Predicting Employee Attrition

1. DATA COLLECTION AND PREPROCESSING:

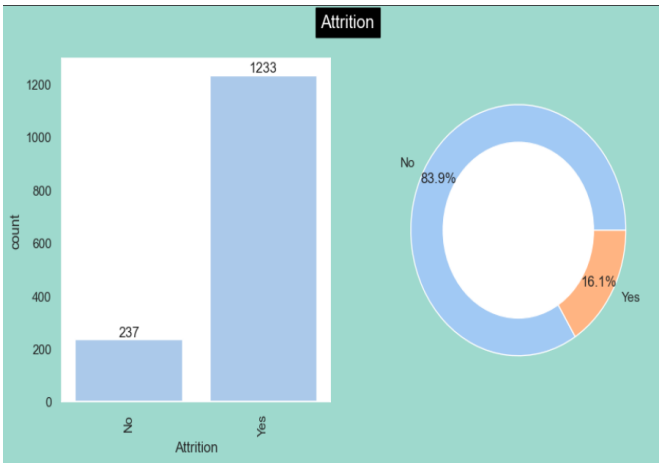
   I collected the IBM HR Analytics Employee Attrition & Performance dataset. After seeing the dataset, I found out that some of the columns were redundant, so I removed them from the dataset.

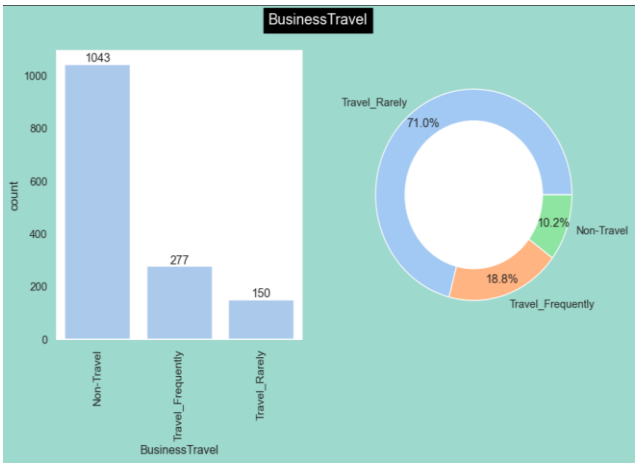
   1. Column 'EmployeeCount' is all 1s which indicate every observation is linked with 1 employee only, thanks for this info and we will drop it.
   2. Column 'StandardHours' is all 80s which means everyone in this dataset works as a fulltime employee and we could definitely drop it as well.
   3. Column 'Over18' is another interesting column which tells us every employee in this dataset is over 18 and we will drop it as well.

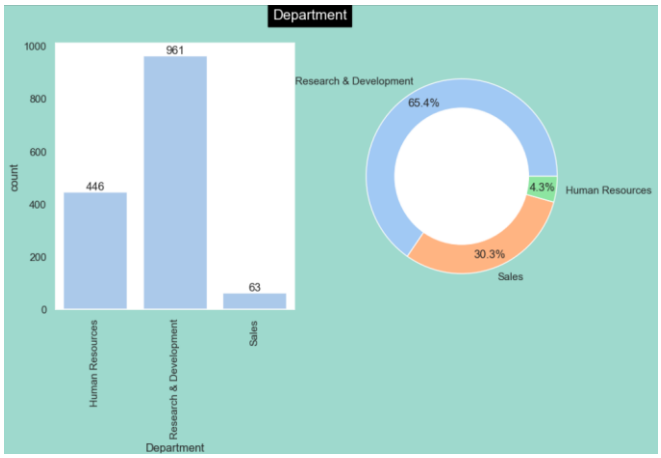
2. DATA VISUALISATION:

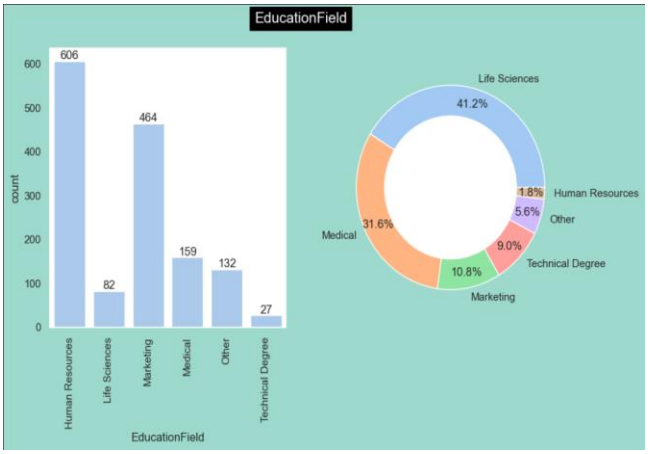
   I drew various types of plots using seaborn to represent the data and draw inferences from it.

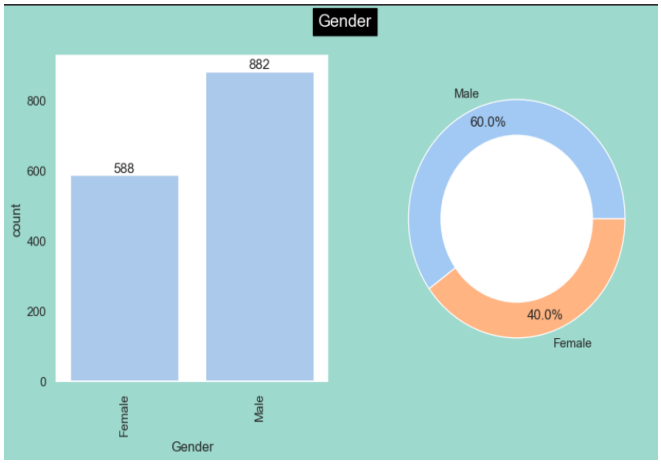
Some of the plots are given below.

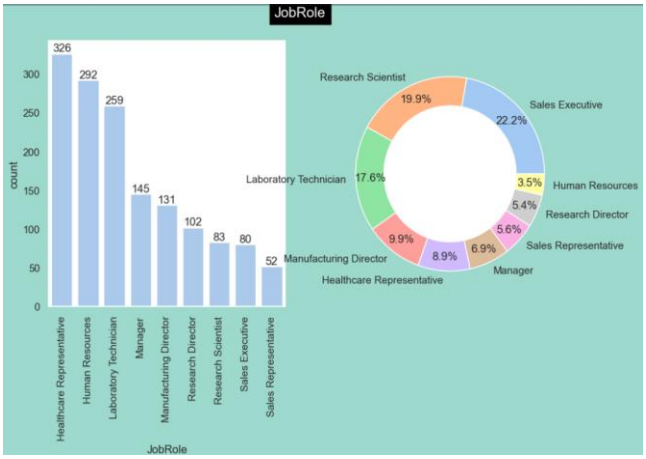


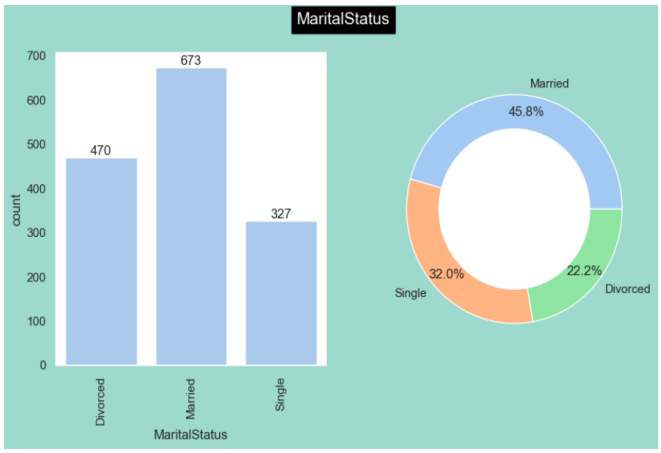


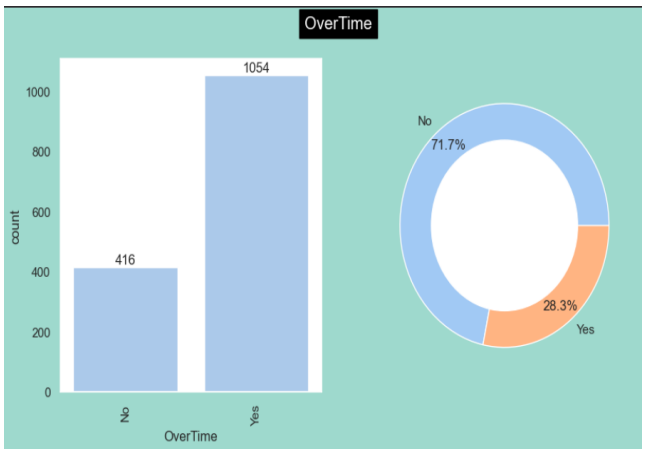

Department
count
1000
800
600
400
200
0
446
961
63
Human Resources
Research & Development
Sales
Department
Research & Development
65.4%
4.3%
Human Resources
30.3%
Sales

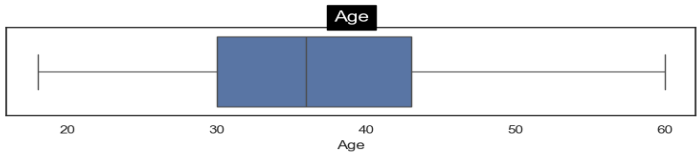

EducationField
count
600
500
400
300
200
100
0
606
82
464
159
132
27
Human Resources
Life Sciences
Marketing
Medical
Other
Technical Degree
EducationField
Life Sciences
41.2%
1.8%
Human Resources
5.6%
Other
Medical
31.6%
9.0%
Technical Degree
10.8%
Marketing

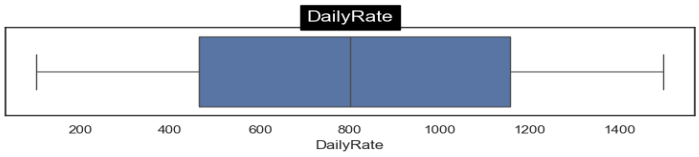

Gender
count
800
600
400
200
0
588
882
Female
Male
Gender
Male
60.0%
40.0%
Female

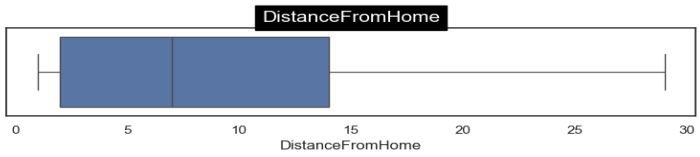

JobRole
count
300
250
200
150
100
50
0
326
292
259
145
131
102
83
80
52
Healthcare Representative
Human Resources
Laboratory Technician
Manager
Manufacturing Director
Research Director
Research Scientist
Sales Executive
Sales Representative
JobRole
Research Scientist
19.9%
Sales Executive
22.2%
Laboratory Technician
17.6%
3.5%
Human Resources
5.4%
Research Director
5.6%
Sales Representative
9.9%
6.9%
Manager
8.9%
Manufacturing Director
Healthcare Representative

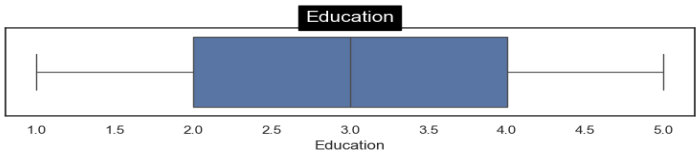

MaritalStatus
count
700
600
500
400
300
200
100
0
470
673
327
Divorced
Married
Single
MaritalStatus
Married
45.8%
Single
32.0%
22.2%
Divorced


OverTime
count
1000
800
600
400
200
0
416
1054
No
Yes
OverTime
No
71.7%
28.3%
Yes

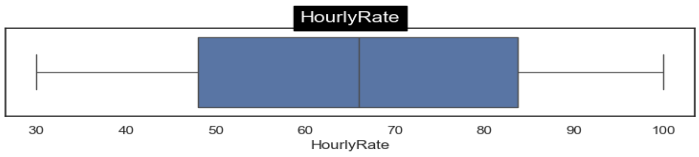

Age
20
30
40
50
60
Age

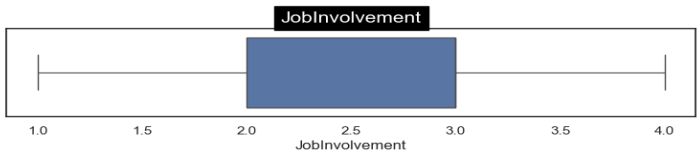

DailyRate
200
400
600
800
1000
1200
1400
DailyRate

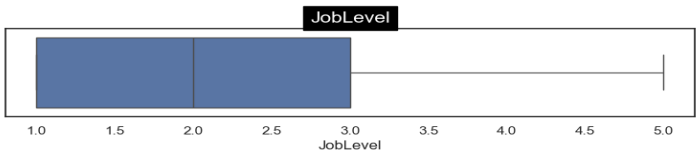

DistanceFromHome
0
5
10
15
20
25
30
DistanceFromHome


Education
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Education

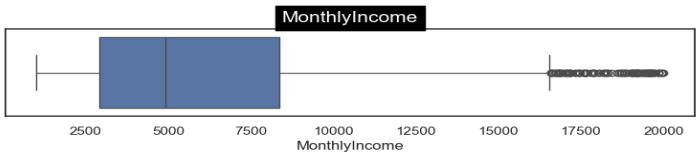

EnvironmentSatisfaction
1.0
1.5
2.0
2.5
3.0
3.5
4.0
EnvironmentSatisfaction

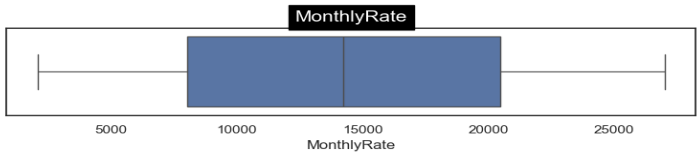

HourlyRate
30
40
50
60
70
80
90
100
HourlyRate

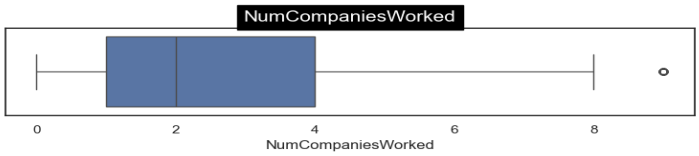

JobInvolvement
1.0
1.5
2.0
2.5
3.0
3.5
4.0
JobInvolvement


JobLevel
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
JobLevel

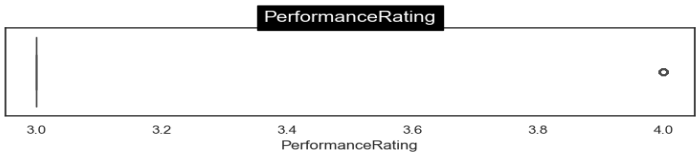

JobSatisfaction
1.0
1.5
2.0
2.5
3.0
3.5
4.0
JobSatisfaction


MonthlyIncome
2500
5000
7500
10000
12500
15000
17500
20000
MonthlyIncome


MonthlyRate
5000
10000
15000
20000
25000
MonthlyRate

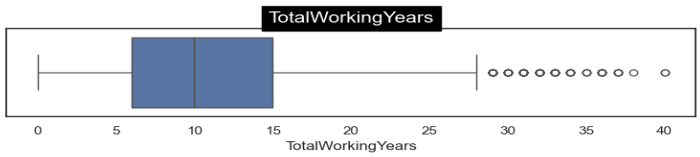

NumCompaniesWorked
0
2
4
6
8
NumCompaniesWorked


PercentSalaryHike
12
14
16
18
20
22
24
PercentSalaryHike


PerformanceRating
3.0
3.2
3.4
3.6
3.8
4.0
PerformanceRating

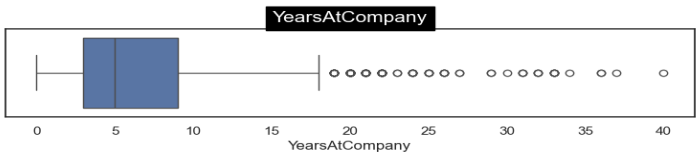

RelationshipSatisfaction
1.0
1.5
2.0
2.5
3.0
3.5
4.0
RelationshipSatisfaction

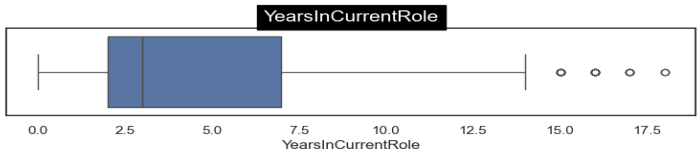

StockOptionLevel
0.0
0.5
1.0
1.5
2.0
2.5
3.0
StockOptionLevel

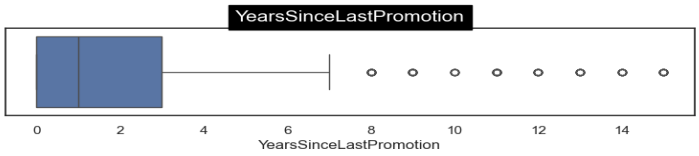

TotalWorkingYears
0
5
10
15
20
25
30
35
40
TotalWorkingYears

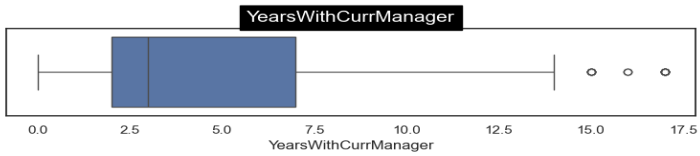

TrainingTimesLastYear
0
1
2
3
4
5
6
TrainingTimesLastYear

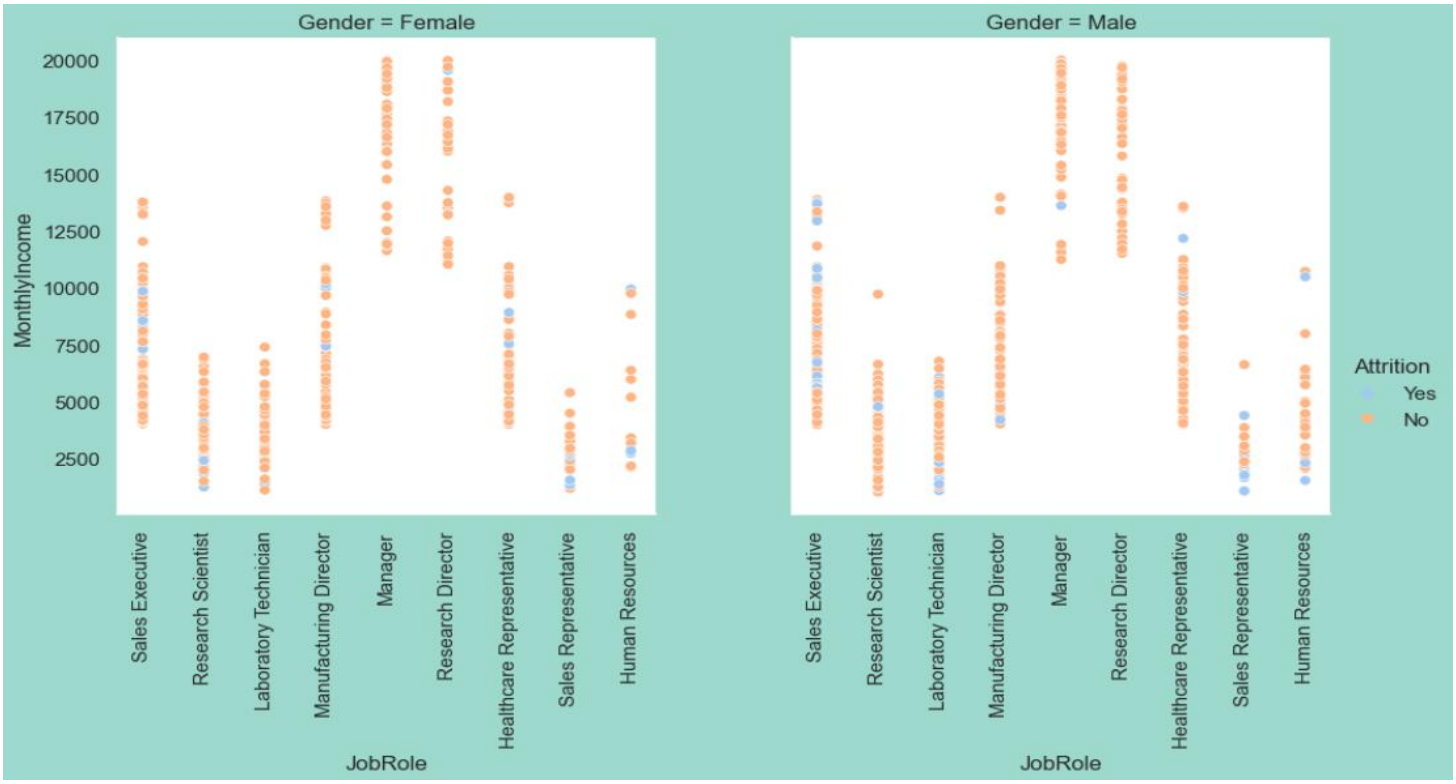

WorkLifeBalance
1.0
1.5
2.0
2.5
3.0
3.5
4.0
WorkLifeBalance


YearsAtCompany
0
5
10
15
20
25
30
35
40
YearsAtCompany


YearsInCurrentRole
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
YearsInCurrentRole


YearsSinceLastPromotion
0
2
4
6
8
10
12
14
YearsSinceLastPromotion

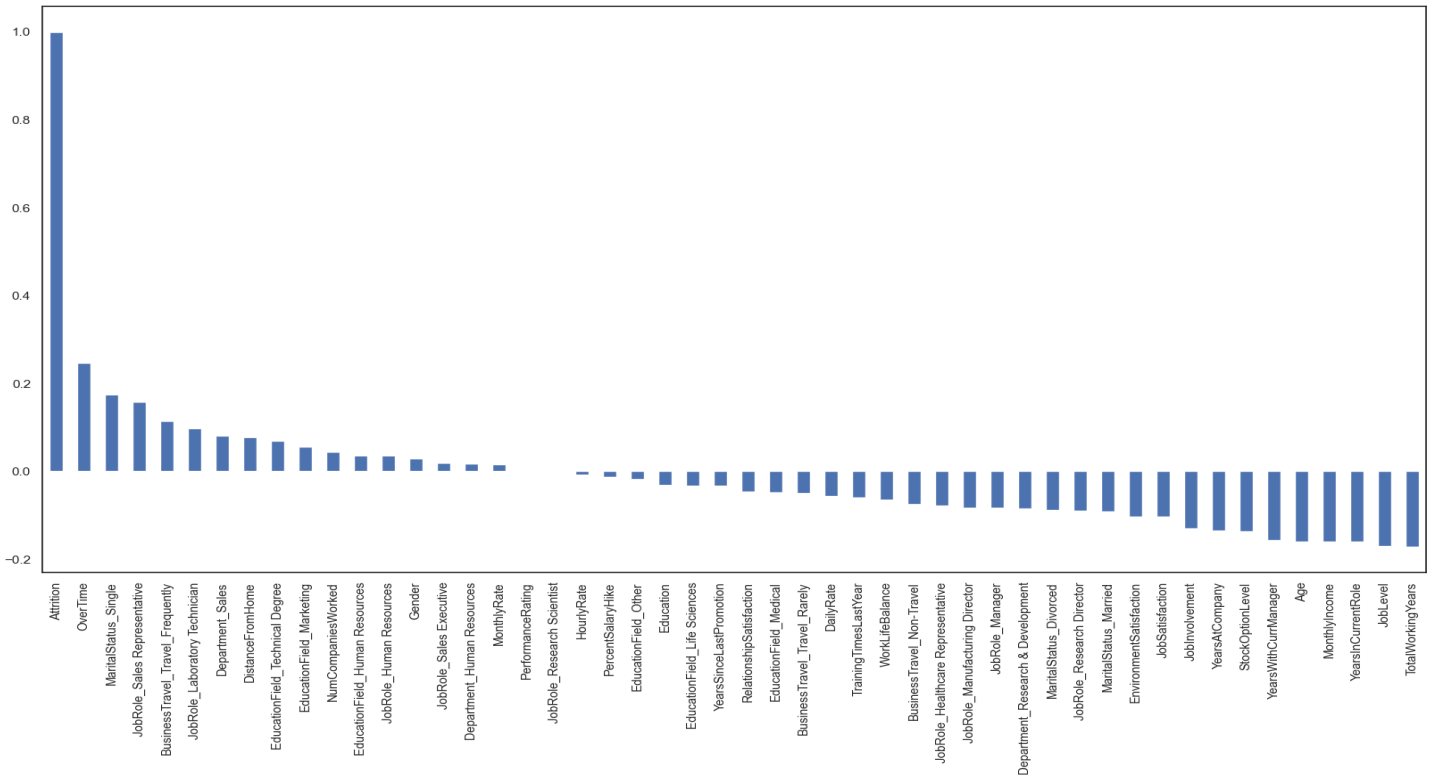

YearsWithCurrManager
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
YearsWithCurrManager

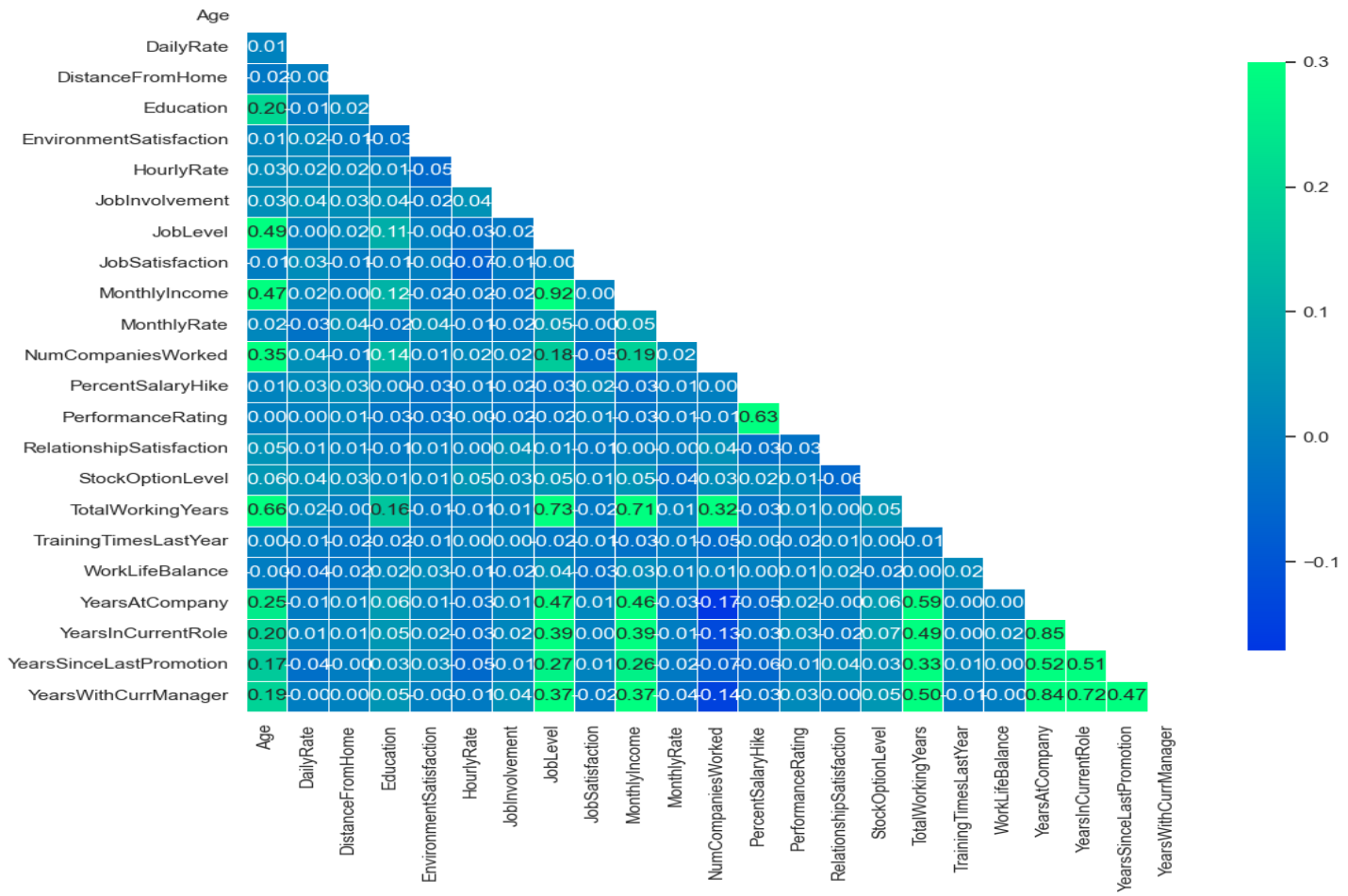


💡 Analysis of graphs

- **Age and Attrition:**
  - Attrition is highest for both men and women between **18 and 35 years of age**, gradually decreasing afterward.
- **Income and Attrition:**
  - As **income increases**, attrition decreases.
- **Marital Status and Attrition:**
  - Attrition is **much lower** for divorced women.
- **Travel Frequency and Attrition:**
  - Employees who **travel frequently** have higher attrition rates, especially among women.
- **Job Level and Attrition:**
  - Attrition is highest for employees in **level 1 jobs**.
- **Specific Job Positions and Attrition:**
  - Women in positions such as **manager, research director, and laboratory technician** experience almost no attrition.
  - Men in the **sales expert** position have a significant attrition rate.

FEATURE CORELATIONS AND HEATMAPS:





- **High Correlations:**
  - **Monthly Income** and **Job Level**
  - **Year in Current Role**, **Year at Company**, and **Year with Current Manager** with **Year in Current Role**

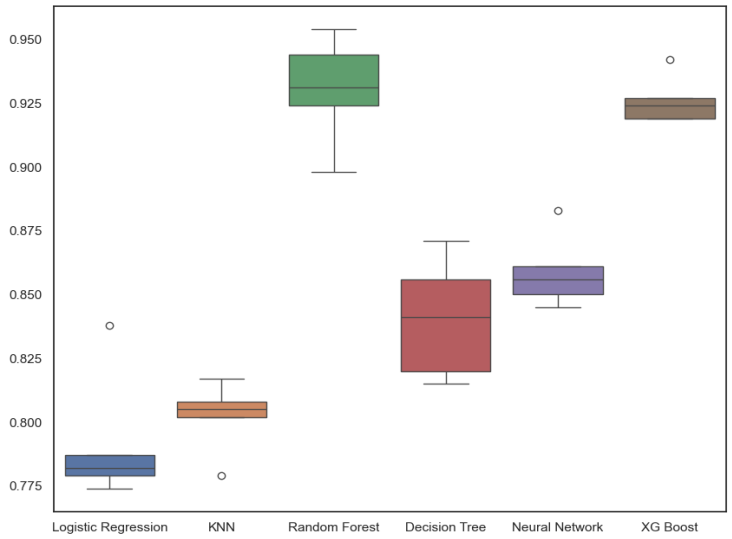
We see a greate imbalance in the attrition label of the data. Standard ML techniques such as Decision Tree and Logistic Regression have a bias towards the majority class, and they tend to ignore the minority class. They tend only to predict the majority class, hence, having major misclassification of the minority class in comparison with the majority class. In more technical words, if we have imbalanced data distribution in our dataset then our model becomes more prone to the case when minority class has negligible or very lesser recall.

To solve this we use oversampling. In this case we will use **SMOTE(synthetic minority oversampling technique)**

MODEL SELECTION:

We now fit this dataset into multiple models in a naive way, i.e., without hyperparameter tuning.
After fitting and having a 5-fold cross validation, we get these metrics:

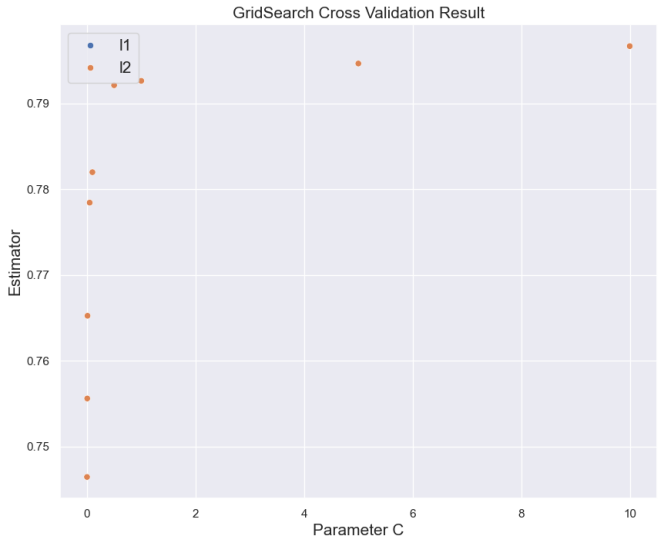
```
============================================================
The Score is listed below

      Logistic Regression    KNN  Random Forest  Decision Tree
cv_1                0.782  0.808          0.954          0.841
cv_2                0.838  0.805          0.924          0.856
cv_3                0.787  0.817          0.944          0.815
cv_4                0.779  0.779          0.898          0.871
cv_5                0.774  0.802          0.931          0.820

     Neural Network  XG Boost
cv_1          0.856     0.927
cv_2          0.861     0.919
cv_3          0.883     0.942
cv_4          0.850     0.924
cv_5          0.845     0.919
============================================================
```



We can infer that the logistic regression model here isn't doing that well.

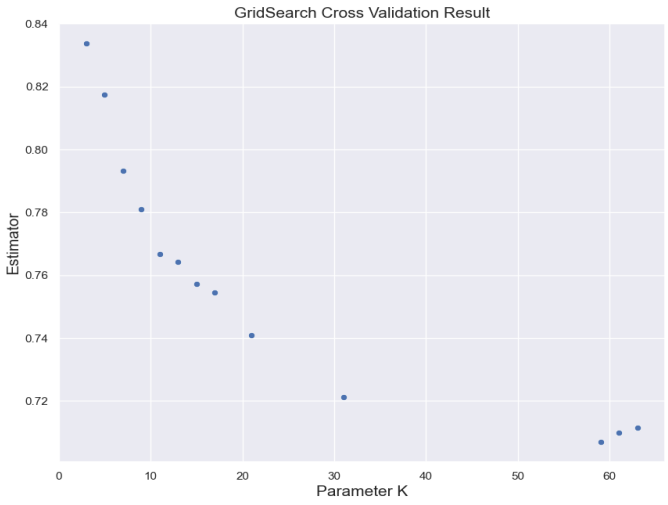
Now we use Grid-Search to further optimize the hyperparameters of these models.

- Logistic regression after tuning:



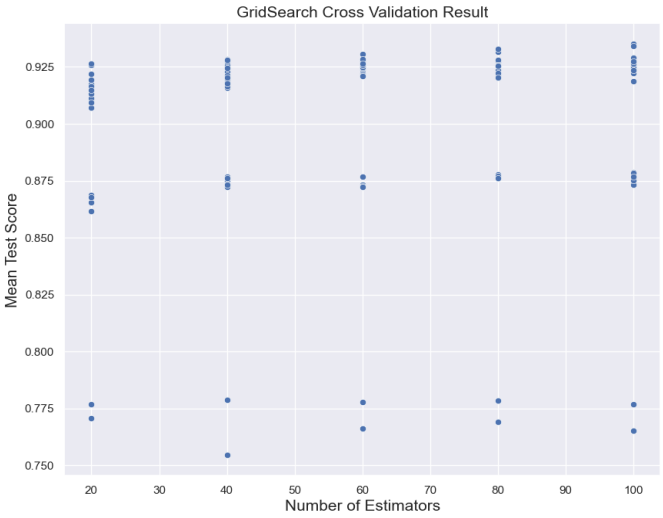

```
Best Score: 0.797
Best Parameters:
C : 10
penalty : l2
```

- KNN after tuning:



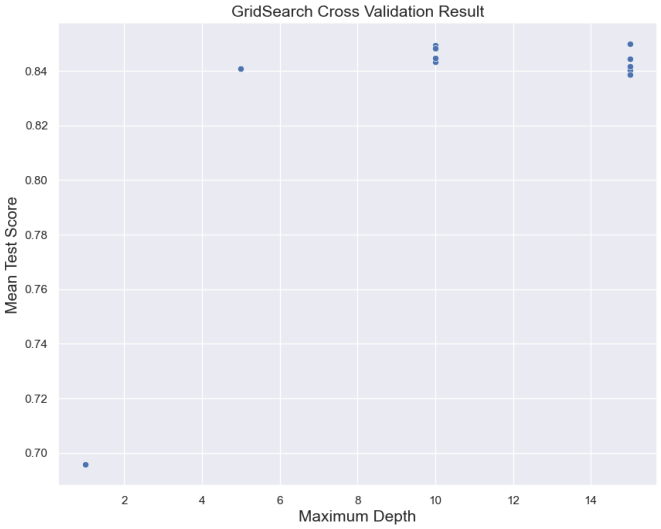

```
Best Score: 0.834
Best Parameters:
n_neighbors : 3
```

- Random Forest after tuning:



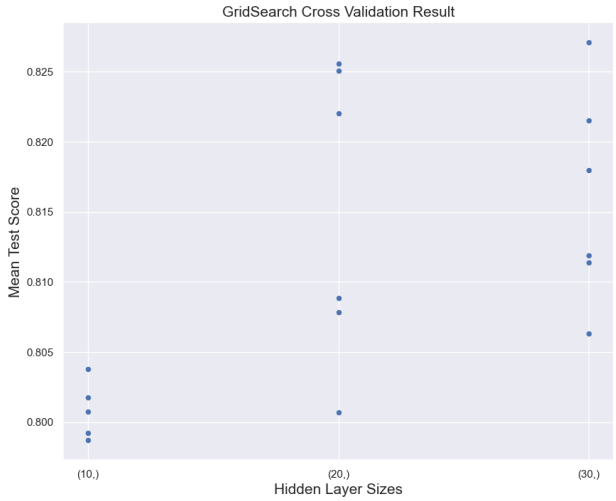

```
Best Score: 0.935
Best Parameters:
max_depth : 15
max_features : log2
min_samples_split : 4
n_estimators : 100
```

- Decision Trees after tuning:



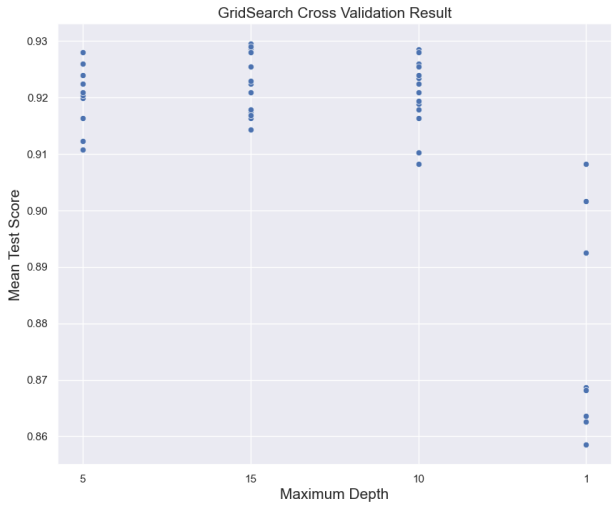

```
Best Score: 0.850
Best Parameters:
max_depth : 15
min_samples_split : 2
```

- MLP after tuning:



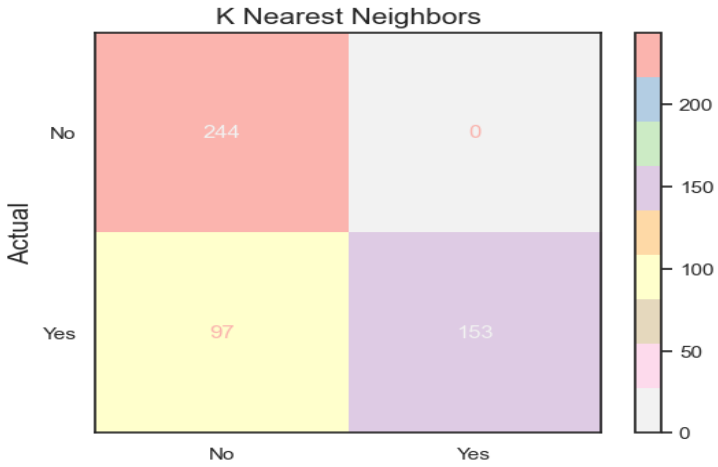

```
Best Score: 0.827
Best Parameters:
activation : relu
alpha : 0.1
hidden_layer_sizes : (30,)
```

- XGBoost after tuning:

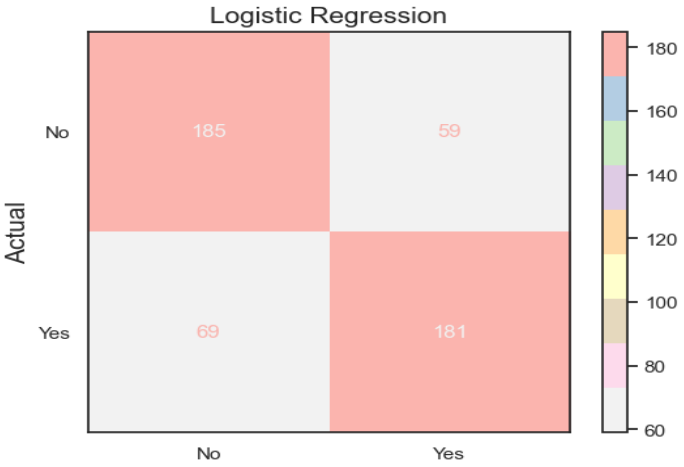



```
Best Score: 0.930
Best Parameters:
subsample : 0.6
reg_lambda : 0.01
reg_alpha : 0.01
n_estimators : 60
max_depth : 15
learning_rate : 0.3
eta : 0.8
```

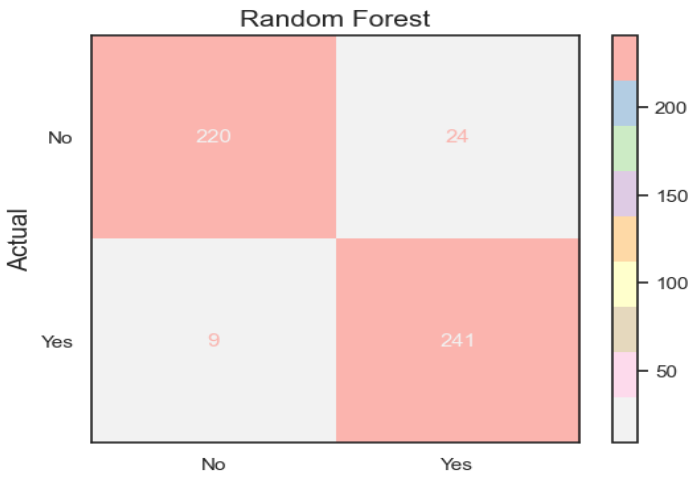
We can now check the metrics of these tuned models:



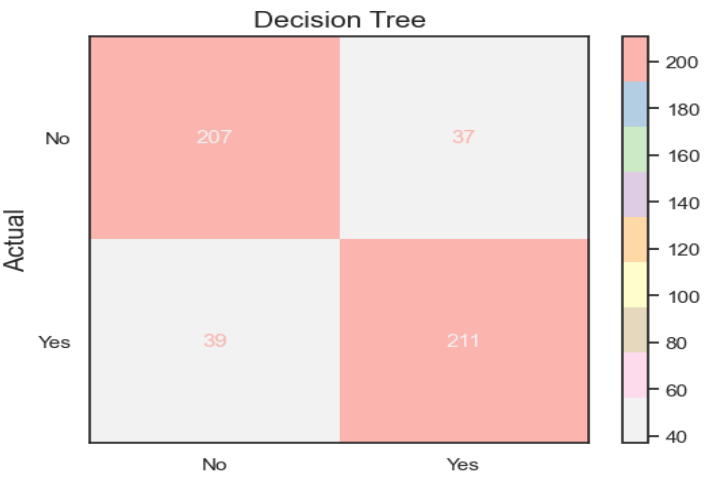
Accuracy = 0.804
Precision = 1.0
Recall = 0.612
F1 = 0.759



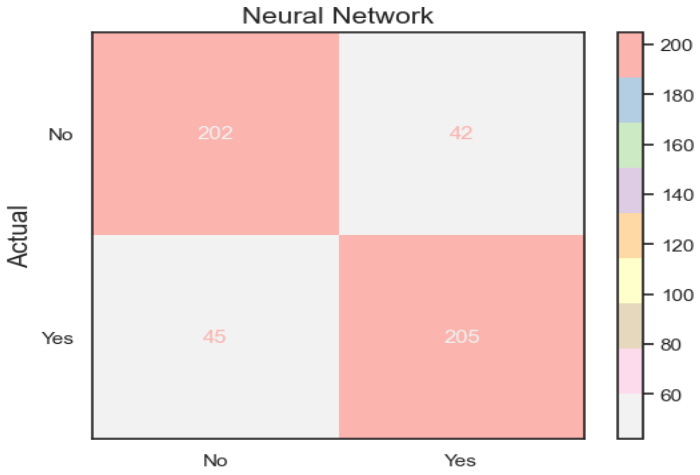
Accuracy = 0.741
Precision = 0.754
Recall = 0.724
F1 = 0.739



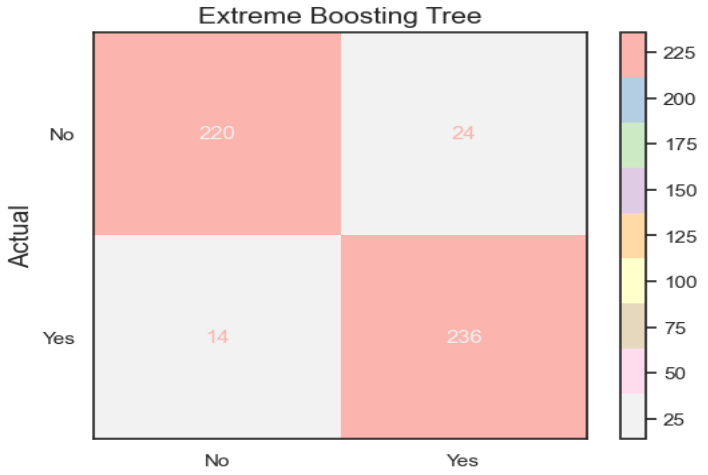
Accuracy = 0.933
Precision = 0.909
Recall = 0.964
F1 = 0.936



Accuracy = 0.846
Precision = 0.851
Recall = 0.844
F1 = 0.847



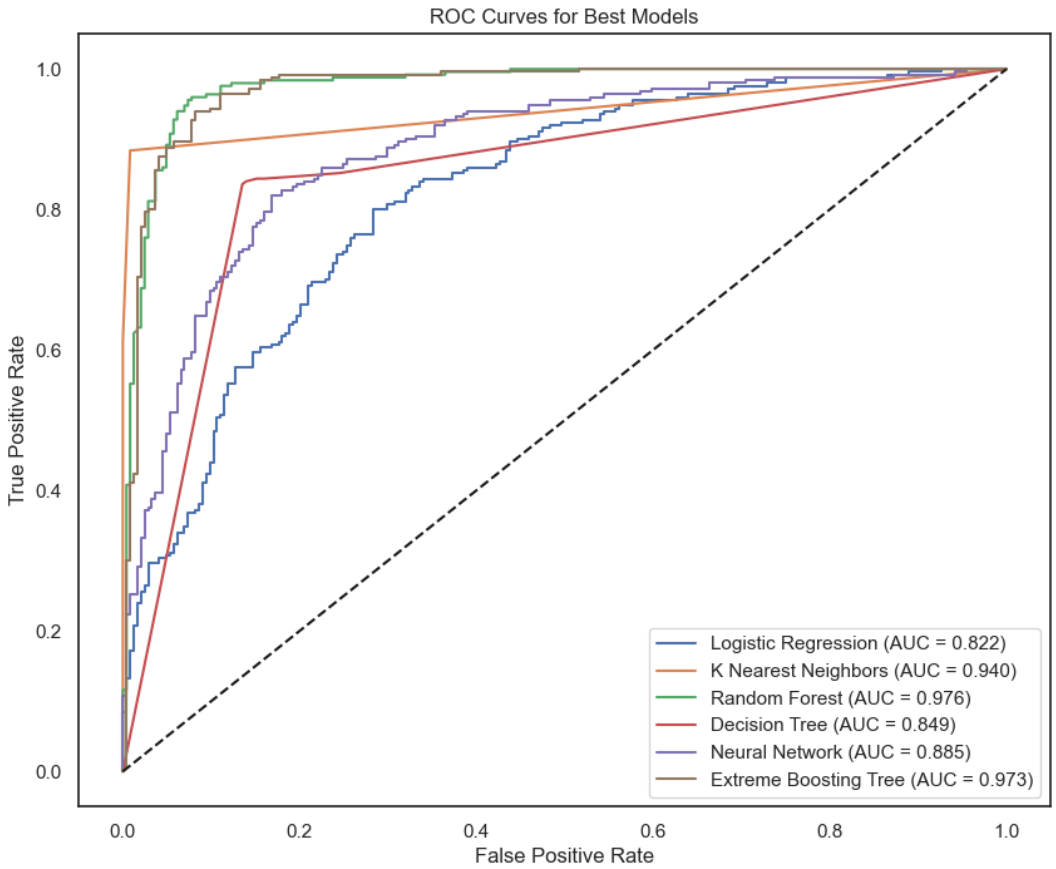
Accuracy = 0.824
Precision = 0.83
Recall = 0.82
F1 = 0.825



Accuracy = 0.923
Precision = 0.908
Recall = 0.944
F1 = 0.925

From these confusion matrices we can observe that the best performing model is the Random Forest model.
Since we did SMOTE oversampling our overall metrics are much better than the models available online.

We can also observe its AUC scores:



CONCLUSION:

The employee attrition analysis on the IBM HR analytics dataset successfully developed a high-performing binary classification model leveraging the Random Forest algorithm. After addressing the significant class imbalance through SMOTE oversampling, a rigorous model evaluation process involving cross-validation and hyperparameter tuning was conducted across multiple machine learning classifiers. The Random Forest model emerged superior, achieving an **accuracy of 93.3%, recall of 96.4%, F1-score of 93.6%, and precision of 90.9%.** This robust model demonstrates exceptional capability in accurately identifying attrition cases while maintaining a desirable balance *between precision and recall*. It also has a very high **AUC score of 97.6%** so its predictions are very stable and non-biased. Its deployment can enable organizations to proactively mitigate employee turnover and optimize workforce management strategies through targeted retention initiatives.