for classification. However by doing so would also increase the false positive rate. we need to find the optimum value of cut-off for classification.

From the classification report,



```
             precision    recall  f1-score   support

          0       0.87      0.93      0.90       370
          1       0.45      0.30      0.36        71

avg / total       0.80      0.83      0.81       441
```

Because of imbalanced data we got less f1 score