Why? Employees Left Company, HR Attrition Analysis Using Machine Learning Algorithms To Find Ratio Of Attrition

Ankit Dadarwala – October 19, 2021

Intermediate, Machine Learning, Algorithms, Data structure, Visualization, Jupyter Notebook

Introduce To Problem Statement:

Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees.

HR Analytics

- where HR Analytics fit in this? and is it just about improving the performance of employees?
- Today era of Data science and 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 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 In HR

- Attrition in human resources refers to the gradual loss of employees overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

- 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.

Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

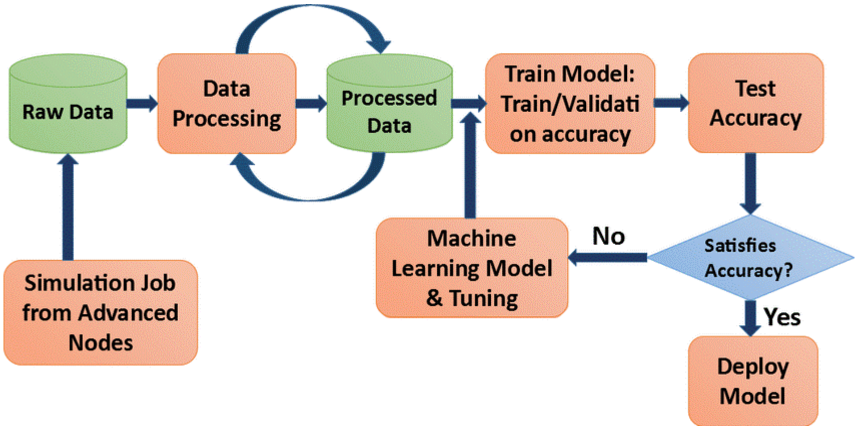
How does Attrition affect companies? and how does HR Analytics help in analysing attrition? We will write the code and try to understand the process step by step.

Here We Use LOGISTIC REGRESSION Why?

In our problem definition people left company in that answer is “**YES”** either “**NO”** so in our problem dependant variable are categorical for that we can’t use Linear Regression, So we have to justify our problem by Logistic Regression

Methodology Used:

2. **Data Collection**
3. **Data Exploration**
4. **Data Cleaning and Transformation**
5. **Balancing Data**
6. **Split Data Into Train and Test**
7. **Built Model Base on Training Dataset**
8. **Check Accuracy Of Models**
9. **Select Best Model for Hyper tuning**
10. **Check Cross Validation**
11. **Plot AUC ROC Curve**


Raw Data
Data Processing
Processed Data
Train Model: Train/Validati on accuracy
Test Accuracy
Simulation Job from Advanced Nodes
Machine Learning Model & Tuning
No
Satisfies Accuracy?
Yes
Deploy Model

Features In Dataset:

In Dataset 1470 Rows and 35 Columns
There is no Missing values present in Dataset

```
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):

     Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
```

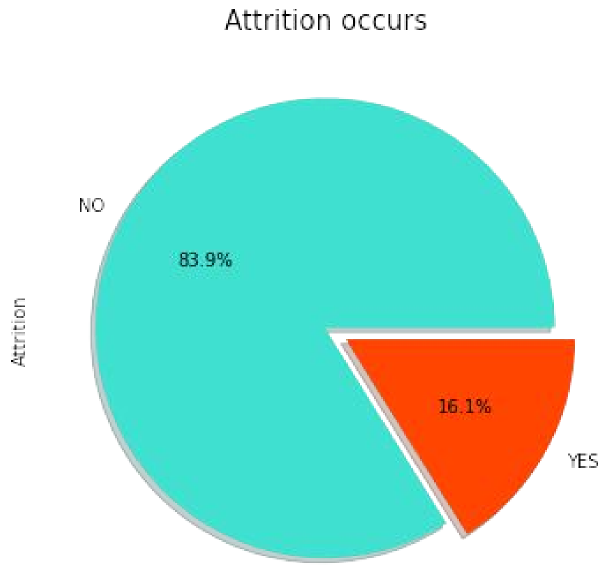
Data Exploration:

That columns not give any information logical think 18+ persons are allowed for job, Standard hours also have same value and working hours are fixed for all organization, Remove Other Columns that Can’t Give Proper information about Dependant features.

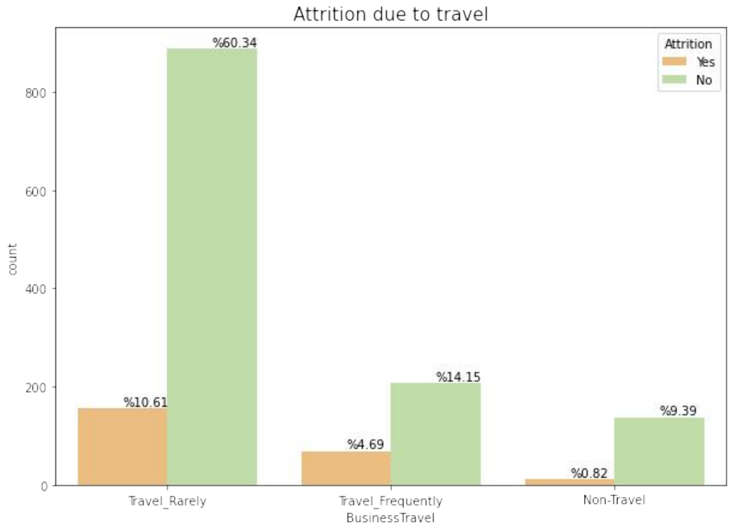
```
df.drop(['Over18','EmployeeCount','EmployeeNumber','StandardHours'],axis=1,i
nplace=True)
```

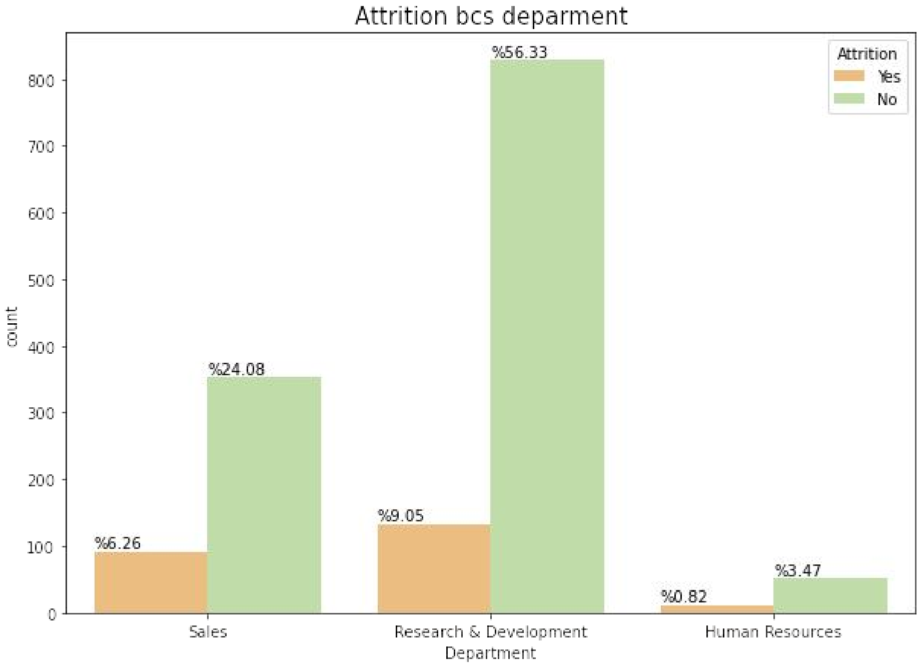
Data Visualization Analysis:

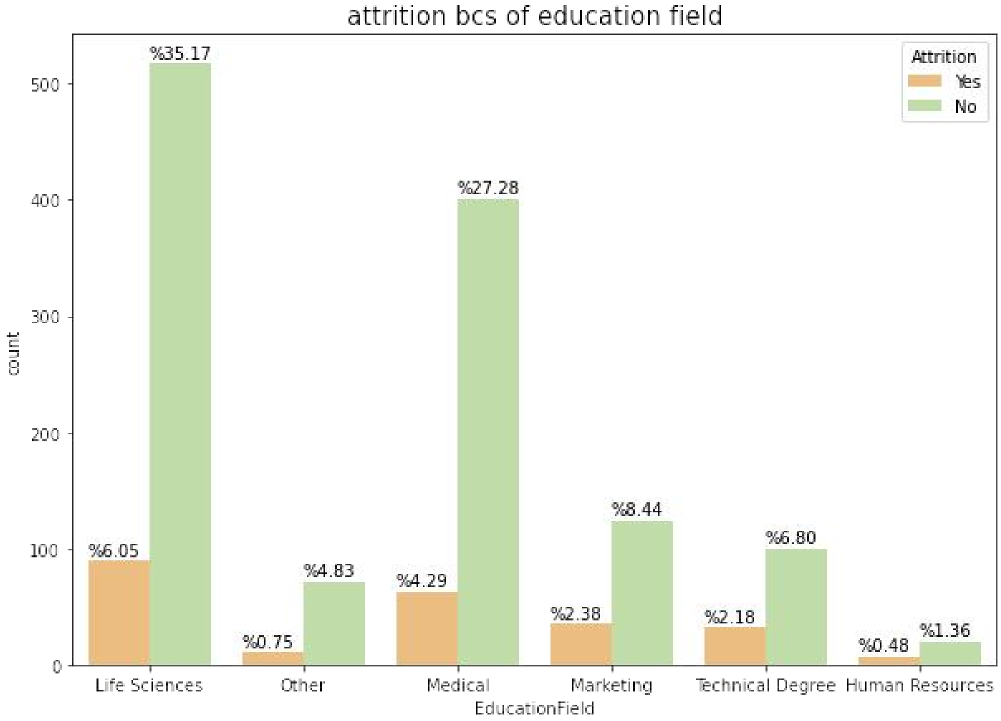
Here we can see that attrition ratio of employee is about to 85% Not left to 15% employee left company.
Also Give information about Imbalanced Dataset of Target column



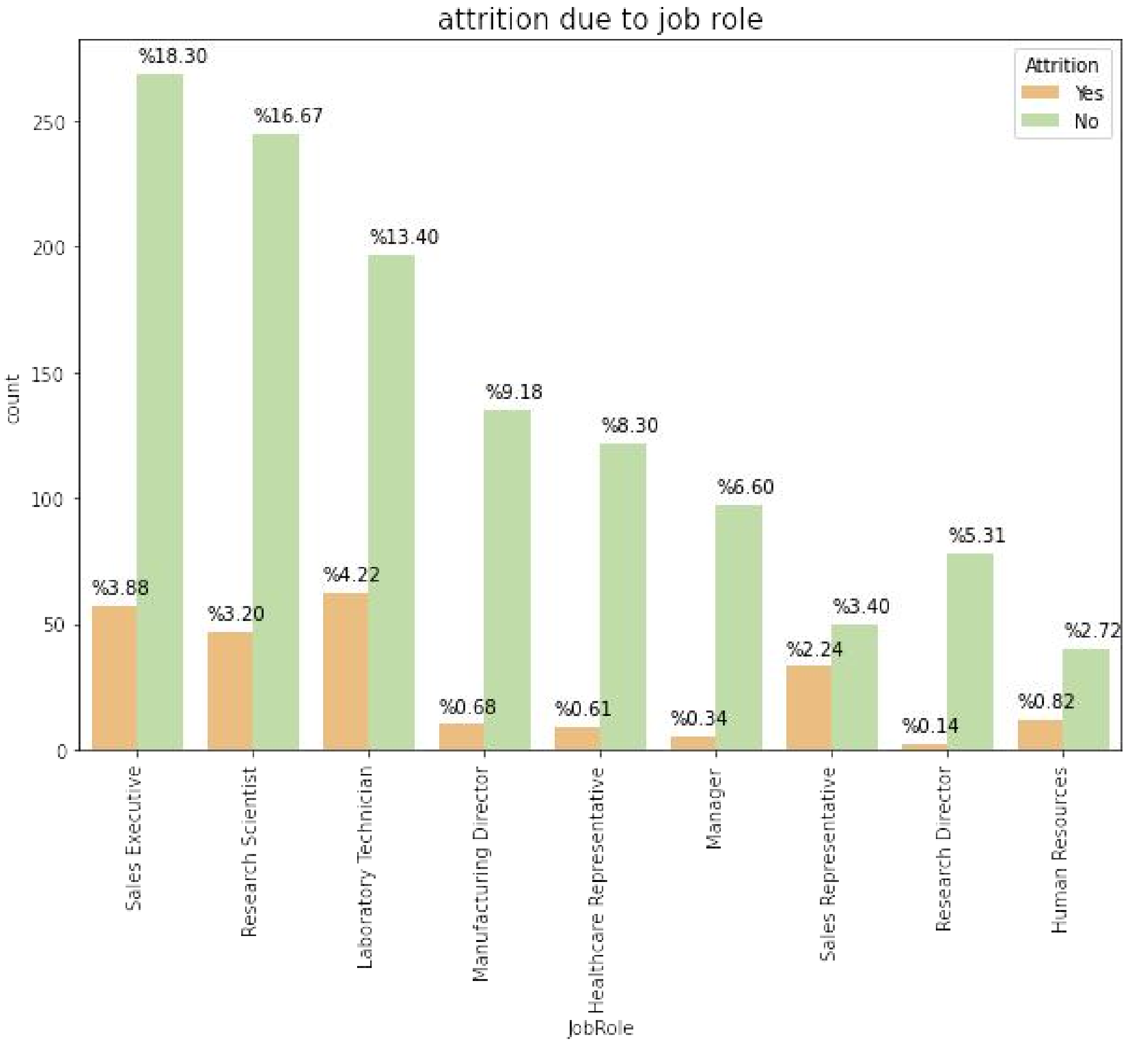
1. Attrition factors are Business Travel, Working department, Job role, Married Status, Age, Over Time
2. In travelling also major factor affect in attrition most marital persons because leave stays away from family travel frequently have 50% attrition of employee
3. Working Department also have affecting factor for employee attrition in that Sales and R&D departments have most employee leave
4. Job role also affect employees attrition because of not satisfied job or Hight workload as compare to salary etc…
5. Married status also role in attrition in single persons leave Jobs quickly as compare to married
6. Age of 18 to 35 years persons change job mostly frequently for salary hike or other beneficial reasons

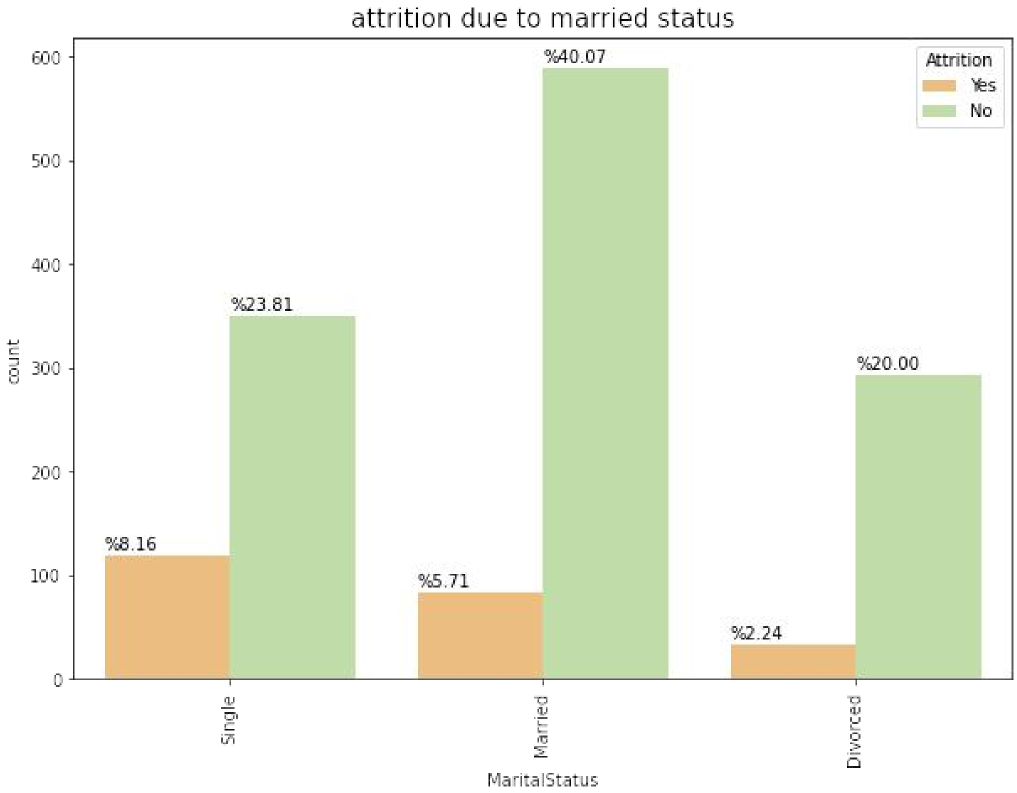

Attrition due to travel
Attrition
Yes
No
count
0
200
400
600
800
%60.34
%10.61
%14.15
%4.69
%9.39
%0.82
Travel_Rarely
Travel_Frequently
Non-Travel
BusinessTravel

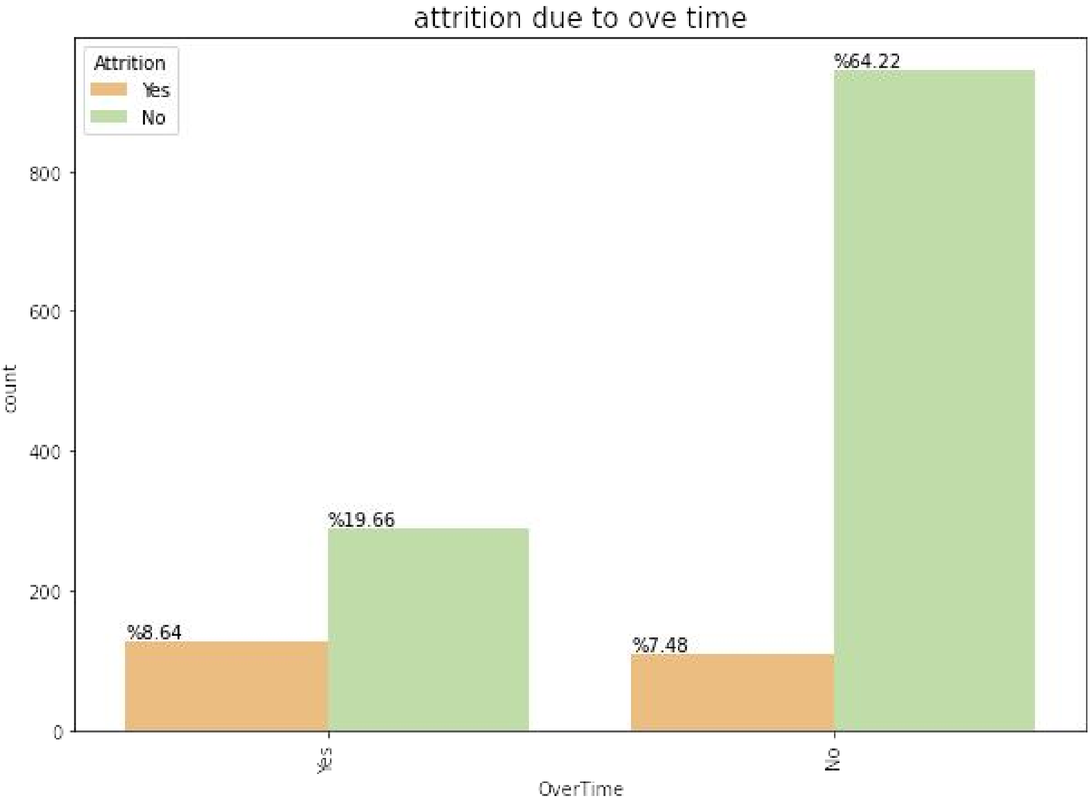

Attrition bcs deparment
Attrition
Yes
No
count
0
100
200
300
400
500
600
700
800
%56.33
%24.08
%6.26
%9.05
%0.82
%3.47
Sales
Research & Development
Human Resources
Department

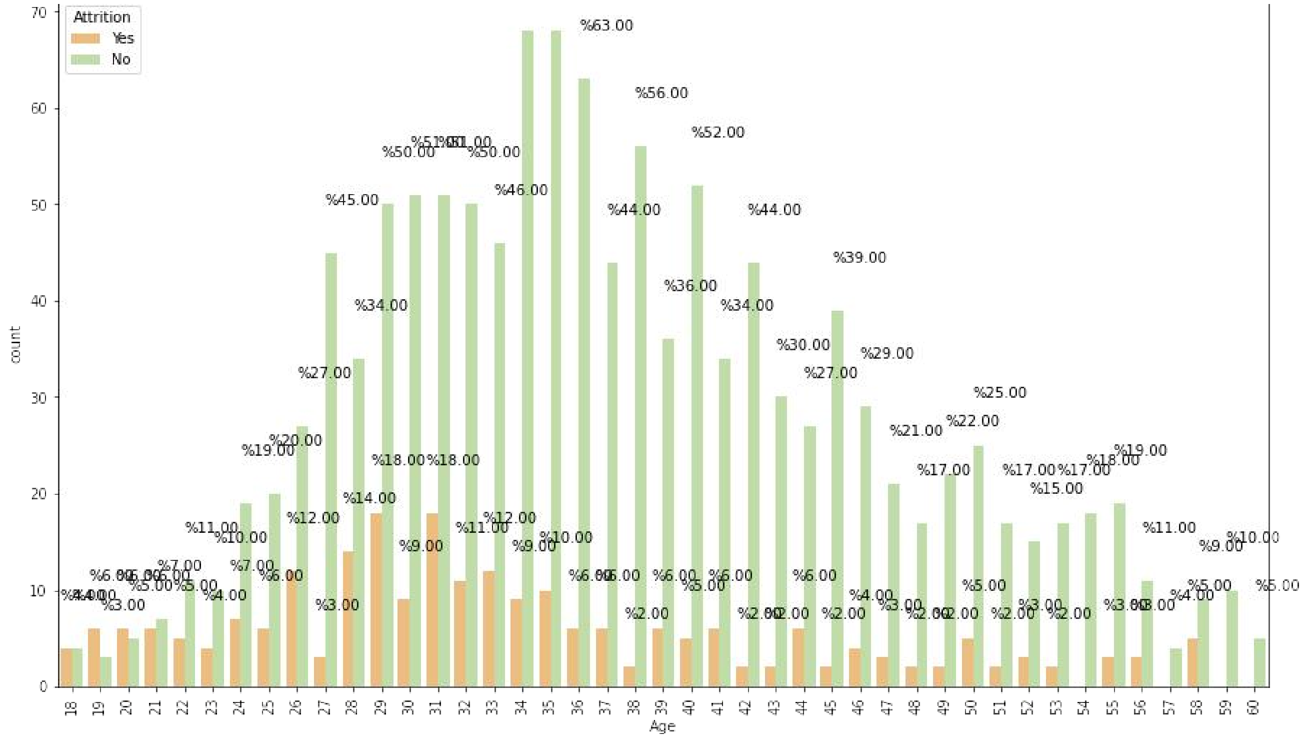


attrition bcs of education field
Attrition
Yes
No
count
0
100
200
300
400
500
%6.05
%35.17
%0.75
%4.83
%4.29
%27.28
%2.38
%8.44
%2.18
%6.80
%0.48
%1.36
Life Sciences
Other
Medical
Marketing
Technical Degree
Human Resources
EducationField

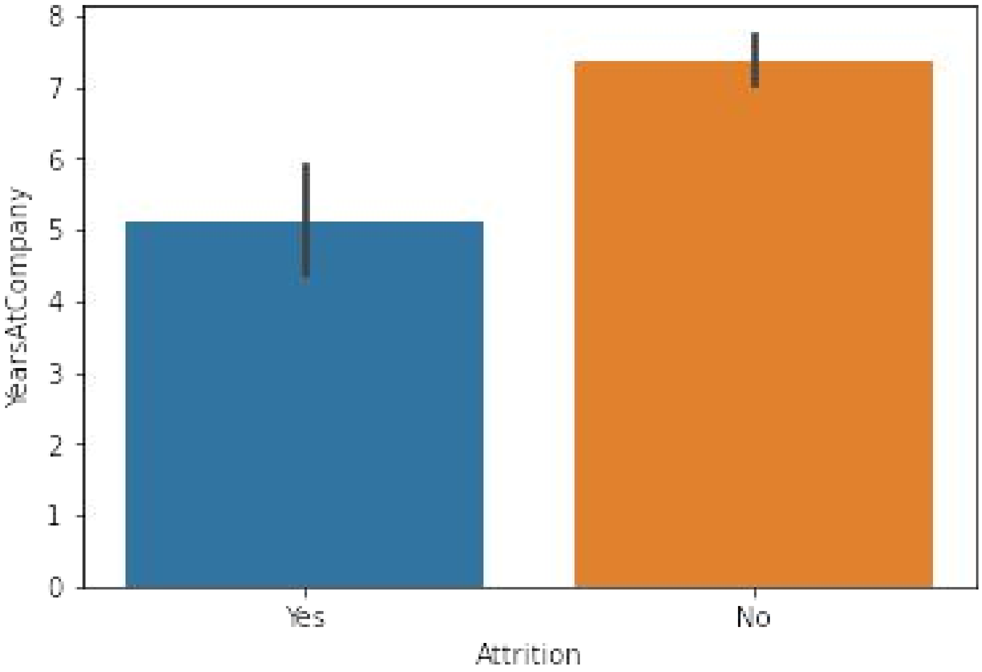

attrition due to job role
Attrition
Yes
No
count
0
50
100
150
200
250
%18.30
%16.67
%13.40
%9.18
%8.30
%6.60
%5.31
%3.40
%2.72
%3.88
%3.20
%4.22
%0.68
%0.61
%0.34
%2.24
%0.14
%0.82
Sales Executive
Research Scientist
Laboratory Technician
Manufacturing Director
Healthcare Representative
Manager
Sales Representative
Research Director
Human Resources
JobRole


attrition due to married status
Attrition
Yes
No
count
0
100
200
300
400
500
600
%8.16
%23.81
%5.71
%40.07
%2.24
%20.00
Single
Married
Divorced
MaritalStatus

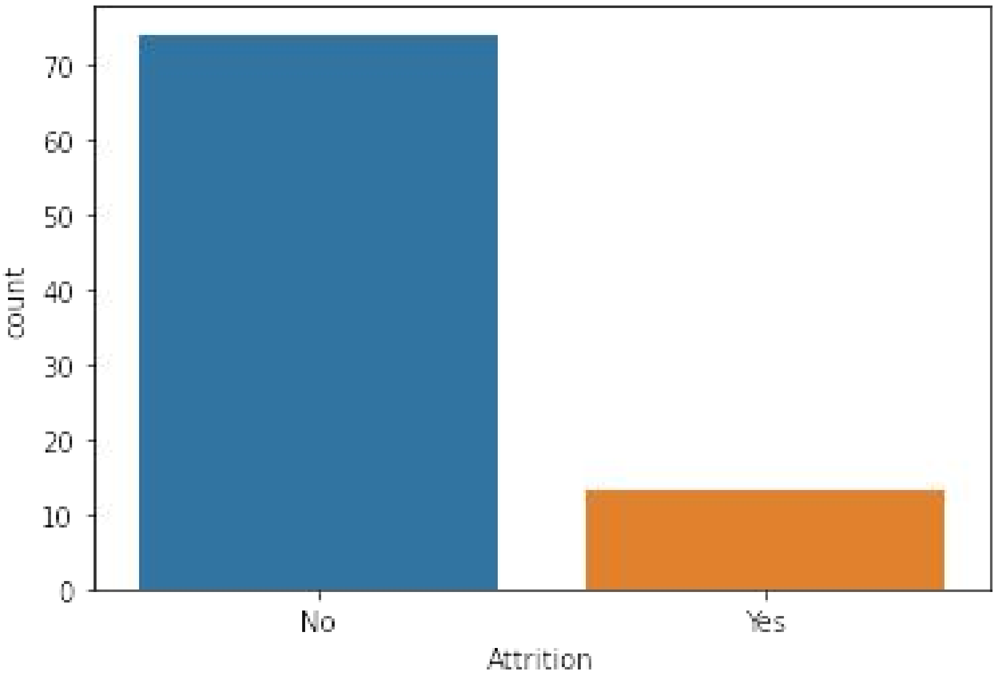

attrition due to ove time
Attrition
Yes
No
count
0
200
400
600
800
%8.64
%19.66
%7.48
%64.22
Yes
No
OverTime



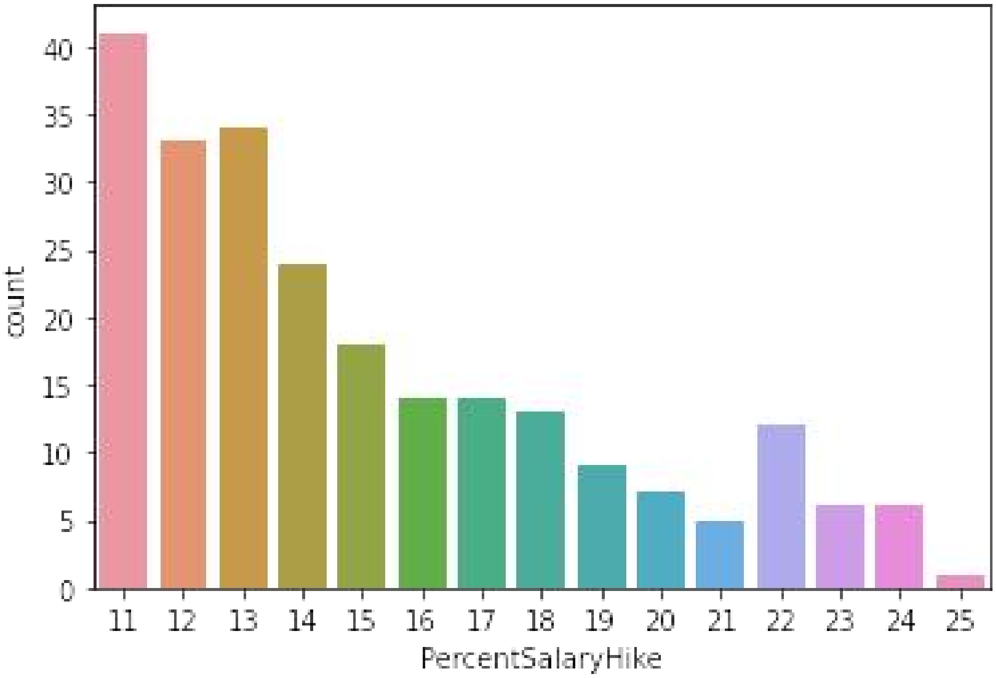
1. In Performance rating most of employees get **3 Rating** have left company most.
2. Working max up to 5 years at single company employee change job for better option.
3. Working distance from the also mater for employee attrition long distance travel avoid by employees
4. Environment satisfaction also role in attrition working environment not good ,behaviour of management not well, problem solving skills are not satisfied of management etc.., all are reasons behind employees leave company.
5. Hike in salary is most import role in attrition because employees not get best reward against work or partiality in hike are some are reasons for attrition,
6. 11% to 14% hike increasing employee left company earlier

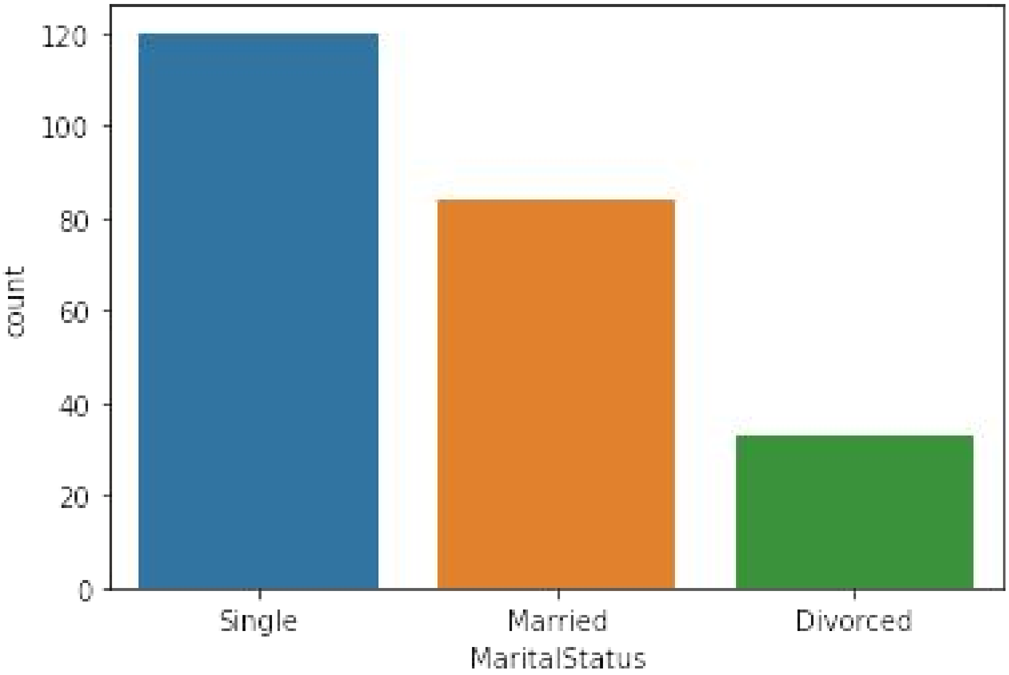


8
7
6
5
4
3
2
1
0
YearsAtCompany
Yes
No
Attrition



70
60
50
40
30
20
10
0
count
No
Yes
Attrition


count
0
5
10
15
20
25
30
35
40
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
PercentSalaryHike

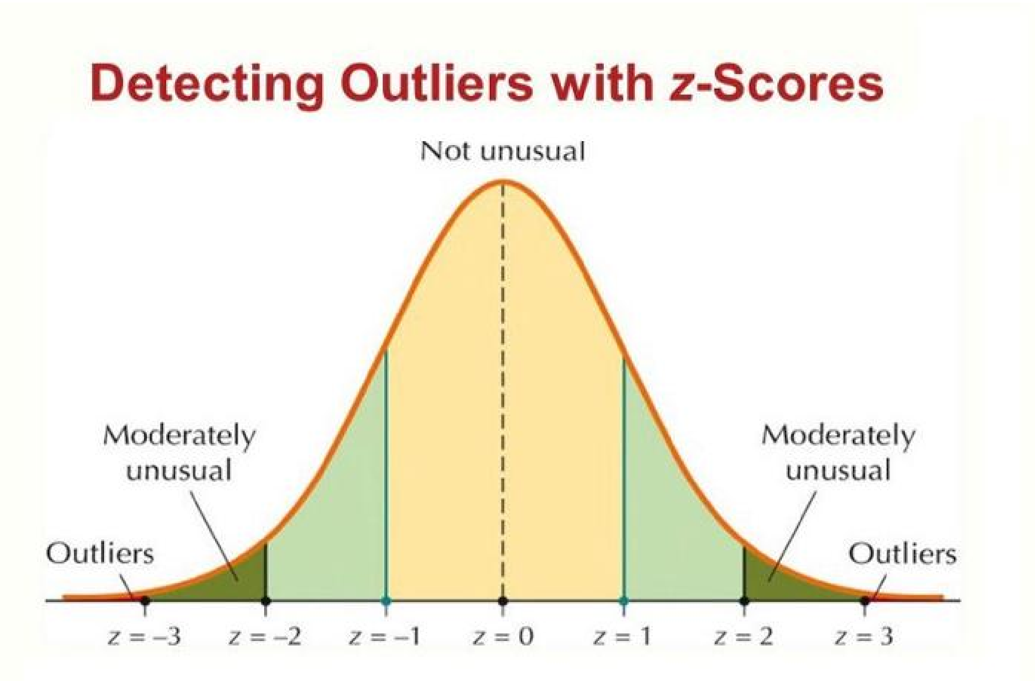

count
0
20
40
60
80
100
120
Single
Married
Divorced
MaritalStatus

- Counting the No. of year working and get promoted in that not promotion getting employees are most left company 50% of employees left who get one promotion

	Attrition counts
0	110
1	49
2	27
3	9
4	5
5	2
6	6
7	16
9	4
10	1
11	2
13	2
14	1
15	3

Data Cleaning and Tranformation:

Remove outliers by **z-SCORE** method Outliers Data that Far away from data points that can be happens by human error, data entry error or etc... from here we take threshold value as **3** in Normal distribution plot as ±3rd Standard deviation data point in that 99.7% data lies between them and reaming 0.3% data consider as Outliers



Picture taken From AnalyticsVidhya

Here code:

```
from scipy.stats import zscorez
z=np.abs(zscore(df))
print(np.where(z>3))
df_n=df[(z<3).all(axis=1)]
```

Transforming all categorical values in Numerical by “LabelEncoder”

Here code:

```
for i in categorical_features:
    df[i]=lb.fit_transform(df[i])
```

Checking Skewness:

CODE:

```
df.skew()
```

Checking Multicollinearity by Variance inflation factor (VIF)

What is VIF

- Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.
- Code for that:

```
    vif= pd.DataFrame()
    vif['vif'] = [variance_inflation_factor(x.values,i ) for i
in range(x.shape[1])]
    vif['columns']= x.columns
```

1. Remove multicollinearity columns that are below
2. from the heat map and relation with target column total working years and years in company are correlated indicated employee how many years employee worked
3. so, drop total working years in company column
4. Environment satisfaction & job satisfaction are indicated same thing so drop one column because both equally correlated with target
5. salary hike and performance rating are dependant thing so taking performance rate because salary hike negative with target also depend on performance
6. year in current role that indicate employee present in company so related with Years with manager so drop one that column
7. monthly rate also include daily rate because days in month daily rate negative with target so drop
8. Job level increase with performance rating it id depend on that
9. Work life balance also involve job involvement how employee balancing with work and personal life

Code:

```
x1=x.drop(columns=['TotalWorkingYears','EnvironmentSat
isfaction','PercentSalaryHike','YearsWithCurrManager',
'DailyRate','JobLevel','JobInvolvement'],axis=1)
```

Use Standard scaler for scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way.

Remove skewness from numerical dataset using log method for normally distribution of data

Split Data into Train & Test Sets:

Here we split data in two parts as train set for training models and test set testing models we divide into 70% train set data and 30% test set data

Resampling the data is an import thing in Classification problem because it is improving accuracy and quantify the uncertainty of Population parameters,

So, we use SMOTE up sampling method for balancing data set

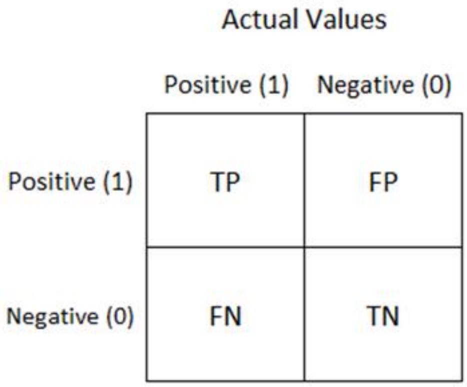
Prediction by Using Different Models:

1. Logistic Regression:
2. DecisionTree Classifier:
3. Support Vector Classifier:
4. KNeighbors Classifiers:
5. Stochastic Gradient Decent Classifier:
6. XGB Classifier:
7. RandomForest Classifier:
8. AdaBoost Classifiers:
9. GradientBoosting Classifier:
10. Bagging Classifier:

Base on Different matrics like accuracy score, f1-score, Confusion matrix, classification report we select out best model on high scores from that metrics

Understand Matrics:-

1. **Confusion Matrix: -** We get out best score in Random Forest Classifier as Accuracy score is 93.52 %, here model predict as True Positive, 28 False Positive , 17 False Negative and 328 True Negative.



Actual Values

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

2. **Accuracy: -** It can be defined as the ratio of total number of correct classifications divided by total number of classifications.
   Accuracy=(TP+TN)/(TP+FP+TN+FN)

3. **Recall:–** It is measure of correctly identified positive cases from all the actual positive cases.it is useful when cost of False Negative is high.
   Recall=TP/(TP+FN)

4. **Precision:** - It is measure of all the positive predictions how many of them actually positive.
5. Precision=TP/(TP+FP)

6. **F1-Score: -** It give the combine result of Recall and Precision, F1-score=2*(Precision*Recall)/ (Precision + Recall)

7. **Hyper Parameter Tunning:-** Hyper parameter optimisation in machine learning is used to find parameters of given machine learning algorithm that perform best as measured on validation. I used GridSearchCV for hyper tunning.

Final Model select with Gridsearch parameters using:-

```
        RandomForestClassifier(class_weight='balanced',
max_features='log2',n_estimators=2000)

        Train score: 100.0

        F1 score: 93.58059914407988

        Accuracy score: 93.5251798561151

        Confusion_matrix:
        [[322  28]
        [ 17 328]]
```

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

                              0       0.95      0.92      0.93       350
                              1       0.92      0.95      0.94       345

                       accuracy                           0.94       695
                      macro avg       0.94      0.94      0.94       695
                   weighted avg       0.94      0.94      0.94       695
```

Cross Validation:

This technique is used to check weather out data set is over fitting or under fitting. If model score is high and cv score is less it means model perform well in train dataset but did not perform well in unseen or test dataset. Feature selection is the best way to overcome the overfitting problem.

There are 3 ways for the validation. KFold Cross validation score, Hold Out Methods and LOOCV.

```
for i in range(2,15):
crs_score= cross_val_score(rf,x3,y3,cv=i)
score= crs_score.mean()
print('cv value:',i)
print('cross value score:',score*100)
print('actual score:',f1_score(y_test,pred_rf)*100)
```

```
cv value: 2
cross value score: 89.68048359240069
```

```
actual score: 93.58059914407988
cv value: 3
cross value score: 90.32815198618307
actual score: 93.58059914407988
cv value: 4
cross value score: 90.71675302245251
actual score: 93.58059914407988
cv value: 5
cross value score: 90.98225590228644
actual score: 93.58059914407988
cv value: 6
cross value score: 91.62348877374782
actual score: 93.58059914407988
cv value: 7
cross value score: 91.75396607420777
actual score: 93.58059914407988
cv value: 8
cross value score: 92.27583223958955
actual score: 93.58059914407988
cv value: 9
cross value score: 92.3214590936969
actual score: 93.58059914407988
cv value: 10
cross value score: 91.93032542170474
actual score: 93.58059914407988
cv value: 11
cross value score: 92.44927268623951
actual score: 93.58059914407988
cv value: 12
cross value score: 92.44386873920553
actual score: 93.58059914407988
cv value: 13
cross value score: 92.36780732292628
actual score: 93.58059914407988
cv value: 14
cross value score: 92.40833463725029
actual score: 93.58059914407988
```

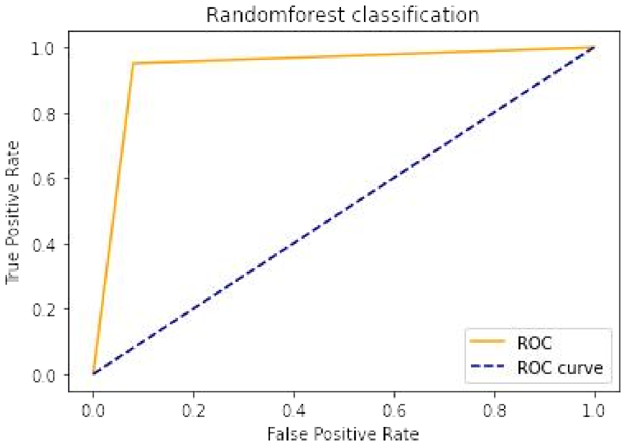
```
scores=cross_val_score(rf,x3,y3,cv=11,scoring='roc_auc')
      score= np.mean(scores)
      std= np.std(scores)
      print('CV mean',score)
      print('std:',std)

      CV mean 0.9790396844844285
```

```
std: 0.028610986968785227
```

AUC ROC Curve:

```
        fpr, tpr, threshold = roc_curve(y_test,pred_rf)
        print('AUC roc score: ',roc_auc_score(y_test,pred_rf))
        plt.plot(fpr, tpr, color ='orange', label ='ROC')
        plt.plot([0, 1], [0, 1], color='darkblue', linestyle='--',
label ='ROC curve')
        plt.xlabel('False Positive Rate')
        plt.ylabel('True Positive Rate')
        plt.title('Randomforest classification')
        plt.legend()
        plt.show()
```

```
AUC roc score:  93.53623188405799
```



Conclusion: -

1. We learned gleefully to analyse HR attrition Rate using “RANDOMFOREST CLASSIFIER” with help of Jupyter Notebook, Use of Couples of Codes and Visualization we find affecting areas that consideration for employees’ attrition for that improve working environment and make Human resource empower.
2. We used different classifiers like Logistic Regression, Decision tree Classifiers and Random Forest classifiers. And also used the data Balanced process and also hyper parameter tunning for improving score.
3. We use SMOTE method for class balancing.
4. We get good score in Random Forest Classifiers: - F1 score is 93.58 % and Roc AUC Score is 93.53 % the model performance is Well.