Project Interim Report

Batch details	DSE – AUG 2022 (BENGALURU)
Team members	Hari Haran Radhika P Shivangi Bharadwaj Vishwesh Yash Gehlot
Domain of Project	**HR (Attrition)**
Proposed project title	Application of Data Science to Reduce Employee Attrition
Group Number	Team 08
Team Leader	Yash Gehlot
Mentor Name	Siddharth Koshta

Date: 27/12/2022

Siddhart Koshta

Signature of the Mentor

Yash Gehlot

Signature of the Team Leader

Table of Contents

Project Details

1.BUSINESS UNDERSTANDING:

The problem understanding for an attrition machine learning project is to identify and analyze the factors that contribute to employee attrition and develop a model that can predict employee attrition rates. The goal of the project is to use machine learning techniques to identify patterns and correlations between employee characteristics, job roles, and other factors that may contribute to employee attrition. The model should be able to accurately predict the likelihood of an employee leaving the organization.

1.1 BUSINESS PROBLEM STATEMENT

The objective of this project is to identify the factors that are contributing to employee attrition and develop strategies to reduce it. The goal is to reduce employee attrition by understanding the underlying causes and implementing strategies to address them. This will help the organization retain its valuable employees and improve overall productivity.

Collect and analyze data related to employee attrition, such as employee demographics, job satisfaction, job performance, and other relevant factors. Identify patterns in the data that may be contributing to employee attrition. Develop strategies to address the identified patterns and reduce employee attrition. Monitor the effectiveness of the strategies implemented and adjust as needed.

1.2 TOPIC SURVEY:

1. Problem understanding:

The goal of this Attrition ML project is to use machine learning techniques to predict employee attrition. The data set contains information about employee characteristics such as age, gender, job role, salary, and other factors that may influence attrition. The project will use supervised learning algorithms to build a model that can accurately predict whether an employee is likely to leave the company or not. The model will be evaluated using metrics such as accuracy, precision, recall, and F1 score. The results of the model will be used to identify which factors are most influential in predicting employee attrition and to provide insights into how the company can better retain its employees

2. Current Solution to the problem:

A. Improve employee engagement: Create a culture of open communication and collaboration, provide meaningful recognition and rewards, and offer opportunities for professional development.
B. Enhance job satisfaction: Offer competitive salaries and benefits, provide flexible work arrangements, and ensure job security.
C. Increase job security: Provide job stability through long-term contracts and career paths.
D. Improve work-life balance: Offer flexible work hours, telecommuting options, and other perks that can help employees better manage their personal and professional lives.

3. Proposed Solution to the Problem:

A. Increase Employee Engagement: Implementing activities and initiatives that promote employee engagement, such as team-building activities, recognition programs, and open communication channels, can help to reduce attrition.
B. Improve Workplace Culture: Creating a positive work environment that encourages collaboration and respect can help to reduce attrition. This can be done by implementing policies that promote diversity and inclusion, providing flexible work schedules, and offering competitive salaries and benefits.

C. Offer Professional Development Opportunities: Offering employees the opportunity to develop their skills and knowledge can help to reduce attrition. This can be done by providing training courses, mentorship programs, and other professional development opportunities.

D. Improve Job Satisfaction: Implementing measures to improve job satisfaction can help to reduce attrition. This can be done by offering competitive salaries, providing job security, and offering opportunities for career advancement.

1.3. CRITICAL ASSESSMENT OF TOPIC SURVEY:

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. Nowadays, with the existence of Data Science and prediction techniques, this task can be automatically done, which allows the managers of the companies to obtain the information they require from the employees in a much faster and efficient way than it was obtained in the past when the task was done manually by the Human Resources department. This results in a significant decrease of the costs associated with employee attrition, maximising the revenue of the company.

What key gaps are you trying to solve?

The aim of an Attrition ML project is to develop a model that can accurately predict employee attrition and identify the factors that contribute to it. The model should be able to identify which employees are likely to leave the company, and what factors are driving their decision. This will help the company to take proactive measures to reduce employee attrition and retain valuable employees.

2. DATA UNDERSTANDING:

2.1 Data Dictionary

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
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

2.2 VARIABLE CATEGORIZATION:

- Independent variables:

  Numerical column - 27Categorical column - 9

- Target variable:

  Categorical - 1

- Total columns - 37

2.3 DISTRIBUTION OF VARIABLES:

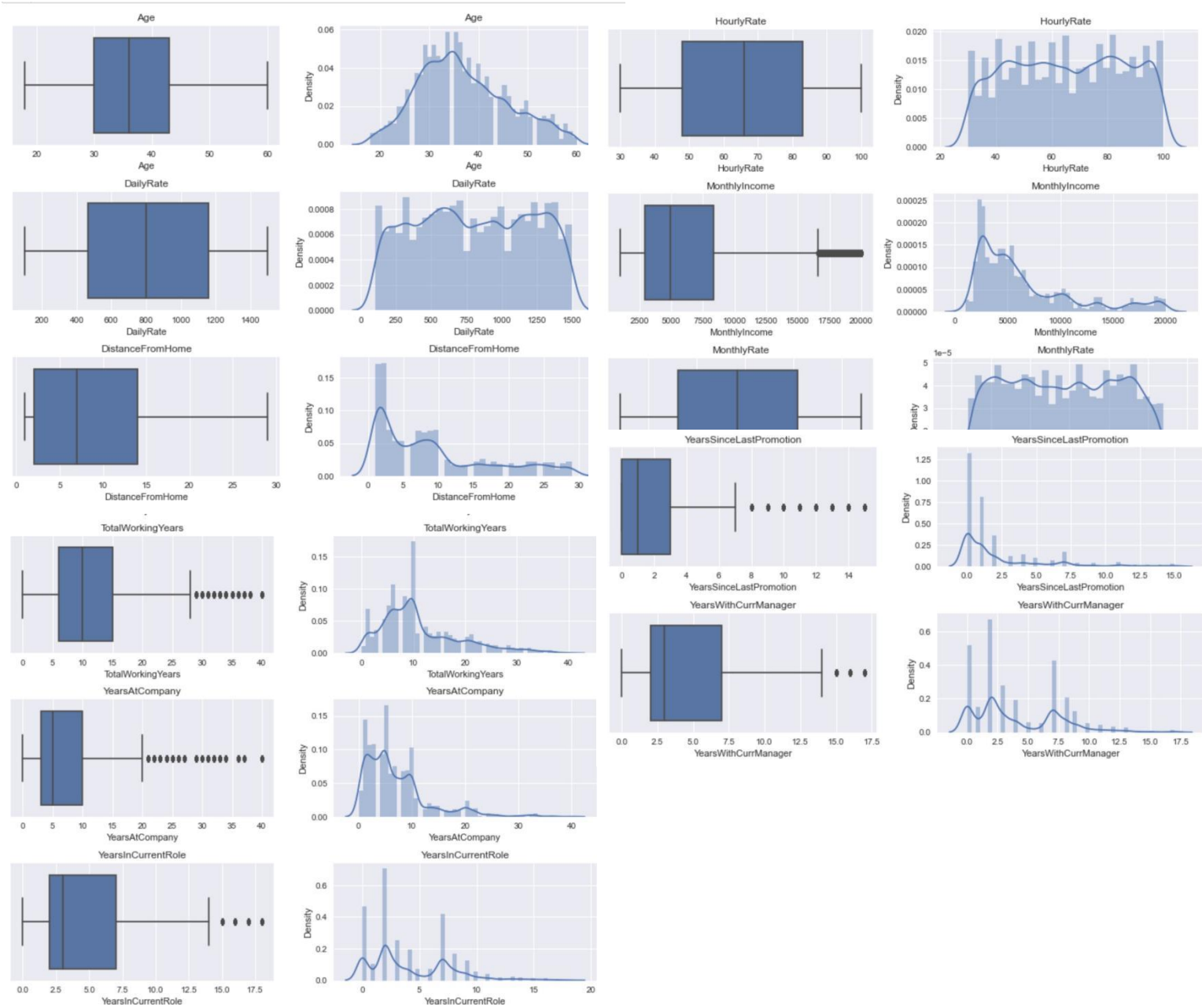
The data given to us is in shape a 23436,37 data frame. After droping the duplicate the shape of the data is 23422,37 The data consists of Numerical and Categorical data. While further analysing the data, we realize that the target data is , and our categorical data being Business Travel , Department, Education Field, Gender, job Role, Marital Status, Over 18, Over Time, and Employee Source.

The numerical features have different scales, which may be a problem for some machine learning algorithms. The features should be rescaled to have similar scale.

Skewness

Column	Skewness
Age	0.450000
DailyRate	-0.020000
DistanceFromHome	0.960000
HourlyRate	-0.030000
MonthlyIncome	1.540000
MonthlyRate	0.030000
TotalWorkingYears	1.030000
YearsAtCompany	1.240000
YearsInCurrentRole	0.730000
YearsSinceLastPromotion	1.750000
YearsWithCurrManager	0.690000

Box Plot and Distribution of Numerical Variables :



Inferences :

- HR Team should focus more on young employees whose age is less 35 particularly.Carefulattention should be given to employees with age 18,19,20 as those ages attrition rate is more than the current employee rate. Hence, company is facing loss as the company is investing so much for the candidates training but the candidates are still leaving the job.

- Employees having Job level 1 and 2 are not satisfied and are having a high attrition rate. So,HR should focus on those set of employees who are having lower job levels.
- HR should focus more on employees who are singles as their attrition rate is higher.
- HR Team should take attentive counselling of employees whose job involvement is lessi.e.,1. Another important observation can be seen that irrestpective of Job Involvement,employees whose monthly income is lees have maximum attrition.
- Employees who do business travel are more likely for attrition than the employees who donot do business travel.

Columns name in the order as follows both for x and y axis

[Age, Attrition, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Environment Satisfaction,Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, NumCompanies Worked, Over Time, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years At Company, Years In Current Role, Years Since Last Promotion, Years With Current Manager, Employee Source']

3. DATA PREPROCESSING:

3.1 NULL VALUE TREATMENT:

Null value treatment is essential to building most of the commonly used machine learning classification models such as logistic regression, decision tree, KNN, and others. To infer that we have used isnull() function the null values from the dataset.

	Total	Percentage of Missing Values
YearsInCurrentRole	15	0.064042
PercentSalaryHike	14	0.059773
Attrition	13	0.055503
YearsAtCompany	13	0.055503
MonthlyIncome	13	0.055503
Employee Source	12	0.051234
DailyRate	12	0.051234
Education	12	0.051234
OverTime	12	0.051234
YearsSinceLastPromotion	11	0.046964
TrainingTimesLastYear	11	0.046964
MonthlyRate	11	0.046964
MaritalStatus	11	0.046964
Department	11	0.046964
PerformanceRating	10	0.042695
WorkLifeBalance	10	0.042695
Gender	10	0.042695
StandardHours	10	0.042695
Over18	10	0.042695
JobInvolvement	9	0.038425

DistanceFromHome	9	0.038425
EducationField	9	0.038425
StockOptionLevel	9	0.038425
HourlyRate	9	0.038425
NumCompaniesWorked	9	0.038425
EnvironmentSatisfaction	9	0.038425
JobSatisfaction	9	0.038425
JobRole	9	0.038425
RelationshipSatisfaction	8	0.034156
TotalWorkingYears	8	0.034156
BusinessTravel	8	0.034156
JobLevel	7	0.029886
YearsWithCurrManager	7	0.029886
EmployeeCount	5	0.021347
Application ID	3	0.012808
Age	3	0.012808
EmployeeNumber	1	0.004269

From the above table, we found that the null values for our columns is less than 1 percent and we also found that the values are wrongly imputed and because of that also we have null values in the dataset. So we can drop the null values too on the basis of above codes.

3.2 PRESENCE OF OUTLIERS AND TREATMENT:

There are about 27 numerical variables on which the presence of outliers is to be determined. A distribution was assumed to be skewed when the skewness is outside the range of ±0.5 as it is quite impossible to have a real dataset with skewness of each variable or at least one of the variables with a perfect zero skewness. It is also understood that as the value of skewness increases, the farther the outlier is. Out of the 27 numerical variables, it was found that some of them are negatively skewed. The skewness values for those 16 numerical variables are shown below.

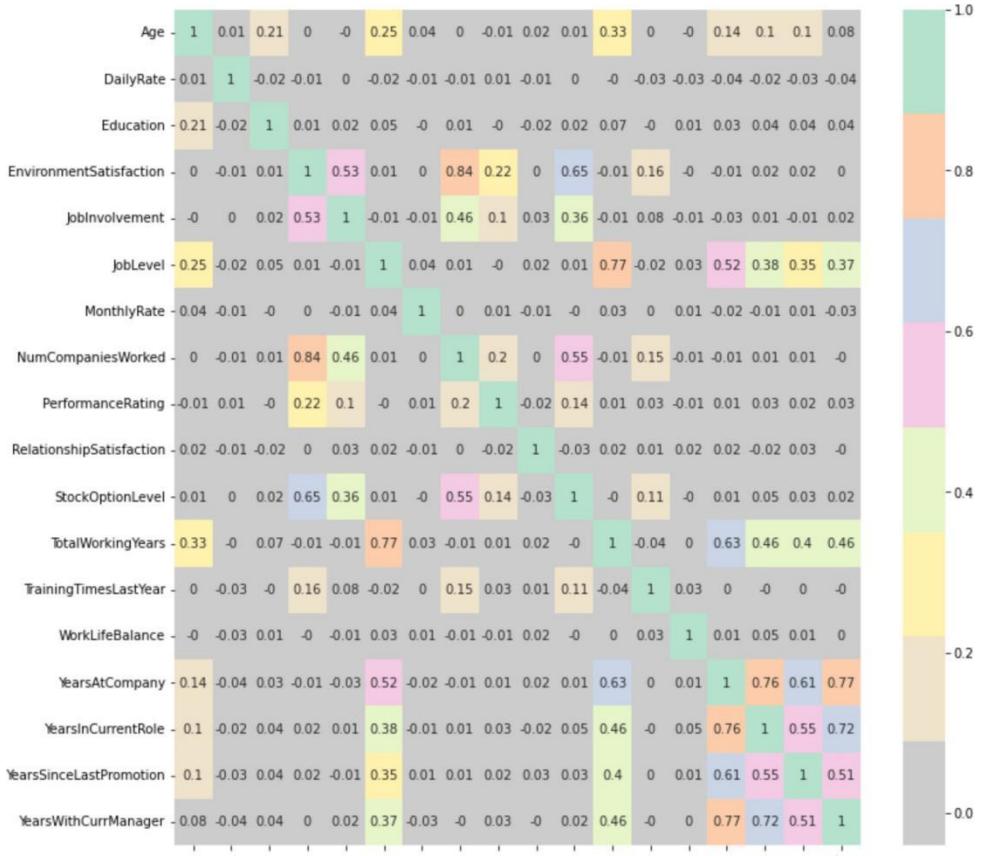
```
data.skew()
```

```
Age                          0.410358
DailyRate                   -0.004403
Education                   -0.284463
EnvironmentSatisfaction    108.202971
JobInvolvement              15.843901
JobLevel                     1.022543
MonthlyRate                  0.019818
NumCompaniesWorked         144.555534
PerformanceRating            2.960828
RelationshipSatisfaction    -0.303302
StandardHours             -108.194271
StockOptionLevel            30.408968
TotalWorkingYears            1.117029
TrainingTimesLastYear        1.045084
WorkLifeBalance             -0.551744
YearsAtCompany               1.758409
YearsInCurrentRole           0.918293
YearsSinceLastPromotion      1.986129
YearsWithCurrManager         0.831309
dtype: float64
```

Once there are outliers outside the acceptable range, it has to be treated. Dropping the rows with outliers or capping outliers are not recommended as the dataset consists of financial records, data with unique detail that can mislead a model when it comes to prediction. Therefore, transformations are the only means to treat outliers. Out of all the transformation techniques, most of them do not deal with negative values and zero values. This dataset contains negative values however, it contains values that are zero and very close to zero which will cause problems when applying the transformation. It was found that the Box-Cox transformation technique handles both negative and zero values appropriately thereby reducing skewness considerably as shown below

3.3 CHECKING FOR MULTI-COLLINEARITY AND TREATMENT:

Correlation Matrix

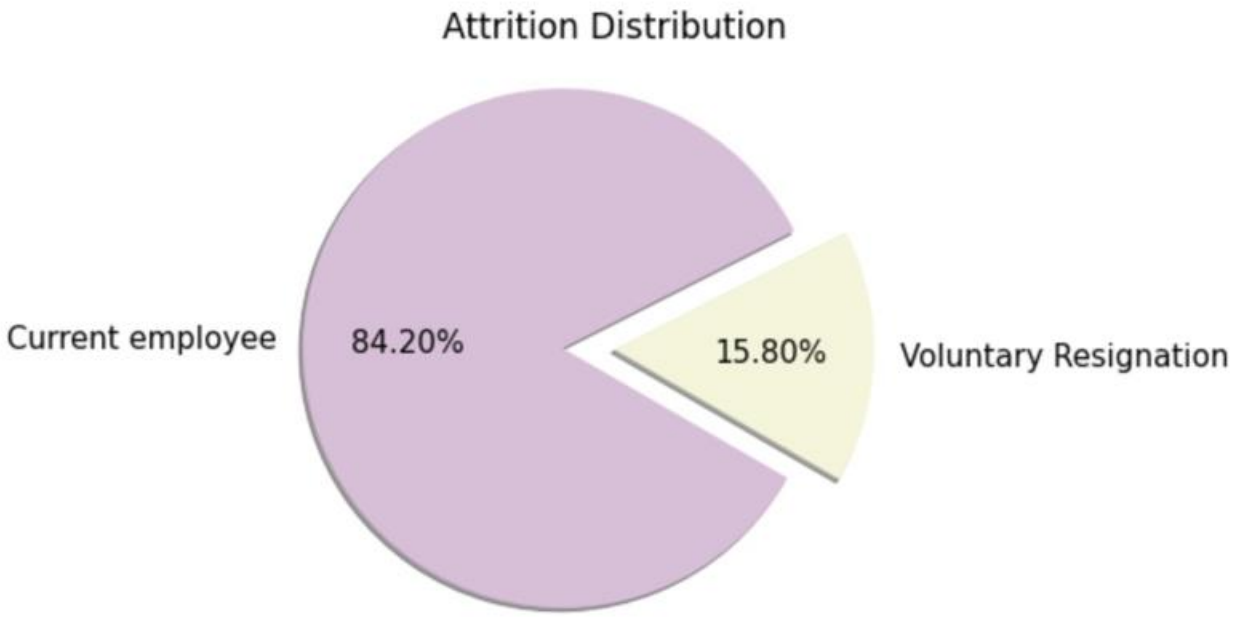


When a pair of independent variables exhibit high correlation (that is when a pair of independent variables can explain one another with strong linear relation either positively or negatively) with each other it is termed a collinear effect. When more than one pair of independent variables exhibita high correlation with each other, it is termed a multi-collinear effect.

A threshold is to be set to the correlation value to categorize it between the collinear effect and non-collinear effect. The dataset taken into consideration for this project has 27 of independent variables and we found that there is no high Multicollinearity.

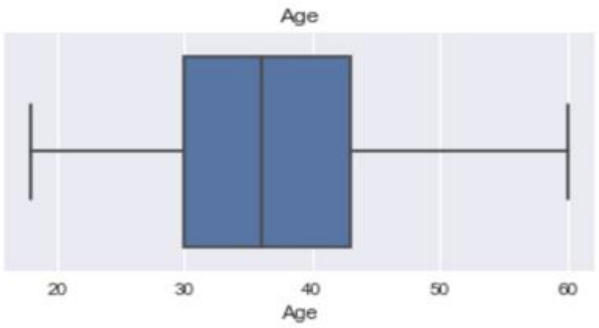
4. EXPLORATORY DATA ANALYSIS:

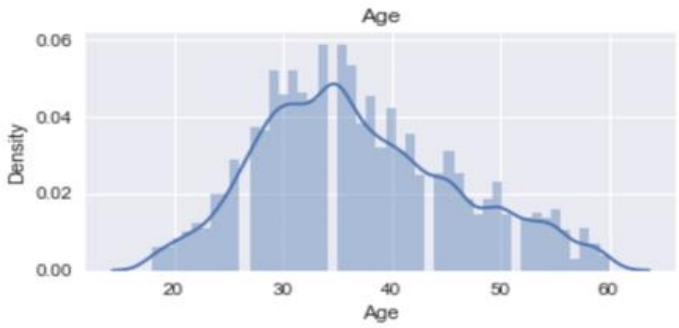
4.1: Target Variable

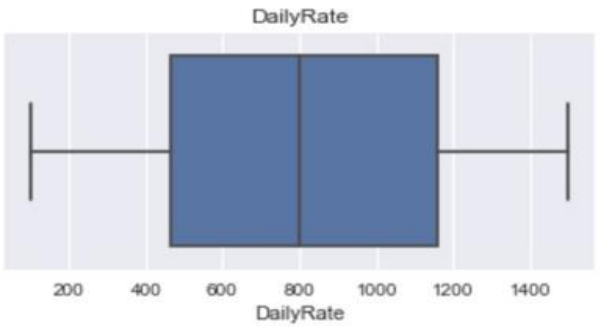


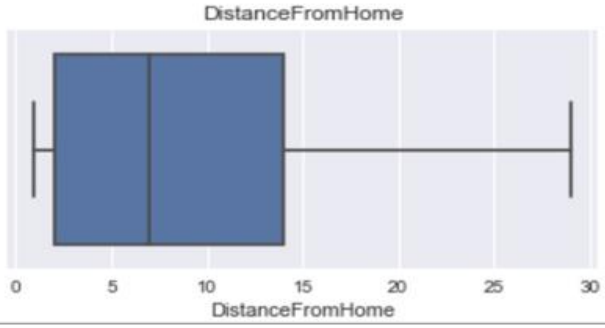
In our data over 15.8% people have resigned, that is around 3700 of them

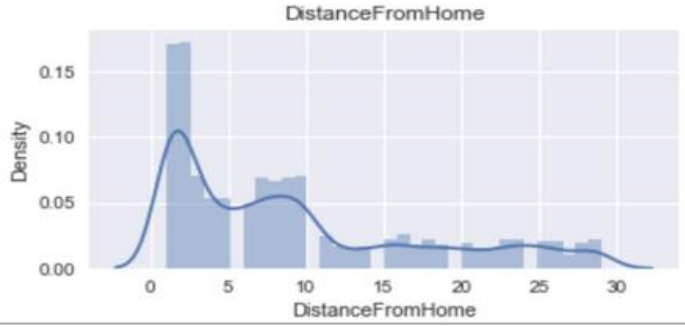
4.2: Quantitative Variable:



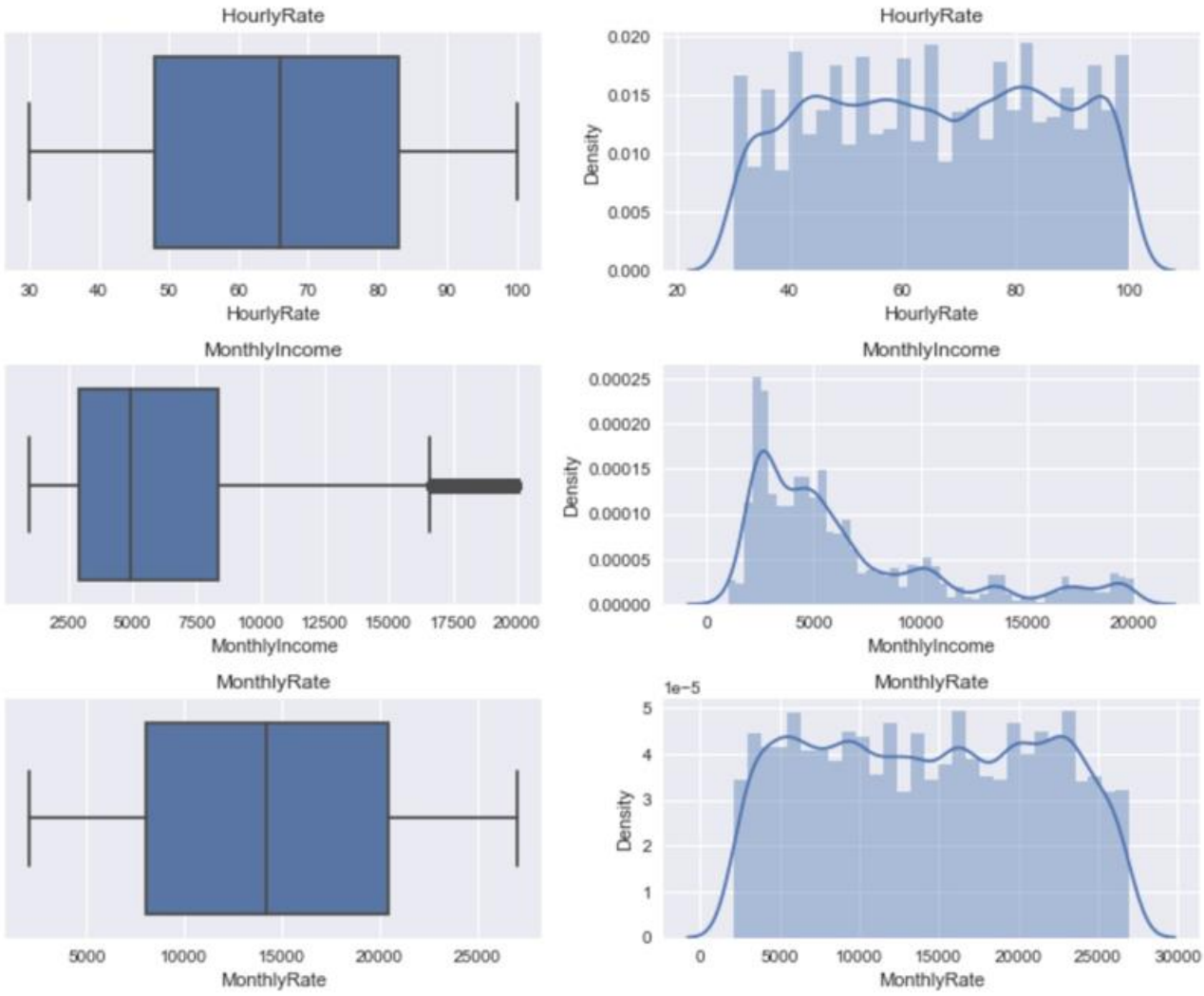








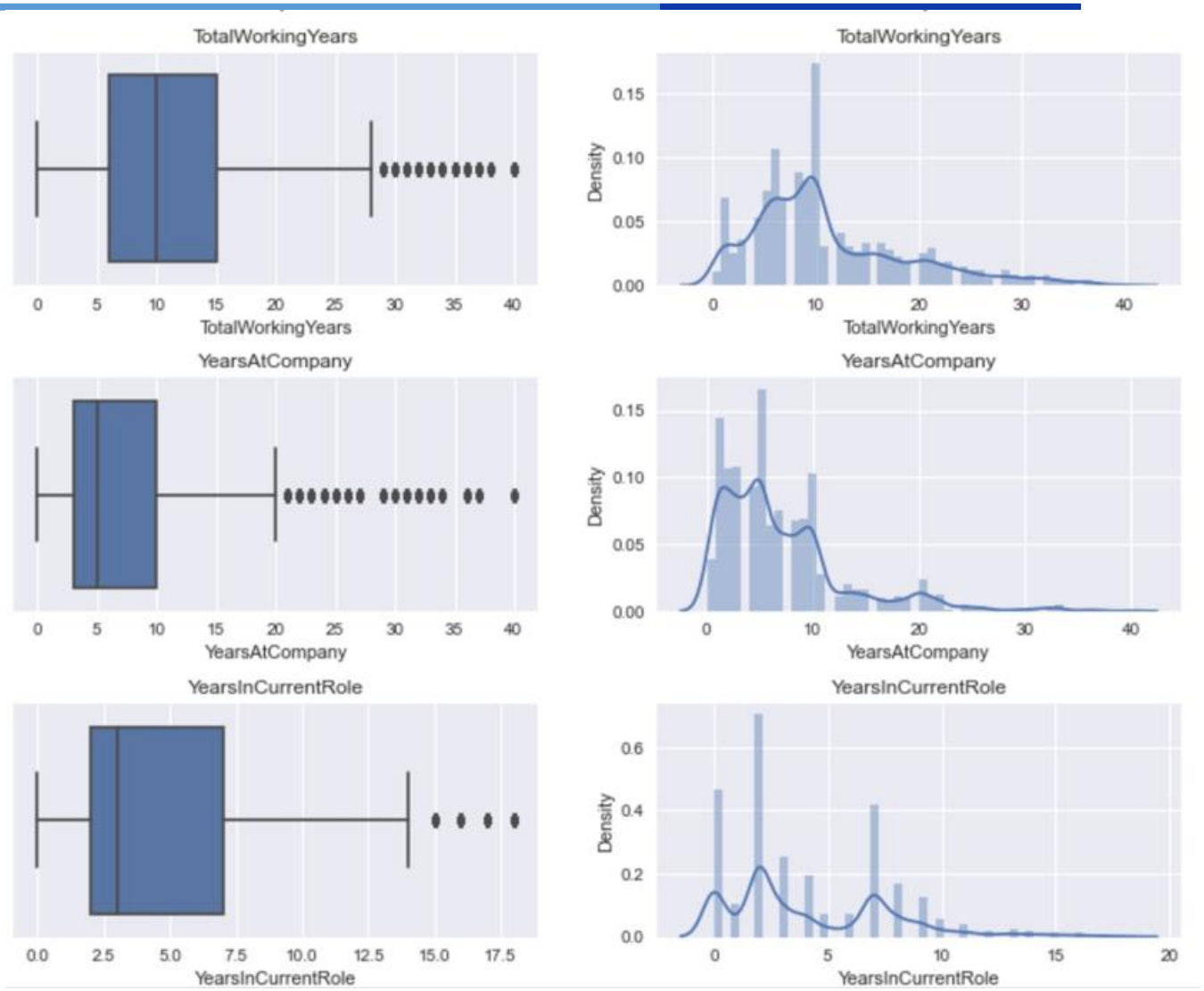
- Age : Majority of employees lie between the age range of 30 - 40 years.
- Daily Rate : The average of daily rate is somewhere around $802, minimum is $102 and maximum is $1499.
- Distance From Home : We can see that the average distance from home is around 9 Km, minimum is 1 Km and maximum is 29 Km.



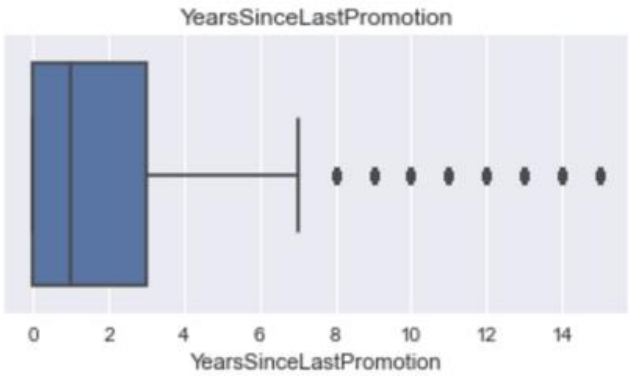
- Hourly Rate : The average of hourly rate is somewhere around $65, minimum is $30 and maximum is $100.

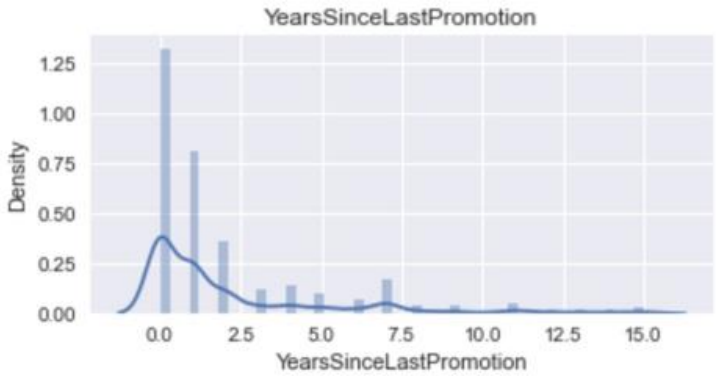
- Monthly Income : Minimum monthly income of employees is $1009 and maximum monthly income of employees is $19999 and average monthly income of employees is $6507.

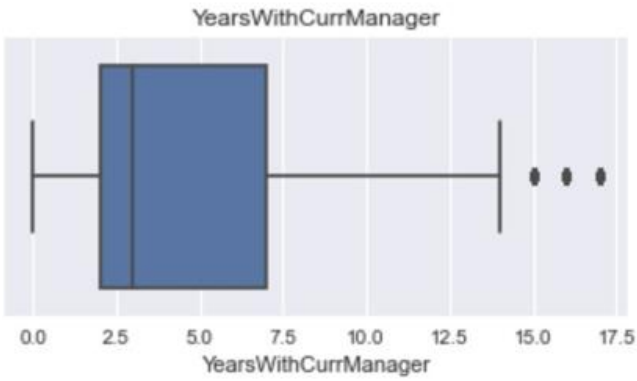
- Monthly Rate : Minimum monthly income of employees is $2094 and maximum monthly income of employees is $26999 and average monthly income of employees is 14302. Majority of employees are having monthly income greater than 5000.



- 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.







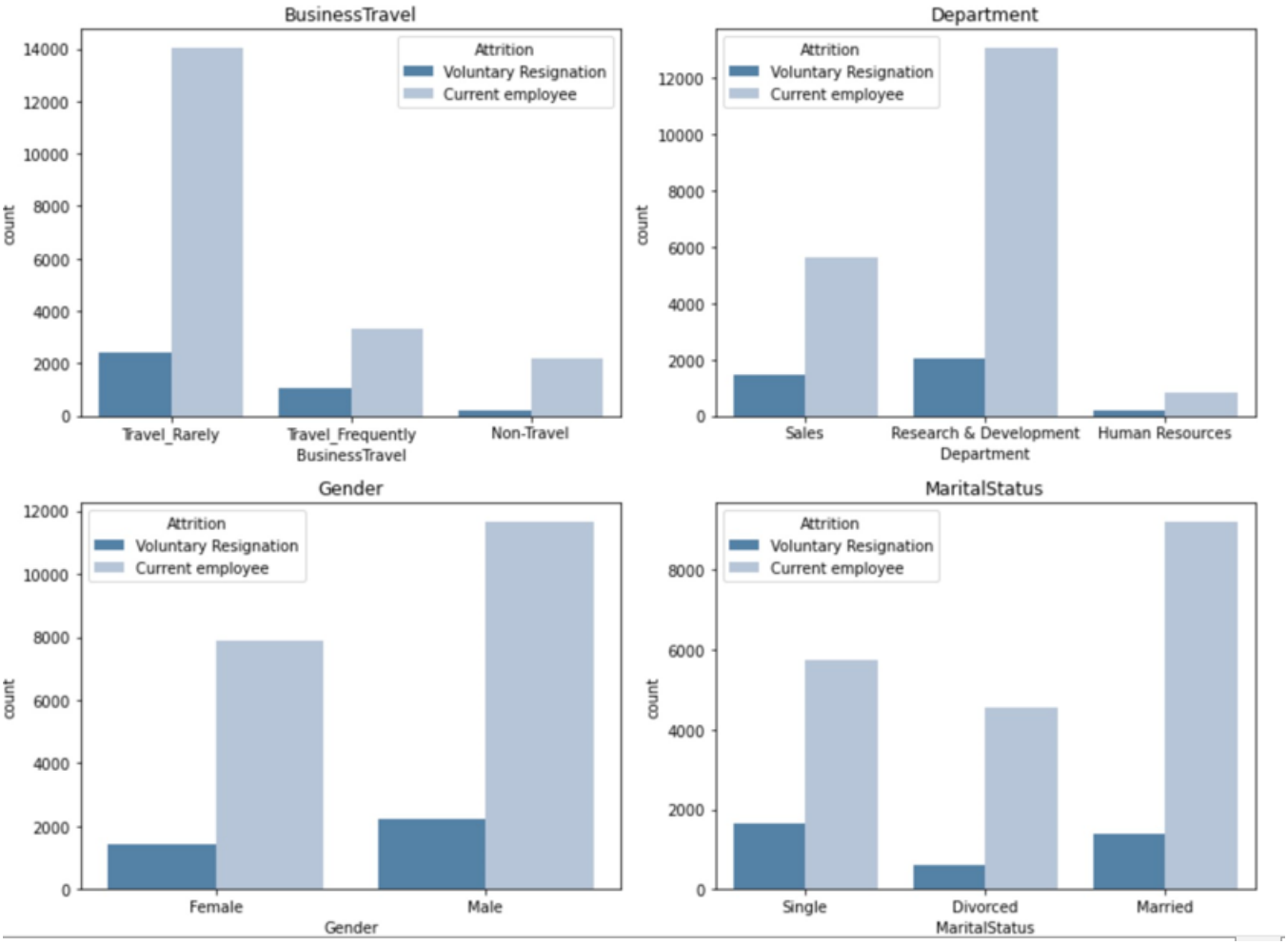
➢ Years Since Last Promotion :

- 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.

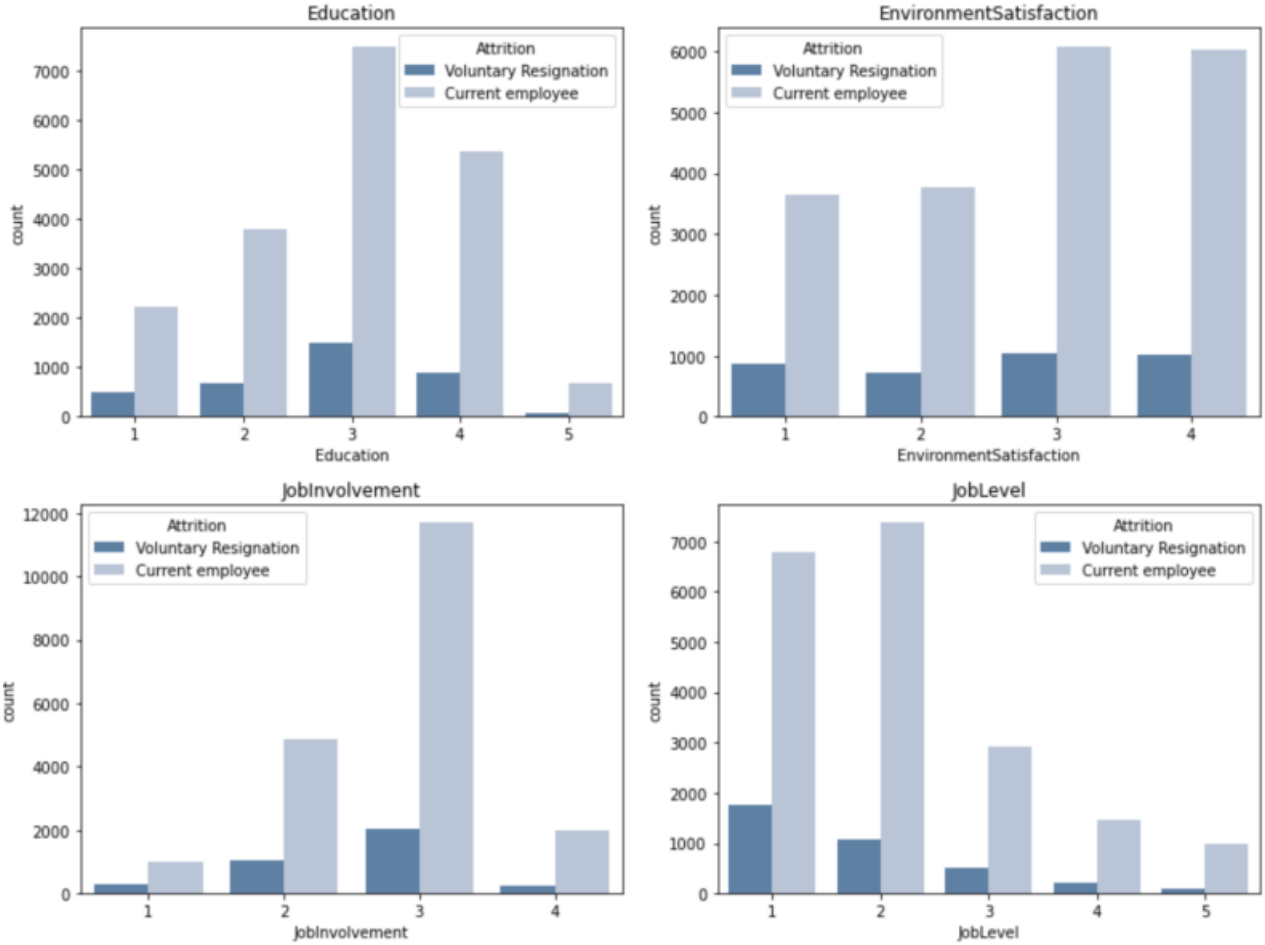
➢ Years With Current Manager :

- 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.

4.3: 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.

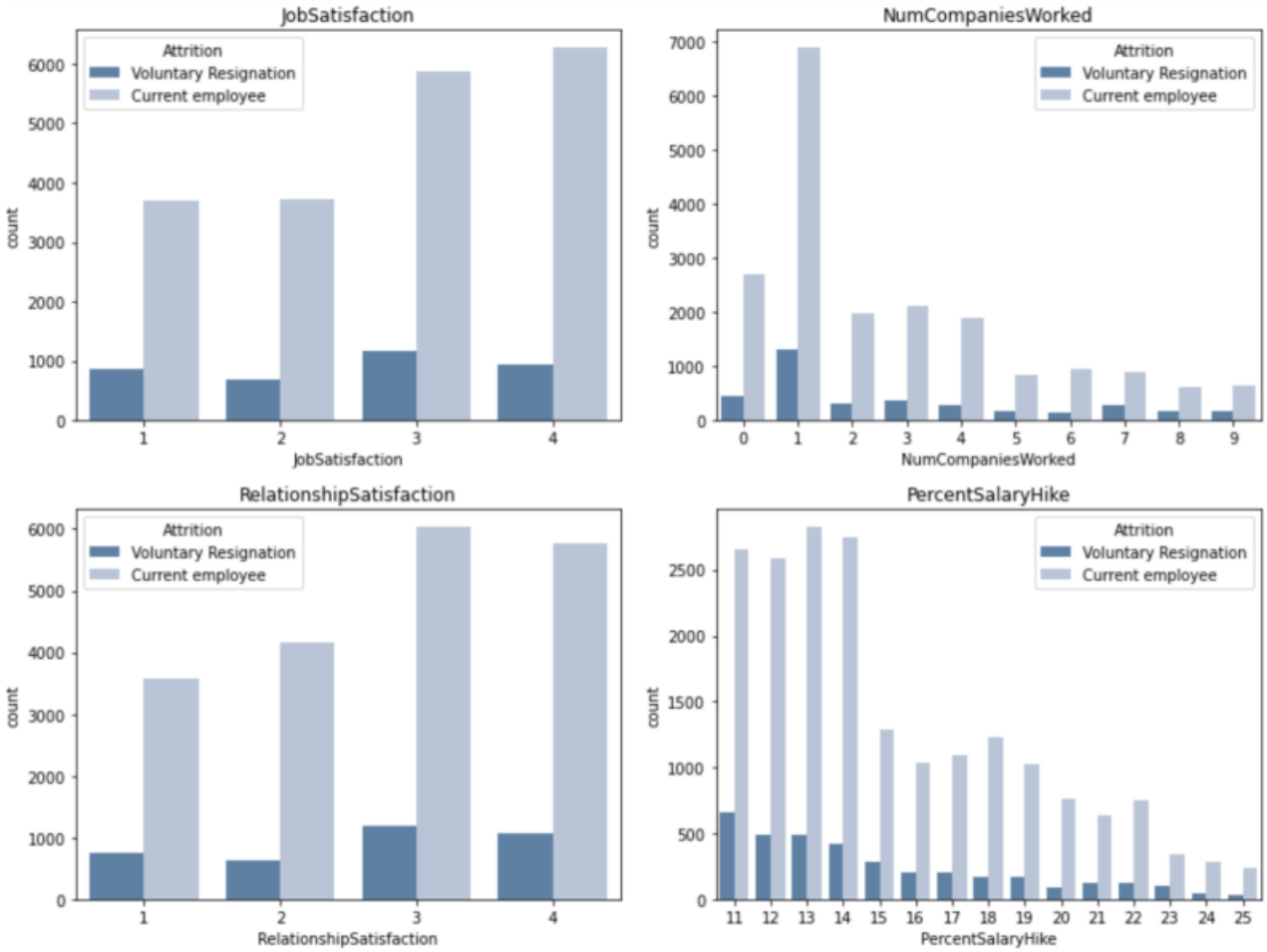


- Education: Around 30% of employees have education level of 3. For both male and female, attrition rate is higher for education level 1,2 and 3.

- 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.

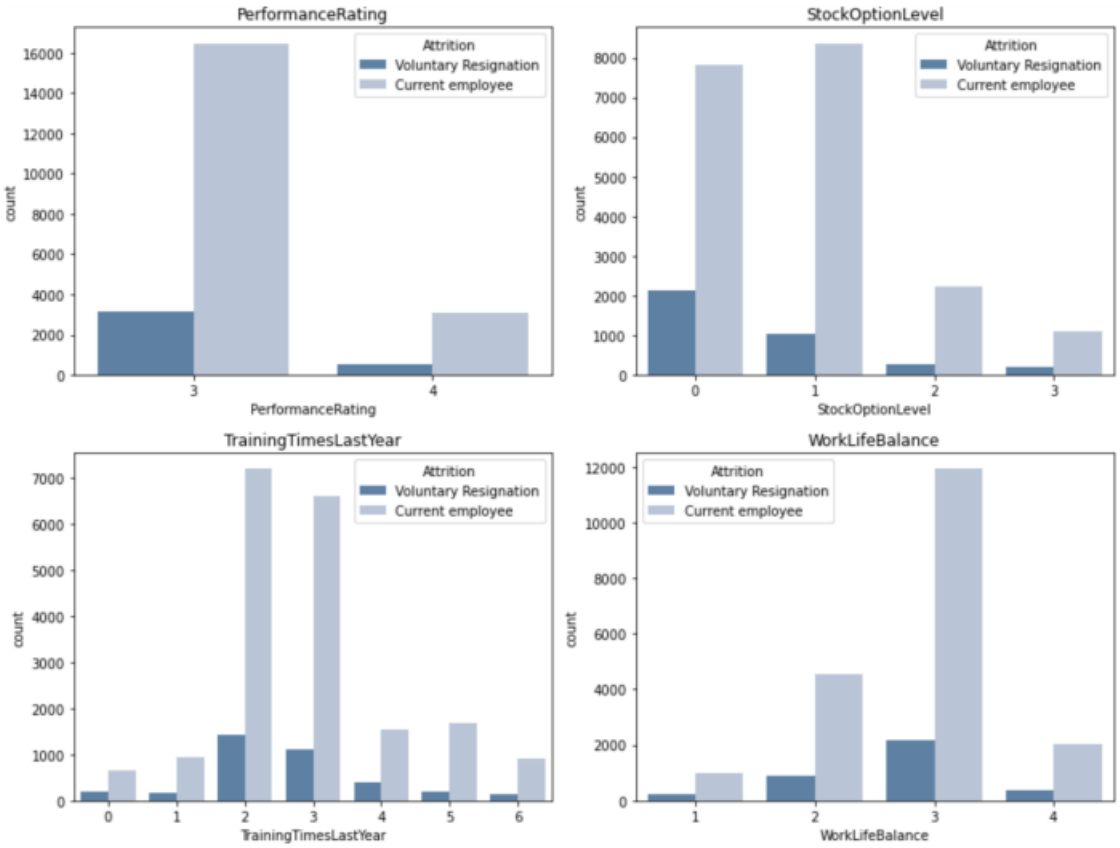


- Job Satisfaction : Job Satisfaction count for 3 and 4 are more than 1 and 2. Higher attrition rate can be seen in Job Satisfaction level 1 and 2.

- 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.

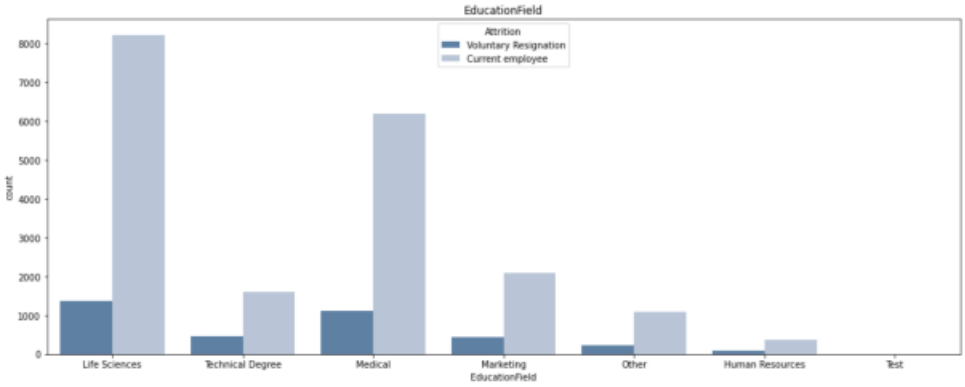


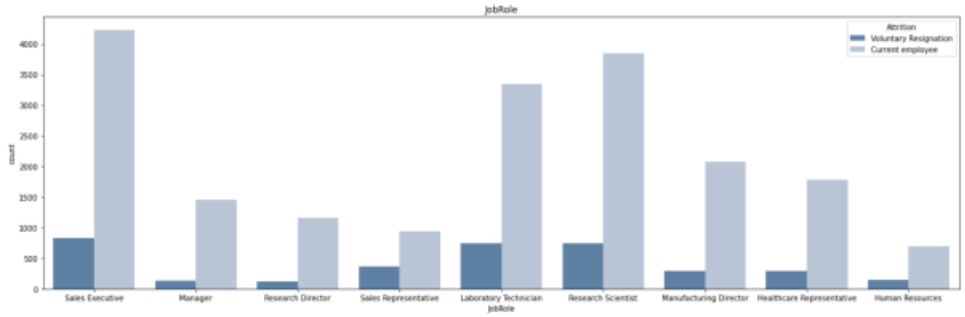
- Performance Rating: There are very few employees who have performance rating 4. Performance Rating 3 has higher rate of attrition for both male and female.

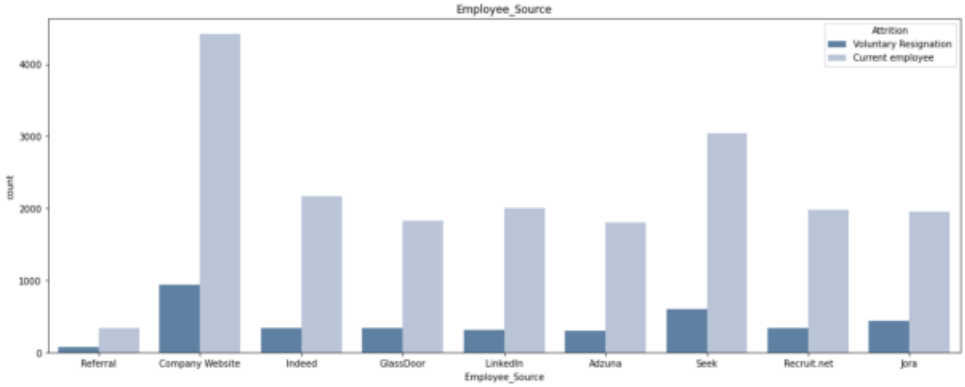
- Stock Option Level : There are many employees who does not have stock options level, As the stock options level increases the count of employees reduces. Higher attrition rate is observed in lower stock options level for both genders.

- 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.

- 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.





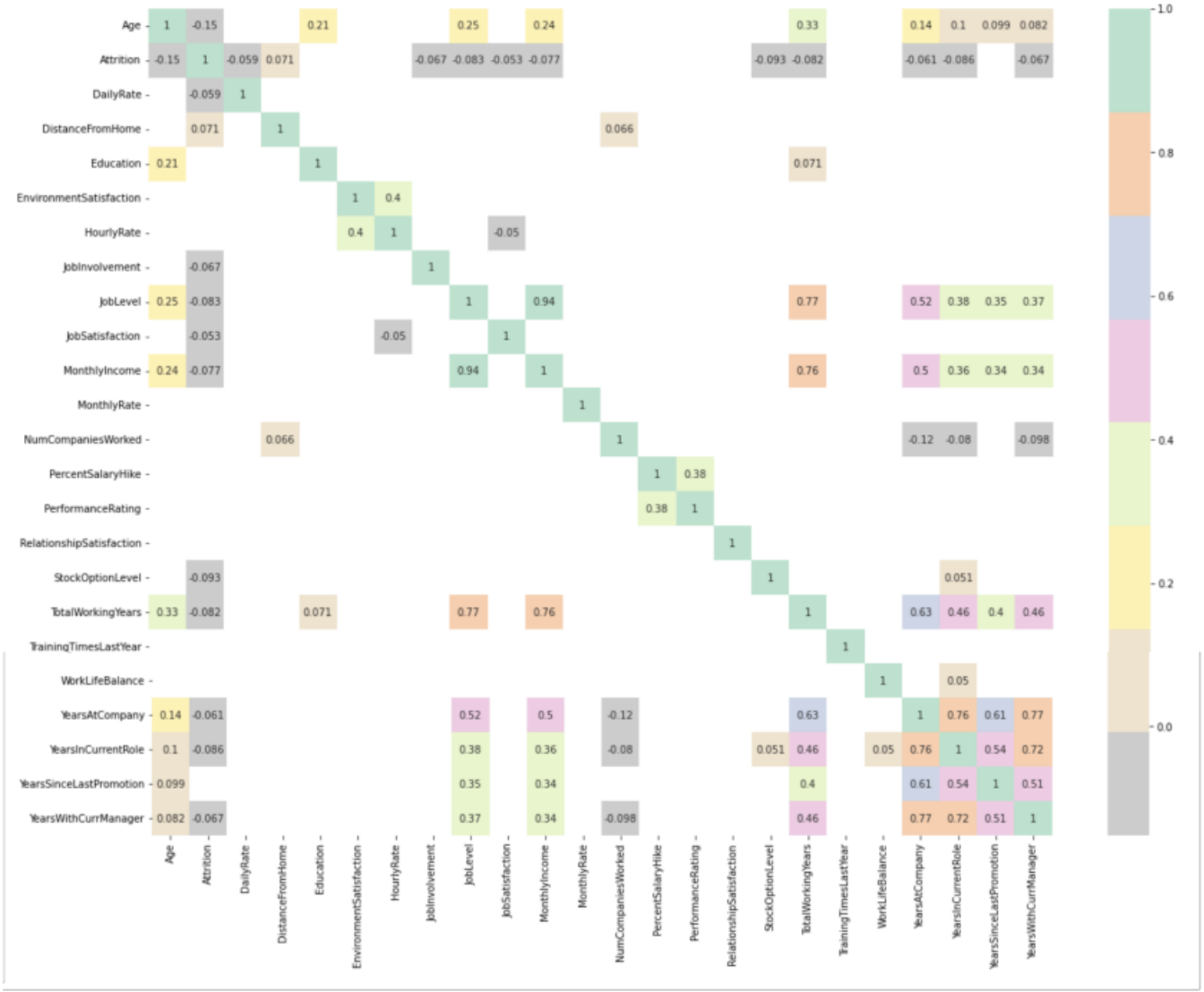


- Education Field : Around 70% of employees are having 'Life Sciences' and 'Medical' education field. Attrition rate of female in 'HR' education field is less when compared to male. Attrition rate of female in 'Life Sciences' and 'Medical' is more when compared to male.

- 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

4.4: Correlation Matrix



- 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

5: Base Model: Logistic Regression

Logit Regression Results

Dep. Variable:	Attrition	**No. Observations:**	16230
Model:	Logit	**Df Residuals:**	16199
Method:	MLE	**Df Model:**	30
Date:	Wed, 28 Dec 2022	**Pseudo R-squ.:**	0.1085
Time:	16:17:33	**Log-Likelihood:**	-6279.7
converged:	True	**LL-Null:**	-7044.3
Covariance Type:	nonrobust	**LLR p-value:**	2.454e-303

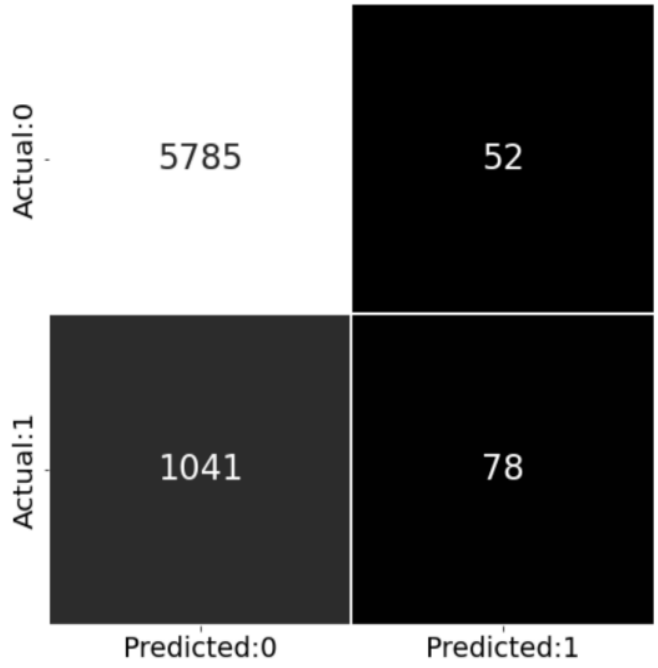
Inferences :

- The Pseudo R-squ. obtained from the above model summary is the value of McFadden's R-squared. • The LLR p-value is less than 0.05, implies that the model is significant.

- The maximum of Cox & Snell R-squared is always less than 1. By above model Cox & Snell R-squared is less than 1 i.e. (0.1085).

Great Learning
POWER AHEAD

Confusion Matrix And Classification Report:



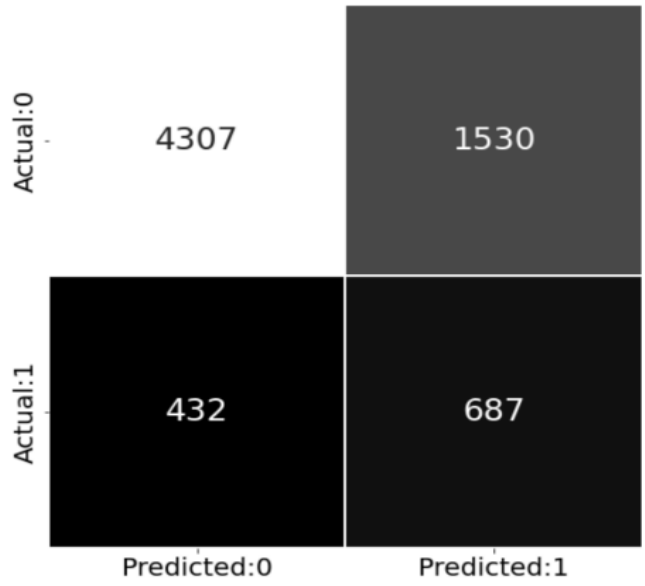
Classification Report :

	precision	recall	f1-score	support
0	0.85	0.99	0.91	5837
1	0.60	0.07	0.12	1119
accuracy			0.84	6956
macro avg	0.72	0.53	0.52	6956
weighted avg	0.81	0.84	0.79	6956

Inference :

- We can infer that the recall of the positive class is known as sensitivity and the recall of the negative class is specificity.
- Accuracy of model is 0.84 for 0.5 as the Threshold.

Confusion Matrix For Best Cut Off Value:



Classification Report :

	precision	recall	f1-score	support
0	0.91	0.74	0.81	5837
1	0.31	0.61	0.41	1119
accuracy			0.72	6956
macro avg	0.61	0.68	0.61	6956
weighted avg	0.81	0.72	0.75	6956

Inference :

- Base model performance is good, but it can be improved by passing different values and obtain a cut-off based on the passed values (Youden's Index, Cost-based Method).

- We have to implement various classification machine learning algorithms and take feedback for them. This will take us to the best suited model.

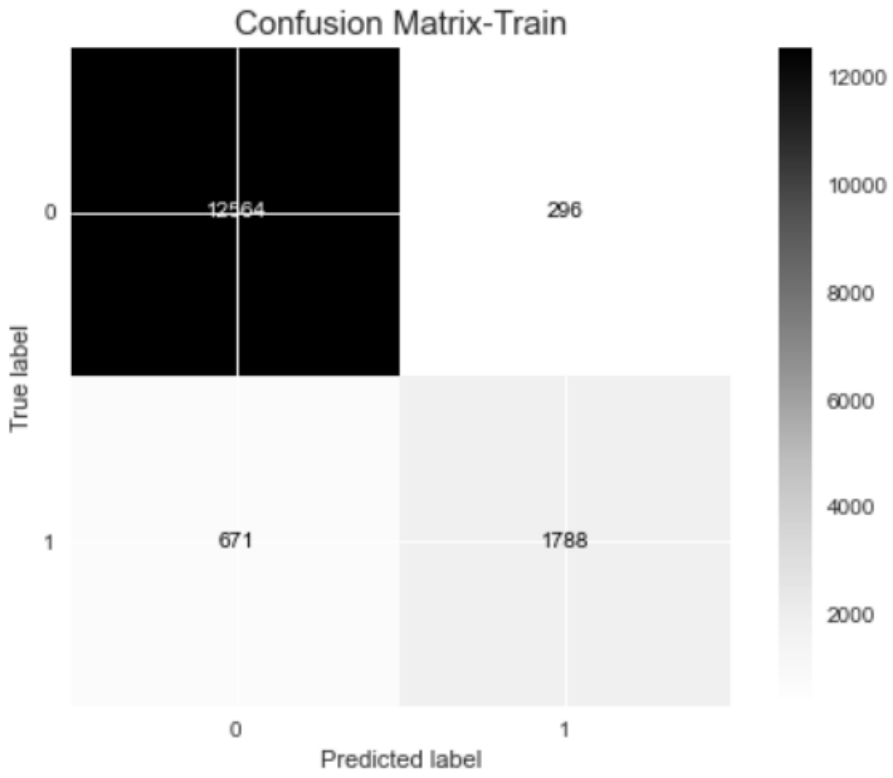
6.1: KNN:

K-NN algorithm stores all the available data and classifies a new data point based on the similarity.

- Train

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

           0       0.95      0.98      0.96     12860
           1       0.86      0.73      0.79      2459

    accuracy                           0.94     15319
   macro avg       0.90      0.85      0.88     15319
weighted avg       0.93      0.94      0.93     15319
```

The decision tree model gives us the Accuracy of 0.93, AUC score as 0.97 and f1 score as 0.78
As we already know our data is highly imbalanced we will do build a model on test data

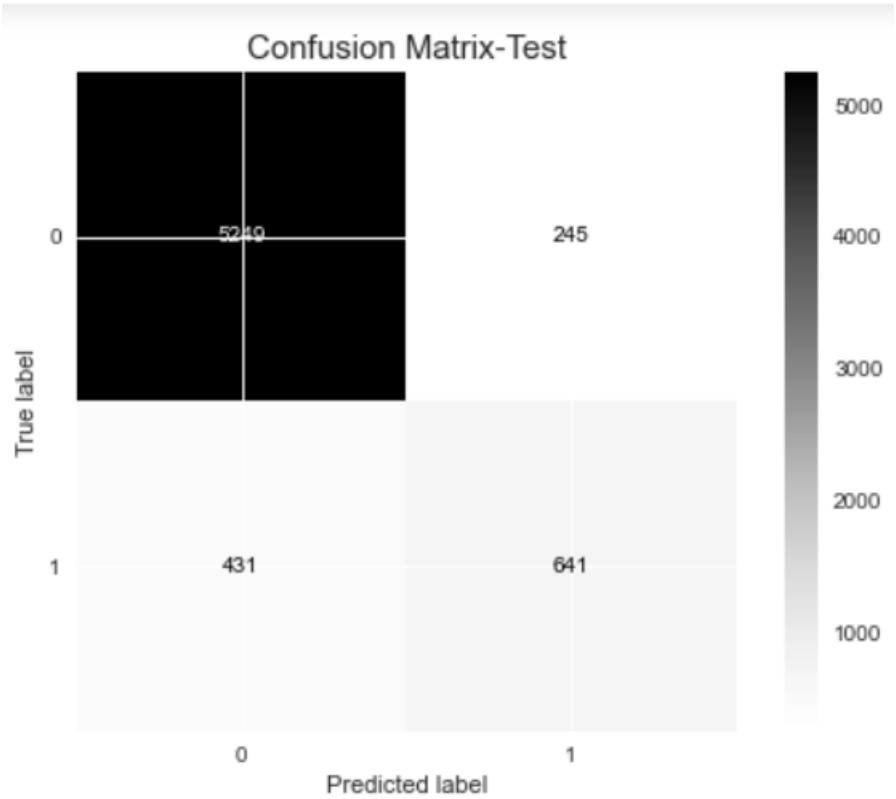


The confusion matrix shows the separation clearly of the true positives, true negatives, false positives and false negatives. Here the false negatives and false positives classifications are minimal, this could be because of the over fit. We will have to build a model with the test data to verify

- Test:

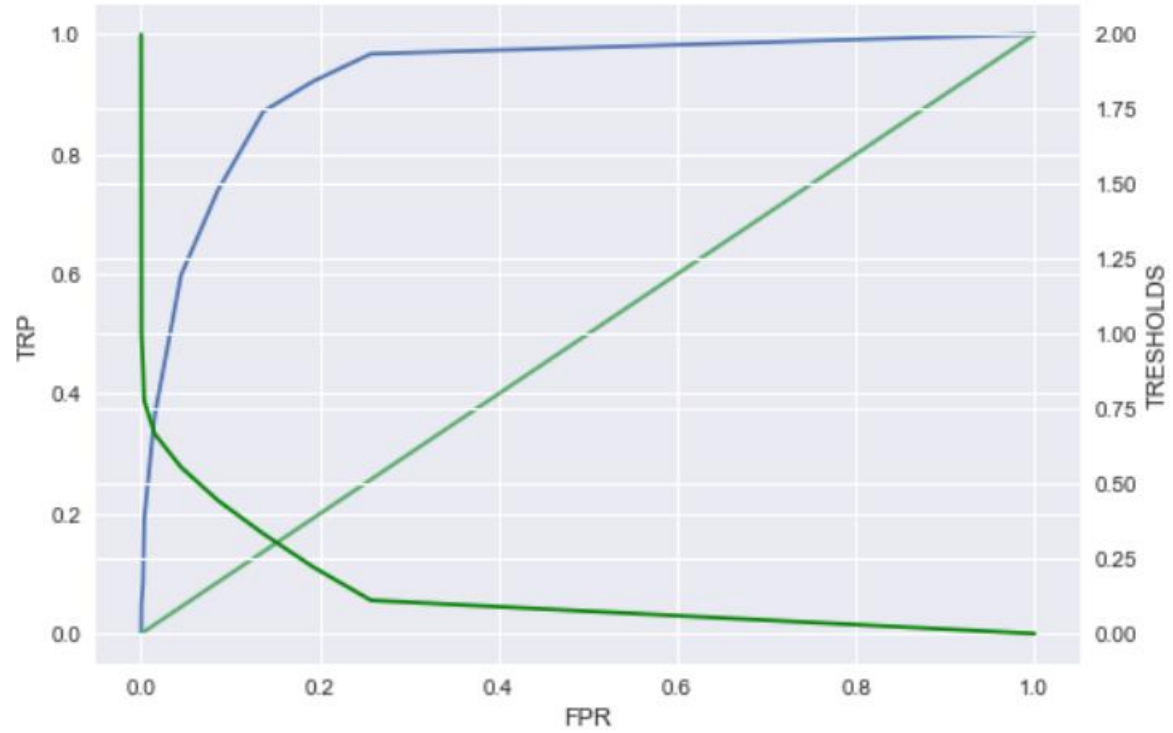
```
Classification Report-Test :
              precision    recall  f1-score   support

           0       0.92      0.96      0.94      5494
           1       0.72      0.60      0.65      1072

    accuracy                           0.90      6566
   macro avg       0.82      0.78      0.80      6566
weighted avg       0.89      0.90      0.89      6566
```

The accuracy in test model is 0.89 however in train model it is around 0.93 so there is no case of overfitting in our model.



It shows huge imbalance in true positive and true negative compare to our train model

Plot : AUC-ROC Curve



6.2 Decision Tree Classifier:

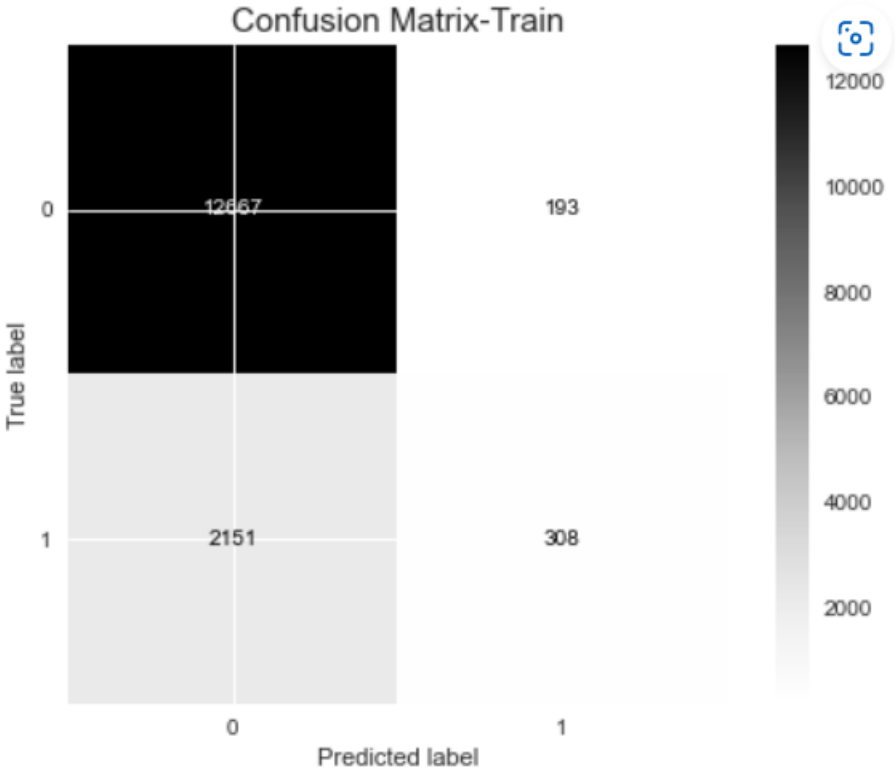
Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes.

Train:

Classification Report-Train :

	precision	recall	f1-score	support
0	0.85	0.98	0.92	12860
1	0.61	0.13	0.21	2459
accuracy			0.85	15319
macro avg	0.73	0.56	0.56	15319
weighted avg	0.82	0.85	0.80	15319

As we know That the data is imbalanced in our target variable so the precision, recall, accuracy, and f1 score are 0.61, 0.13, 0.84, 0.21. Next we will build a model on test data.

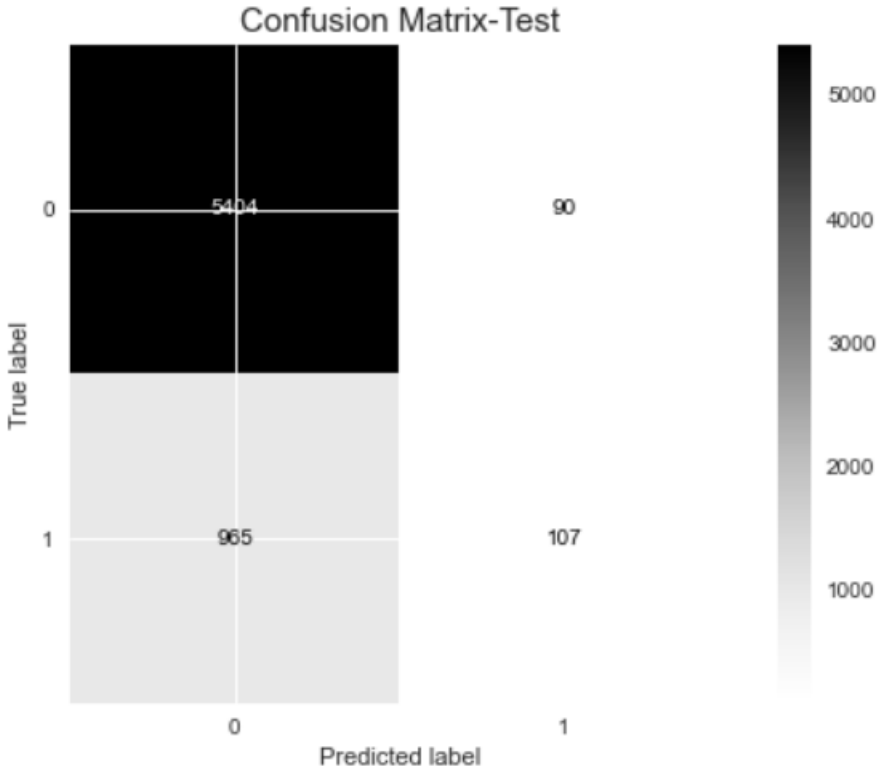


The confusion matrix shows the separation clearly of the true positives, true negatives, false positives and false negatives. Now we will build a model with the test data.

Test:

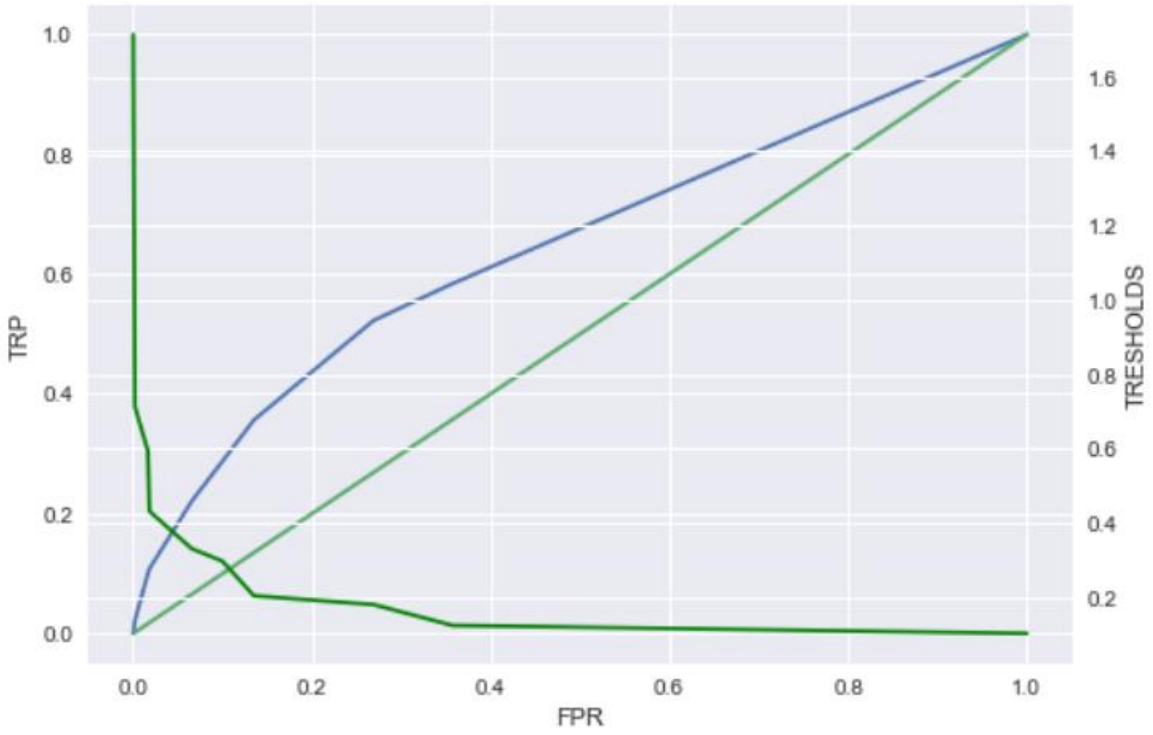
```
Classification Report-Test :
              precision    recall  f1-score   support

           0       0.85      0.98      0.91      5494
           1       0.54      0.10      0.17      1072

    accuracy                           0.84      6566
   macro avg       0.70      0.54      0.54      6566
weighted avg       0.80      0.84      0.79      6566
```

The precision recall f1-score and accuracy of the test model is similar to train model



Shows imbalance in false negative and false positive as compared to train model.

Plot : AUC-ROC Curve



6.3: Random Forest Classifier

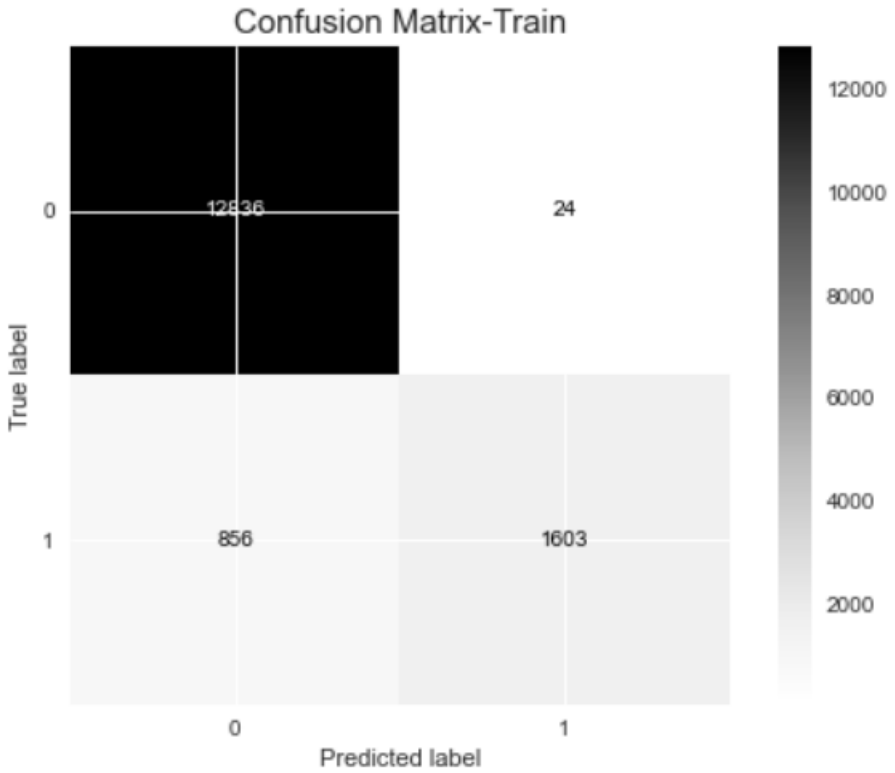
Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Train:

Classification Report-Train :

	precision	recall	f1-score	support
0	0.94	1.00	0.97	12860
1	0.99	0.65	0.78	2459
accuracy			0.94	15319
macro avg	0.96	0.83	0.88	15319
weighted avg	0.95	0.94	0.94	15319

For random forest classifier Accuracy of our train model is 0.94 and f1-score is 0.78 Precision is 0.99 and recall is 0.65. Now we will build a model with the test data.

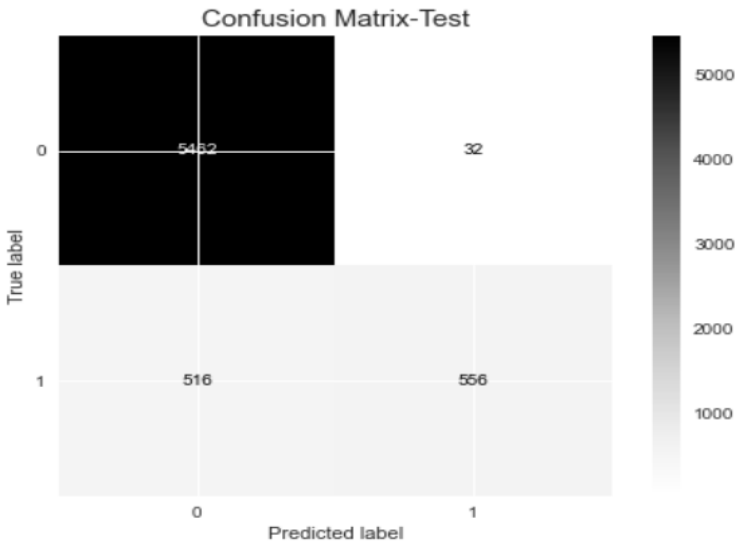


The confusion matrix shows the separation clearly of the true positives, true negatives, false positives and false negatives. Now we will build a model with the test data.

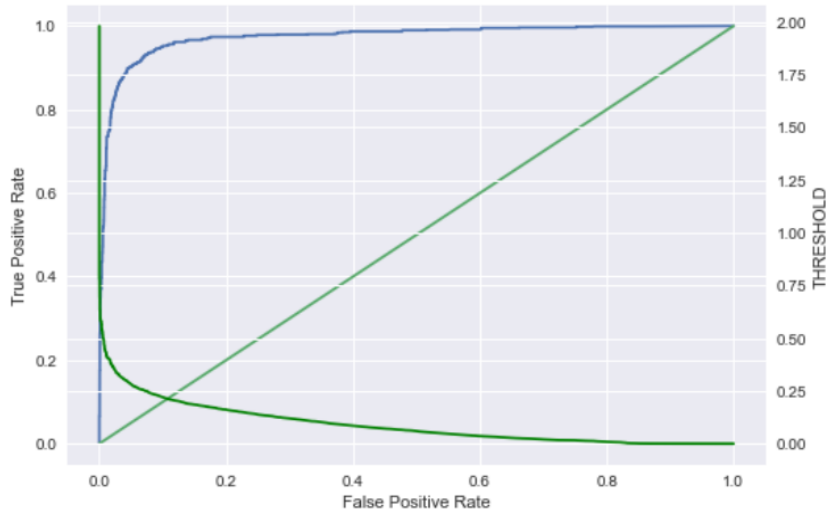
Test:

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

           0       0.91      0.99      0.95      5494
           1       0.95      0.52      0.67      1072

    accuracy                           0.92      6566
   macro avg       0.93      0.76      0.81      6566
weighted avg       0.92      0.92      0.91      6566
```

And the for our test model the accuracy score is 0.91 and f1-score is 0.67 which is similar to our train model.



Plot : AUC-ROC Curve



We see significant difference in the train and test models.

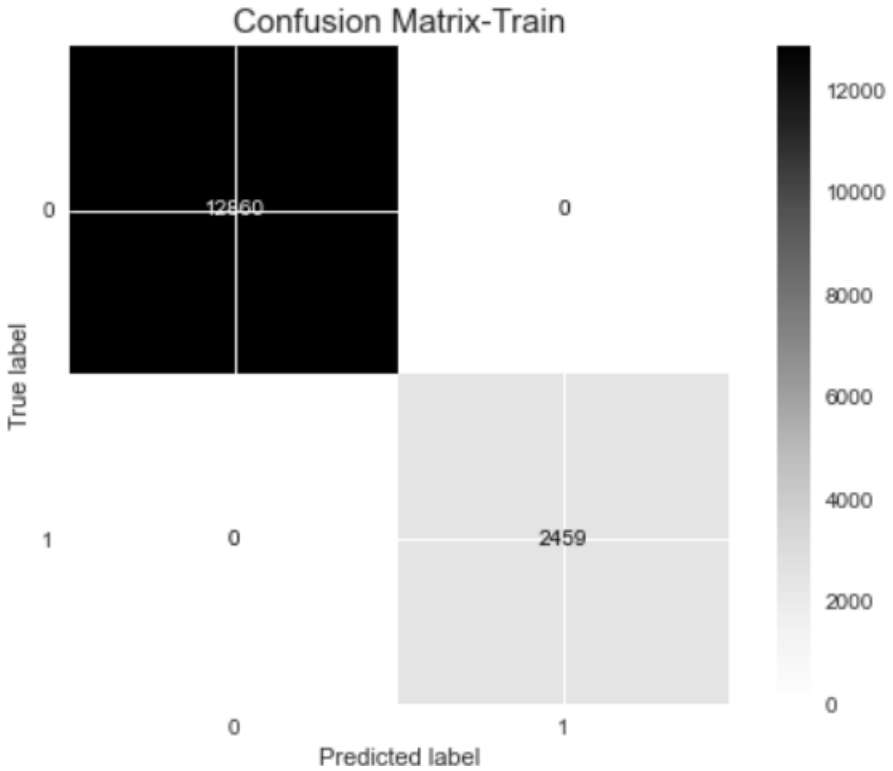
6.4: XGB Classifier

Boosting is an ensemble modeling, technique that attempts to build a strong classifier from the number of weak classifiers.

Train:

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

           0       1.00      1.00      1.00     12860
           1       1.00      1.00      1.00      2459

    accuracy                           1.00     15319
   macro avg       1.00      1.00      1.00     15319
weighted avg       1.00      1.00      1.00     15319
```

The XGB Classifier model gives us an over fit model as we know that the dataset is imbalanced in the target variable. The accuracy, precision, recall and f1-score are all 100% in the train data depicting a clear over fit of the model. Next we will build a model for the test data to cross verify whether there is an over fit in the model or not.

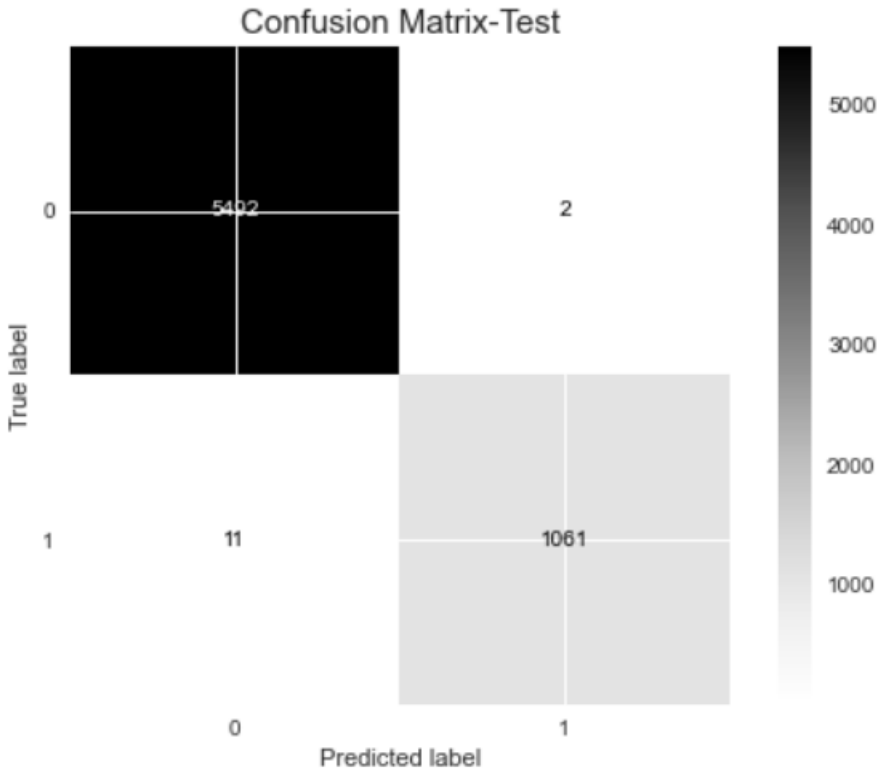


The confusion matrix shows the separation clearly of the true positives, true negatives, false positives and false negatives. Here the false negatives and false positives classifications are 0, this could be because of the over fit. We will have to build a model with the test data to verify.

Test:

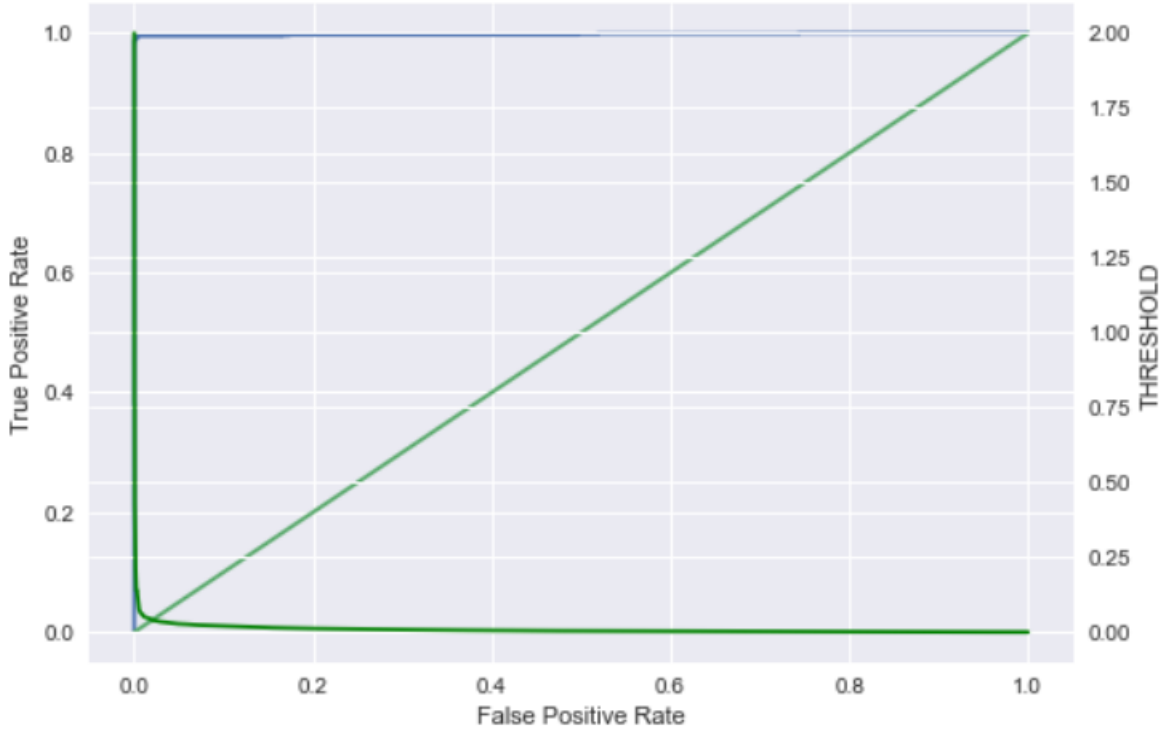
```
Classification Report-Test :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      5494
           1       1.00      0.99      0.99      1072

    accuracy                           1.00      6566
   macro avg       1.00      0.99      1.00      6566
weighted avg       1.00      1.00      1.00      6566
```

The accuracy for the test model is 100% where in the train data, it was too 100%. We also see that the precision for class 1 is 100% where as in the train model, it was 100%. It shows our model is not over fit.



The confusion matrix for the test model is almost same as our train model.

Plot : AUC-ROC Curve

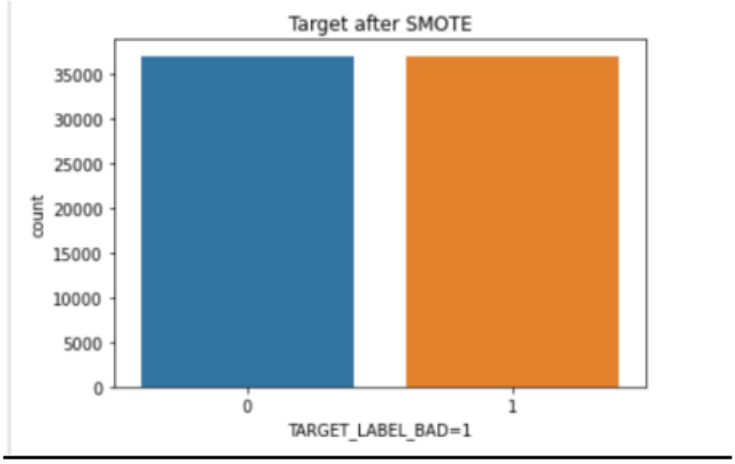


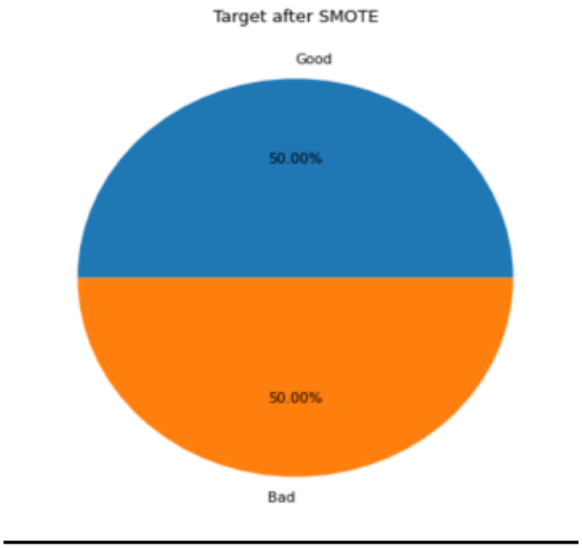
Inference from Base Model:

From all the above models, we can see that there is a severe drop down in the train prediction and test prediction scores. This is mainly because of the imbalance in the target variable. This is evident from the classification report as the 0 class is predicted very well in some models, where as the 1 class is not predicted as expected. This is because there is a huge imbalance of 84.5% in the dataset, where the 0 class has 85% of the data points and the 1 class only has 15% of the data points. This can be rectified using the SMOTE function in python.

SMOTE

Smoting is the technique used to rectify any imbalance in the dataset. We use this for our dataset as there is a severe imbalance in the 0 and 1 class





From the above images, we can see that the imbalances are rectified and hence we can use this smoted dataset for out further models.

REBUILDING THE MODELS

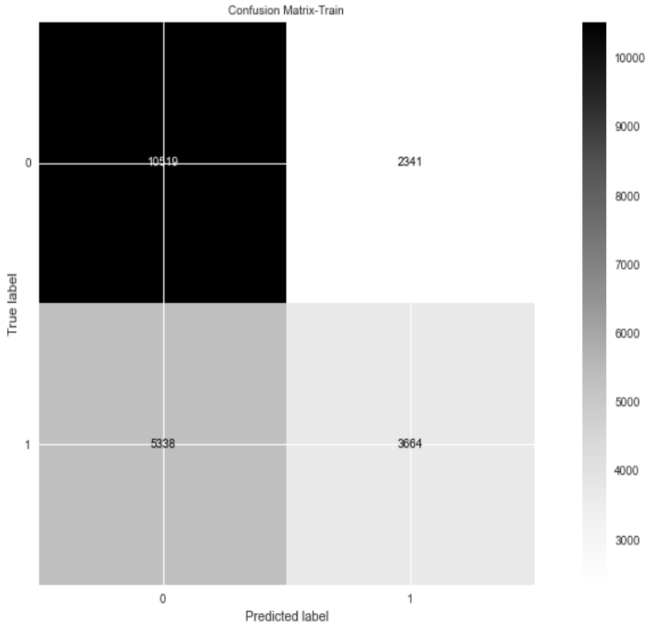
After smoting the data, we will rebuild the models to check the performance of the models.

Logistic Regression:

Train:

Classification Report-Train

	precision	recall	f1-score	support
0	0.66	0.82	0.73	12860
1	0.61	0.41	0.49	9002
accuracy			0.65	21862
macro avg	0.64	0.61	0.61	21862
weighted avg	0.64	0.65	0.63	21862

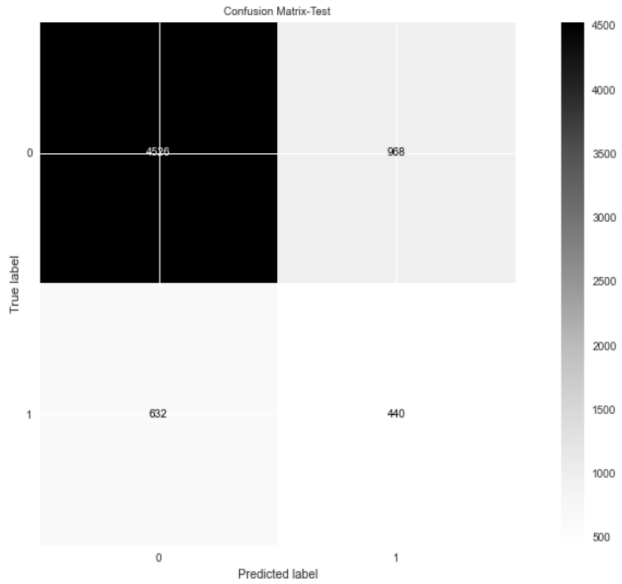


With Smote we are getting the accuracy of 0.64 for the train data now we will build a model with test data.

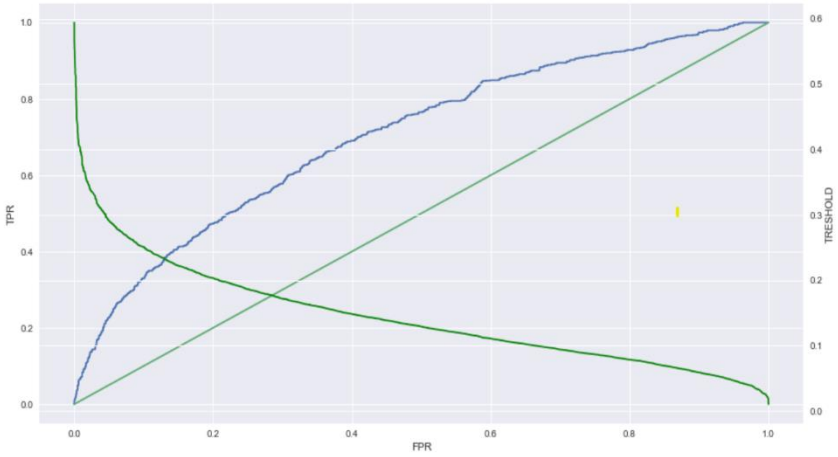
Test:

Classification Report-Test

	precision	recall	f1-score	support
0	0.88	0.82	0.85	5494
1	0.31	0.41	0.35	1072
accuracy			0.76	6566
macro avg	0.59	0.62	0.60	6566
weighted avg	0.79	0.76	0.77	6566



The test accuracy with smote we are getting as 0.75.

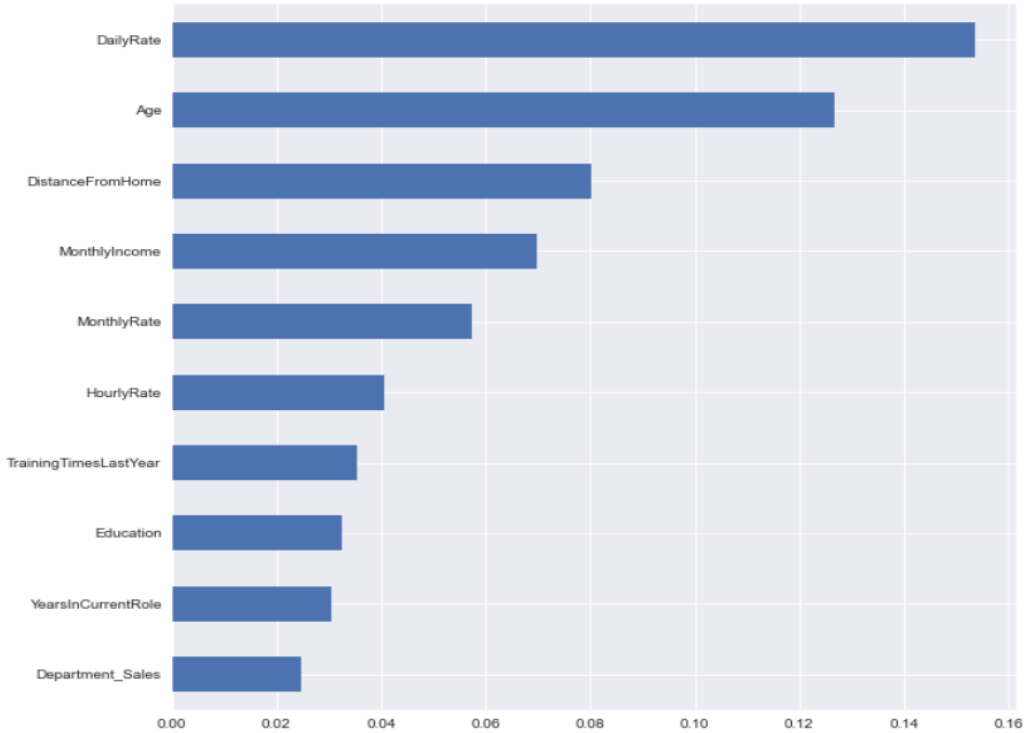


Feature Importance using different classifiers:

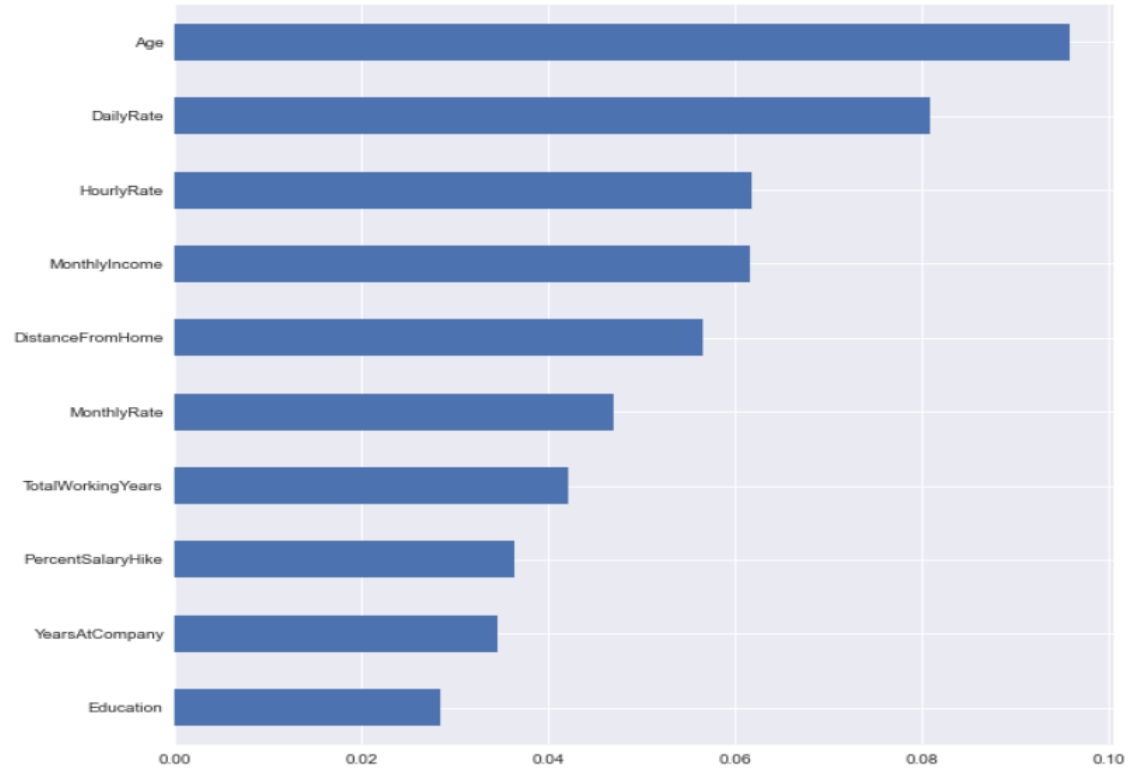
Here we have used three classifiers to get the best features of all ,

```
pipeline=[DecisionTreeClassifier(),RandomForestClassifier(),XGBClassifier()]
```

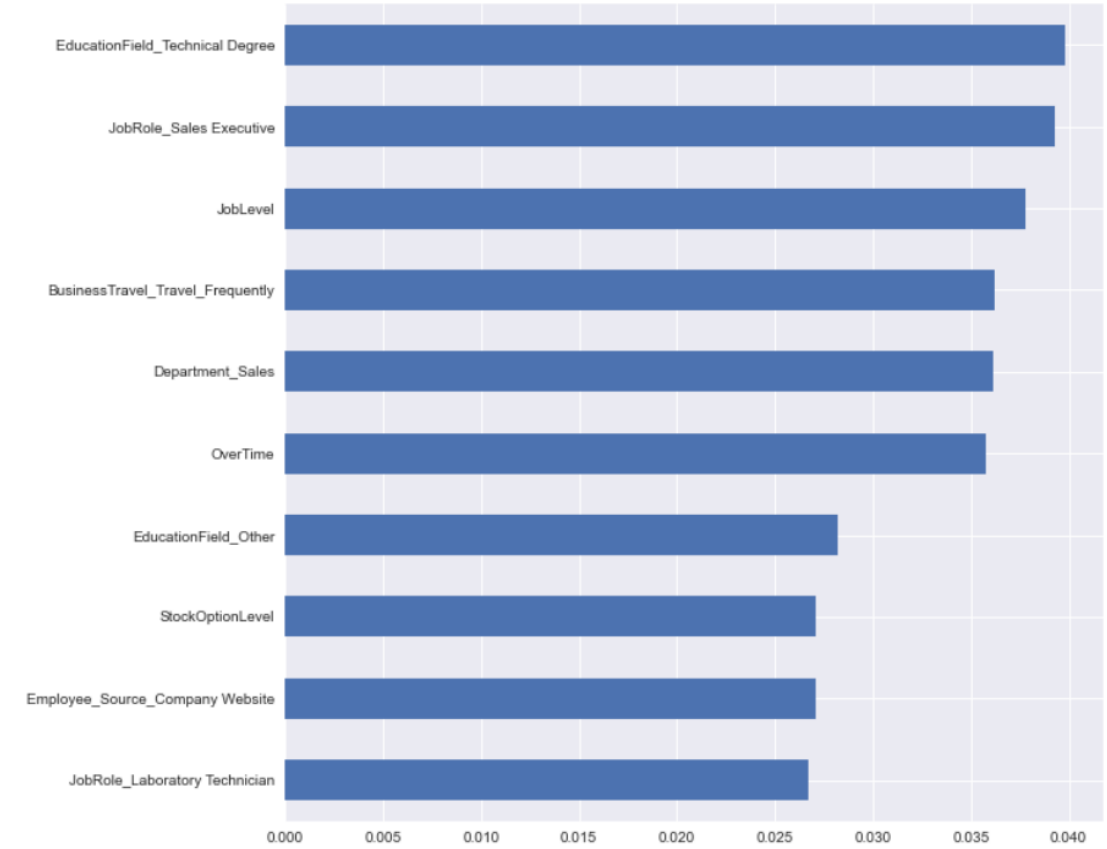
```
DecisionTreeClassifier()
```



```
RandomForestClassifier()
```



XGBoost Classifier ()



	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	Education	MonthlyRate	OverTime
6	HourlyRate	TotalWorkingYears	EducationField_Other
7	TrainingTimesLastYear	PercentSalaryHike	StockOptionLevel
8	PercentSalaryHike	YearsAtCompany	Employee_Source_Company Website
9	YearsInCurrentRole	Education	JobRole_Laboratory Technician

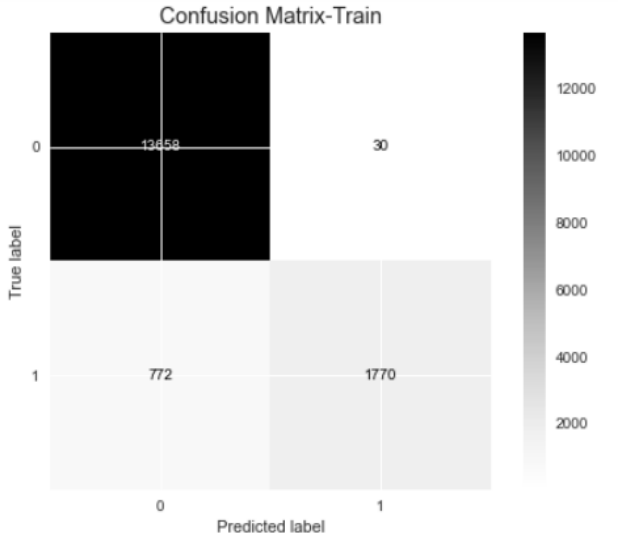
Optimal Features After doing feature selection

	0
0	DailyRate
1	Age
2	DistanceFromHome
3	MonthlyIncome
4	MonthlyRate
5	Education
6	HourlyRate
7	TrainingTimesLastYear
8	PercentSalaryHike
9	YearsInCurrentRole
10	Age
11	DailyRate
12	HourlyRate
13	MonthlyIncome
14	DistanceFromHome
15	MonthlyRate
16	TotalWorkingYears
17	PercentSalaryHike
18	YearsAtCompany
19	Education
20	EducationField_Technical Degree
21	JobRole_Sales Executive
22	JobLevel
23	BusinessTravel_Travel_Frequently
24	Department_Sales
25	OverTime
26	EducationField_Other
27	StockOptionLevel
28	Employee_Source_Company Website
29	JobRole_Laboratory Technician

Model with Optimal Features

Random Forest Classifier

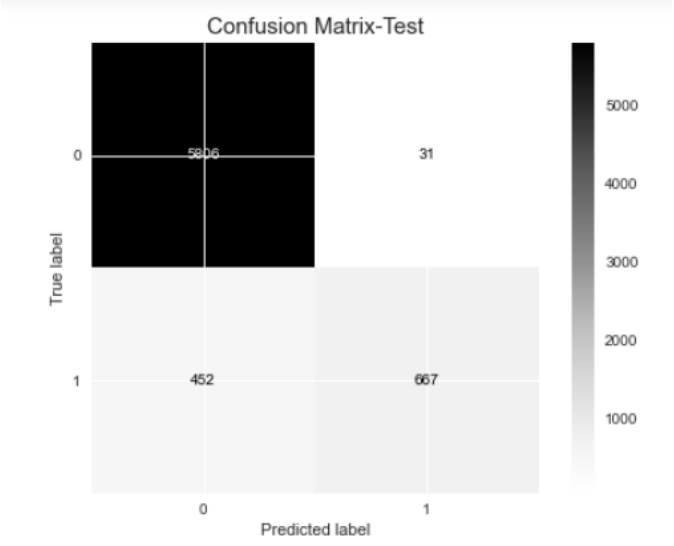
Train




```
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
```

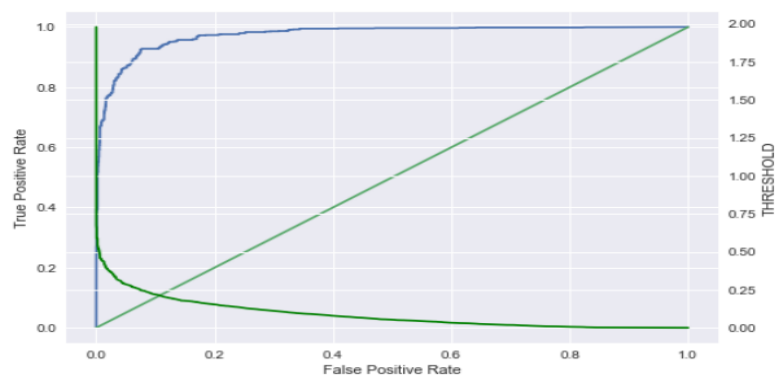
Test




```
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
```

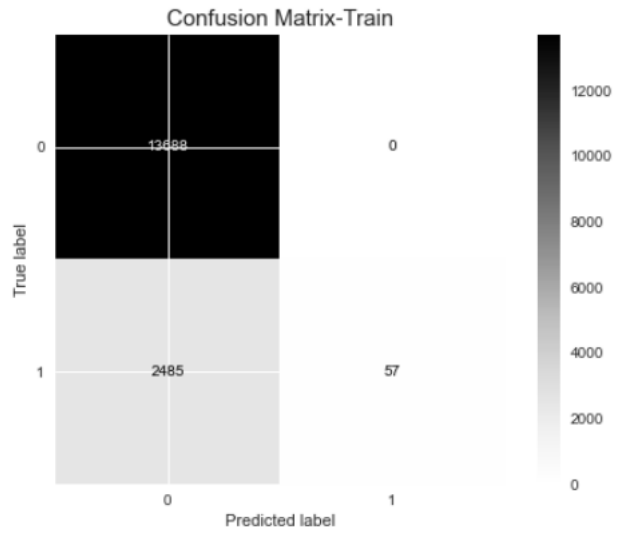
```
Plot : AUC-ROC Curve
```



GridSearchCV:

Random Forest Classifier

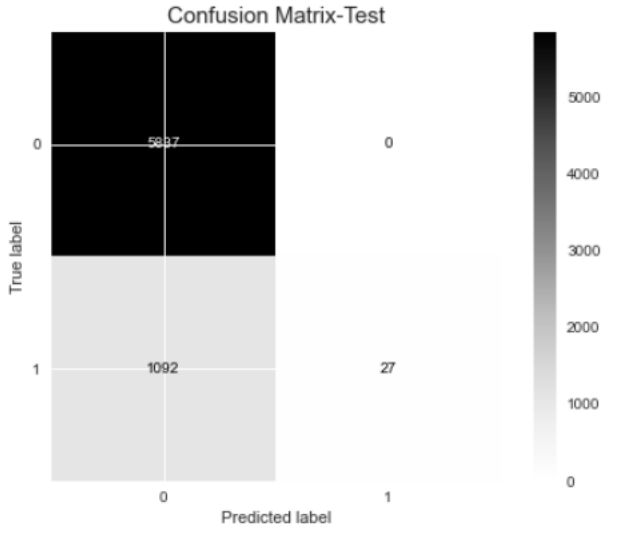
Train




```
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
```

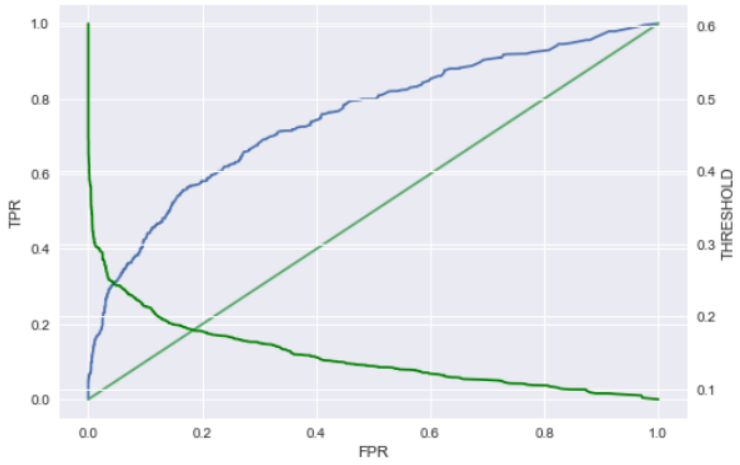
Test




```
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
```

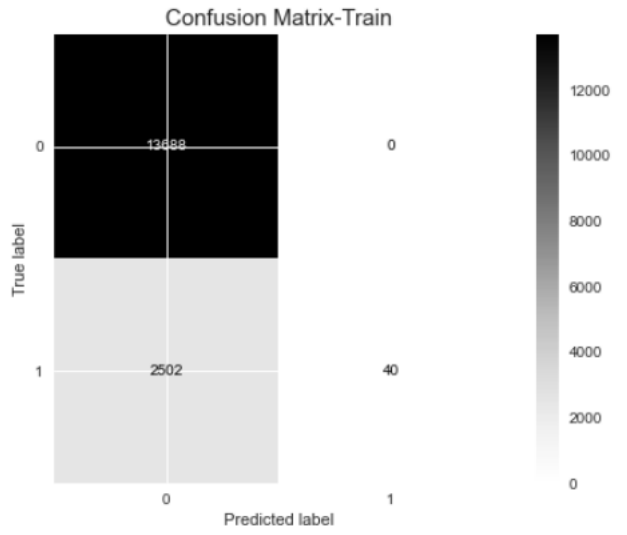
```
Plot : AUC-ROC Curve
```



RandomisedSearchCV

Random Forest Classifier

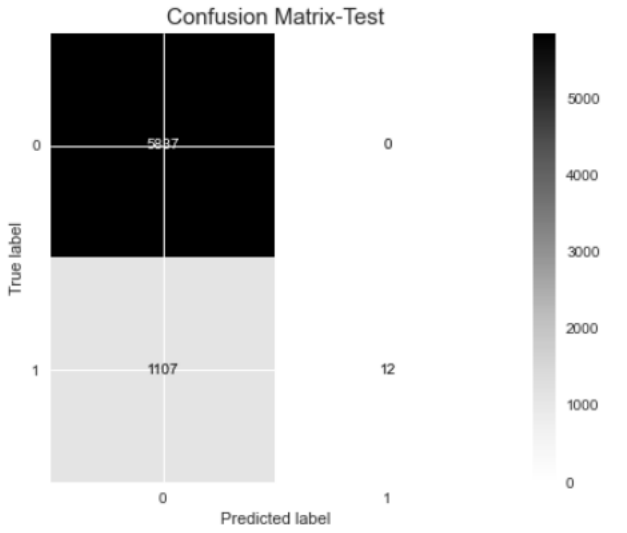
Train




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

           0       0.85      1.00      0.92     13688
           1       1.00      0.02      0.03      2542

    accuracy                           0.85     16230
   macro avg       0.92      0.51      0.47     16230
weighted avg       0.87      0.85      0.78     16230
```

```
Accuracy Score-Train : 0.8458410351201479

AUC Score-Train : 0.7355976290315684
```

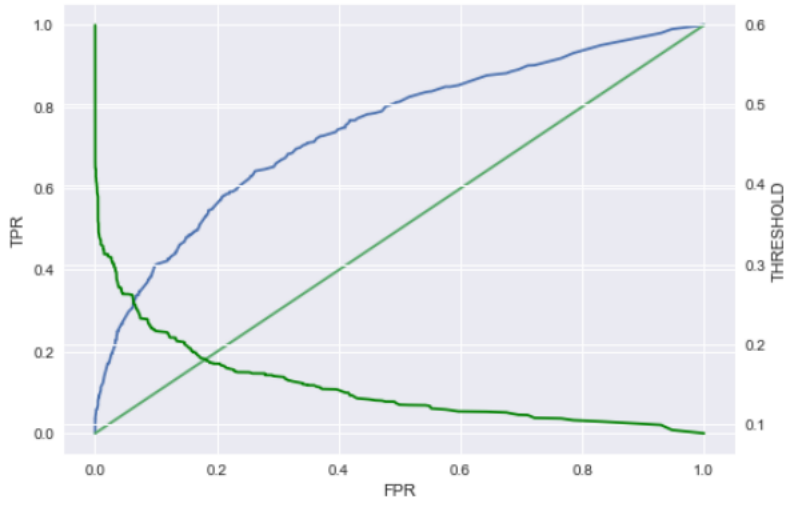
Test




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

           0       0.84      1.00      0.91      5837
           1       1.00      0.01      0.02      1119

    accuracy                           0.84      6956
   macro avg       0.92      0.51      0.47      6956
weighted avg       0.87      0.84      0.77      6956
```

```
Accuracy Score-Test : 0.8408568142610696

AUC Score-Test : 0.7421562516889039
```

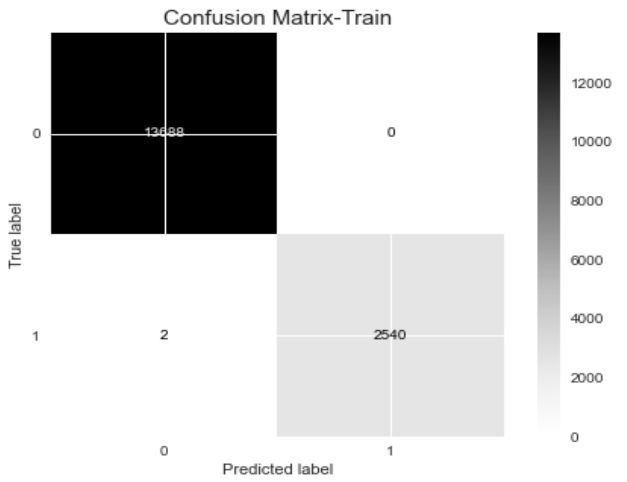
```
Plot : AUC-ROC Curve
```



Final Model with best Parameters

XGB Classifier

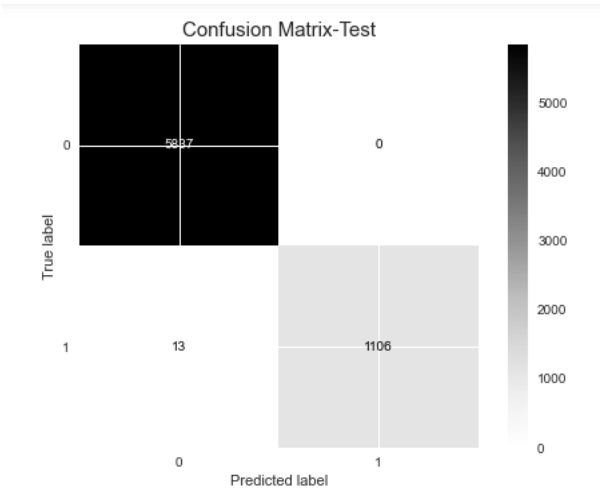
Train




```
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
```

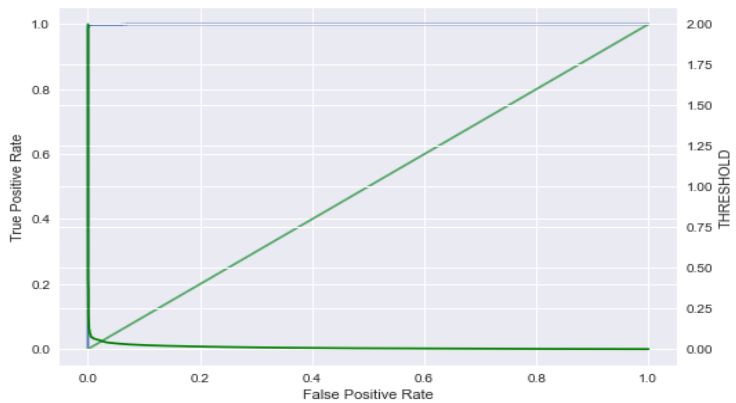
Test




```
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
```

```
Plot : AUC-ROC Curve
```



Cross Validation

```
cross_val = cross_val_score(xgbc,X_train,y_train ,cv = 10,scoring="accuracy")
```

```
cross_val
```

```
array([0.99753543, 0.99815157, 0.99630314, 0.99445471, 0.99691929,
       0.99938386, 0.99876771, 0.99691929, 0.99691929, 0.995687  ])
```

```
cross_val.mean()
```

```
0.9971041281577324
```

```
cross_val.std()
```

```
0.0013791145585704985
```

```
print("Accuracy: %0.2f (+/- %0.5f)" % (cross_val.mean(), cross_val.std() * 2))
```

```
Accuracy: 1.00 (+/- 0.00276)
```

Model Interpretation:

- From all the above models, we can see that the XG-Boost model gives the best fit for the data after smoting as well as removing the insignificant variables. The Random Forest and Decision Tree models also give good results on the smoted data, but the model over fits on the data and hence giving us sub-optimal classification on the test data compared to the train data. Hence we go ahead with the XG-Boost model to avoid over fitting and to keep intact the precision of the model.

Final Accuracy of the model:

- In this project we used 7 algorithms. We started with the base model and took the inferences from them, after evaluating important measures we went for their parameter tuning to increase their performance.
- Since, our data is highly imbalanced we mainly focus on the F1-Score,AUC score because the performance metrics ‘accuracy’ would be affected due to bias.
- The major Limitation of the model we have built is the high imbalance in data and even after handing the imbalance using SMOTE, we are trying to introduce BIAS in the model.
- Train Accuracy - 99.98% Test Accuracy - 99.81%
- XGBclassifier out performed all the other models and have got highest train model accuracy 99% and test model accuracy as 98%
- We have done feature selection and passed the important features into the XGB classifier model.
- Hence, we suggest to use opt XGB Classifier to classify whether the employee will leave the organization or not.

Conclusion:

- By means of Data Science, the problem of employee turnover that every company fears can be analyzed and it also helps in reducing the costs of replacing valuable employees and maximizing profits.
- An EDA (Exploratory Data Analysis) has been carried out to uncover the reasons that lead to an employee's voluntary exit, and prediction algorithms have been used to predict when an employee is going to resign from the company.
- The Decision Tree algorithm also suggests us to reconsider the salary of the employees who earn less and assess the possibility of rising their salary
- 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.
- Invest in the development of those employees in the 21-30 age range and offer them possibilities to learn new skills and to grow inside the company.

This illustrates the great range of possibilities that Data Science offers to help managers make data-driven decisions within the staffing process.

References :

[1] E. Gallardo-Gallardo, N. Dries and T. F- González-Cruz, "What is the meaning of 'talent' in the world of work?," Human Resource Management Review, vol. 23, no. 4, pp. 290-300, 08 2013.

[2] H. Boushey and S. J. Glynn, "Center for American Progress," 16 11 2012. [Online]. Available: https://www.americanprogress.org/wpcontent/uploads/2012/11/CostofTurnover.pdf. [Accessed 27 08 2016]. [3]C. Rudin and S. Elston, "edX," 2015. [Online]. Available: https://courses.edx.org/courses/coursev1:Microsoft+DAT203x+3T2015/courseware/5bbcfcf04e6d49bda1eecb1e1c0bfc 24/a57949170b154fdd87057996c87717c7/. [Accessed 27 08 2016].

[3] A. B. Munir, S. H. M. Yasin and F. Muhammad-Sukki, "Big Data: Big Challenges to Privacy and Data Protection," International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, vol. 9, no. 1, p. 355, 2015.

[4] "REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016," 04 05 2016. [Online]. Available: http://eurlex.europa.eu/legalcontent/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN. [Accessed 03 09 2016].

[5] "Agencia Estatal Boletín Oficial del Estado," 13 12 1999. [Online]. Available: https://www.boe.es/boe/dias/1999/12/14/pdfs/A43088-43099.pdf. [Accessed 03 09 2016].

[6] V. Dhar, "Data Science and Prediction," Communications of the ACM, vol. 56, no. 12, pp. 64-73, 2013. [8] C.-Y. Zhang and C. P. Chen, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Information Sciences, vol. 275, pp. 314-347, 2014

[7] O. Maimon and L. Rokach, "Introduction to Knowledge Discovery in Databases," in Data Mining and Knowledge Discovery Handbook, New York, Springer, 2010, p. 1.

[8] H. J. Watson and B. H. Mixon, "The Current State of Business Intelligence," Computer, vol. 40, no. 9, pp. 96-97, 2007.

[9] R. Schutt and C. O'Neil, Doing Data Science: Straight Talk from the Frontline, Sebastopol: O'Reilly Media, 2013.

[10] C. Massey, M. Haas and M. Bidwell, "Coursera," 2016. [Online]. Available: https://www.coursera.org/learn/wharton-people-analytics. [Accessed 30 08 2016].

[11] «Watson Analytics,» [En línea]. Available: http://www03.ibm.com/software/products/en/watson-analytics. [Último acceso: 27 08 2016].

[12] "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2010. [Online]. Available: http://www.kdnuggets.com/polls/2010/data-mininganalytics-tools.html. [Accessed 30 08 2016].

[13] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining," Journal of Data Warehousing, vol. 5, no. 4, p. 13, 2000.

[14] "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2002. [Online]. Available: http://www.kdnuggets.com/polls/2002/methodology.htm. [Accessed 16 08 2016].

[15] "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2004. [Online]. Available: http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm. [Accessed 16 08 2016].

[16] "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2016. [Online]. Available: http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm. [Accessed 16 08 2016]

[17] "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2016. [Online]. Available: http://www.kdnuggets.com/polls/2014/analytics-datamining-data-science-methodology.html. [Accessed 16 08 2016].

[18] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crispdm.eu/reference-model/. [Accessed 16 08 2016].

[19] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crispdm.eu/business-understanding/. [Accessed 16 08 2016].

[20] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crispdm.eu/data-understanding/. [Accessed 16 08 2016].

[21] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crispdm.eu/data-preparation/. [Accessed 16 08 2016].

[22] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crispdm.eu/modelling/. [Accessed 16 08 2016].

[23] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crispdm.eu/evaluation/. [Accessed 16 08 2016]

[24] R. Kohavi, «A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,» de International Joint Conference on Artificial Intelligence (IJCAI), Montreal, 1995.