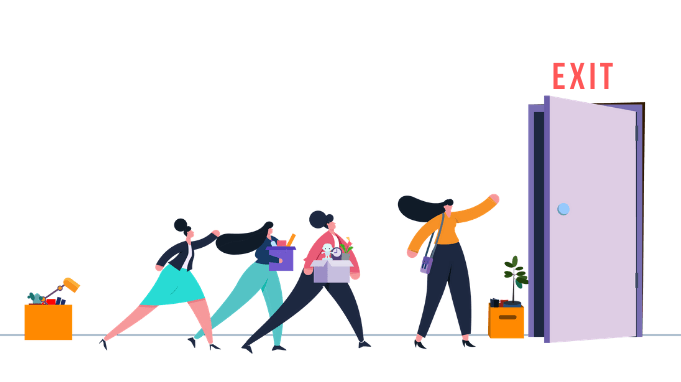
**Problem Definition**

**HR Analytics**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes



**Attrition**  
Employee attrition is when an employee leaves the company through any method, including voluntary resignations, layoffs, failure to return from a leave of absence, or even illness or death.



**Goal of the project**

The goal is to help the HR Analytics department of a company, by understanding the type of employees who have undergone attrition.Thereby making a model to understand the employees that may undergo attrition in the future.

**Problems caused to company due to Attrition**

1. the company has invested time and energy in building up human resource and when the employee leaves, the efforts value would diminish. Regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time.
2. if the business is customer-facing, as customers often prefer to interact with familiar people
3. Errors and issues are more likely if you constantly have new workers.
4. A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them.

**Buisness aim of the project**

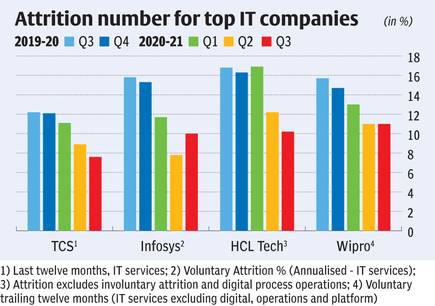
By understanding if a person will undergo attrition:

1.The HR can recruit those who will in future not undergo attrition, thereby avoiding the loss

2.The HR can focus on the employees that might show attrition by providing them benefits and also encouraging them through various means.

**Statistics of Attrition In India**

Attrition for different Indian companies varies between 7 per cent and 17 per cent. “Voluntary attrition for IT services calculated on an annualised basis increased to 15.2 per cent, as demand for talent increased,” Pravin Rao, COO, Infosys, told analysts while discussing the March quarter results 2021



**Models aim**

The aim of the model built will predict the Employee will undergo attrition or not.The model built to predict the Attrition will be aiming to obtain the highest Accuracy score. An accuracy score is fraction of predictions our model got right.

The model will also aim towards obtaining the highest ROC AUC score, precision score and recall.

**Methodology for execution of the modelling**

The data was cleaned and exploratory analysis was done

The data from the data set was split into the train and test data. The train data was then be checked for the best sample state. Using the sample state, the data was split again into the test and train data.

The data was tried to be balanced and it was balanced using under sampling.

This train data was modelled through various algorithms.The algorithms used here are Logistic regression, DecisionTreeClassifier, RandomForestClassifier and SVC. The model with least difference between accuracy score and Cross validation score was used for hyperparameter tuning and then the model was saved. The area under curve of the ROC (ROC AUC) will also be taken into consideration in model selection as a secondary criterion as it is important to distinguish between fraud and legit claims.

**Challenges in the project**

1. The dataset dependent variable is imbalanced
2. The sample size is small. Statistical models are more stable when data sets are larger. It also generalizes better as it takes a bigger proportion of the actual population.
3. the data only capture 3 department

**Criteria for success:**

The **model should be able to classify if attrition would occur or not on a data set**. The independent variables would be about the employee, through which the model would accurately predict if attrition is likely to occur or not. The model is successful if it can predict if employee/candidate will undergo attrition accurately.

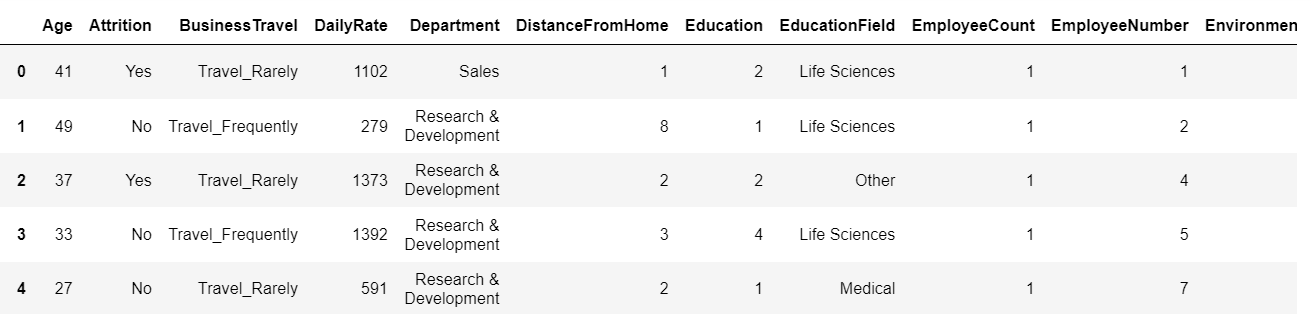
**Data Analysis**

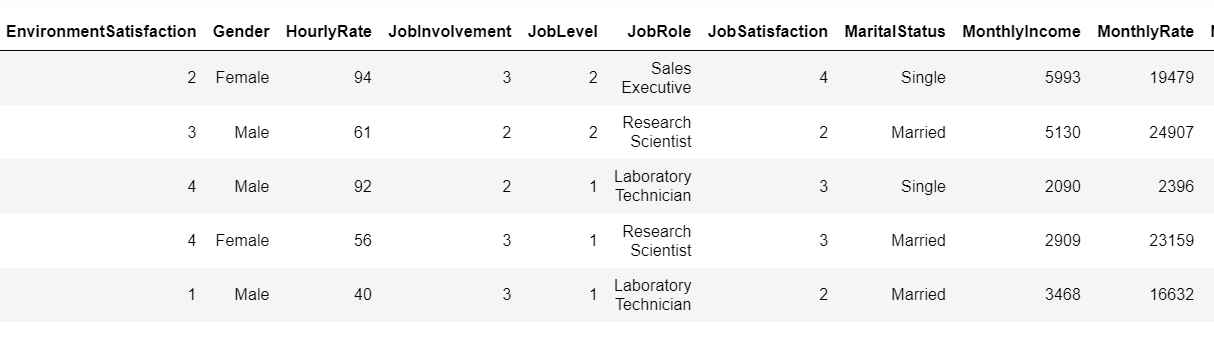
**Dataset Analysis**

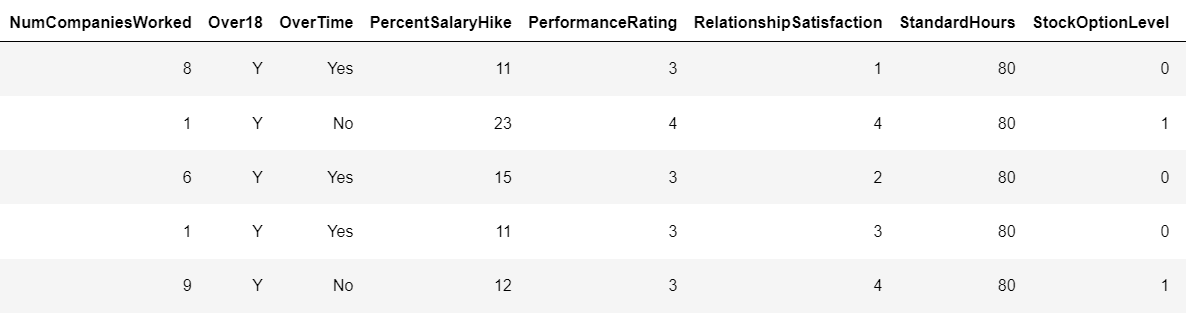
a.The dataset is obtained from:

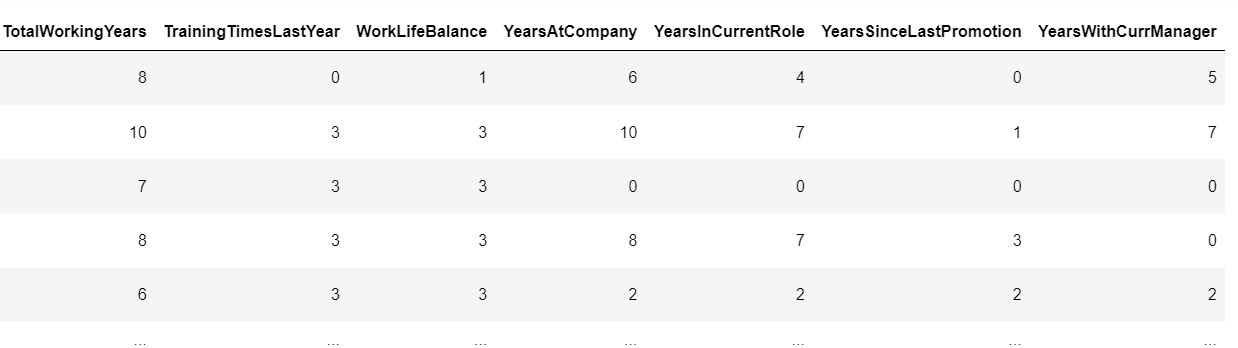
[**https://github.com/dsrscientist/IBM\_HR\_Attrition\_Rate\_Analytics**](https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics)

b.The dataset obtained is in CSV format

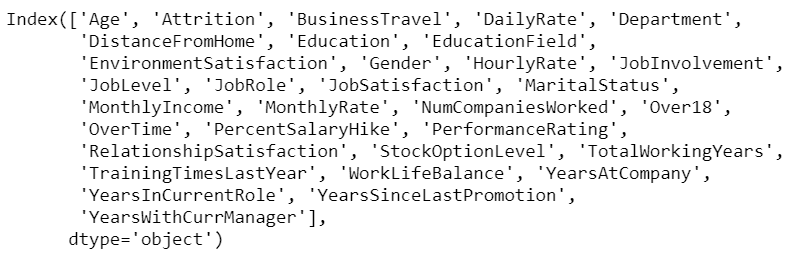








c.The column names/independent variables are :



d.Dataset has both categorical and numerical data. The label/dependent variable is categorical in nature.

e. There are 35 columns and the Meaning of each of the column

|  |  |
| --- | --- |
| Age | Age of employee |
| Attrition | Yes/no |
| BusinessTravel | Non Travel/Travel Frequently/Travel Rarely |
| DailyRate | Wage rate per day |
| Department | Department to which employee belongs  i.e, Human Resources,Sales and Research & Development |
| Distance From Home | Distance between office and employees house |
| Education Field | Education done by employee  i.e Human Resources,Other,Technical Degree,Marketing,Medical and Life Sciences |
| EnvironmentSatisfaction | Rating of environment satisfaction by employee between 1-4 |
| Gender | Employees gender |
| HourlyRate | Wage obtained per hour |
| Job Involvement | Involvement of employee in the job(rated between 1-4) |
| Job Level | Level of the job the employee performing(level 1-5) |
| Job Role | Role in the job, the employee is performing Human Resources/Research Directo/Sales Representative/Manager/Healthcare Representative /Manufacturing Director/Laboratory Technician/Research Scientist/Sales Executive |
| MaritalStatus | Divorced/Single/Married |
| MonthlyIncome | Monthly income of employee |
| NumCompaniesWorked | Number of companies employee worked before |
| Over18 | If employee is above 18years of age |
| OverTime | If the employee works overtime |
| PercentSalaryHike | The percentage of salary hike given to the employee |
| PerformanceRating | Rating of the employee |
| Relationship Satisfaction | Relationship Satisfaction of employee |
| StockOptionLevel | The level the employee is to buy stocks |
| TotalWorkingYears | Number of years employee has been working anywhere |
| TrainingTimesLastYear | Number of times employee took training last year |
| WorkLifeBalance | Employees rating on work life balance |
| YearsAtCompany | Number of years employee has been working for this company |
| YearsInCurrentRole | Number of years the employee has been in current role |
| YearsSinceLastPromotion | Number of years since employee got promoted |
| YearsWithCurrManager | Number of years the employee has been with the current manager |

**Statistical analysis of dataset**

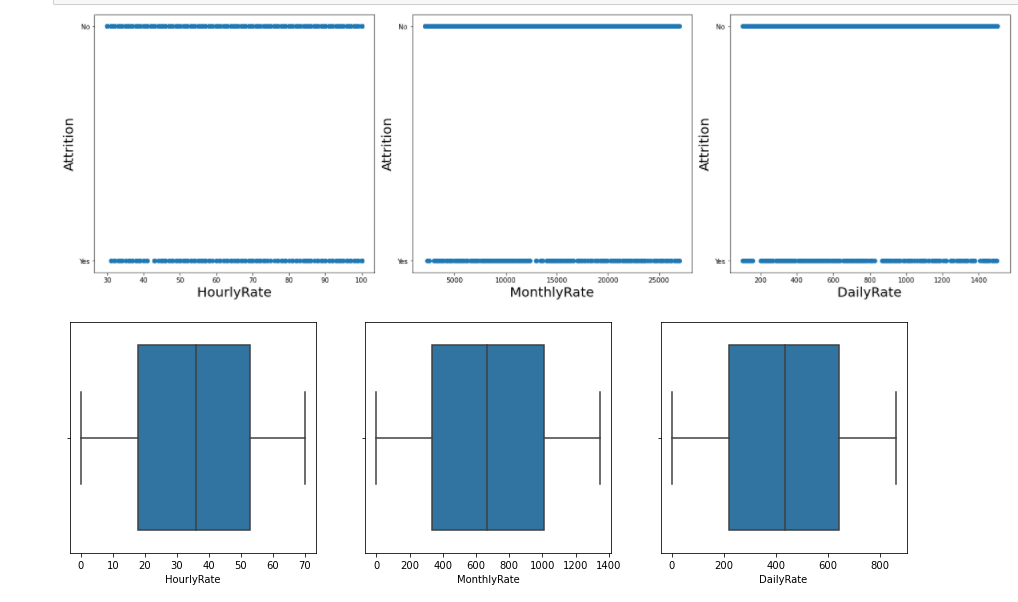
1.There is no column with single unique values

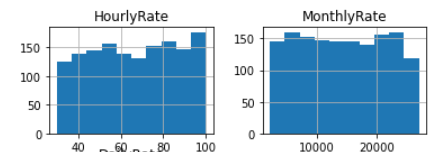
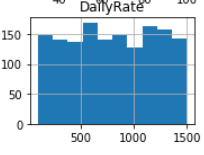
2.Null values is not present as all of them have 1000 datas

3.Outliers are seen in many columns as there is large difference between the mean and 50% data

**Graphical analysis of dataset**

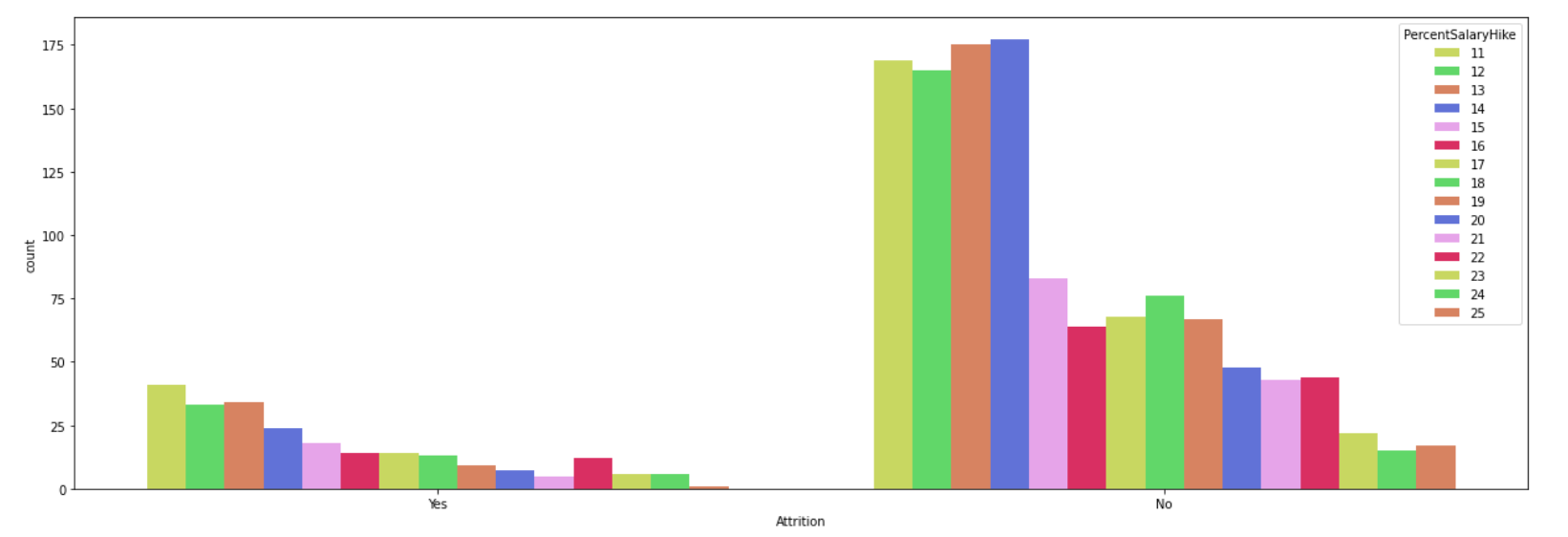
**1.Income of employee**

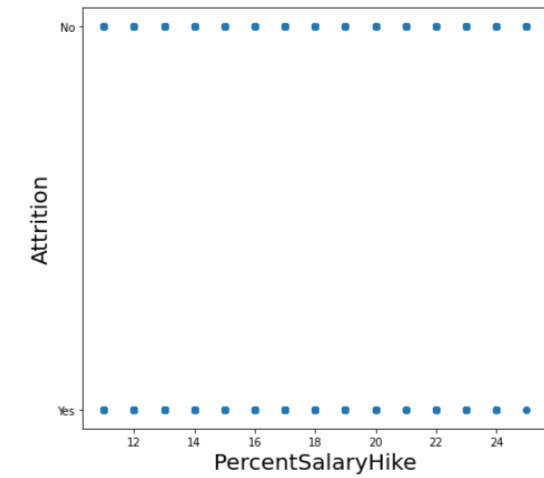


The attrition doesn’t depend much on the hourly, monthly and daily rate

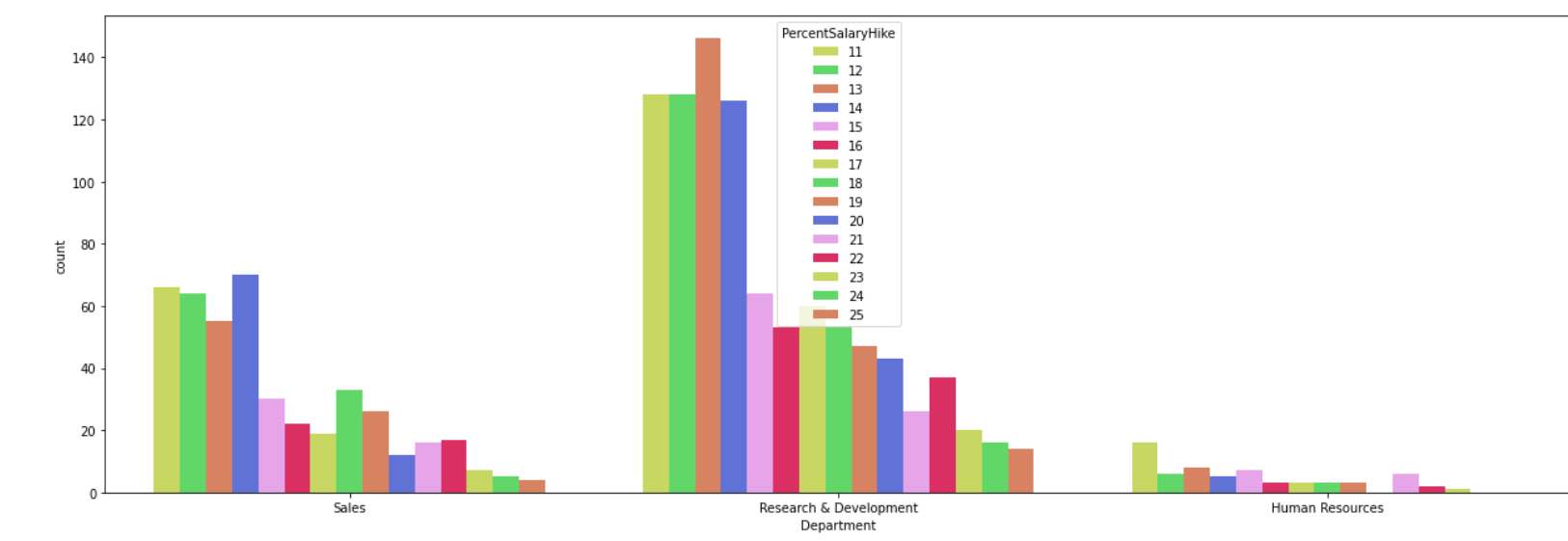
**Salary Hikes**

****

****

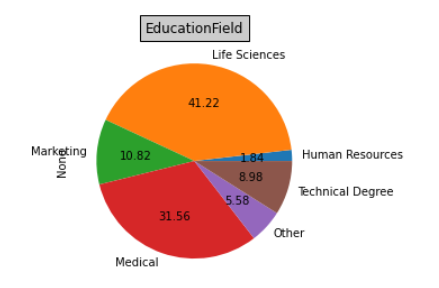
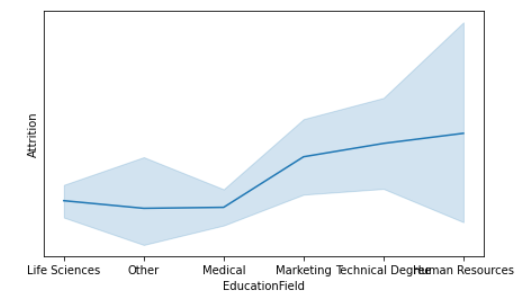
Salary hikes have been in the range of 11 to 25%. Where mostly employees were given 11,12,13 and 14% hike.

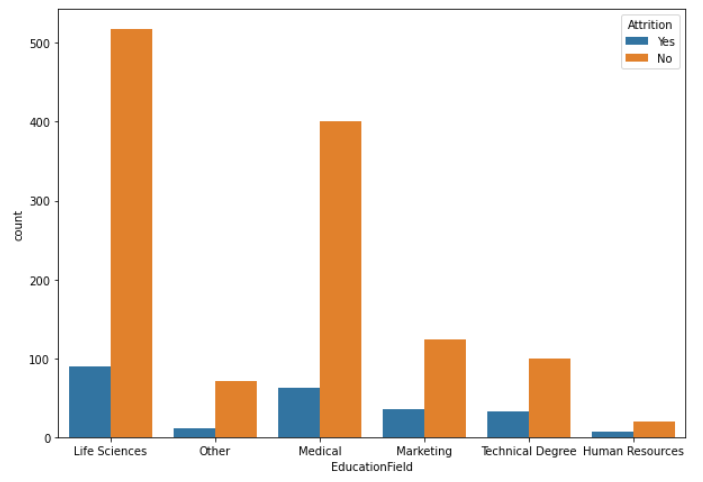
Attrition is not much impacted by salary hike



**2.Background of the employee**

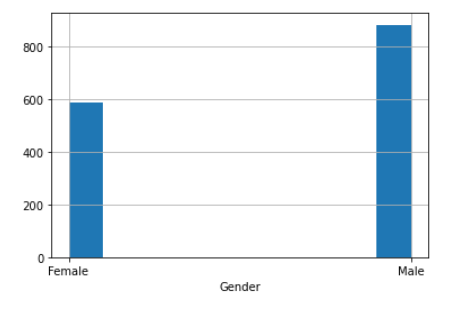
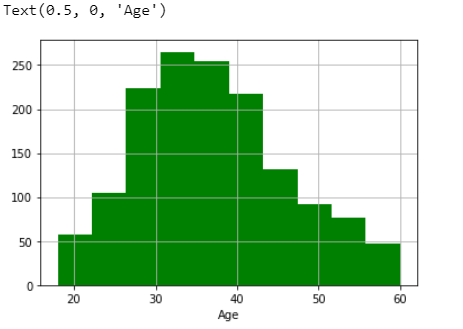
**A.EDUCATIONAL FIELD**



The largest employees have done Life sciences and large number of attrition can be seen in life science educational field. But in proportion to the employees present, Medical has shown the highest percentage of attrition

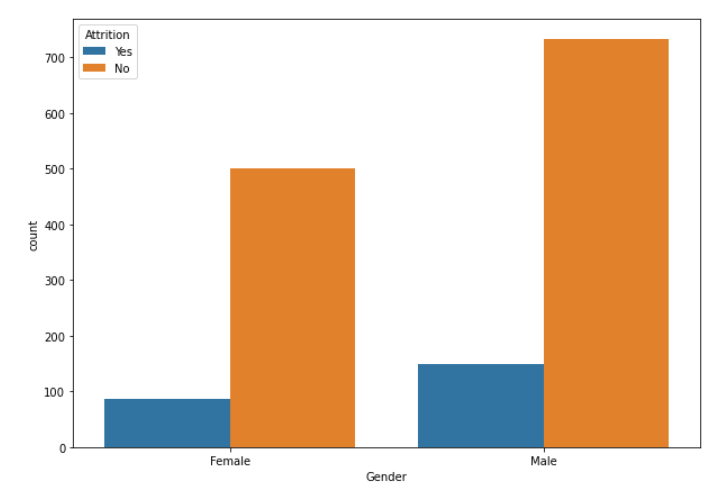
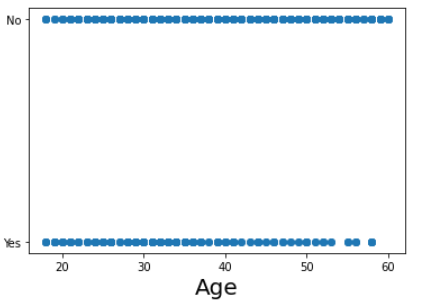
B.**Gender and Age of the employee**

**** ****

There are more males than females

The age of the employees maximum between 30 and 40

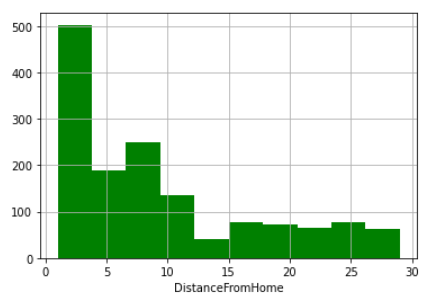
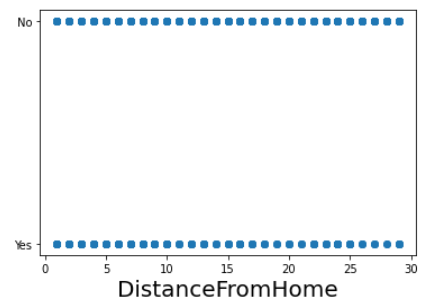
The maximum age at which employees retire is 60years

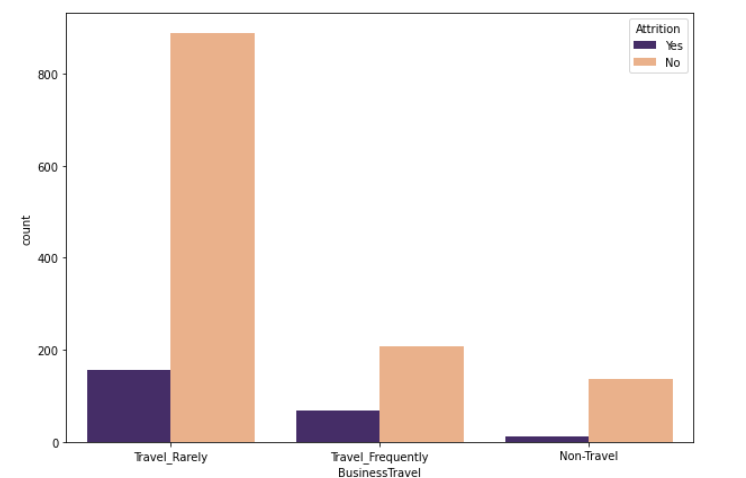
**** 

Both male and female have shown equal percentage chances of attrition

Attrition is not dependent on age

**2.TRAVELLING DONE BY EMPLOYEE**



1.Employees have to travel home and also for business.

2.Most of the employees travel within 10km

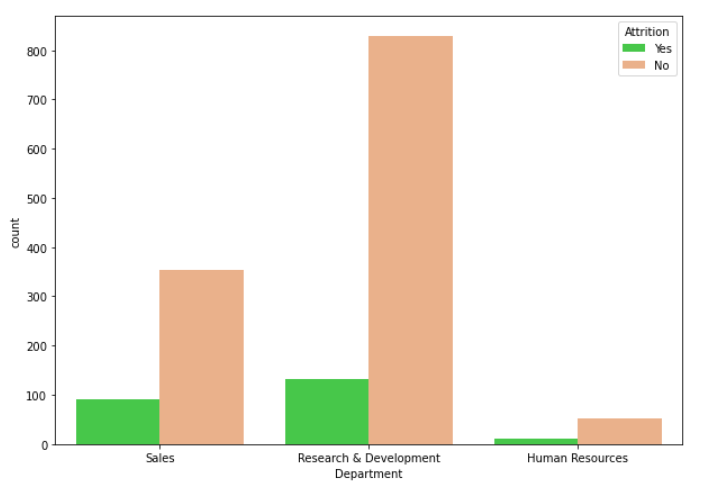
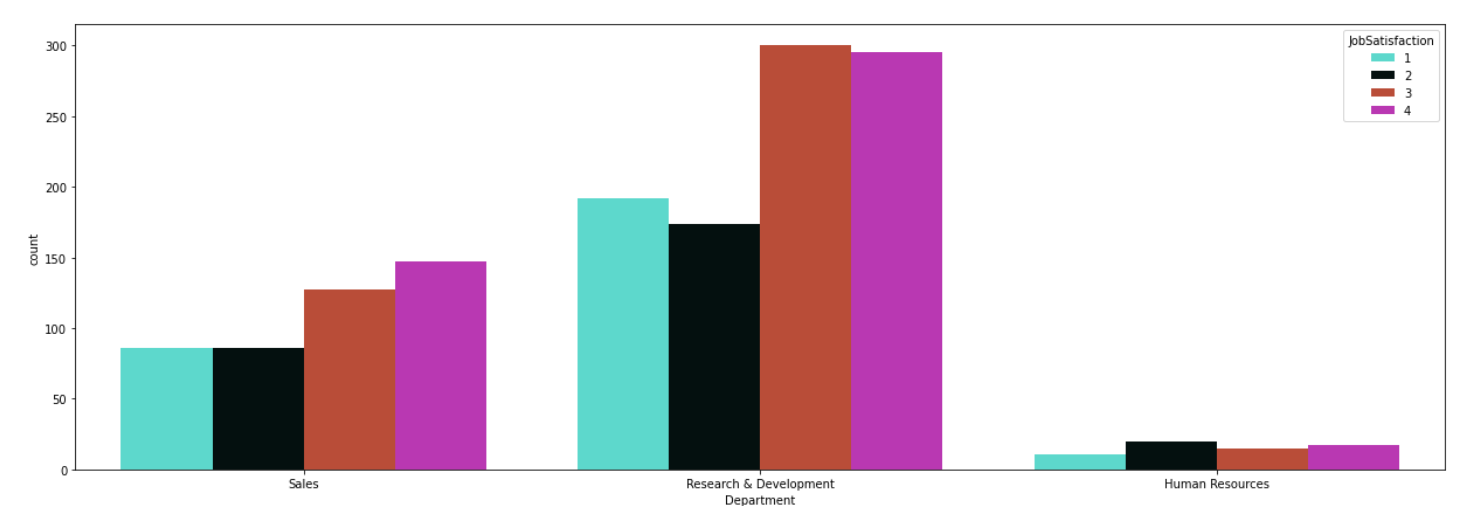
3.Distance from home is not impacting attrition,however we can also see most of them are staying within 10km radius

4.The percentage of those who travel frequently is shown to have more attrition

5.However most of the employees travel rarely

**3.Job details**

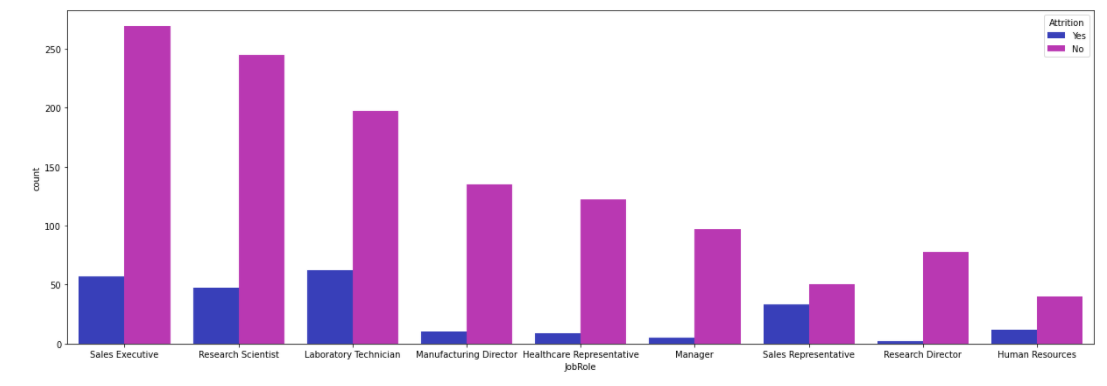
**a.department**

**** 

1.Maximum employees are in the R&D department, but sales department shows higher percentage of attrition

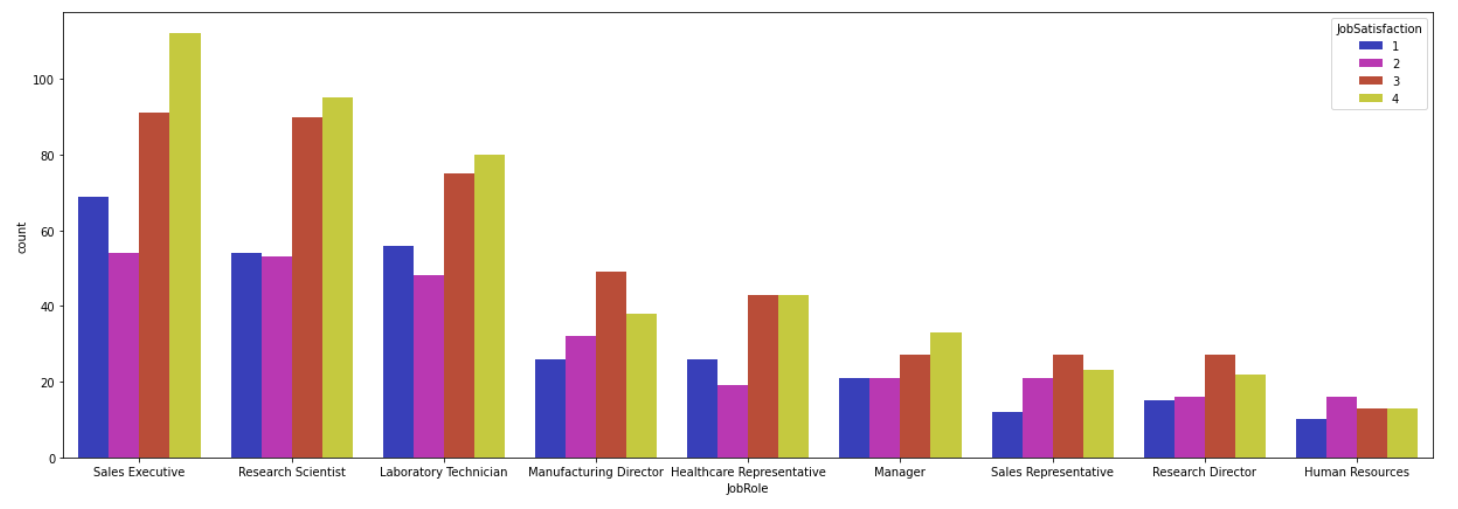
2.Job satisfaction can be seen less in the Sales department

**b.Job role**

****

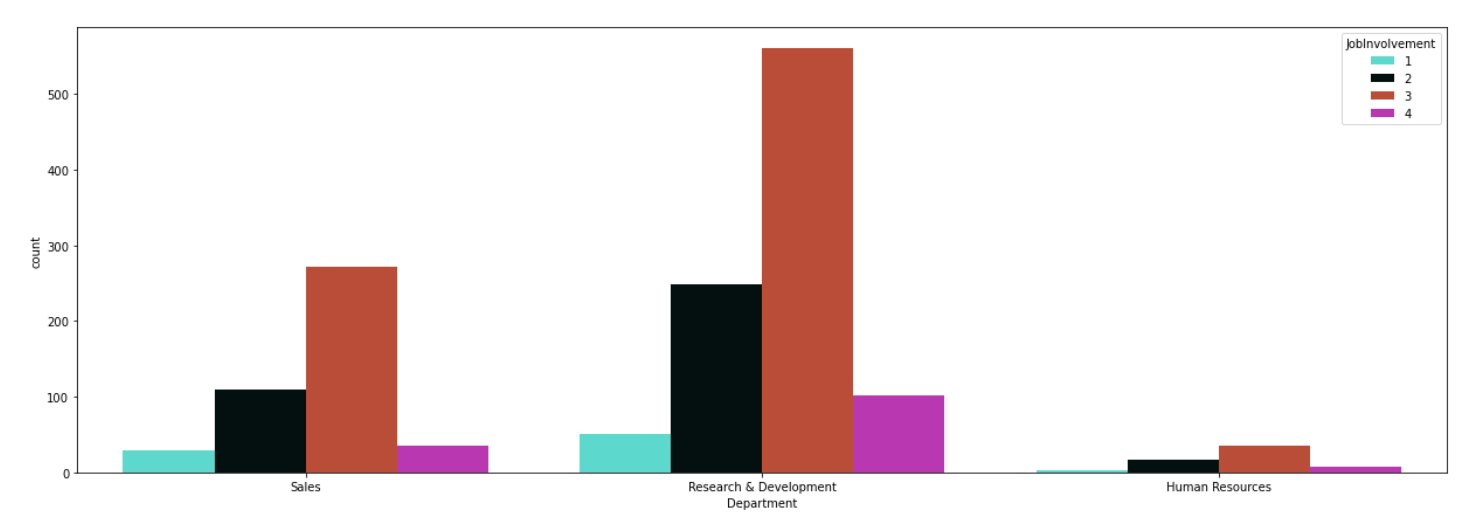
1. and maximum job roles of employees are Sales Executive and Research Scientist

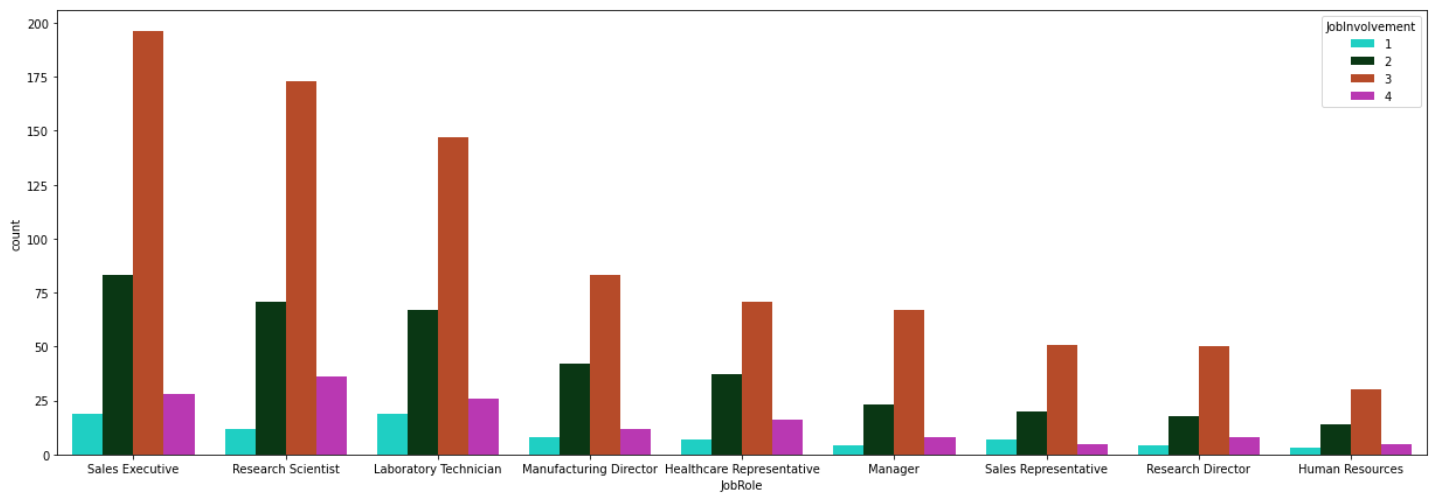
2.Maximum attrition is seen among Laboratory Technicians



It can be seen that Sales executive, Research Scientist and Laboratory technicians show high job unsatisfaction. Hence chances to attrition

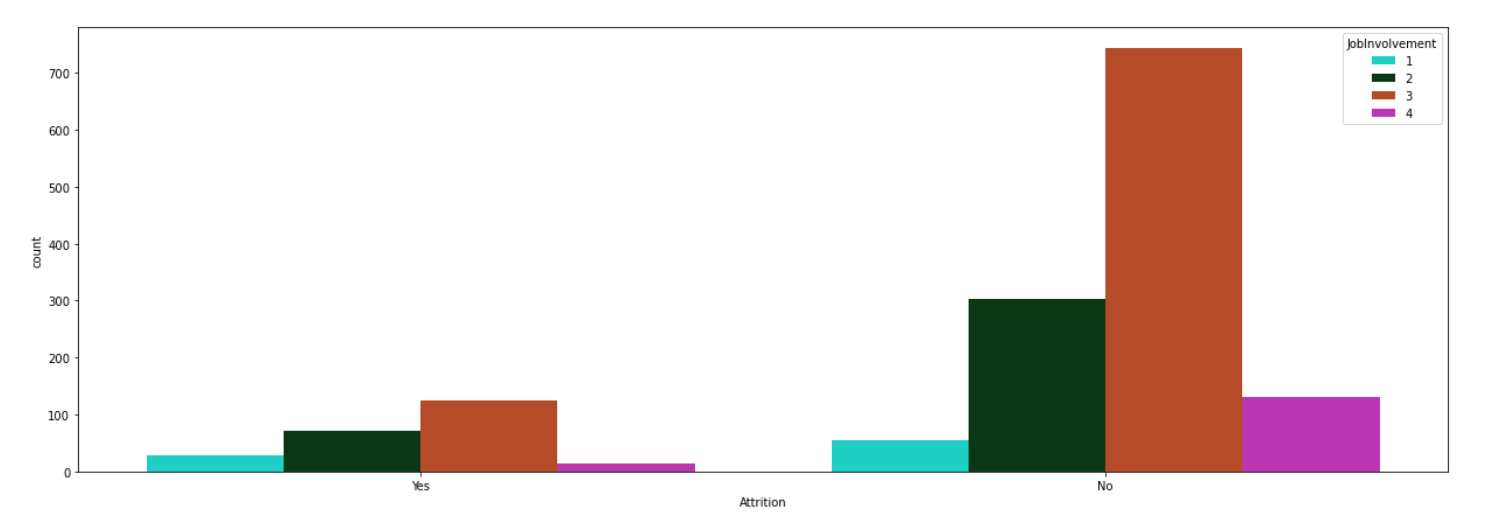
**c.Job involvement**

****

****

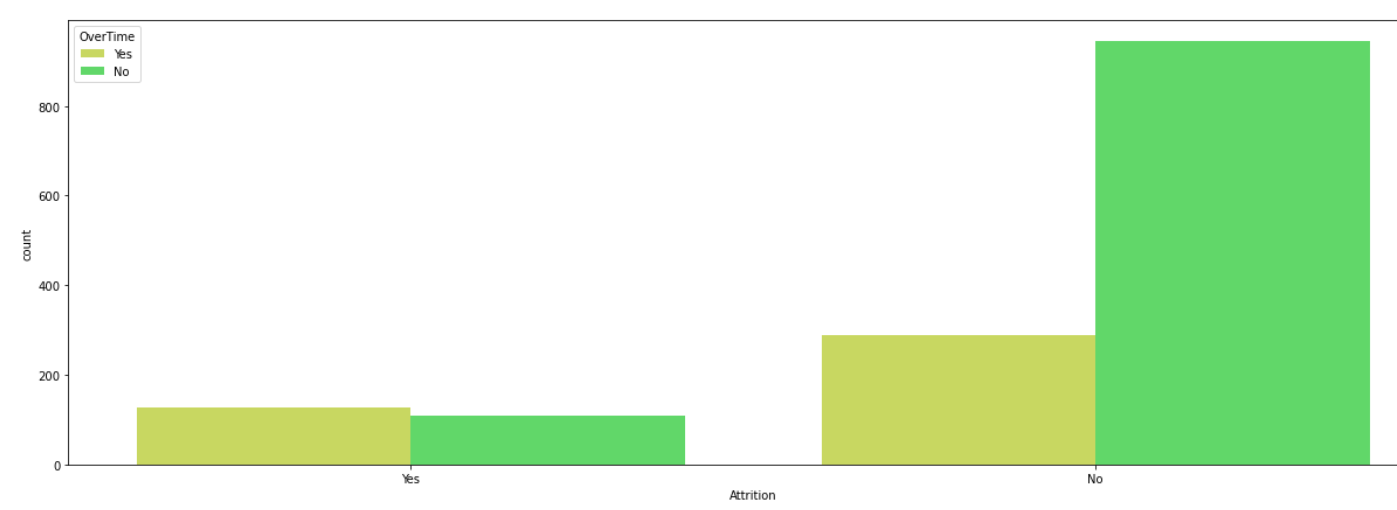
Highest job involvement seen among R&D research scientist department and Job roles of Sales executive and Research scientist

Human resources show very less Job involvement



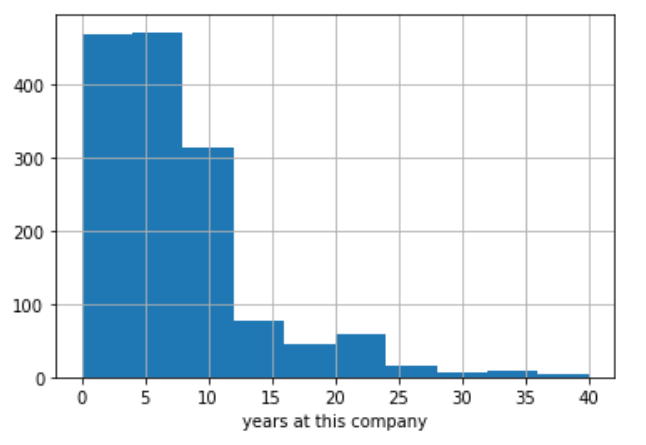
Attrition is lower with more job involvement

**d.Overtime**



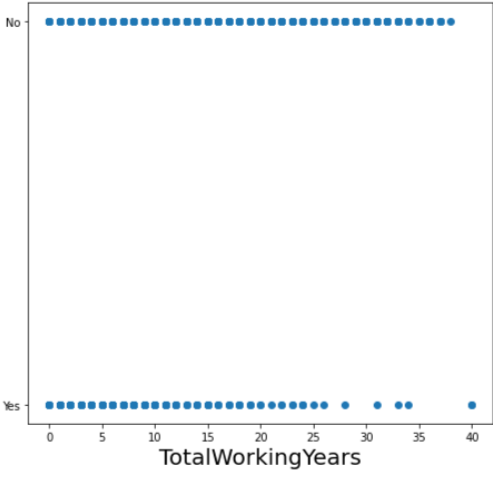
Overtime does not impact the Attrition much

**e.Years at Company**



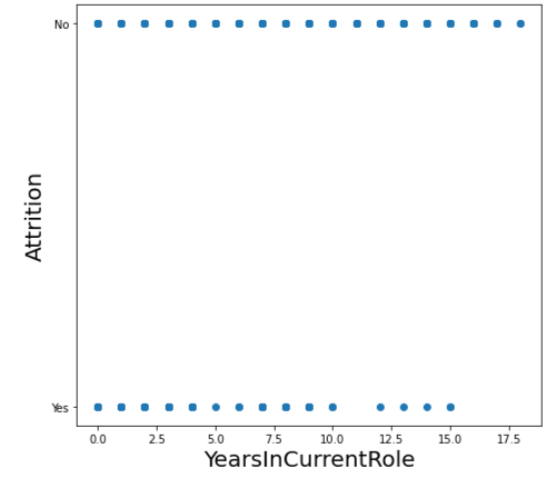
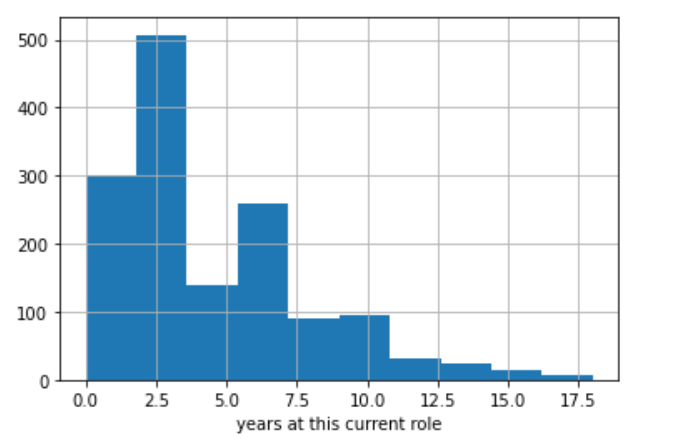
Most of the employee have been in the company since 8years

**Total working years(in company and outside company)**



As Total working years Increase, the chance of attrition decreases

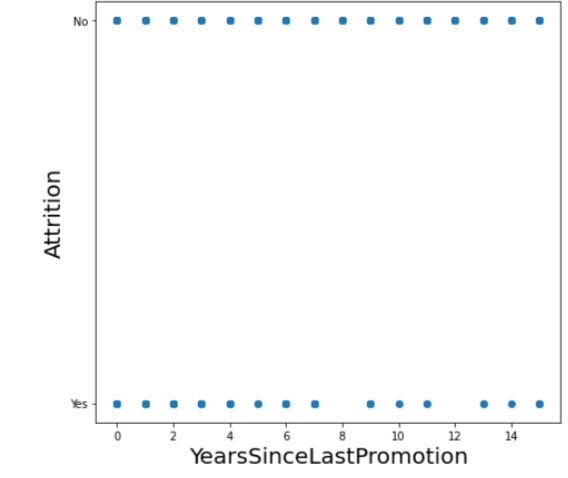
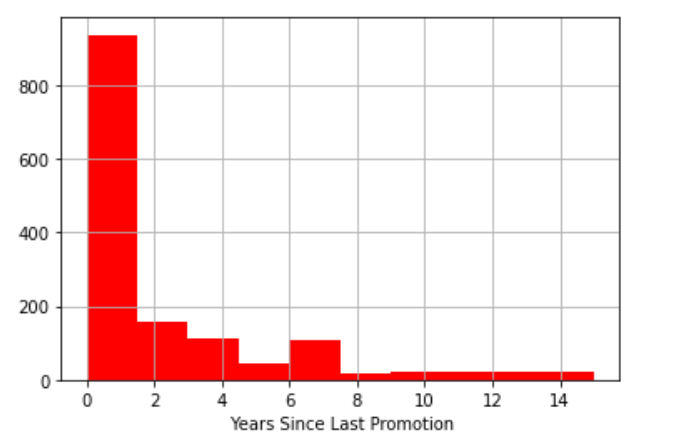
**f.Years the employees were in this current role**



As years in current role increase, the chance of attrition decreases

The maximum employees has been in current role since 2.5years

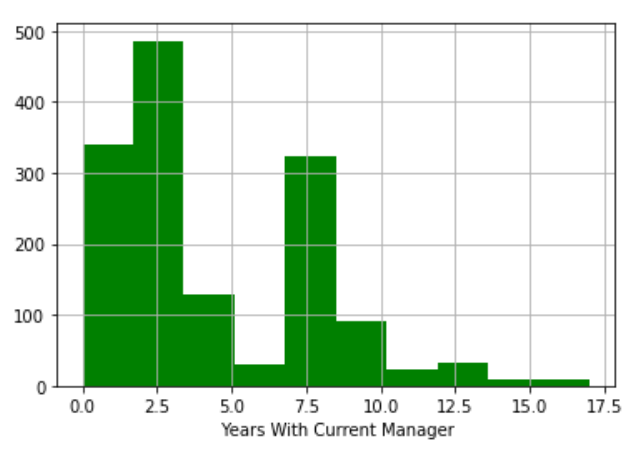
**g.Years since last promotion**



Most of the employees were promoted inside last 2 years

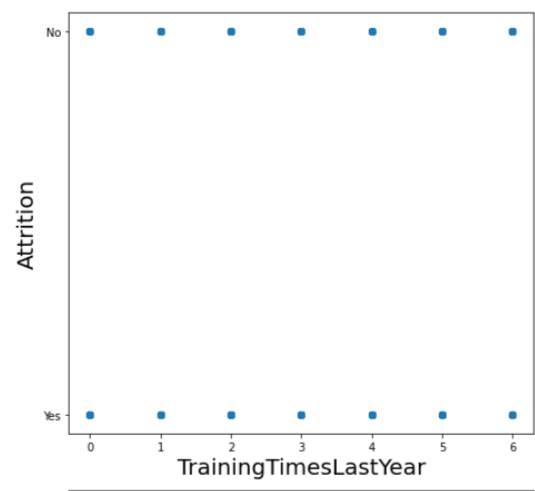
The years since last promotion doesn’t have much impact on attrition

**h.Years with current manager**

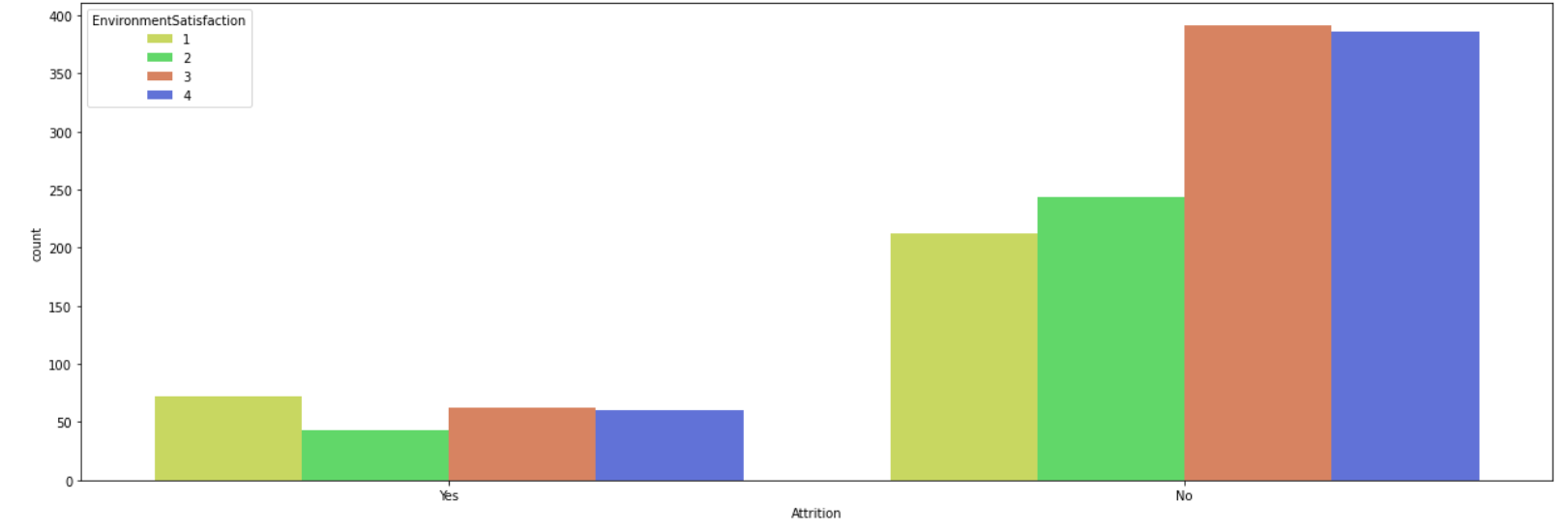


Most of the employess have had 1-3 or 7.5years with their current manager

**i.Training times in last year**

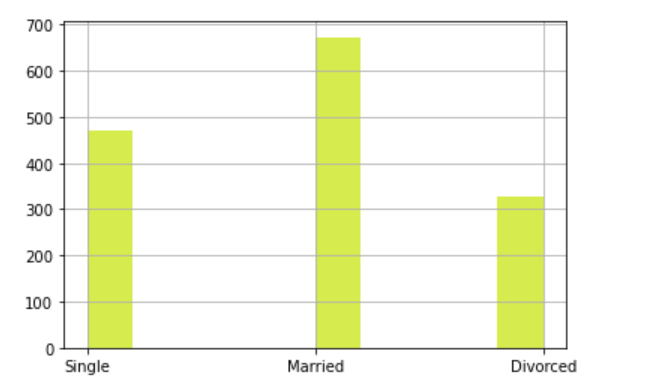
 It doesn’t impact the attrition

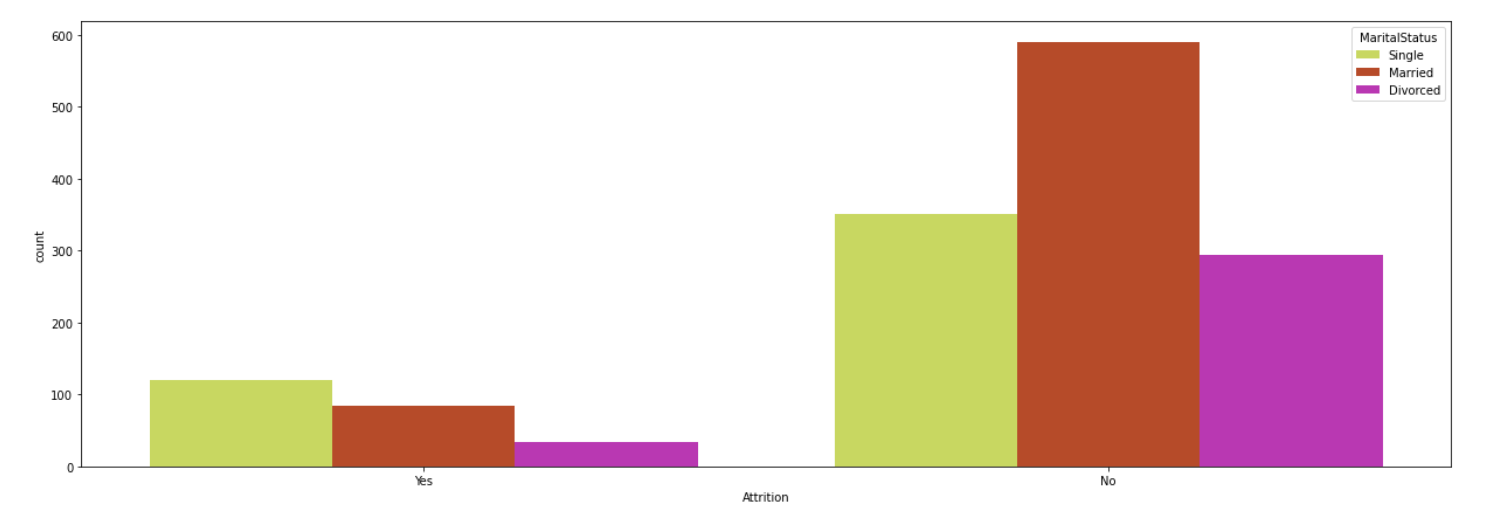
**j.Environment satisfaction**

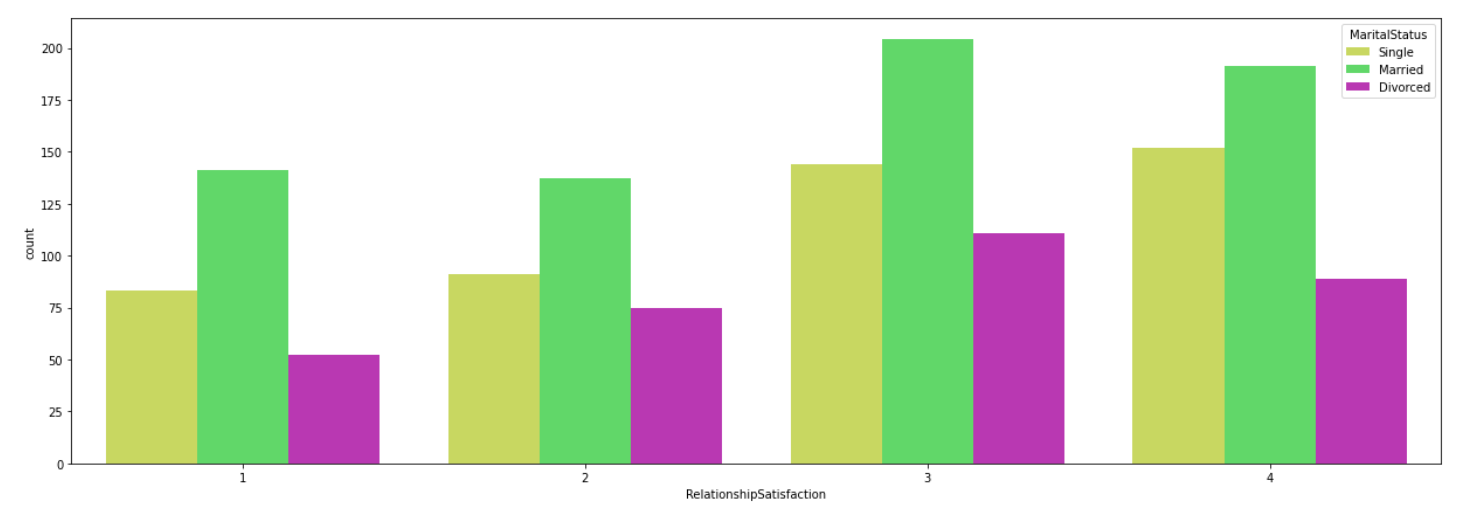
****

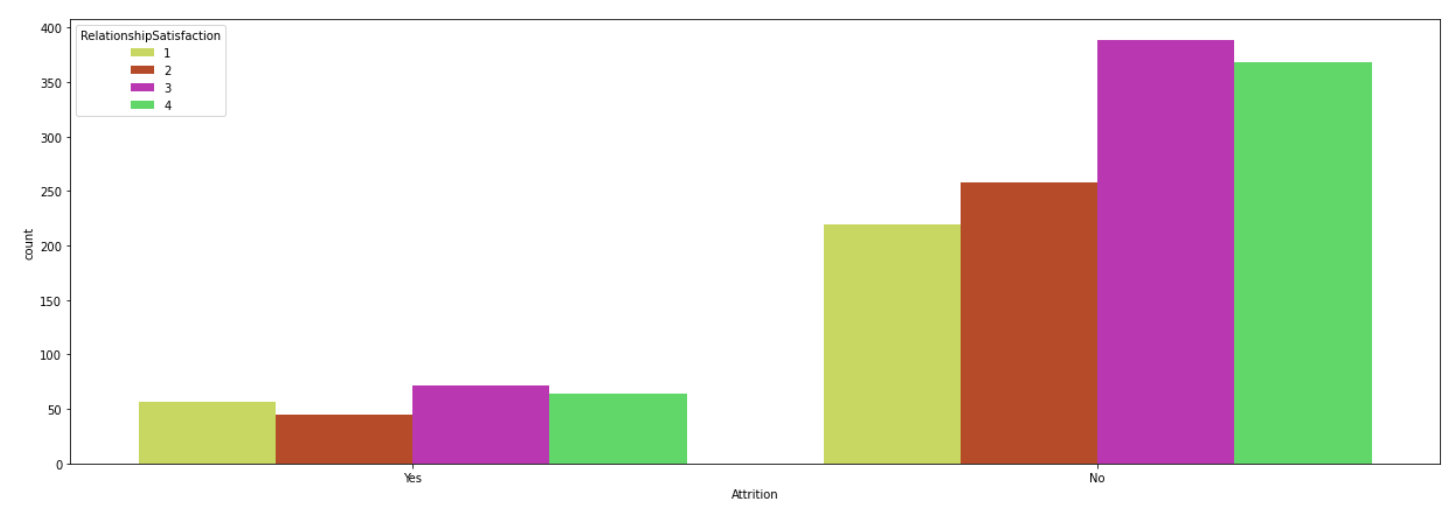
The Environment satisfaction does not affect the employee attrition much

**4.Relationship details**

****

****

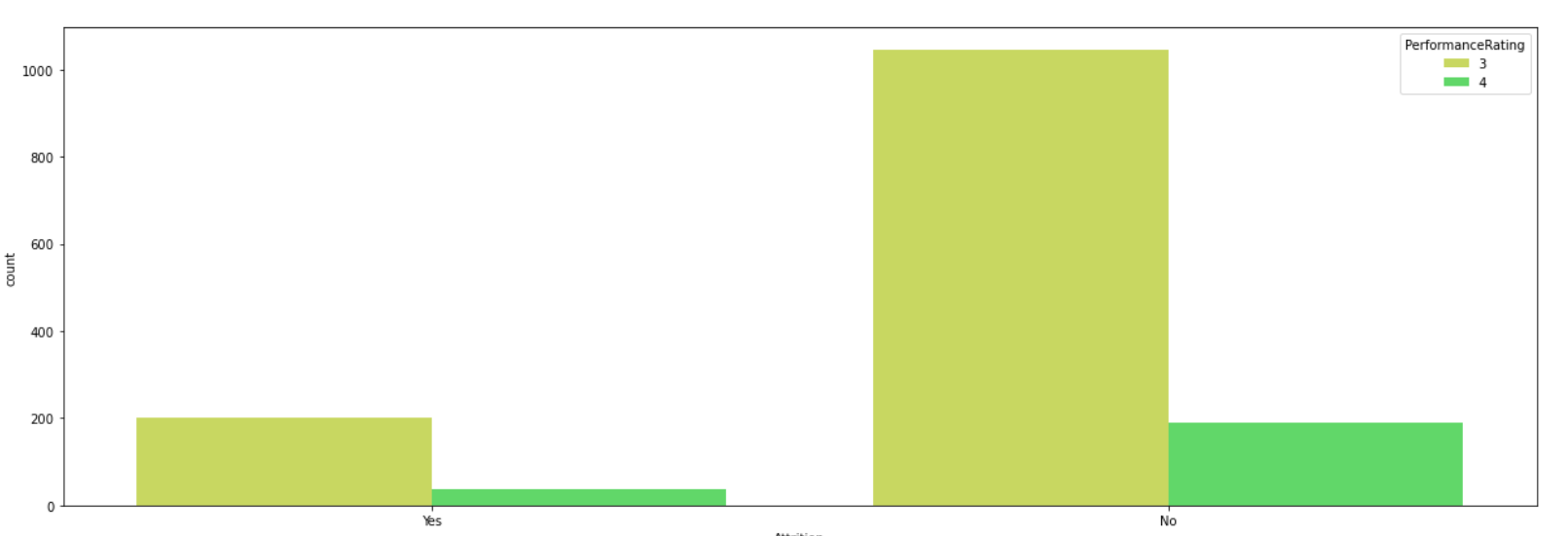
****

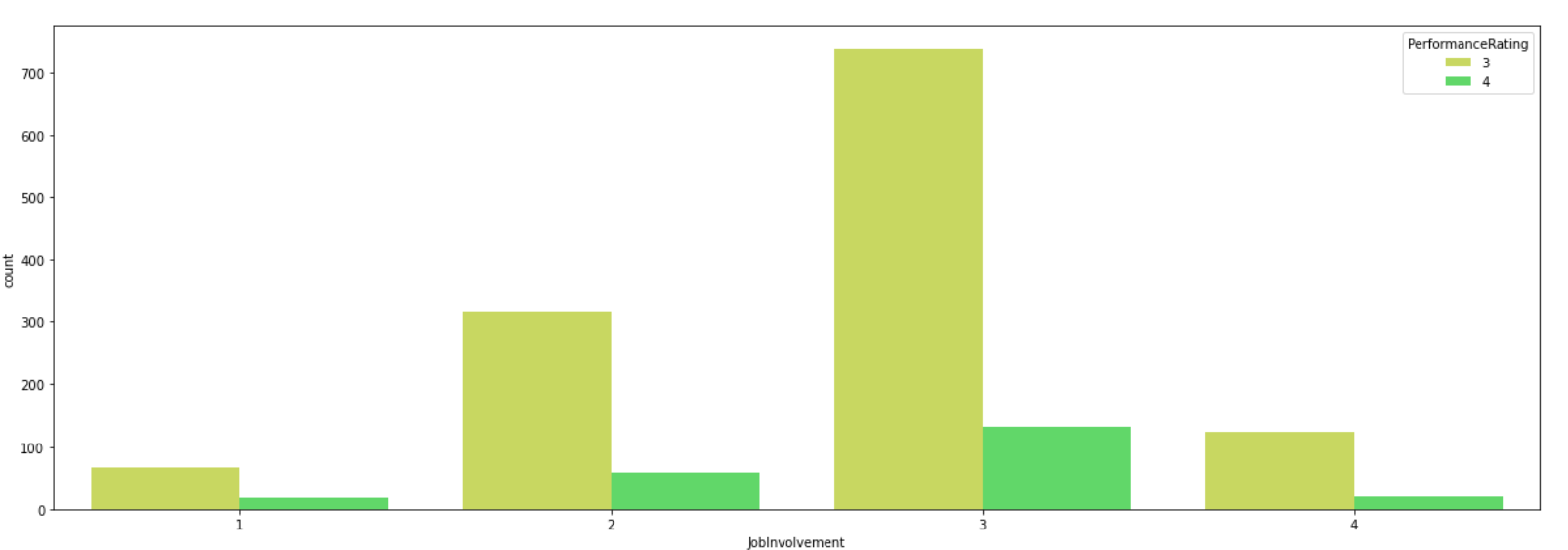
****

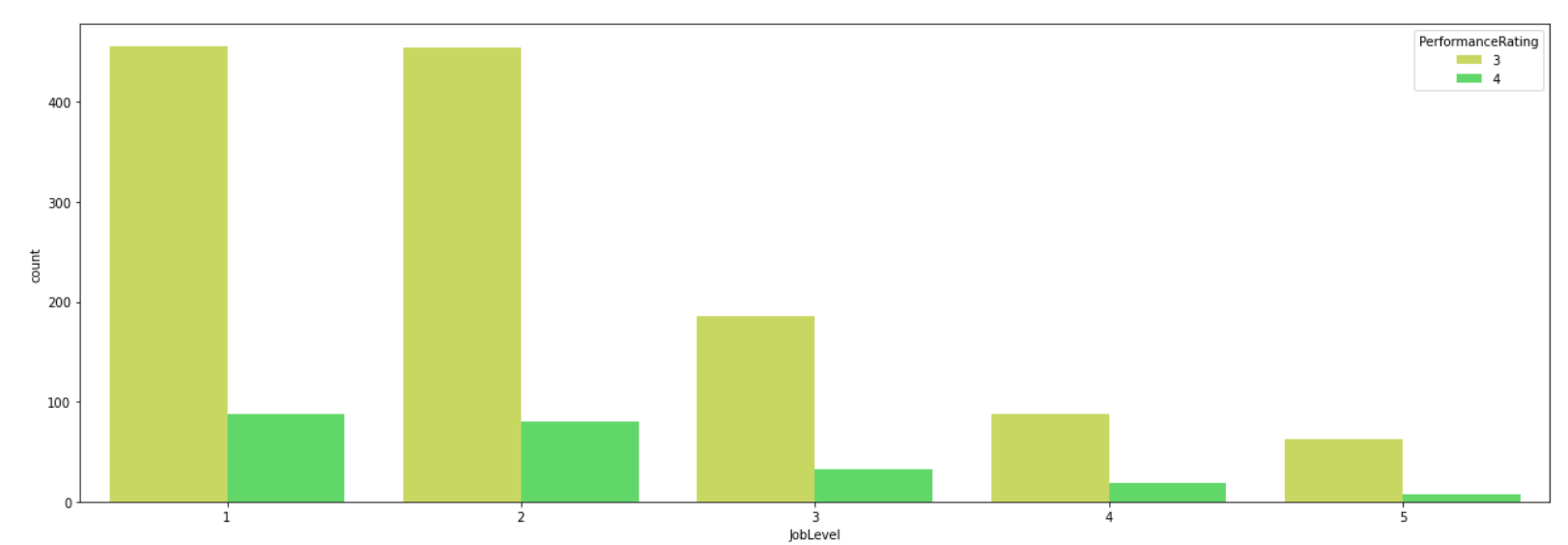
1.Eventhough there are more Married employees than Single employees, the single employees have shown the highest attrition

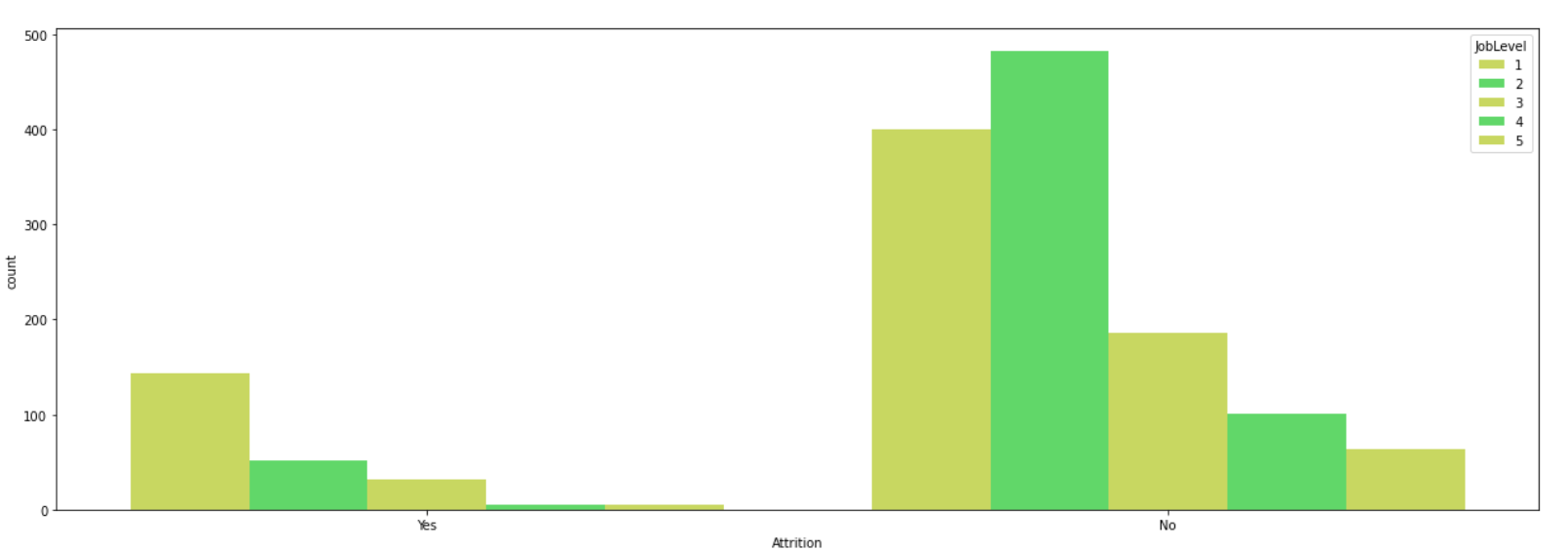
2.Relationship satisfaction doesn’t affect attrition much

**5.Employees job level and performance level**



****

****

****

Attrition is seen higher for employees whose performance rating was 3

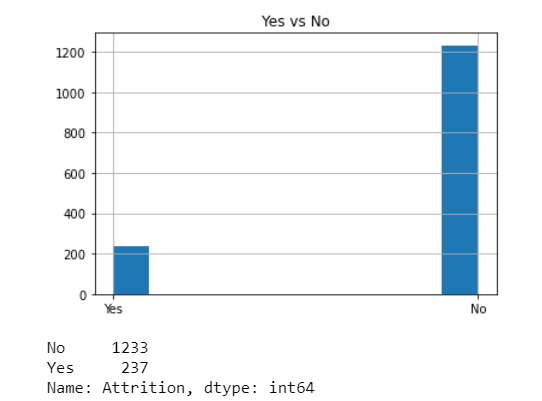
Employee whose job involvement was high have also seen a low performance rating, showing that the employees wasn’t efficient enough

Job level 1 and 2 are highest employees

Attrition is highest for Job level 1 and very low for Job level 4 and 5

**EDA Concluding Remark**

1.Dependent/label:



There were 1233 cases were employees did not undergo attrition and 237 cases where employees underwent attrition. 83.87 % of the employees did not undergo attrition while 16.12% underwent attrition.

The label is **imbalanced**.

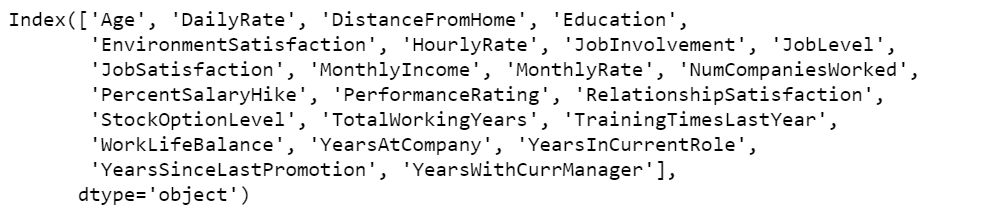
2.There is **no null values** present

3.There are columns which has **single unique value**, which need to be dropped

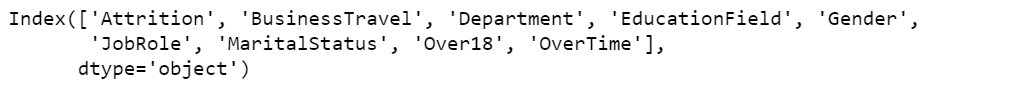
4.Numerical and categorical columns are separated.

Outliers are removed from numerical coloumns and label encoding done on categorical columns

Numerical columns consists of:



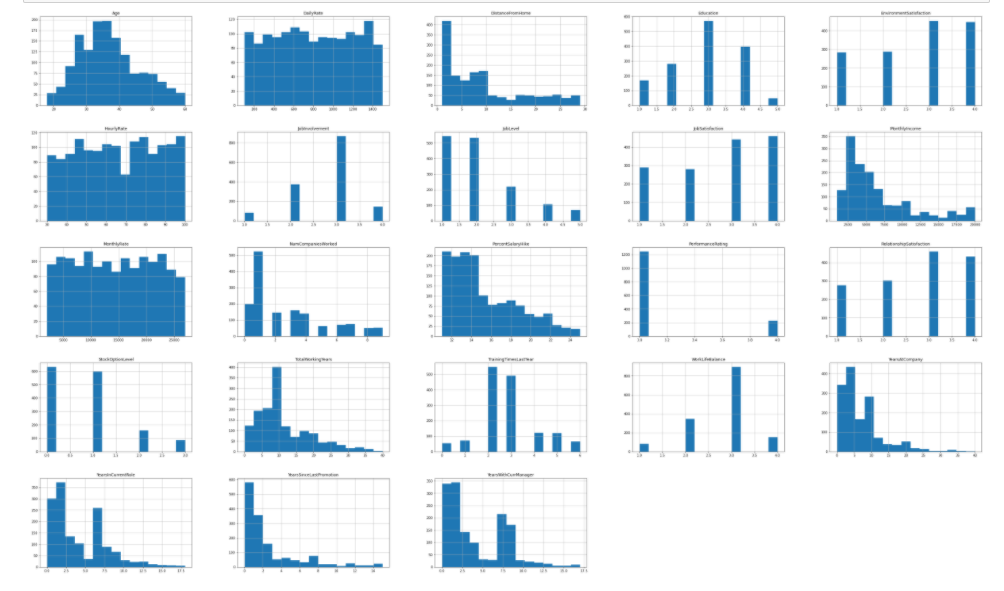
Categorical columns consists of:



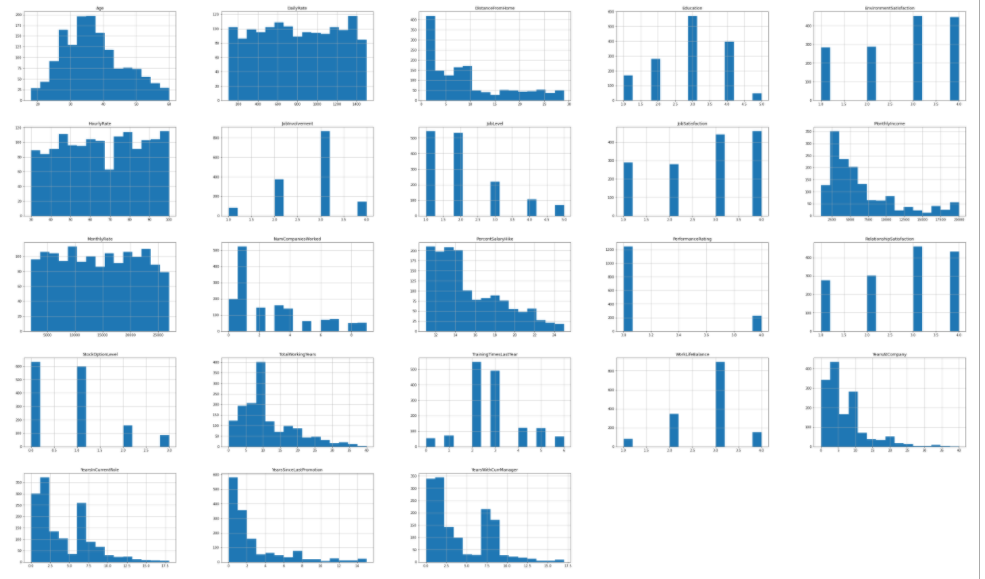
**4.Removing the outliers using zscore**

The zscore is applied for standard deviation of 3. After the zscore was applied the data became more normalised

Dataset before applying zscore:



Dataset after applying zscore:



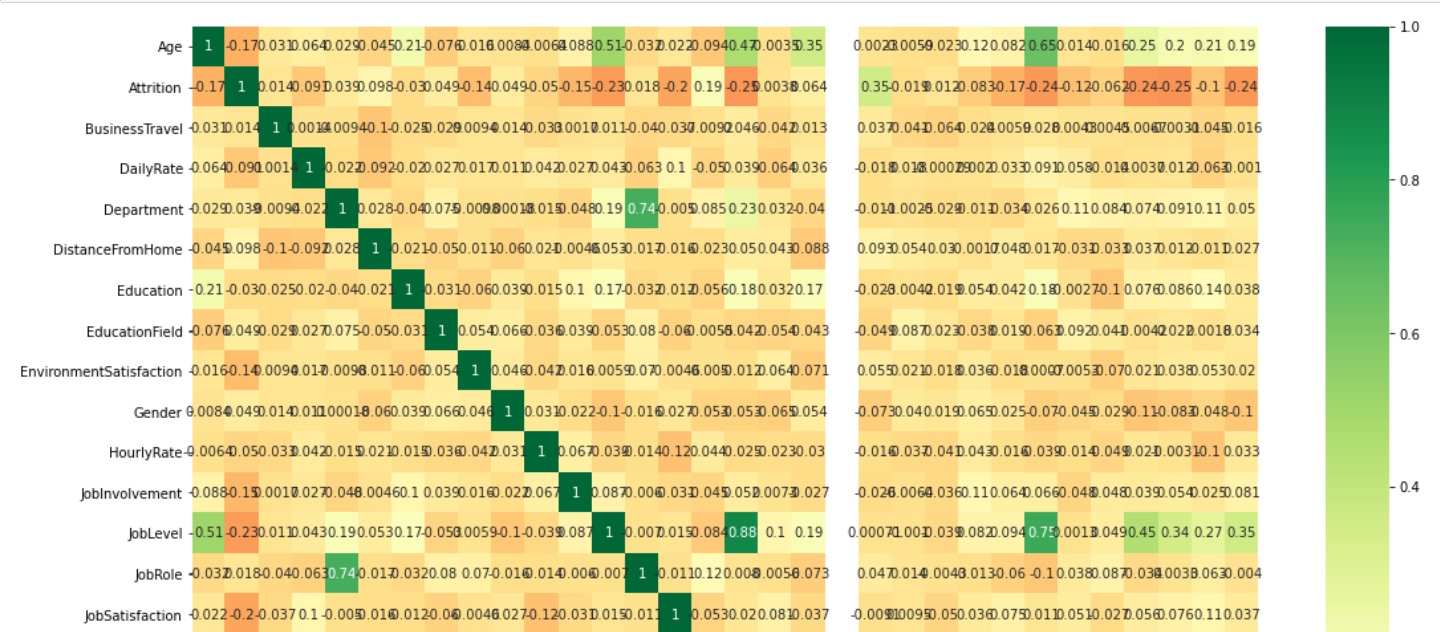
5.**Power transformation is not done** as the data is not skewed

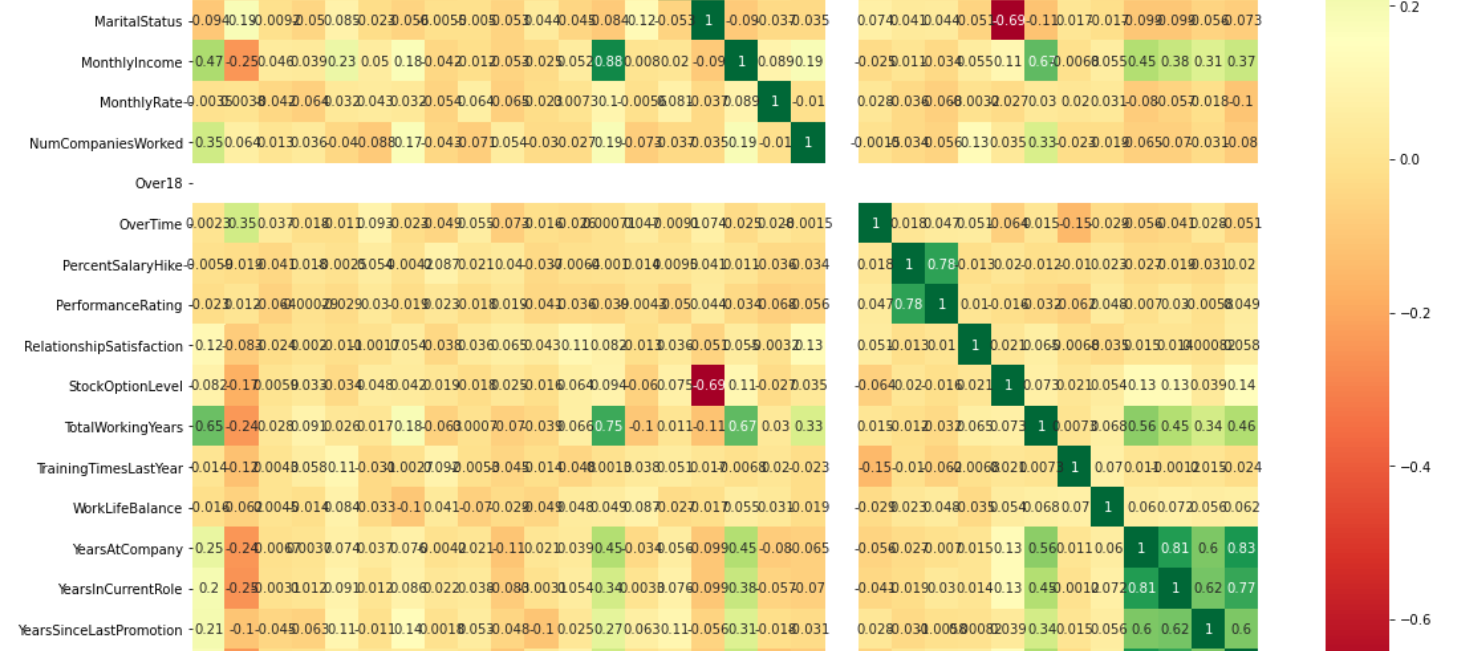
6.Data after EDA, was less skewed and more normalised.

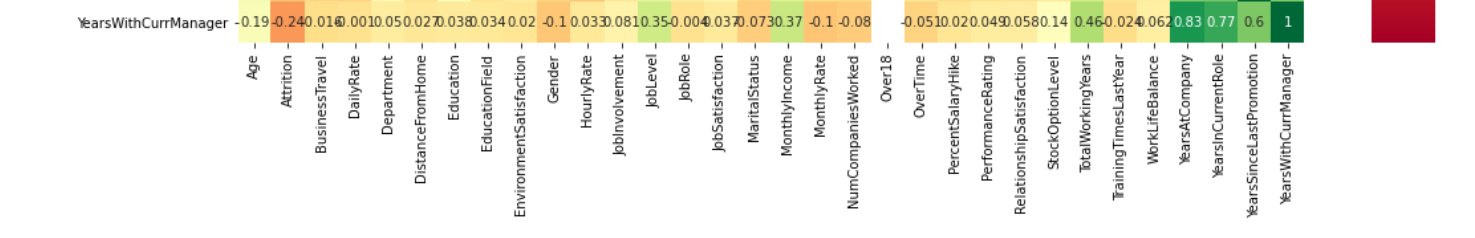
7.The **categorical data is label encoded**

**Preprocessing Pipeline**

**feature Selection**







**Heatmap was plotted for variables**:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable 1 | Variable 2 | Multicollinearity score | High/medium/low multicollineairty |
| department | job role | 0.75 | Medium |
| Monthly Income | Job level | 0.85 | high |
| Total working years | job level | 0.75 | medium |
| Performance rating | Percent salary hike | 0.78 | medium |
| Years with current manager | years with company | 0.85 | high |
| Years in current role | years with company | 0.81 | medium |

Apart from that, there don’t seem to be much correlations in the data.

There seems to be multicollinearity problem, but since accuracy score will decrease and we will lose data which is precious, we will not be dropping columns

**Feature engineering**

1.The label encoding was done on the entire categorical data and the label coded 1 for No and 0 for Yes

2.There is single unique values in certain columns and hence they are dropped







Hence employeeCount,StandardHours and Over18 is dropped

3.Employee Number is an unnecessary column which will be dropped

4.**Label encoding** was applied on the independent variables.

5.The label/Attrition column was balanced using Undersampling technique

6.Only 32 columns, hence PCA is not done

**Baseline Score**

Since the class is balanced using Undersampling, we could use **either Accuracy score or ROC AUC score** to find the best fit model for our Attrition dataset.

**Building machine learning model**

**Modeling**

**1.Split the data into test and train**

a.Both the dependent variable and independent variable was split into test and train data(75% train and 25% test)

b. Best sample state was found to be 109 and the data was split as per it.

2. Four different classifiers were used in this project:

- logistic regression

- K-nearest neighbours

- Random forest

- DecisionTreeClassifier

i.Accuracy score,f1 score,precision,classification matrix ,f1 score and recall was obtained for each of it:

|  |  |
| --- | --- |
| **a.Logistic regression** |  |
| **b.Decision tree classifier** |  |
| **c.RandomForest**  **Classifier** |  |
| **d. SVC** |  |

**ii.Cross validation score of each of the algorithm:**

|  |  |
| --- | --- |
| Logistic Regression | 0.8363452198529984 |
| Decision Tree Model | 0.7851388203516609 |
| Random forest Model | 0.8572448900085707 |
| SVC Model | 0.8348959821312626 |

**iii.ROC AUC Score**

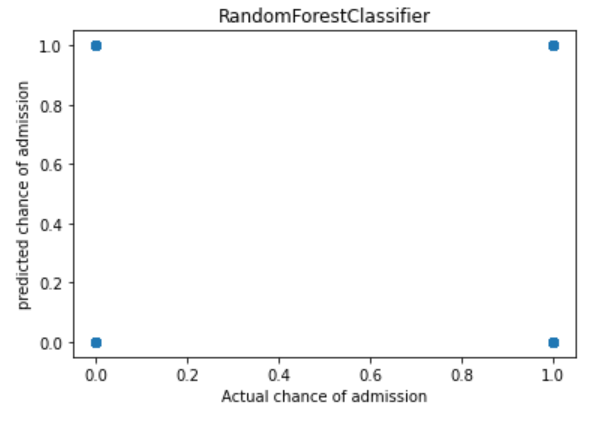
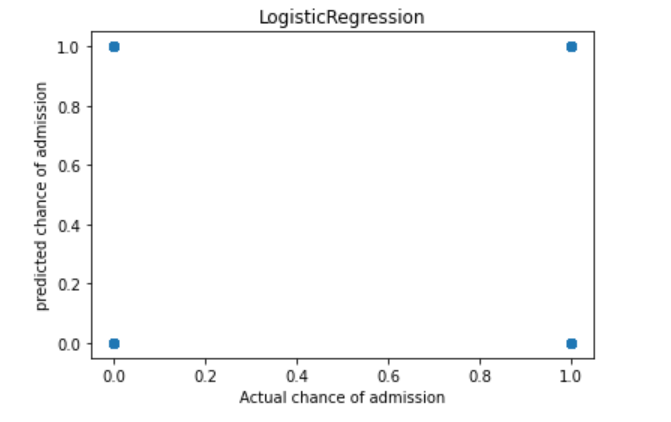
The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’.

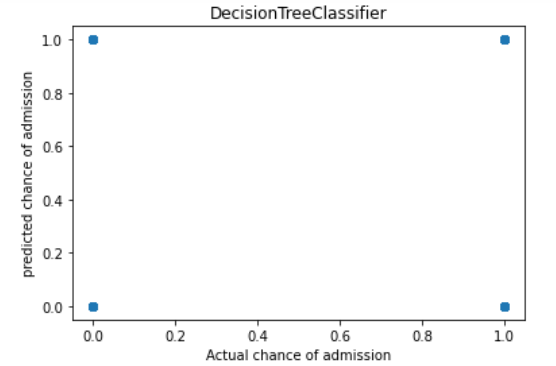
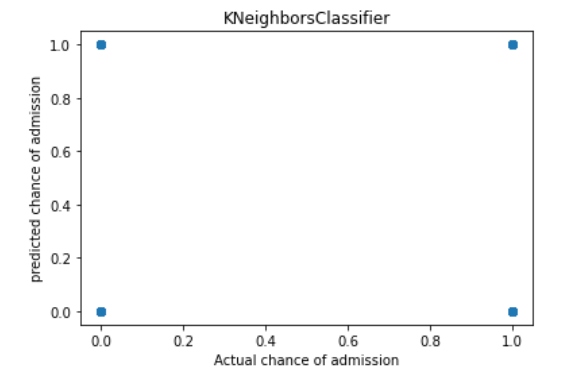
|  |  |
| --- | --- |
| Logistic Regression | 0.732894370122631 |
| Decision Tree Model | 0.6568770903010033 |
| Random forest Model | 0.5904055183946488 |
| SVC Model | 0.5574832775919732 |

**iv.ROC AUC curve**

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.



**3.Chose Logistic regression for hyper parameter tuning**

Reason for choosing Logistic regression model:

1.Least difference is present between the accuracy score and cv score

2.The ROC-AUC score is higher for logistic regression

**4. Hyperparameter tuning on Logistic regression**

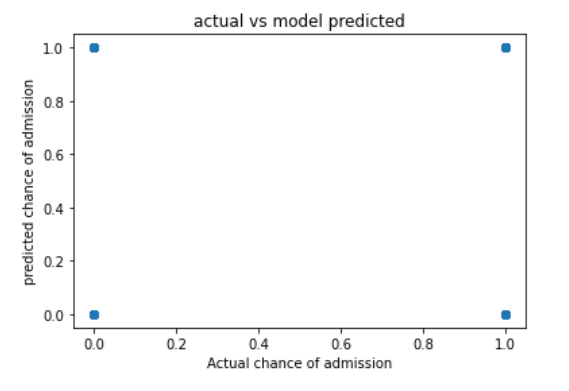
* Hyperparameter tuning and selection was done for Logistic Regression Model using GridSearchCV.
* After a 6-fold GridSearchCV, the model with its selected hyperparameters were fitted on the training set
* Model was created using the best parameters and saved as a pickle file

**5.Model scores obtained are:**

|  |  |
| --- | --- |
| Cross validation score | 83.63 |
| Accuracy score | 91.64 |
| ROC AUC score | 73.2 |

**6.ROC AUC Curve of final model**

The ROC curve below summarizes how well our model is at balancing between the true positive rate(sensitivity) and the false positive rate(1-specificity). Ideally, we want to have a 100% true positive rate of predicting fraud and a 100% true negative rate of predicting non-frauds (or a 0% false positive which is 100% — 100% true negative rate). This means we have perfect prediction for both classes. However, in imbalance class problems, this is extremely hard to achieve in the real world. On top of that, there is a trade of between the true positive rate and the true negative rate and conversely the false positive rate.



**CONCLUDING REMARKS**

This project has built a model that can detect Attrition in HR Analytics. In doing so, the model can reduces loses for companies in terms of Human resource Managment. The challenge behind finding the attrition in machine learning is that attrition are far less common as compared to Employees who stay.

Five different classifiers were used in this project: - logistic regression, K-nearest neighbours, Random forest and DecisionTreeClassifier. The best model for the dataset chosen is Logistic regression model. The Model can predict with an accuracy of 91.9% if the employee will undergo attrition or not

The parameters that influence if the employee will undergo attrition are many.Among which Marital status and Overtime has the highest impact on attrition.

The advantage of the prediction model is that, it makes it easier for a Company to not go through the entire process of recruitment time and again, thereby retaining the assets with less expenditure. Hence it saves resources of time and money for recruitment and also the knowledge base developed by the company

However the study having its own limitation is small sample size, restriction to few departments has its inability to easily adapt to every company.