Capestone Project

TITLE : Application of Data Science To Reduce Employee Attrition

Guided By :
Siddharth Koshta

Submitted By :
Hari Haran
Radhika Pandian
Shivangi Bharadwaj
Vishwesh
Yash Gehlot

Problem Statement

Retaining valuable employees and preventing their resignation is a matter that can make a company save a considerable amount of time and money. Traditionally, this task had been carried out by the Human Resources department of the companies, who would regularly conduct interviews among the employees in order to subsequently analyse them and try to extract conclusions and patterns that could help them understand the reasons why employees leave and thus, prevent the resignation of other employees in the future.

The problem we are trying to solve is to predict employee attrition in the IBM company using a classification model.

Objective

The aim of this project is to provide a solution to this problem by means of Data Science and Data Analytics.

To achieve this goal, IBM datasets have been collected from open data sources in order to be processed with data analytics technologies to extract insights that can help understand the data and can model the profile of the employees that abandon the company. In addition, data mining techniques will be used with the goal of obtaining a prediction that can allow managers to anticipate the employees' attrition in order to prevent it. The objective is to apply some Data Science techniques to analyze employee attrition in two different scenarios.

Data Set Description

(Data source: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.)

- This data set presents an employee survey from IBM, indicating if there is attrition or not. The data set contains approximately 24000 entries. Given the limited size of the data set, the model should only be expected to provide modest improvement in identification of attrition vs a random allocation of probability of attrition.

- While some level of attrition in a company is inevitable, minimizing it and being prepared for the cases that cannot be helped will significantly help improve the operations of most businesses.

- As a future development, with a sufficiently large data set, it would be used to run a segmentation on employees, to develop certain “at risk” categories of employees. This could generate new insights for the business on what drives attrition, insights that cannot be generated by merely informational interviews with employees.

Employee Details

Age	Employee Age
Attrition	Employee leaving the company
BusinessTravel	Travel By Employee
DailyRate	Salary Level
Department	Department of Employee
DistanceFromHome	Distance from work to home
Education	Education Level
EducationField	Education Background Field
EmployeeCount	Employee Count
EmployeeNumber	Employee ID
ApplicationID	Employee Application ID
EnvironmentSatisfaction	Employees Satisfaction with the Environment
Gender	Employee's Gender
HourlyRate	Employee Hourly Salary rate
JobInvolvement	Employee's Dedication Towards Work
JobLevel	Employee's Job Level
JobRole	Employee Job Role in Company
JobSatisfaction	Employee's Satisfaction rating to Job

Data Set Information:

- Number of Features - 36 (Independent Variable)
- Target Variable - Attrition (Dependent Variable)
- Number of Records - 23436
- Reductant Feature - EmployeeCount, EmployeeNumber, ApplicationID, StandardHours, Over18
- Number of Categorical Features - 25
- Number of Quantitative Features - 11

MaritalStatus	Marital Status of Employee
MonthlyIncome	Employee’s Monthly Salary
MonthlyRate	Employee’s Monthly Rate
NumCompaniesWorked	Number of Companies worked at
Over18	Employee Age above 18
OverTime	Overtime(1=Yes, 0=No)
PercentSalaryHike	Employee’s percent of Hike in Salary
PerformanceRating	Employee’s performance towards work
RelationshipSatisfaction	Organization's efforts to create and maintain a positive relationship with its employees
StandardHours	Employee’s working hours every 15days
StockOptionLevel	Type of equity compensation granted by company to employees
TotalWorkingYears	Employee’s total working experience
TrainingTimeLastYear	Employee’s total number of trainings last year
WorkLifeBalance	Ability to manage both personal and professional responsibilities
YearsAtCompany	Total years worked at the comapany
YearsInCurrentRole	Number of years in the current job role
YearsSinceLastPromotion	Number of years since last promotion
YearsWithCurrManager	Number of years spent with the current manger
Employee Source	Source of employee from where he connect with the company

Missing Values

- All features have missing values.
- All the features have less than 0.1% of the missing values.
- If we drop the missing values we have sufficient data and go forward with the actual data.

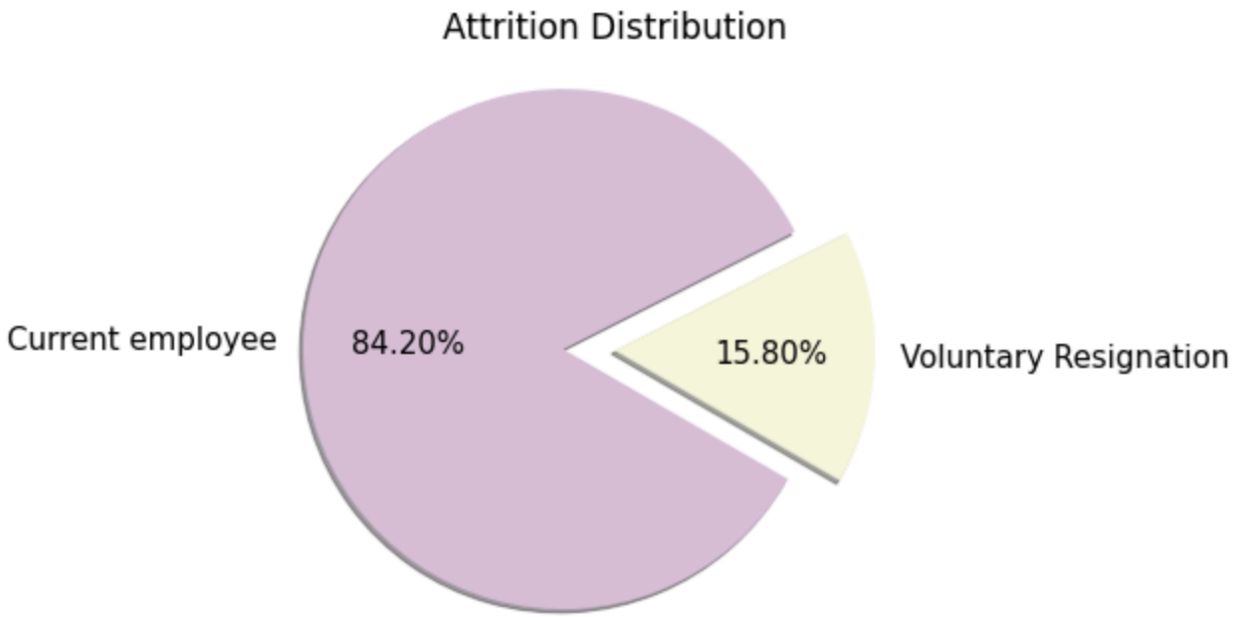
index	No. Missing Values	Percent
Age	3	0.010000
Attrition	13	0.060000
BusinessTravel	8	0.030000
DailyRate	12	0.050000
Department	11	0.050000
DistanceFromHome	9	0.040000
Education	12	0.050000
EducationField	9	0.040000
EmployeeCount	5	0.020000
EmployeeNumber	1	0.000000
Application ID	3	0.010000
EnvironmentSatisfaction	9	0.040000
Gender	10	0.040000
HourlyRate	9	0.040000
JobInvolvement	9	0.040000
JobLevel	7	0.030000
JobRole	9	0.040000
JobSatisfaction	9	0.040000
MaritalStatus	11	0.050000

index	No. Missing Values	Percent
MonthlyIncome	13	0.060000
MonthlyRate	11	0.050000
NumCompaniesWorked	9	0.040000
Over18	10	0.040000
OverTime	12	0.050000
PercentSalaryHike	14	0.060000
PerformanceRating	10	0.040000
RelationshipSatisfaction	8	0.030000
StandardHours	10	0.040000
StockOptionLevel	9	0.040000
TotalWorkingYears	8	0.030000
TrainingTimesLastYear	11	0.050000
WorkLifeBalance	10	0.040000
YearsAtCompany	13	0.060000
YearsInCurrentRole	15	0.060000
YearsSinceLastPromotion	11	0.050000
YearsWithCurrManager	7	0.030000
Employee Source	12	0.050000

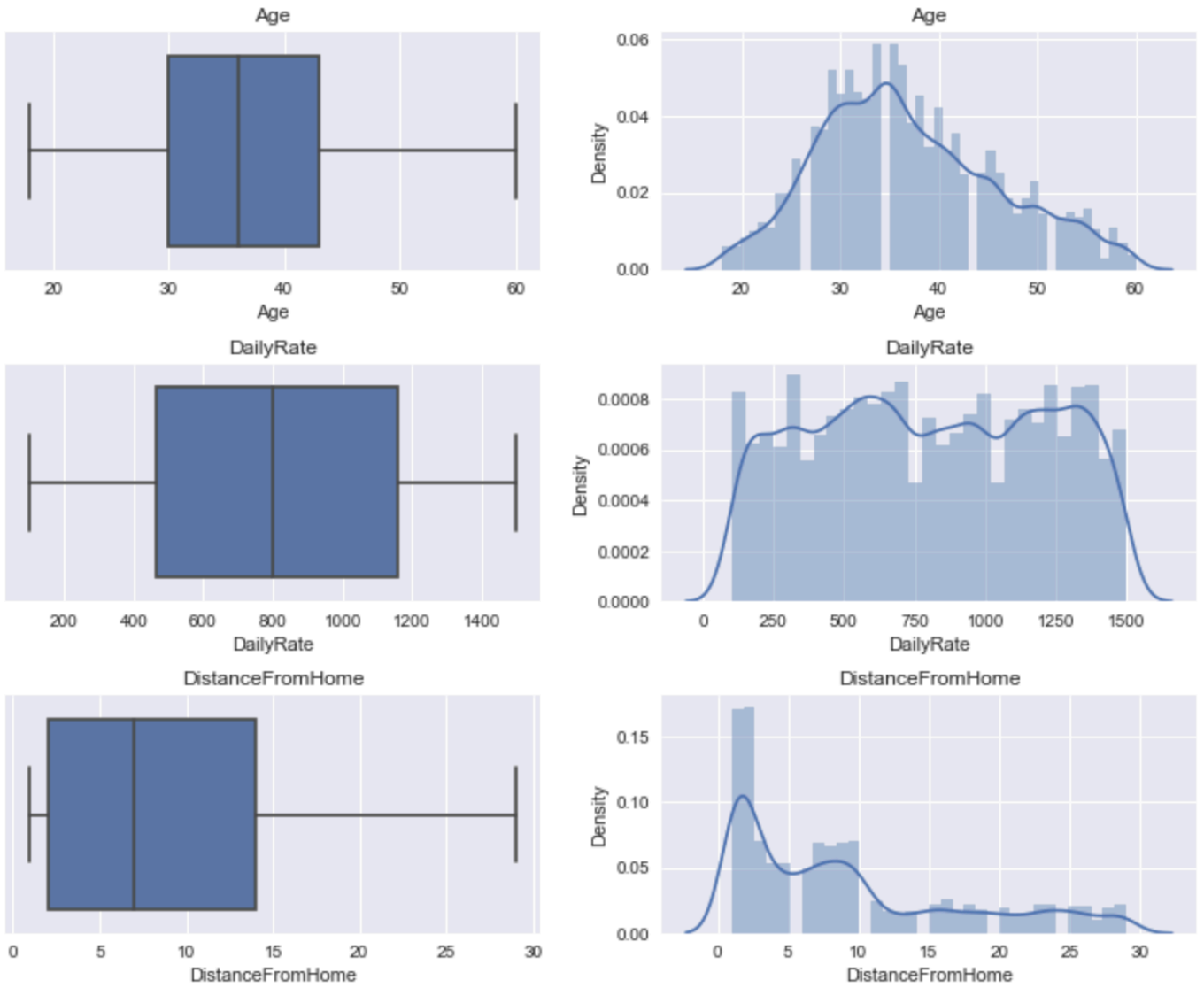
Univariate Analysis

Target Variable :

- In our data over 15.8% people have resigned, that is around 3700 of them.

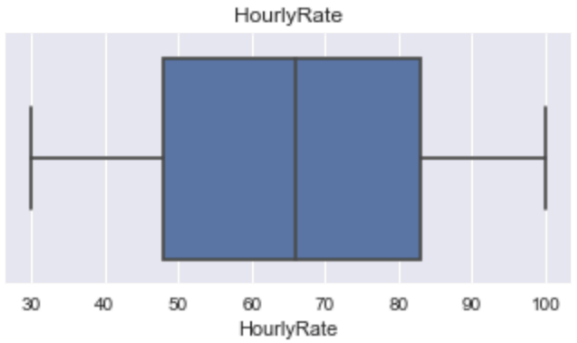


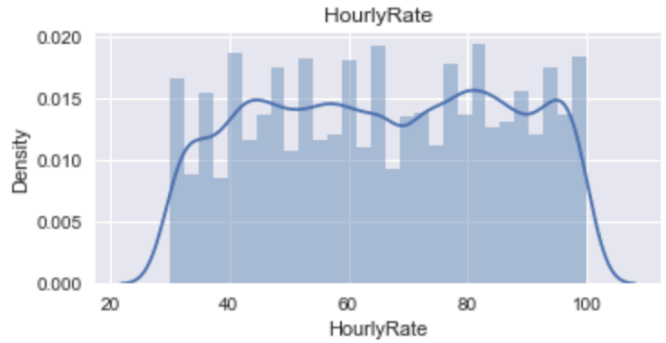
Quantitative Variables

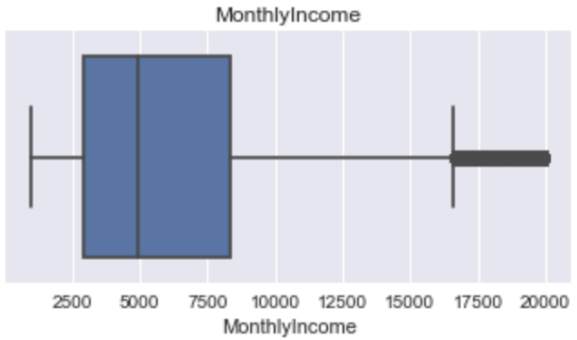


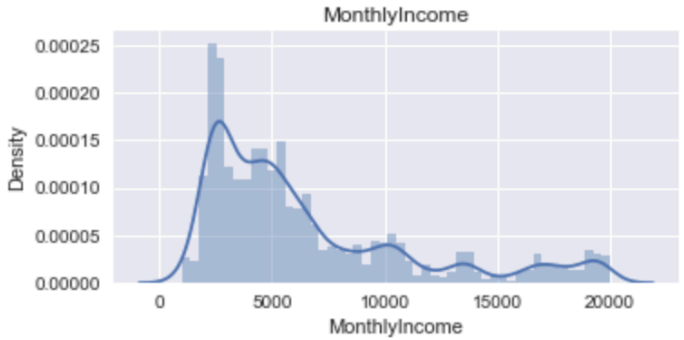
- **Age :** Majority of employees lie between the age range of 30 - 40 years.
- **Daily Rate :** The average of daily rate is somewhere around $802, minimum is $102 and maximum is $1499.
- **Distance From Home :** We can see that the average distance from home is around 9 Km, minimum is 1 Km and maximum is 29 Km.

Quantitative Variables







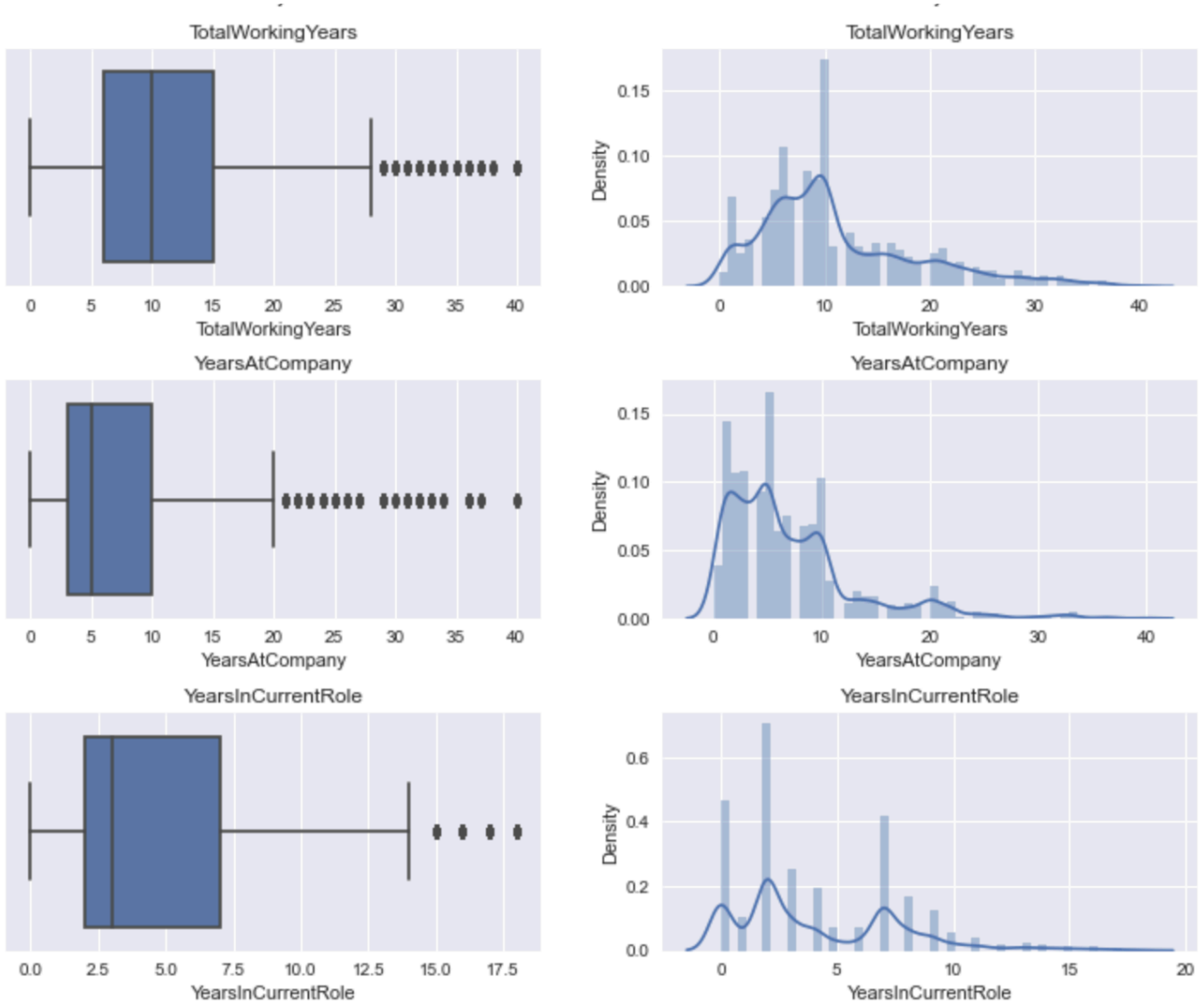


- **Hourly Rate :** The average of hourly rate is somewhere around $65, minimum is $30 and maximum is $100.

- **Monthly Income :** Minimum monthly income of employees is $1009 and maximum monthly income of employees is $19999 and average monthly income of employees is $6507.

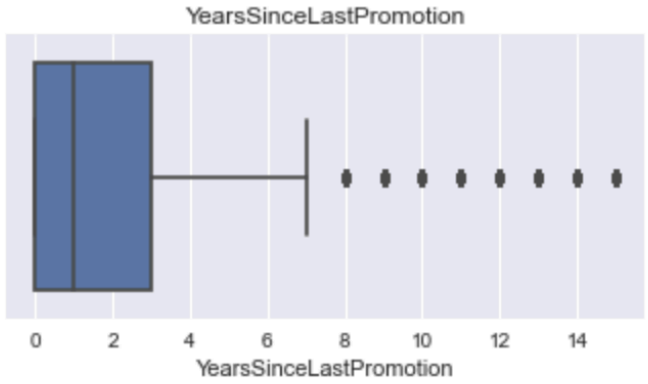
- **Monthly Rate :** Minimum monthly income of employees is $2094 and maximum monthly income of employees is $26999 and average monthly income of employees is 14302. Majority of employees are having monthly income greater than 5000.

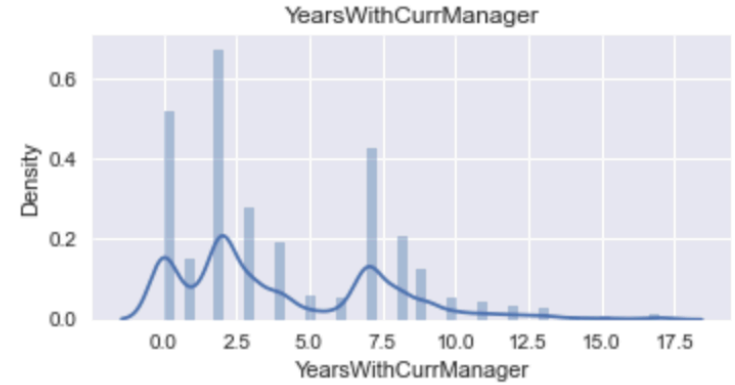
Quantitative Variables



- **Total Working Years :** Majority of employees have a working experience of 10 years.
- **Years At Company :** Majority of employees have been working in the company for 5 years.
- **Years In Current Role :** Majority of employees have been working for their current role in the company for 2 years.

Quantitative Variables





Years Since Last Promotion :

- Most of the employees have promoted like couple of years back.
- Outliers are just some of the employees working with no promotion for long time.

Years With Current Manager :

- Most of the employees are working consistently with the same manager.
- Outliers are few of the employees working with same manager for very long time.

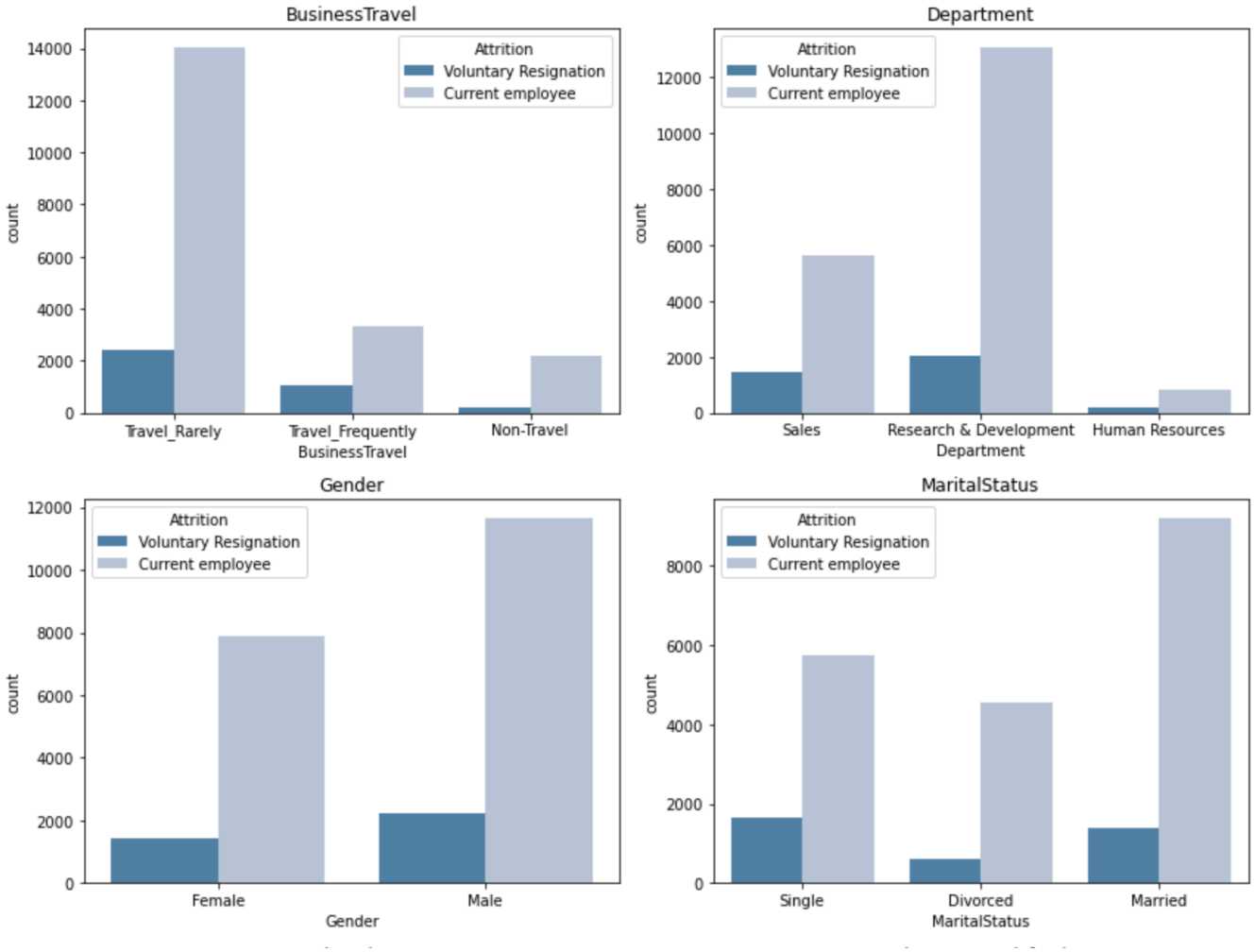
Categorical Vs Target

Business Travel : Employees who do business travel are more likely for attrition than the employees who do not do business travel.

Department : Around 60% employees are working in R&D Department. Sales department has a high attrition rate.

Gender : Approximately female and male ratio is 3:2. For better inference, male and female attrition rate is: Female Attrition Rate is 15.29% and Male Attrition Rate 16.12%.

Marital Status : Count of married employees is more. Attrition rate in singles are higher for both male and female.



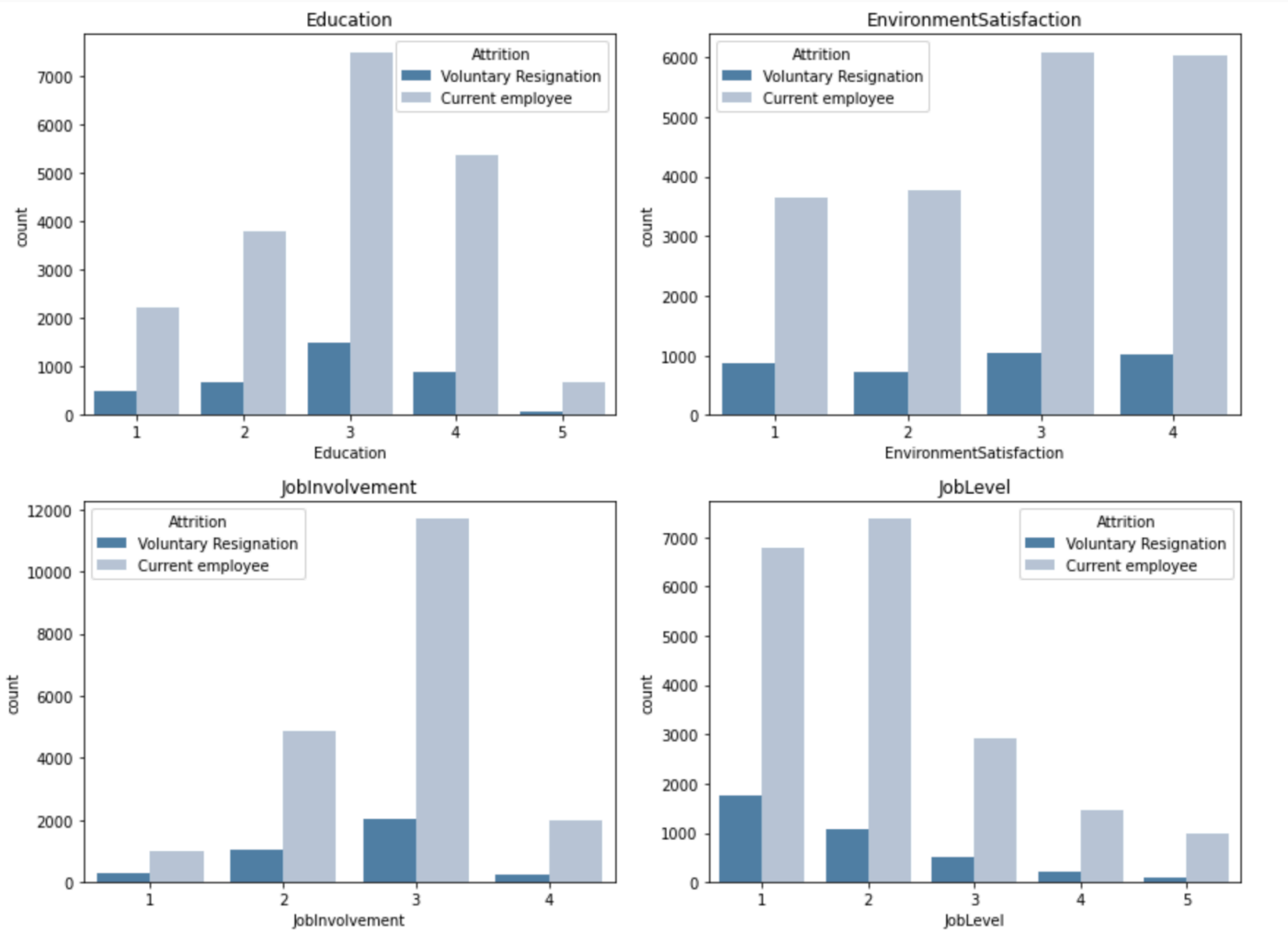
Categorical Vs Target

Education: Around 30% of employees have education level of 3. For both male and female, attrition rate is higher for education level 1,2 and 3.

Environment Satisfaction : Count of environment satisfaction is more towards 3 and 4. For both male and female, attrition rate is high environment satisfaction is 1 and 2.

Job Involvement : Majority of employees lie in the job involvement 2 & 3. Job involvement 3 has slightly more attrition rate than others.

Job Level : Majority of employees lie in the job level 1 and 2 that's why attrition rate is also higher in job level 1 and 2.



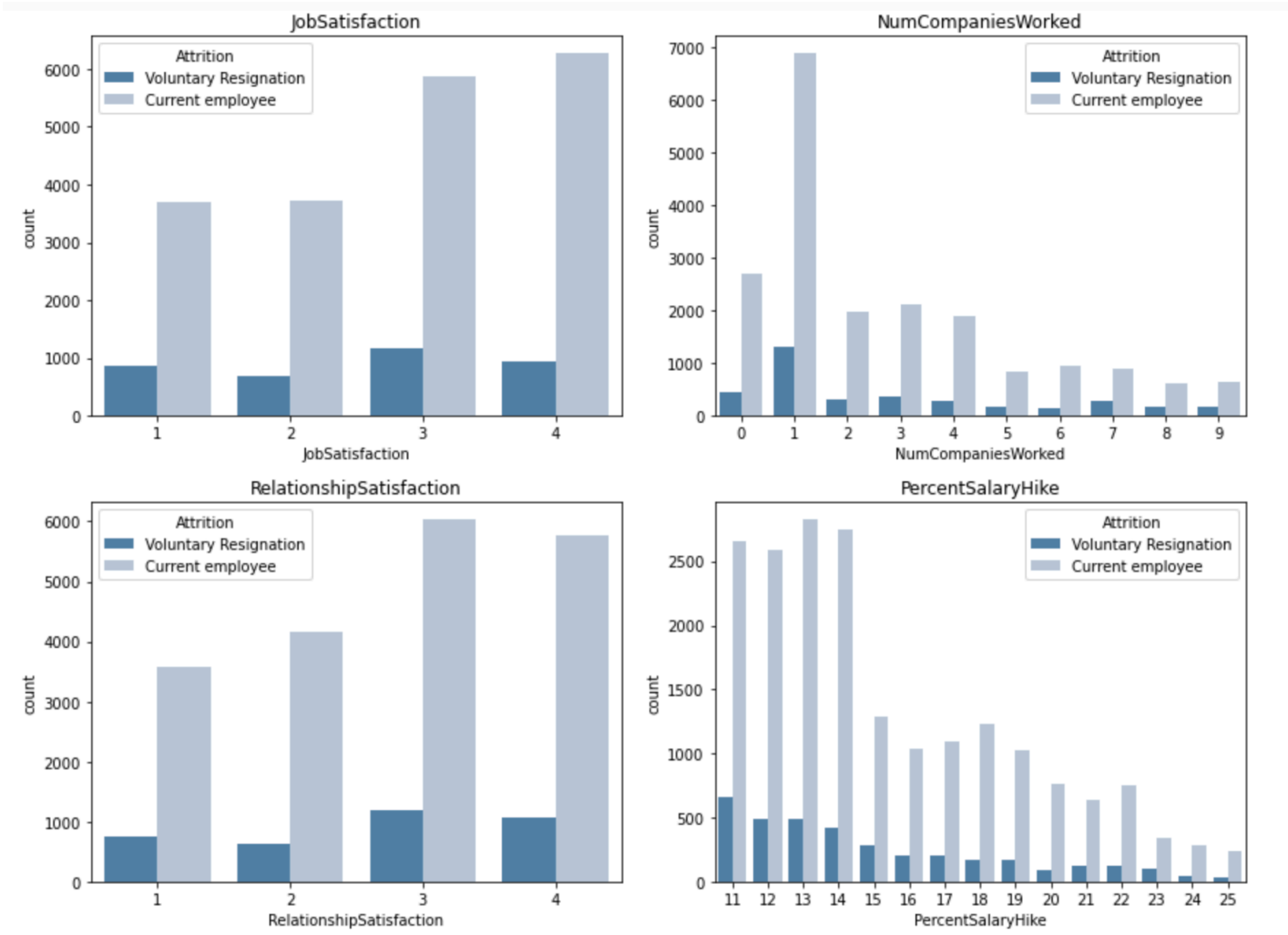
Categorical Vs Target

Job Satisfaction : Job Satisfaction count for 3 and 4 are more than 1 and 2. Higher attrition rate can be seen in Job Satisfaction level 1 and 2.

No. Of Companies worked : Maximum employees have worked in only 1 company. It can be observed that employees who have worked in 1 company have higher attrition rate.

Relationship Satisfaction : Count of employees having relationship satisfaction 3 & 4 are more than 1 & 2. Higher attrition is observed in lower relationship satisfaction for both genders.

Percent Salary Hike : Majority of employees got a salary hike less than 15%. Higher attrition is observed in cases where the salary hike is less than 16% for male when compared to female.



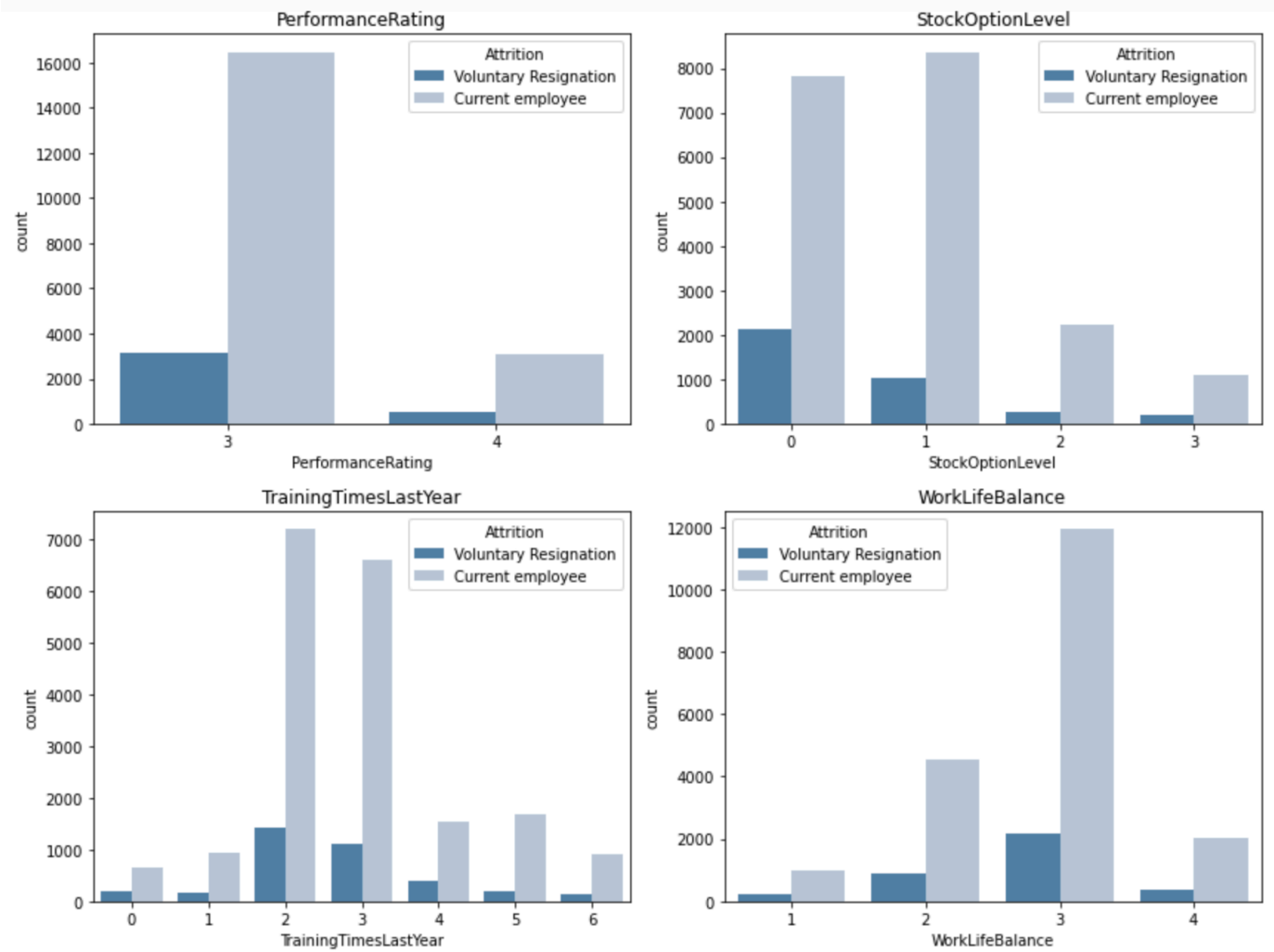
Categorical Vs Target

Performance Rating: There are very few employees who have performance rating 4. Performance Rating 3 has higher rate of attrition for both male and female.

Stock Option Level : There are many employees who does not have stock options level, As the stock options level increases the count of employees reduces. Higher attrition rate is observed in lower stock options level for both genders.

Training Time Last Year : Maximum employees where trained 2 to 3 times since last years. Higher attrition rate can be seen where number of trainings given to employees are less for both gender.

Work Life Balance : Count of employees having worklife balance as 3 is more wrt others. Lower work life balance has somewhat high rate of attrition. HR Department has less attrition rate in any cases of work life balance.

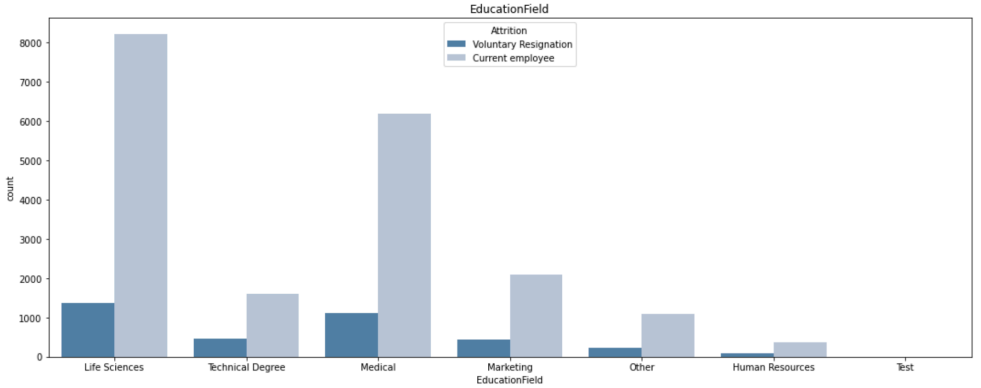


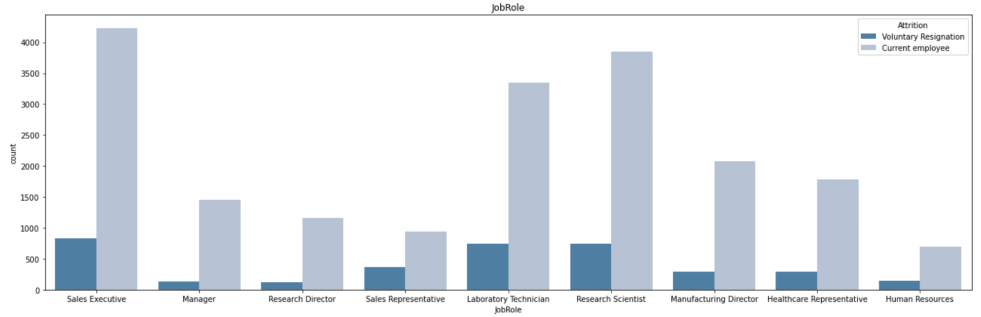
Categorical Vs Target

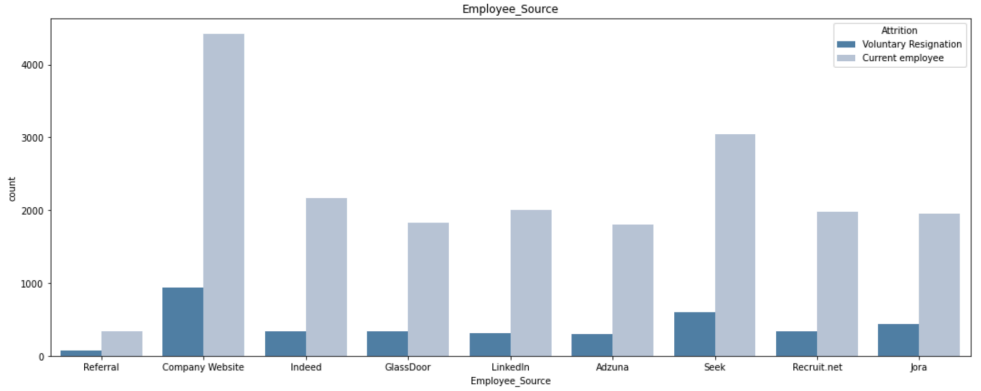
Education Field : Around 70% of employees are having 'Life Sciences' and 'Medical' education field. Attrition rate of female in 'HR' education field is less when compared to male. Attrition rate of female in 'Life Sciences' and 'Medical' is more when compared to male.

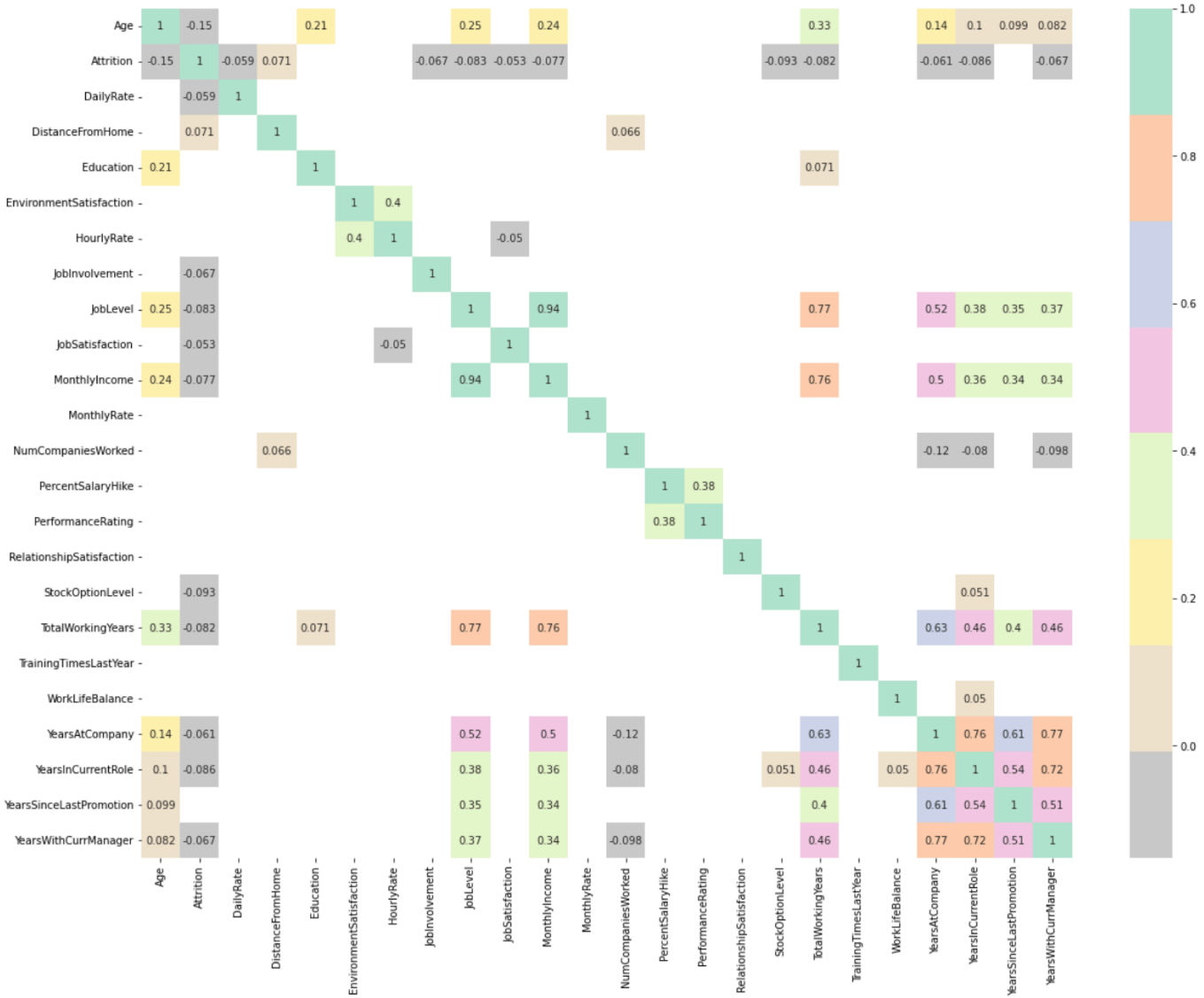
Job Role : Count of employees is more in job role as Sales Executive, Laboratory Technician, Research Scientist. Job role as Sales Representative has the highest attrition rate for both male and female, Job role as HR has high rate of attrition in case of female gender.

Employee Source : Around 25% employee source is Company Website, so we should management to enhance its worth more. At the same time, it is observed that the maximum attrition is taking place for those employees who have joined organization through companies website. Hence, reality check should be done in the website.









Multivariate Analysis

- Monthly Income and Job Role are highly correlated.
- Years At Company, Years in current role and Years with current manager are also positively correlated to each other.
- Total Working Years are correlated with Job Role and Monthly Income.

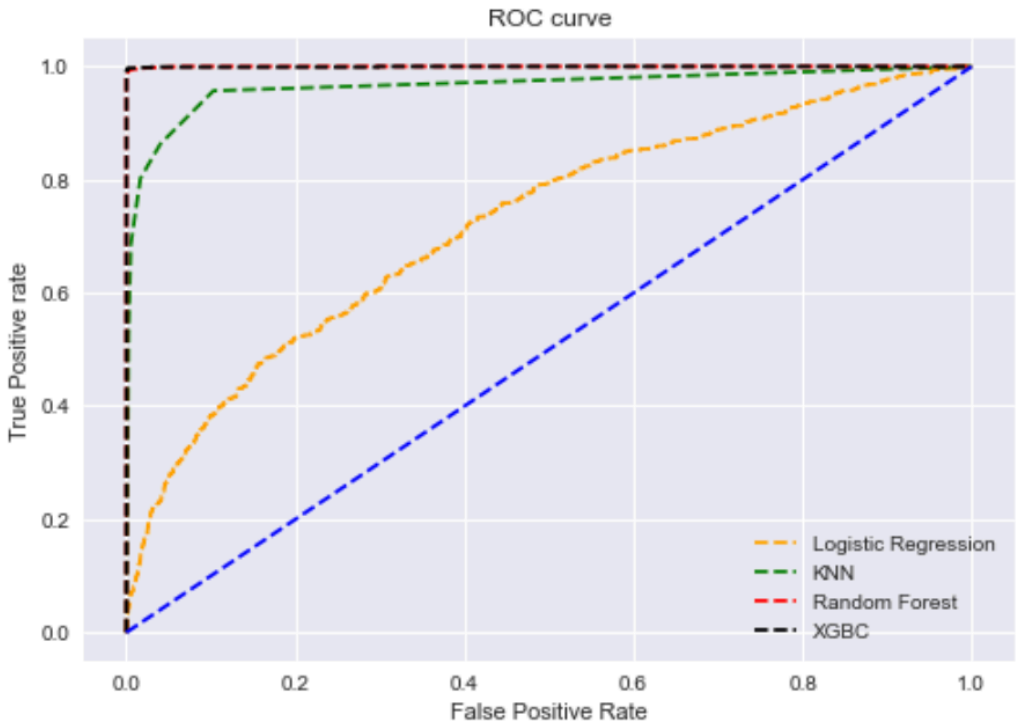
Statistical Results

	Features	P_value
0	Age	0.000000
1	DailyRate	0.000000
2	DistanceFromHome	0.000000
3	Education	0.000000
4	EmployeeCount	NaN
5	EmployeeNumber	0.708986
6	Application ID	0.714334
7	EnvironmentSatisfaction	0.000000
8	HourlyRate	0.059142
9	JobInvolvement	0.000000
10	JobLevel	0.000000
11	JobSatisfaction	0.000000
12	MonthlyIncome	0.000000
13	MonthlyRate	0.210869
14	NumCompaniesWorked	0.000000
15	PercentSalaryHike	0.000109
16	PerformanceRating	0.274755
17	RelationshipSatisfaction	0.427971
18	StandardHours	NaN
19	StockOptionLevel	0.000000
20	TotalWorkingYears	0.000000
21	TrainingTimesLastYear	0.000000
22	WorkLifeBalance	0.000228
23	YearsAtCompany	0.000000
24	YearsInCurrentRole	0.000000
25	YearsSinceLastPromotion	0.003412
26	YearsWithCurrManager	0.000000

- **Insignificant Features :** EmployeeCount, EmployeeNumber, Application ID, StandardHours, Over18.
- **Significant Features :** Age, Attrition, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, MonthlyRate, NumCompaniesWorked, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, Employee_Source.

Multiple Base Model Performance

	Classifier	AUC_Score	Accuracy	f1_score	Precision	Recall
0	LogisticRegression	0.722714	0.844595	0.131727	0.650794	0.073280
1	KNeighborsClassifier	0.965372	0.954284	0.849574	0.902513	0.802502
2	RandomForestClassifier	0.999874	0.997412	0.991906	0.998190	0.985702
3	XGBClassifier	0.999690	0.998706	0.995966	0.999101	0.992851



Final Model using XGBoost Classifier

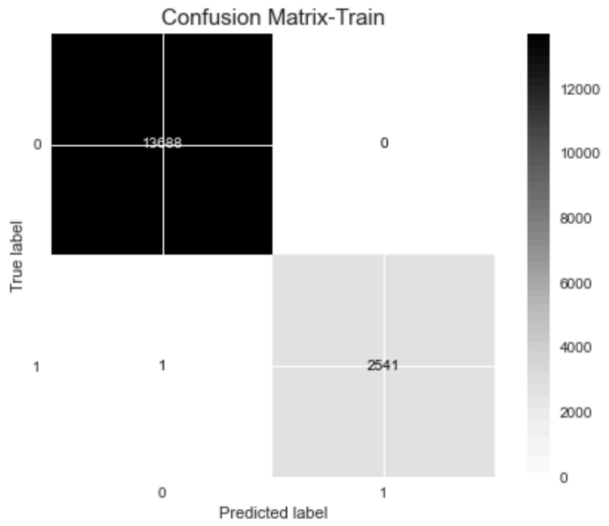
Inference :

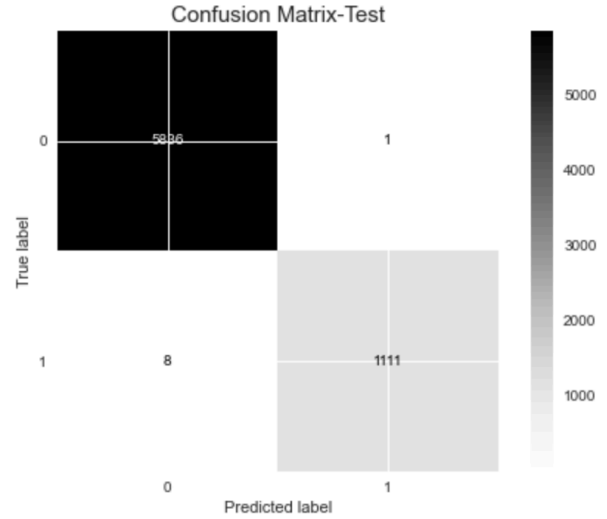
For our problem statement we need to improve on the recall score of the model.

Recall attempts to answer what proportion of actual positives was identified correctly?

Here, we need to identify those employees who are going to leave out correctly.

Recall=TruePositive/TruePositive+FalseNegative





Classification Report-Train :

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13688
1	1.00	1.00	1.00	2542
accuracy			1.00	16230
macro avg	1.00	1.00	1.00	16230
weighted avg	1.00	1.00	1.00	16230

Accuracy Score-Train : 0.9999383857054837

AUC Score-Train : 1.0

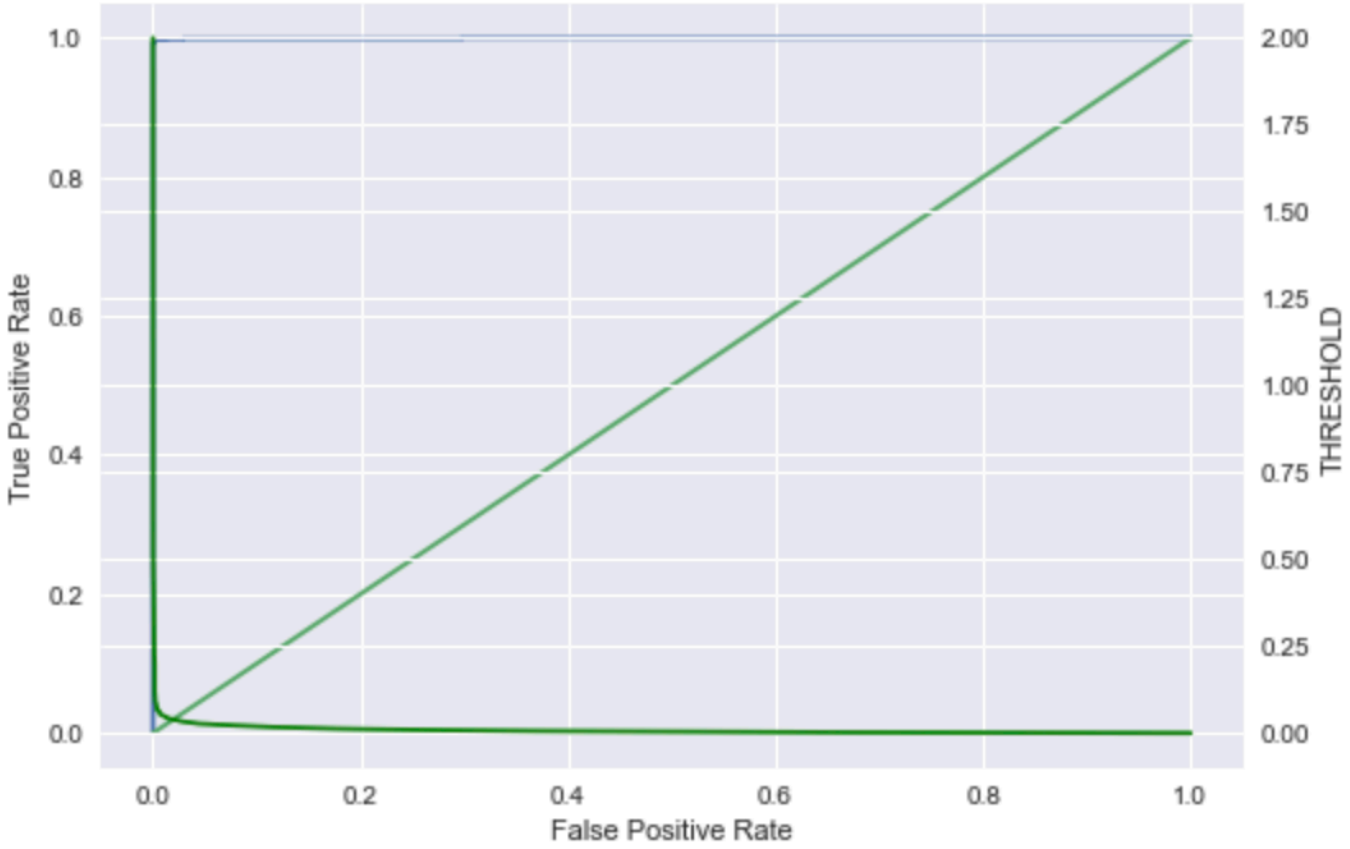
Classification Report-Test :

	precision	recall	f1-score	support
0	1.00	1.00	1.00	5837
1	1.00	0.99	1.00	1119
accuracy			1.00	6956
macro avg	1.00	1.00	1.00	6956
weighted avg	1.00	1.00	1.00	6956

Accuracy Score-Test : 0.9987061529614721

AUC Score-Test : 0.999690122011396

Plot : AUC-ROC Curve



ROC-AUC Curve

Inference :

- We can infer that the AUC Score is 0.9996.

Feature Importance Using Different Classifiers

	DecisionTreeClassifier	RandomForestClassifier	XGBClassifier
0	DailyRate	Age	EducationField_Technical Degree
1	Age	DailyRate	JobRole_Sales Executive
2	DistanceFromHome	HourlyRate	JobLevel
3	MonthlyIncome	MonthlyIncome	BusinessTravel_Travel_Frequently
4	MonthlyRate	DistanceFromHome	Department_Sales
5	HourlyRate	MonthlyRate	OverTime
6	PercentSalaryHike	TotalWorkingYears	EducationField_Other
7	Education	PercentSalaryHike	StockOptionLevel
8	WorkLifeBalance	YearsAtCompany	Employee_Source_Company Website
9	YearsInCurrentRole	Education	JobRole_Laboratory Technician

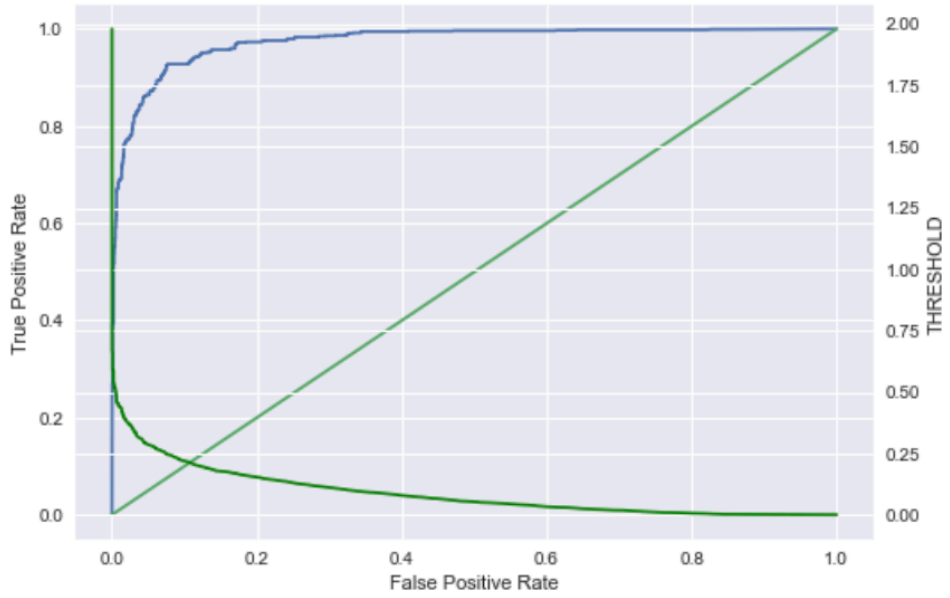
Inference :

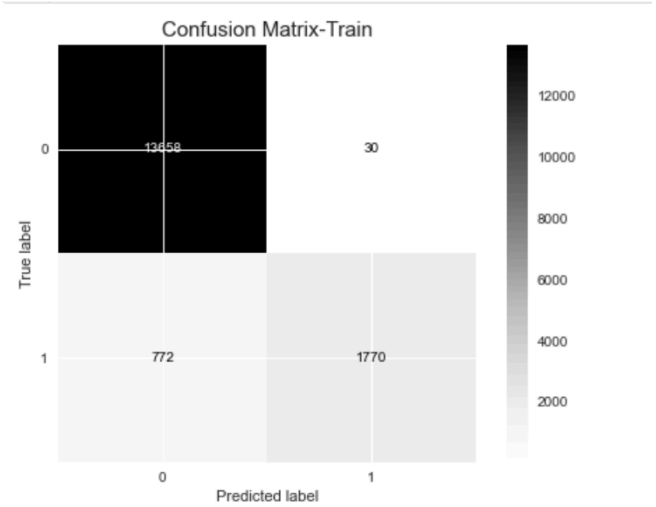
- We filter out the important from different models and pick out a new dataset with these optimal features.
- Using that features we created model and performed hyper parameter tuning.

Model with Optimal Features

Random Forest Classifier

Plot : AUC-ROC Curve





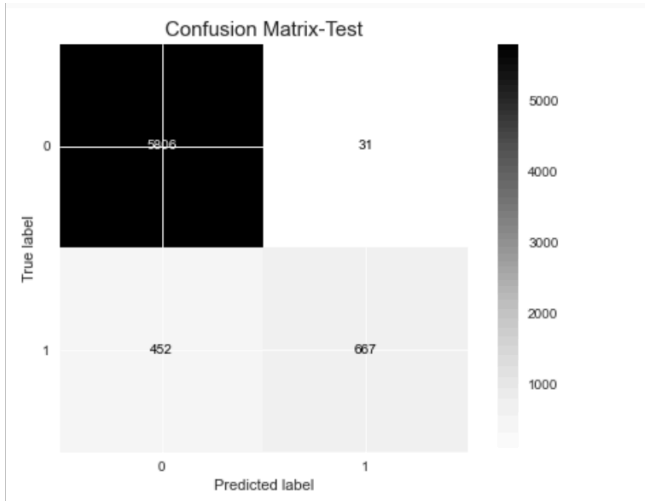
Classification Report-Train :

	precision	recall	f1-score	support
0	0.95	1.00	0.97	13688
1	0.98	0.70	0.82	2542
accuracy			0.95	16230
macro avg	0.96	0.85	0.89	16230
weighted avg	0.95	0.95	0.95	16230

Accuracy Score-Train : 0.9505853357979052

AUC Score-Train : 0.9899162221953473

f1_Score-Train : 0.8152924919391985



Classification Report-Test :

	precision	recall	f1-score	support
0	0.93	0.99	0.96	5837
1	0.96	0.60	0.73	1119
accuracy			0.93	6956
macro avg	0.94	0.80	0.85	6956
weighted avg	0.93	0.93	0.92	6956

Accuracy Score-Test : 0.93056354226567

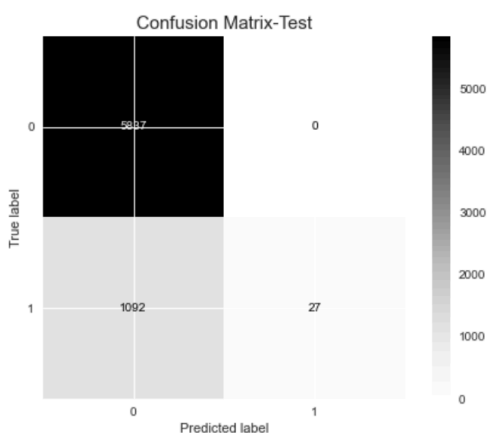
AUC Score-Test : 0.9744926322068257

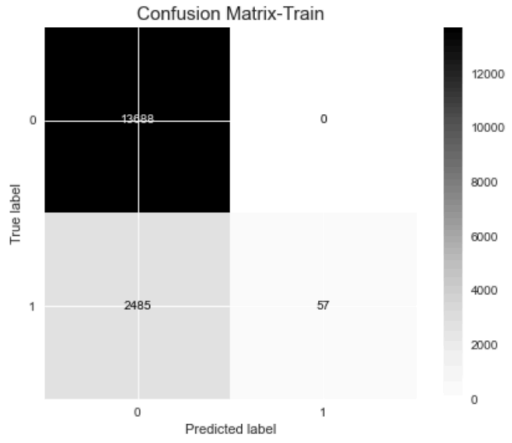
f1_Score-Test : 0.7341772151898733

Inference :

- Using the optimal features we get roc-auc score for train 0.9899 and for test 0.9744.

Hyper Parameter Tuning on Model with Optimal Features






```
Classification Report-Test :
              precision    recall  f1-score   support

           0       0.84      1.00      0.91      5837
           1       1.00      0.02      0.05      1119

    accuracy                           0.84      6956
   macro avg       0.92      0.51      0.48      6956
weighted avg       0.87      0.84      0.77      6956
```

```
Accuracy Score-Test : 0.8430132259919494

AUC Score-Test : 0.7486274043293814
```

```
Classification Report-Train :
              precision    recall  f1-score   support

           0       0.85      1.00      0.92     13688
           1       1.00      0.02      0.04      2542

    accuracy                           0.85     16230
   macro avg       0.92      0.51      0.48     16230
weighted avg       0.87      0.85      0.78     16230
```

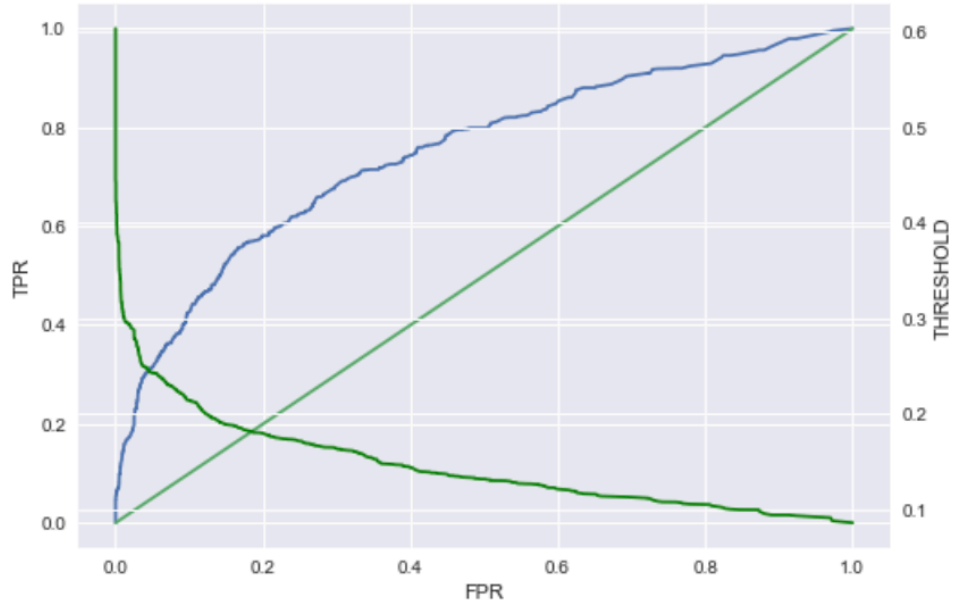
```
Accuracy Score-Train : 0.8468884781269255

AUC Score-Train : 0.7512291889017286
```

GridSearchCV

```
1  rfgrid=GridSearchCV(estimator=RandomForestClassifier(random_state=10),
2                      param_grid=[{'criterion': ['entropy', 'gini'],
3                                   'n_estimators': [5,10],
4                                   'max_depth':[5,10,15],
5                                   'max_features': ['sqrt', 'log2'],
6                                   'min_samples_leaf':[10,50,100],
7                                   'min_samples_split': [20,100,200],
8                                   'max_leaf_nodes': [5, 8]}],
9                      cv=5)
10
11 rfgrid_fit=rfgrid.fit(X_train,y_train)
12
13 print(rfgrid_fit.best_params_)
```

```
{'criterion': 'gini', 'max_depth': 5, 'max_features': 'sqrt', 'max_leaf_nodes': 8, 'min_samples_leaf': 10, 'min_sampl
es_split': 20, 'n_estimators': 10}
```

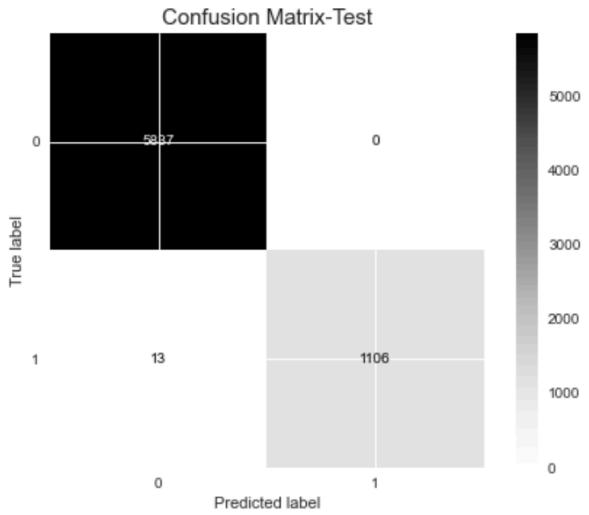
Plot : AUC-ROC Curve

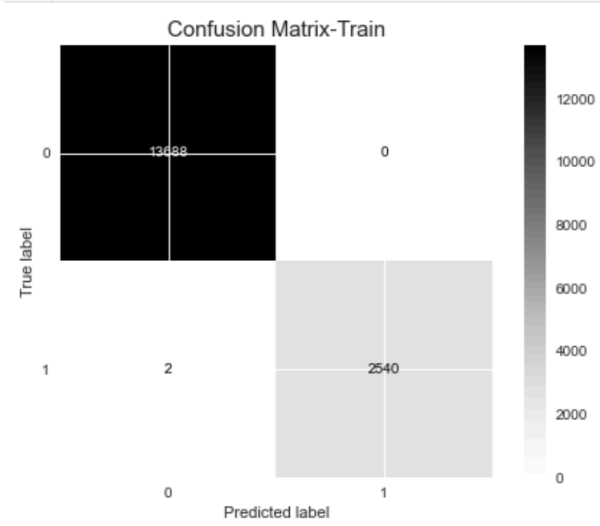


Inference :

- After hyper parameter tuning on the random forest we get roc-auc score for train 0.7486 and for test 0.7512.
- We will go without hyperperameter tuning as it reduces our roc-auc score.

Final Model with best Parameters





Classification Report-Test :

	precision	recall	f1-score	support
0	1.00	1.00	1.00	5837
1	1.00	0.99	0.99	1119
accuracy			1.00	6956
macro avg	1.00	0.99	1.00	6956
weighted avg	1.00	1.00	1.00	6956

Accuracy Score-Test : 0.9981311098332375

AUC Score-Test : 0.9998718538159774

f1_score Train: 0.9941573033707864

Precision Train Score : 1.0

Recall Train Score : 0.9883824843610366

Classification Report-Train :

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13688
1	1.00	1.00	1.00	2542
accuracy			1.00	16230
macro avg	1.00	1.00	1.00	16230
weighted avg	1.00	1.00	1.00	16230

Accuracy Score-Train : 0.9998767714109673

AUC Score-Train : 1.0

f1_score Train: 0.9996064541519087

Precision Train Score : 1.0

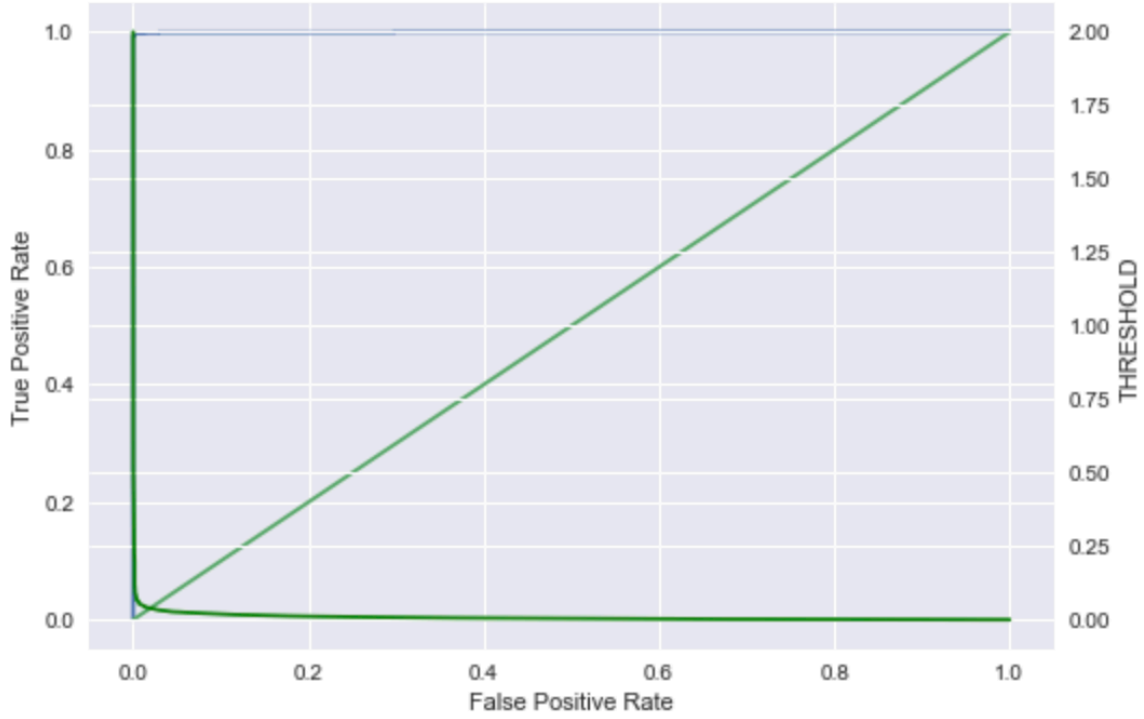
Recall Train Score : 0.999213217938631

Inference :

- Final Model using XGBoost we get roc-auc score for train 0.9998 and for test 1.0.
- From the model we infer that we have good results with XGBoost.

Final Model with best Parameters

Plot : AUC-ROC Curve



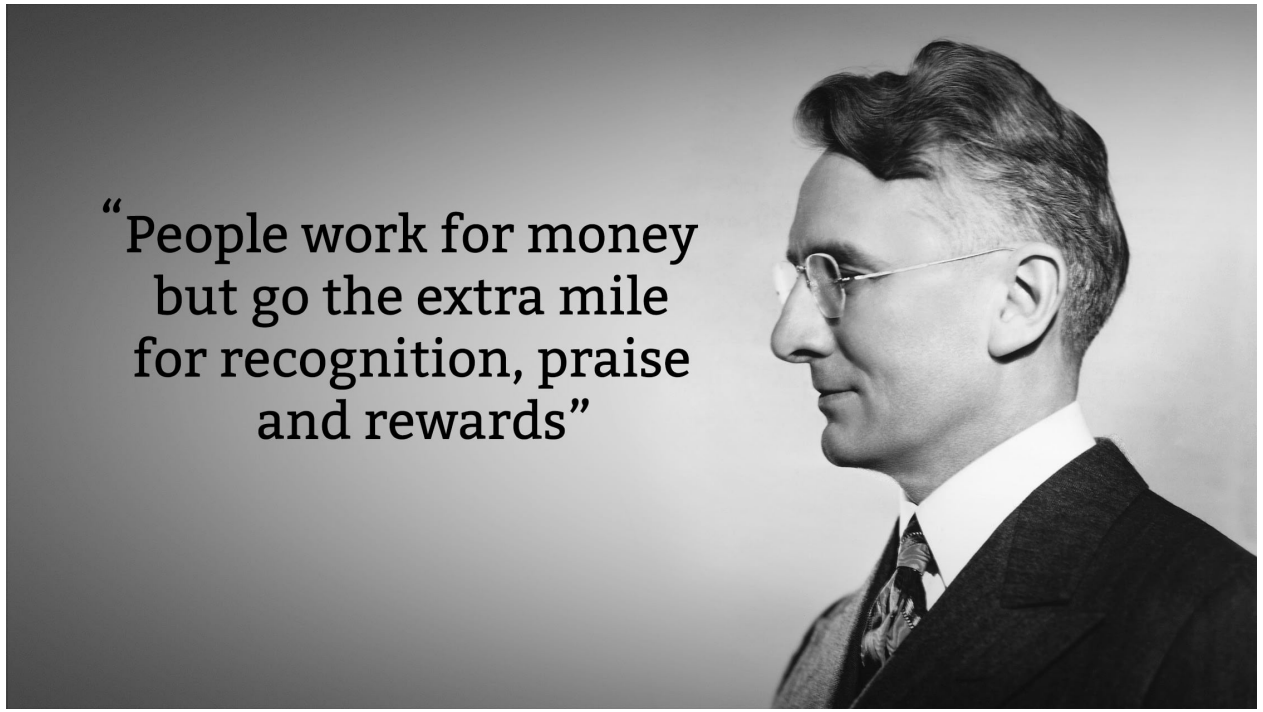
Inference :

- Final Model using XGBoost we get roc-auc score for train 0.9998 and for test 1.0.

Conclusion:

The results of the XGBoost applied to the dataset used in the case study gave the managers of the company very useful information that they could use to reduce their employee turnover rate:

- Reconsider the salary of the employees who earn less than $3.0k and assess the possibility of rising their salary.
- Study the conditions of the employees working in the sales department and determine the reasons for their dissatisfaction with such department.
- Offer incentives and growth possibilities inside the company to those employees younger than 34 who have been working for the company for less than 2 years.



Limitations:

- The data which we have is highly imbalanced this might lead to inaccurate predictions.
- To enhance the data quality and to reduce errors we have transformed the data using power transformer, getting Business insights out of this would be difficult.
- To proceed with Feature Engineering, we need to have domain knowledge

Can Do More :

As future lines of work on People Analytics, some ideas are proposed:

- Extend the study to larger datasets containing real data with a higher number of variables and records.
- Re-do the analysis using programming languages instead of software tools, such as Python or R.
- Deploy an application that automatically returns the attrition prediction for an employee, based on information given to the application.

Great Learning

POWER AHEAD