Employee Resignation Analysis

Group 4

Zhao Zicheng, Huang Xindi, Liu Zihao,
Wang Yifan, Yin Jiake

Content

Background

Employees’ attrition can lead to a lot of negative impacts on the company.

First of all, the employee's resignation may cause the work chaos of other employees in the team, or there may be a collective resignation phenomenon. This situation can lead to a loss of productivity, reputation and so on.

Secondly, when employees leave, the work is still there for someone else to do. At this time, there will be the cost of missing posts. In the case of fewer people, the per capita workload will increase, so they have to work overtime to solve the problem. However, human beings are not machines. They need to rest and work overtime all the time, which will only increase everyone's anxiety and slowly reduce their efficiency.

Thirdly, as employees leave, companies need to hire new employees again, so there will be resources to invest. After the new recruitment is in place, it is necessary to train for the job. After the training, new employees and current employees need to run in with each other. Before the learning curve formed, the efficiency of new employees cannot be too high.

To sum up, the resignation of employees really lead negative impacts to their companies. Therefore, we will build a model to estimate the influence of various factors on employee dismission, and then help the companies to minimize the number of dismission.

Affecting Factors

Since the loss of employees will bring certain losses to the enterprise, we need to find out which factors will lead to the loss of employees, so as to make plans for different factors, so that the enterprise can better retain employees. In fact, there are many reasons for employees to leave, including:

1. The salary of employees is low, and they think that their pay is not proportional to their income. So they choose to leave and find another job.

2. Employees are dissatisfied with the working environment of the company, or their relationship with colleagues is not well handled. Therefore, they are more depressed and unhappy at work.

3. Employees think that there is no room for promotion in their position, and they can't give full play to their work ability, so they choose to leave and change jobs.

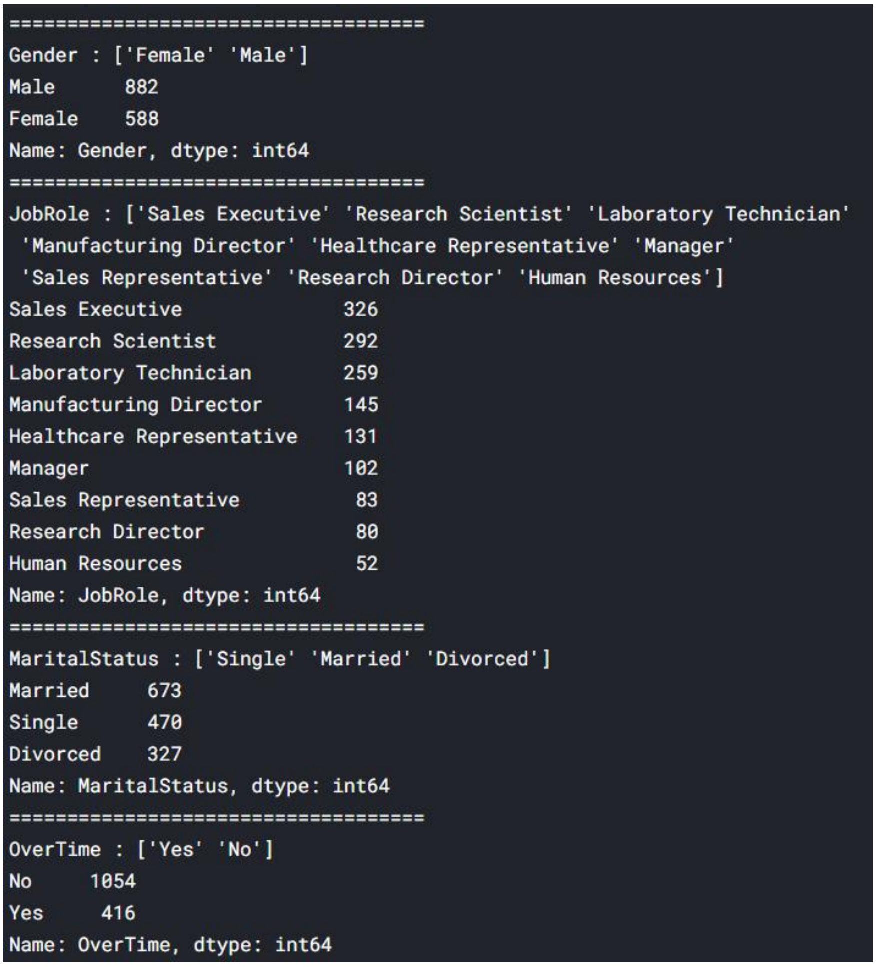
4. Employees are not interested in the tasks assigned by their department and want to find a job that is more suitable for them to give full play to their strengths.

5. Employees are dissatisfied with their leaders. Leaders are slow in thinking, weak in leadership, and fail to fulfill their responsibilities. Employees are not convinced and choose to leave.

6. Employees are required to travel all the year round in their positions. They spend less time with their families. So they want to change their jobs to accompany their families.

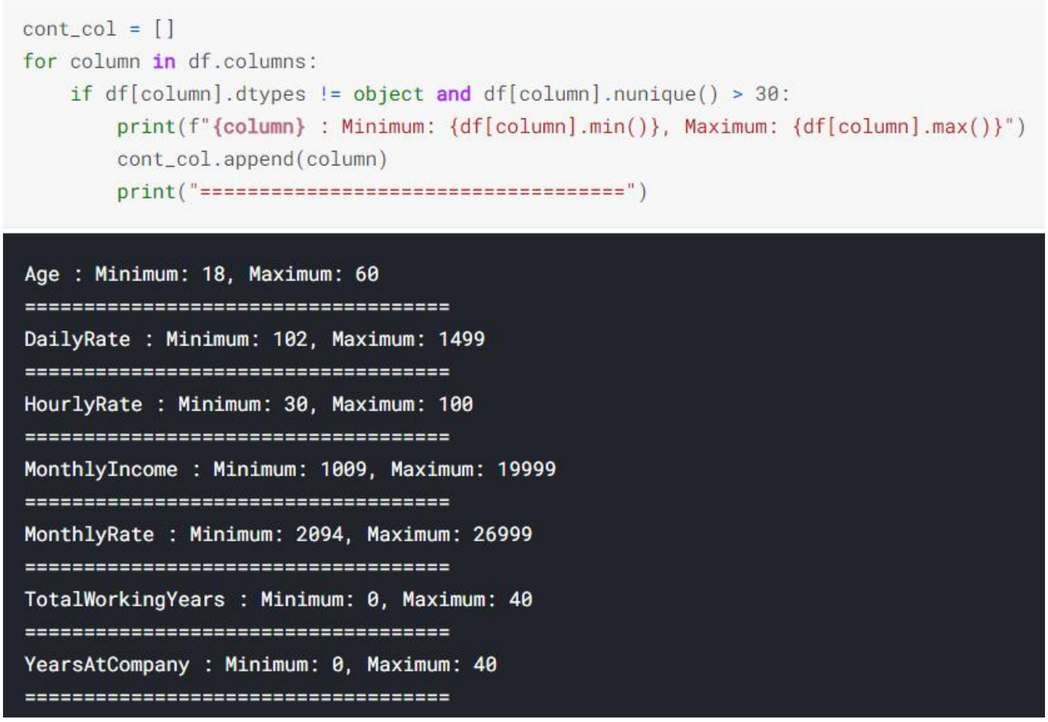
7. Because of old age or physical illness, the employee is not competent for the work and chooses to resign.

Data Analysis



```
====================================
Gender : ['Female' 'Male']
Male      882
Female    588
Name: Gender, dtype: int64
====================================
JobRole : ['Sales Executive' 'Research Scientist' 'Laboratory Technician'
 'Manufacturing Director' 'Healthcare Representative' 'Manager'
 'Sales Representative' 'Research Director' 'Human Resources']
Sales Executive              326
Research Scientist           292
Laboratory Technician        259
Manufacturing Director       145
Healthcare Representative    131
Manager                      102
Sales Representative          83
Research Director             80
Human Resources               52
Name: JobRole, dtype: int64
====================================
MaritalStatus : ['Single' 'Married' 'Divorced']
Married     673
Single      470
Divorced    327
Name: MaritalStatus, dtype: int64
====================================
OverTime : ['Yes' 'No']
No     1054
Yes     416
Name: OverTime, dtype: int64
```

Fig.1 Number of people under different conditions



```
cont_col = []
for column in df.columns:
    if df[column].dtypes != object and df[column].nunique() > 30:
        print(f"{column} : Minimum: {df[column].min()}, Maximum: {df[column].max()}")
        cont_col.append(column)
        print("====================================")
```

```
Age : Minimum: 18, Maximum: 60
====================================
DailyRate : Minimum: 102, Maximum: 1499
====================================
HourlyRate : Minimum: 30, Maximum: 100
====================================
MonthlyIncome : Minimum: 1009, Maximum: 19999
====================================
MonthlyRate : Minimum: 2094, Maximum: 26999
====================================
TotalWorkingYears : Minimum: 0, Maximum: 40
====================================
YearsAtCompany : Minimum: 0, Maximum: 40
====================================
```

Fig.2 Maximum and minimum values of each influencing factor

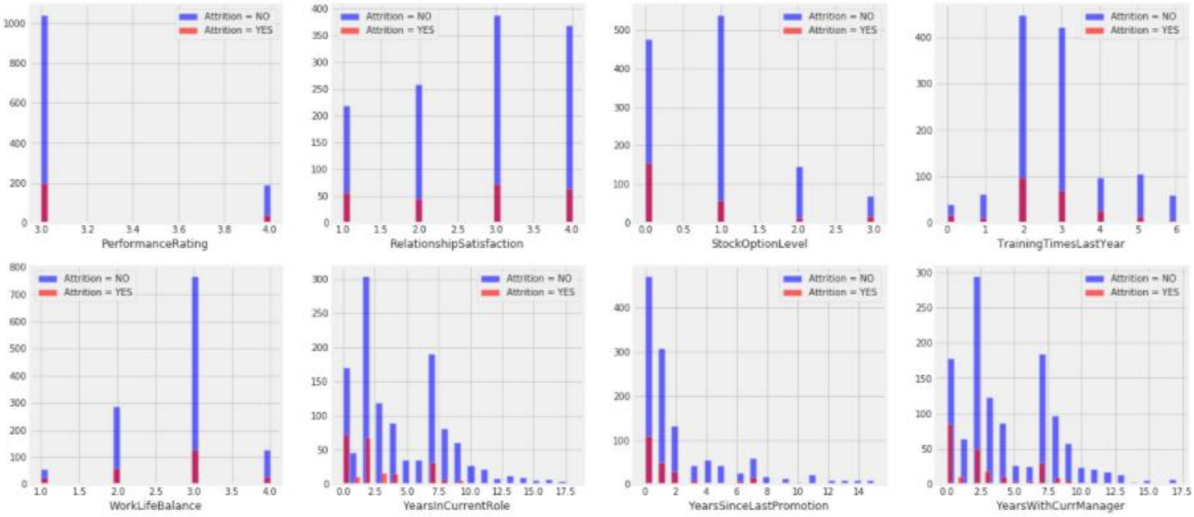


Fig. 3 Ratio of influential factors and the number of attrition

The histogram allows us to see the details of each of these factors. Therefore, we can find that employees with the following characteristics tend to have a higher turnover rate:

Employees with low performance ratings, employees with low relationship satisfaction, employees with a low stock option level, employees who have difficulty balancing their work life, the employee who hasn't been promoted for a long time.

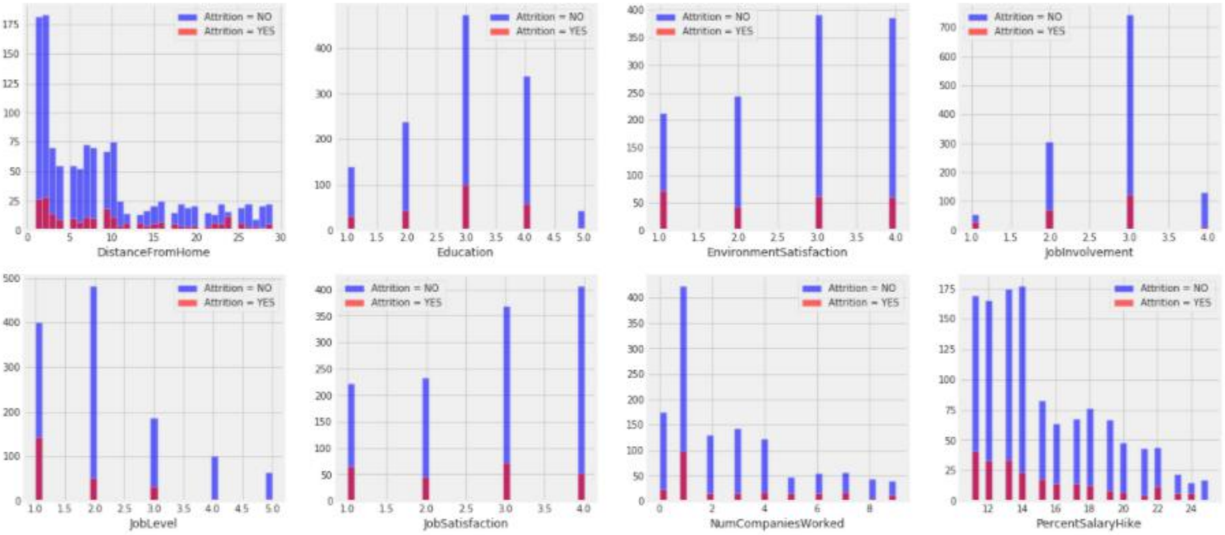


Fig.4 Ratio of influential factors and the number of attrition

From Fig.4, we can see that both Performance Rating, Relationship Satisfaction, Stock Option Level, Training Times Last Year and Work Life Balance can cause empl oyees' resignation. As their Years in Current Role, Years Since Last Promotion and Ye ars with Current Manager goes on, there are fewer employees choose to resign.

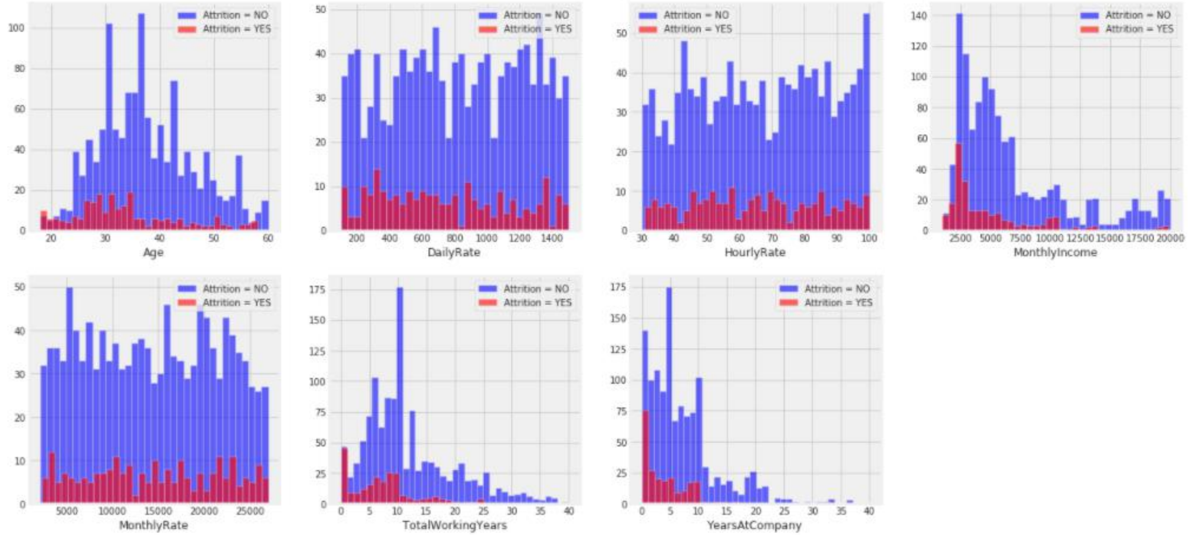
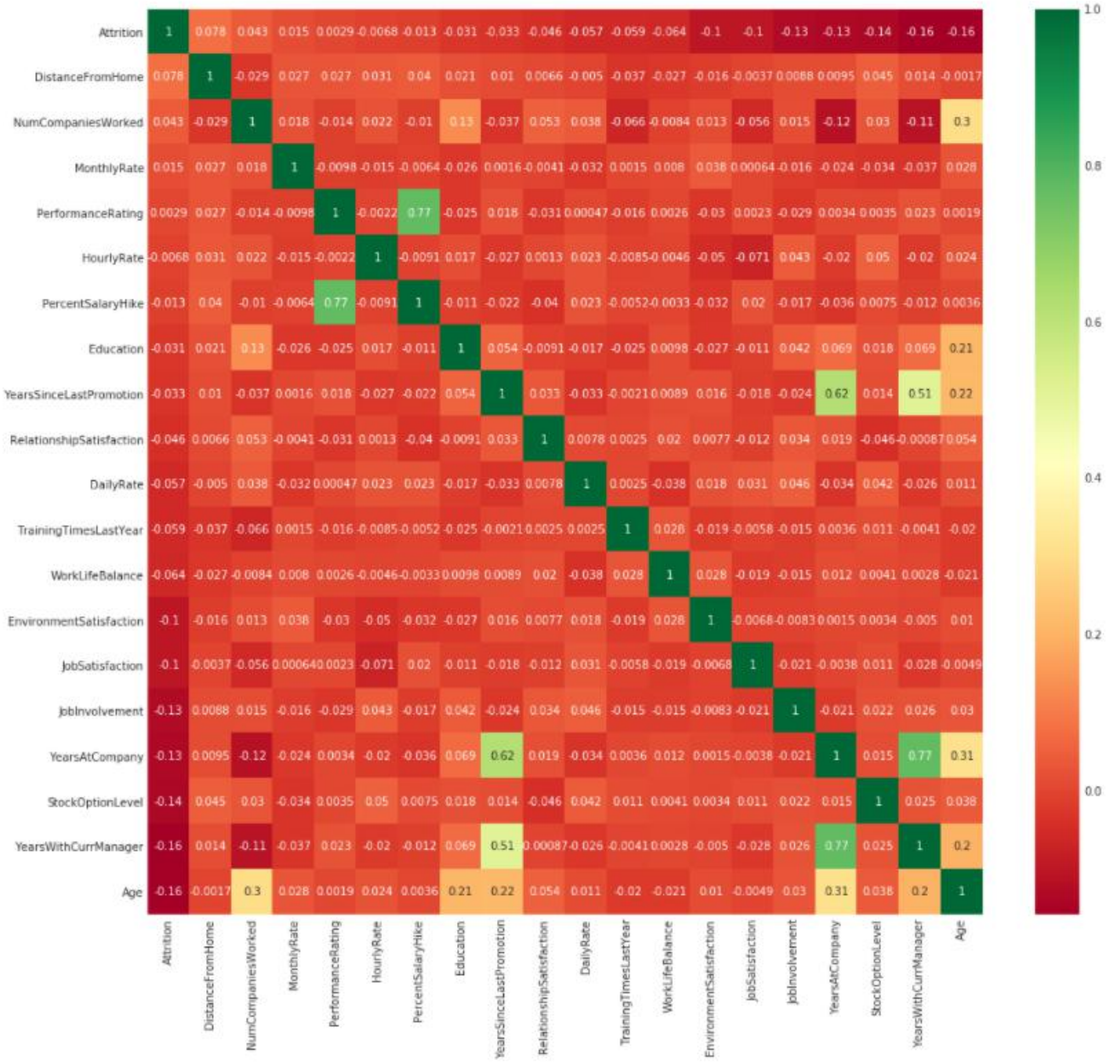


Fig.5 Ratio of influential factors and the number of attrution

From Fig.5, it is obvious that employees of different situations have different choices. Daily Rate, Hourly Rate and Monthly Rate of employees didn’t have specially influences on resignation. As we can see from the histograms, these factors are changing stably to the result.However, as Age, Monthly Income, Total Working Years and Year At Company increased, there are less employees quit from their work.



	Attrition	DistanceFromHome	NumCompaniesWorked	MonthlyRate	PerformanceRating	HourlyRate	PercentSalaryHike	Education	YearsSinceLastPromotion	RelationshipSatisfaction	DailyRate	TrainingTimesLastYear	WorkLifeBalance	EnvironmentSatisfaction	JobSatisfaction	JobInvolvement	YearsAtCompany	StockOptionLevel	YearsWithCurrManager	Age
Attrition	1	0.078	0.043	0.015	0.0029	-0.0068	-0.013	-0.031	-0.033	-0.046	-0.057	-0.059	-0.064	-0.1	-0.1	-0.13	-0.13	-0.14	-0.16	-0.16
DistanceFromHome	0.078	1	-0.029	0.027	0.027	0.031	0.04	0.021	0.01	0.0066	-0.005	-0.037	-0.027	-0.016	-0.0037	0.0088	0.0095	0.045	0.014	-0.0017
NumCompaniesWorked	0.043	-0.029	1	0.018	-0.014	0.022	-0.01	0.13	-0.037	0.053	0.038	-0.066	-0.0084	0.013	-0.056	0.015	-0.12	0.03	-0.11	0.3
MonthlyRate	0.015	0.027	0.018	1	-0.0098	-0.015	-0.0064	-0.026	0.0016	-0.0041	-0.032	0.0015	0.008	0.038	0.00064	-0.016	-0.024	-0.034	-0.037	0.028
PerformanceRating	0.0029	0.027	-0.014	-0.0098	1	-0.0022	0.77	-0.025	0.018	-0.031	0.00047	-0.016	0.0026	-0.03	0.0023	-0.029	0.0034	0.0035	0.023	0.0019
HourlyRate	-0.0068	0.031	0.022	-0.015	-0.0022	1	-0.0091	0.017	-0.027	0.0013	0.023	-0.0085	-0.0046	-0.05	-0.071	0.043	-0.02	0.05	-0.02	0.024
PercentSalaryHike	-0.013	0.04	-0.01	-0.0064	0.77	-0.0091	1	-0.011	-0.022	-0.04	0.023	-0.0052	-0.0033	-0.032	0.02	-0.017	-0.036	0.0075	-0.012	0.0036
Education	-0.031	0.021	0.13	-0.026	-0.025	0.017	-0.011	1	0.054	-0.0091	-0.017	-0.025	0.0098	-0.027	-0.011	0.042	0.069	0.018	0.069	0.21
YearsSinceLastPromotion	-0.033	0.01	-0.037	0.0016	0.018	-0.027	-0.022	0.054	1	0.033	-0.033	-0.0021	0.0089	0.016	-0.018	-0.024	0.62	0.014	0.51	0.22
RelationshipSatisfaction	-0.046	0.0066	0.053	-0.0041	-0.031	0.0013	-0.04	-0.0091	0.033	1	0.0078	0.0025	0.02	0.0077	-0.012	0.034	0.019	-0.046	-0.00087	0.054
DailyRate	-0.057	-0.005	0.038	-0.032	0.00047	0.023	0.023	-0.017	-0.033	0.0078	1	0.0025	-0.038	0.018	0.031	0.046	-0.034	0.042	-0.026	0.011
TrainingTimesLastYear	-0.059	-0.037	-0.066	0.0015	-0.016	-0.0085	-0.0052	-0.025	-0.0021	0.0025	0.0025	1	0.028	-0.019	-0.0058	-0.015	0.0036	0.011	-0.0041	-0.02
WorkLifeBalance	-0.064	-0.027	-0.0084	0.008	0.0026	-0.0046	-0.0033	0.0098	0.0089	0.02	-0.038	0.028	1	0.028	-0.019	-0.015	0.012	0.0041	0.0028	-0.021
EnvironmentSatisfaction	-0.1	-0.016	0.013	0.038	-0.03	-0.05	-0.032	-0.027	0.016	0.0077	0.018	-0.019	0.028	1	-0.0068	-0.0083	0.0015	0.0034	-0.005	0.01
JobSatisfaction	-0.1	-0.0037	-0.056	0.00064	0.0023	-0.071	0.02	-0.011	-0.018	-0.012	0.031	-0.0058	-0.019	-0.0068	1	-0.021	-0.0038	0.011	-0.028	-0.0049
JobInvolvement	-0.13	0.0088	0.015	-0.016	-0.029	0.043	-0.017	0.042	-0.024	0.034	0.046	-0.015	-0.015	-0.0083	-0.021	1	-0.021	0.022	0.026	0.03
YearsAtCompany	-0.13	0.0095	-0.12	-0.024	0.0034	-0.02	-0.036	0.069	0.62	0.019	-0.034	0.0036	0.012	0.0015	-0.0038	-0.021	1	0.015	0.77	0.31
StockOptionLevel	-0.14	0.045	0.03	-0.034	0.0035	0.05	0.0075	0.018	0.014	-0.046	0.042	0.011	0.0041	0.0034	0.011	0.022	0.015	1	0.025	0.038
YearsWithCurrManager	-0.16	0.014	-0.11	-0.037	0.023	-0.02	-0.012	0.069	0.51	-0.00087	-0.026	-0.0041	0.0028	-0.005	-0.028	0.026	0.77	0.025	1	0.2
Age	-0.16	-0.0017	0.3	0.028	0.0019	0.024	0.0036	0.21	0.22	0.054	0.011	-0.02	-0.021	0.01	-0.0049	0.03	0.31	0.038	0.2	1

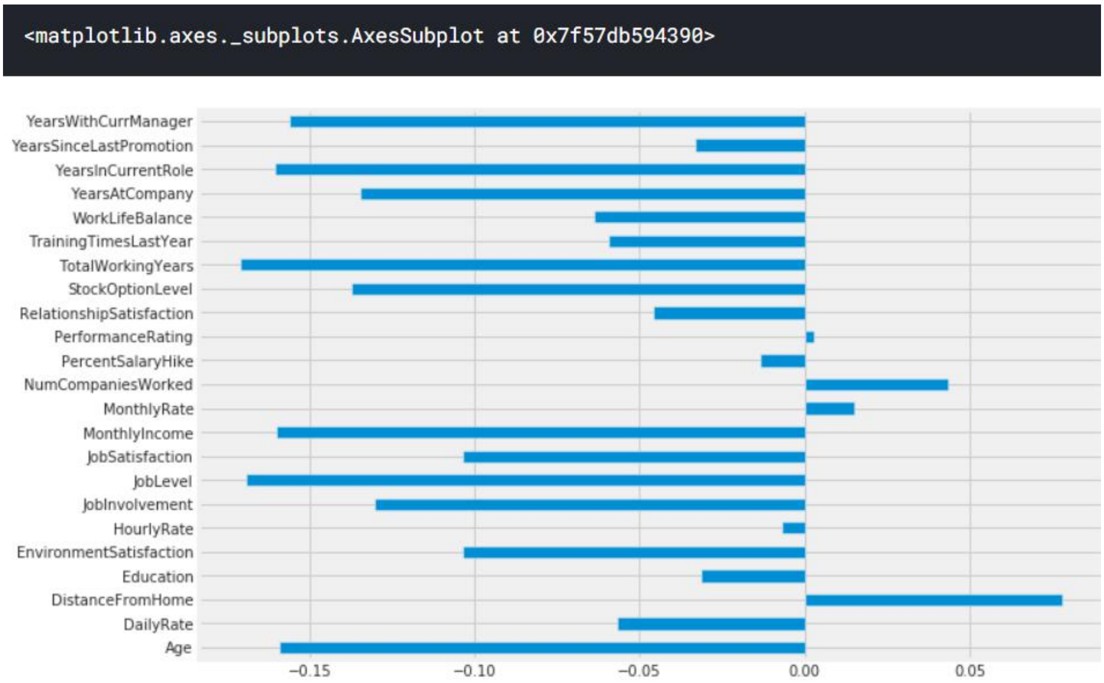
Fig.6 The Correlation Matrix

According to the correlation matrix, we can see the degree of correlation between each two influencing factors.

There's a strong correlation between how long you've worked with your current manager and how long it's been since your last promotion. Therefore, the ability and working style of the manager is closely related to the promotion of the employee. If the employee's manager is not good at identifying the talent and giving the employee promotion, the employee may leave the company.

There's a strong correlation between years at the company and the time since your last promotion. So there are employees who are constantly promoted, who are loyal, and there are new employees who are not yet promoted. But if an employee has been with the company for a long time without a promotion, they may leave.

There is a strong correlation between employee performance and salary increase percentage. The better the employee's performance, the bigger the increase will be. However, if the company does not give an appropriate raise to the outstanding employee, the employee may leave the company.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f57db594390>
```

Fig. 7 Histogram of positive and negative correlation of various factors

This chart shows the positive and negative correlation between various influencing factors and resignation. We can clearly see that most factors are negatively correlated, except the performance rating, the number of worked companies, monthly rate and distance from home. Among them, the number of companies that employees have worked in before and the distance from home have a relatively obvious positive correlation with resignation.

Summary:

The workers with low Job Level, Monthly Income, Year At Company, and Total Working Years are more likely to quit their jobs.

Business Travel: The workers who travel a lot are more likely to quit than other employees.

Department: The worker in Research & Development are more likely to stay than the workers on other department.

Education Field : The workers with Human Resources and Technical Degree are more likely to quit than employees from other fields of educations.

Gender: The Male are more likely to quit.

Job Role: The workers in Laboratory Technician, Sales Representative, and Human Resources are more likely to quit the workers in other positions.

Marital Status: The workers who are bachelor are more likely to quit than Married, and Divorced.

Over Time: The workers who work more hours are likely to quit then others.

Model Analysis

For each model we have used some parameters for evaluation. Initially, we explain the TP、FP、FN、TN. TP-true positive means predicts the positive class correctly as a positive class number. FP-false positive example means predicting a negative class error as a positive class number. FN-false negative example where the positive class is incorrectly predicted as the number of negative classes. TN-true negative example means the negative class is correctly predicted as the number of negative classes.

Real situation	Predict outcomes

	Positive example	Negative example
Positive example	TP	FN
Negative example	FP	TN

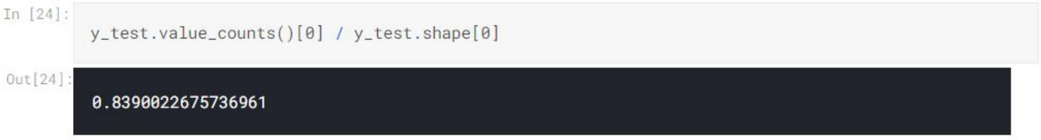
Fig.8 TP/FP/FN/TN

Also, we use the ROC curve. The full name of the ROC curve is receiver operating characteristic which use the goodness of the sort itself that determines the generalization performance of the models. Its vertical axis represents the true case rate of the model ($TPR = \frac{TP}{TP+FN}$) and the horizontal axis represents the false positive case rate of the model($FPR = \frac{FP}{TN+FP}$). When comparing models, if the ROC curve of one model is completely encapsulated by the curve of the other model, it can be asserted that the latter outperforms the former, but not if the ROC curves of the two models cross. Similarly, if two models are to be compared, a more reasonable criterion is to compare the area under the ROC curve.

We compare not only the ROC curves between the different models, but also the PR curves. The full name of PR curve is precision and recall curve which means there are two parameters in PR curve. Precision is the number of positive samples correctly classified as a percentage of the number of samples determined to be positive by the classifier. ($Precision = \frac{TP}{TP+FP}$) Recall is the number of correctly classified positive samples as a proportion of the number of true positive samples. ($\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$)

What defines success?

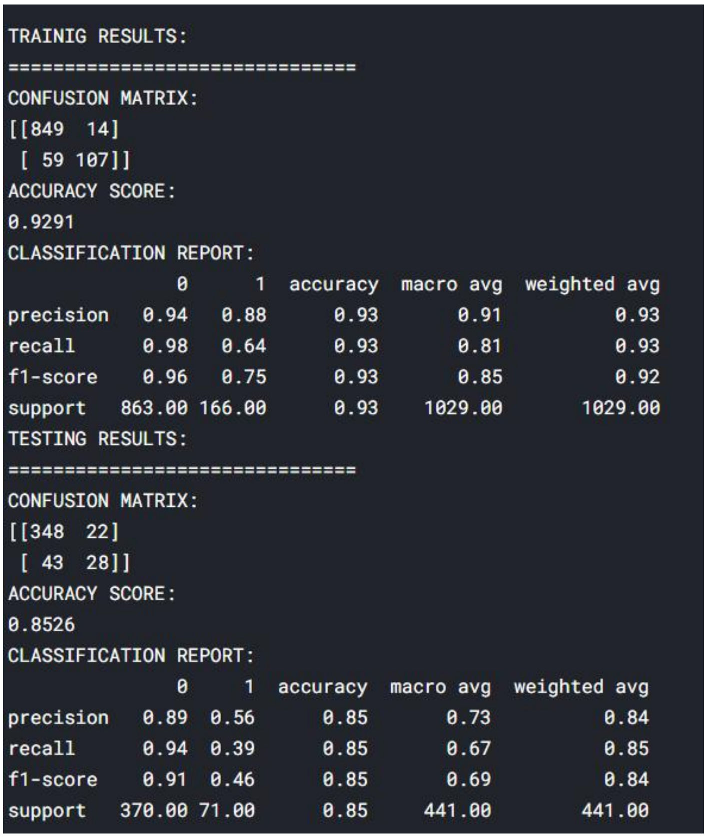
We have an imbalanced data, so if we predict that all our employees will stay, we'll have an accuracy of 83.90%.



```
In [24]: y_test.value_counts()[0] / y_test.shape[0]

Out[24]: 0.8390022675736961
```

Fig.9 Accuracy of results



```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[849  14]
 [ 59 107]]
ACCURACY SCORE:
0.9291
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision   0.94   0.88      0.93       0.91          0.93
recall      0.98   0.64      0.93       0.81          0.93
f1-score    0.96   0.75      0.93       0.85          0.92
support   863.00 166.00      0.93    1029.00       1029.00
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[348  22]
 [ 43  28]]
ACCURACY SCORE:
0.8526
CLASSIFICATION REPORT:
               0     1  accuracy  macro avg  weighted avg
precision   0.89  0.56      0.85       0.73          0.84
recall      0.94  0.39      0.85       0.67          0.85
f1-score    0.91  0.46      0.85       0.69          0.84
support   370.00 71.00      0.85     441.00        441.00
```

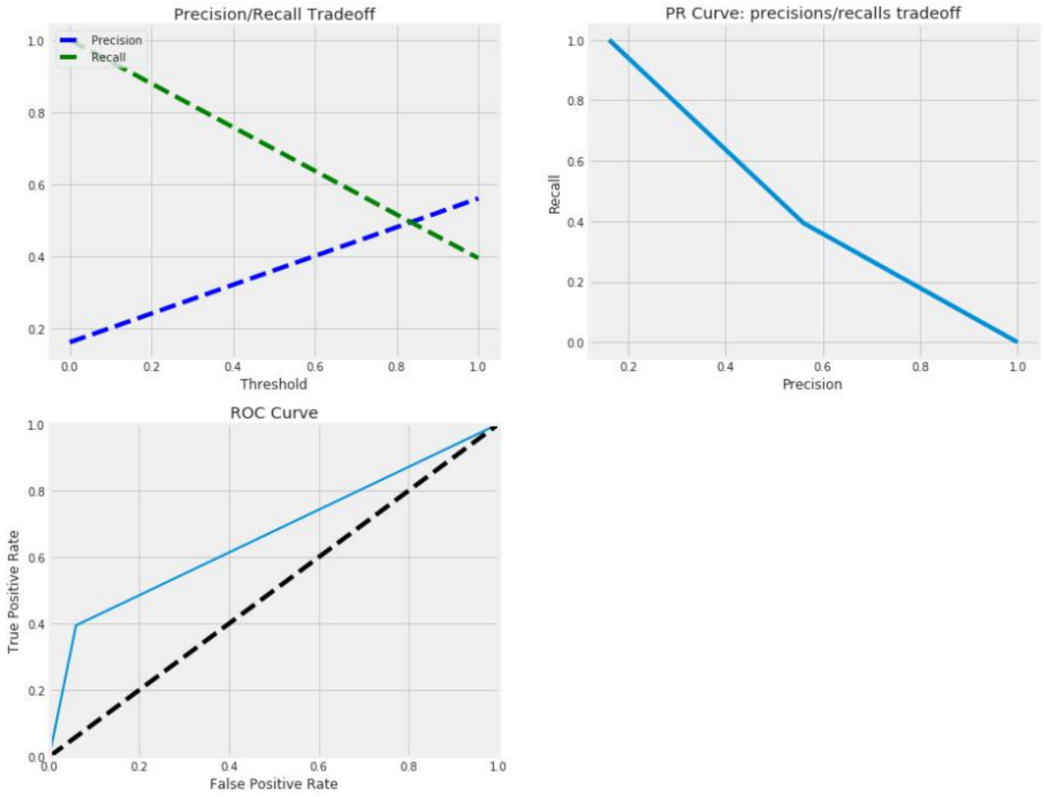
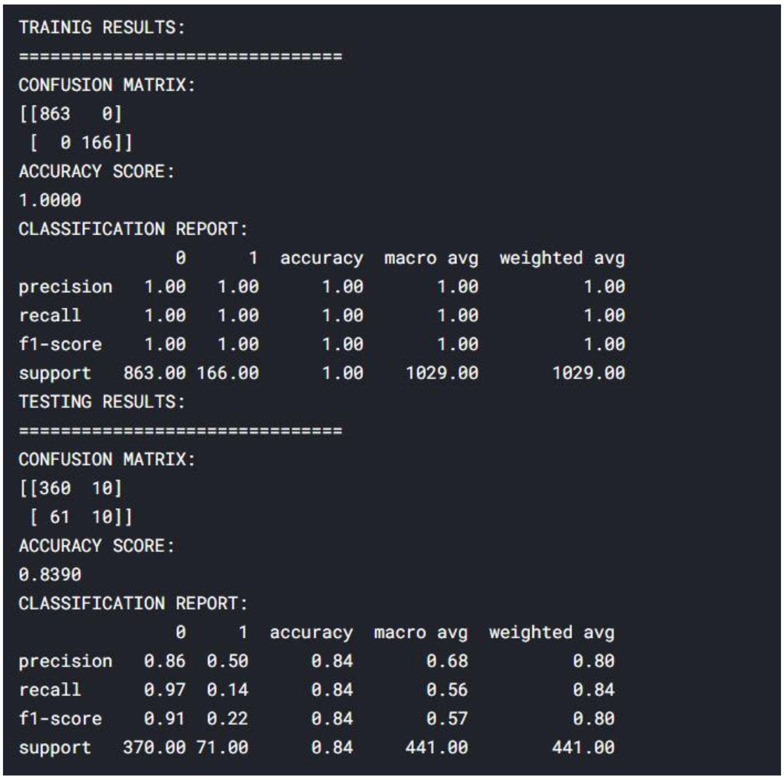


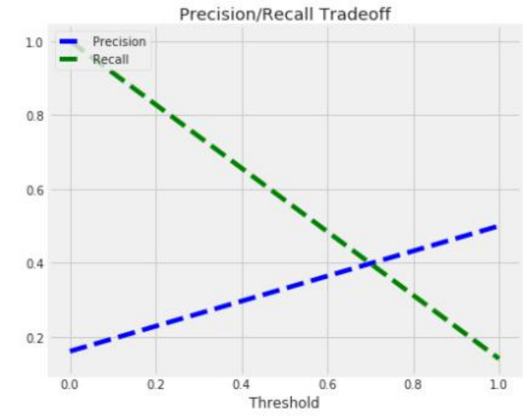
Fig.10 The regression model

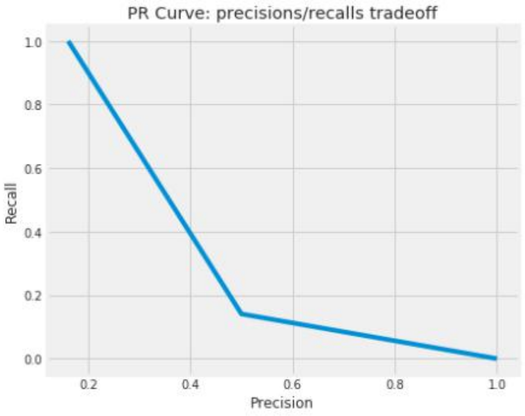
The regression model

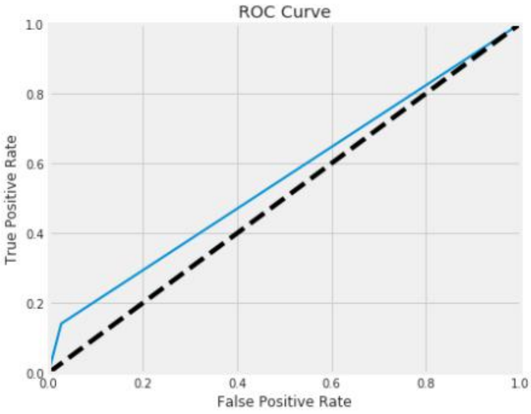
Although called regression, it is actually a classification model and is often used in dichotomy. Logistic Regression is favored by the industry for its simplicity, parallelization, and strong interpretation.



```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[863   0]
 [  0 166]]
ACCURACY SCORE:
1.0000
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision   1.00   1.00      1.00       1.00          1.00
recall      1.00   1.00      1.00       1.00          1.00
f1-score    1.00   1.00      1.00       1.00          1.00
support   863.00 166.00      1.00    1029.00       1029.00
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[360  10]
 [ 61  10]]
ACCURACY SCORE:
0.8390
CLASSIFICATION REPORT:
               0     1  accuracy  macro avg  weighted avg
precision   0.86  0.50      0.84       0.68          0.80
recall      0.97  0.14      0.84       0.56          0.84
f1-score    0.91  0.22      0.84       0.57          0.80
support   370.00 71.00      0.84     441.00        441.00
```







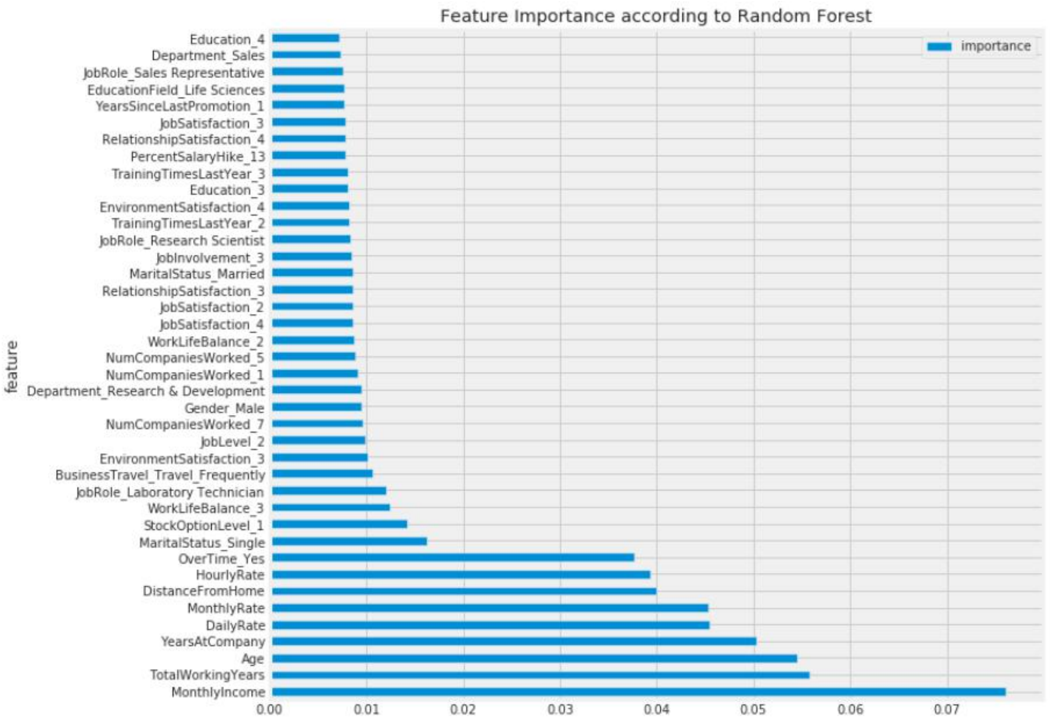
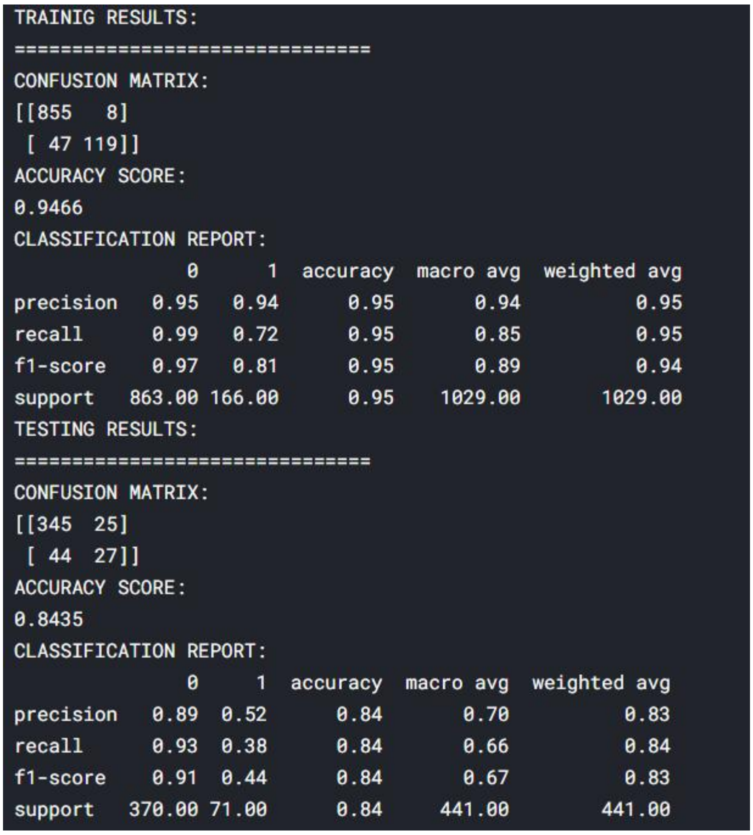


Fig.11 The Random Forest

Random forest classifier

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.



```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[855   8]
 [ 47 119]]
ACCURACY SCORE:
0.9466
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision   0.95   0.94      0.95       0.94          0.95
recall      0.99   0.72      0.95       0.85          0.95
f1-score    0.97   0.81      0.95       0.89          0.94
support   863.00 166.00      0.95    1029.00       1029.00
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[345  25]
 [ 44  27]]
ACCURACY SCORE:
0.8435
CLASSIFICATION REPORT:
               0     1  accuracy  macro avg  weighted avg
precision   0.89  0.52      0.84       0.70          0.83
recall      0.93  0.38      0.84       0.66          0.84
f1-score    0.91  0.44      0.84       0.67          0.83
support   370.00 71.00      0.84     441.00        441.00
```

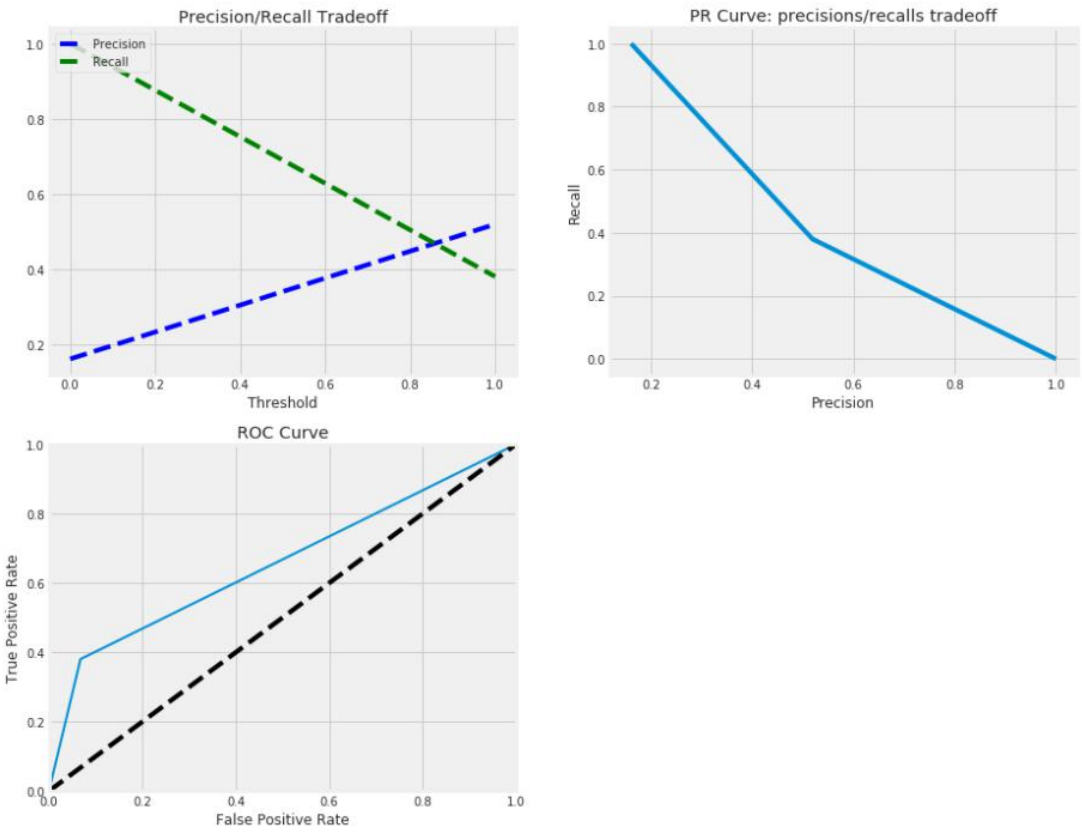
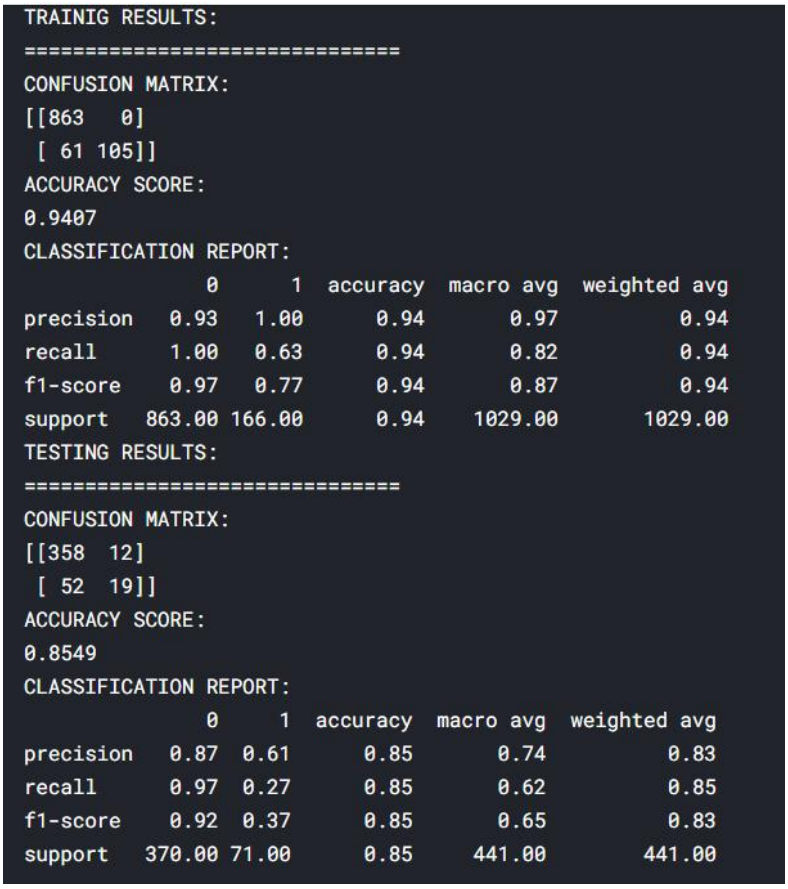


Fig.12 The SVM

SVM

Support vector machines construct hyperplanes or sets of hyperplanes in high-dimensional or infinite dimensional Spaces, which can be used for classification, regression, or other tasks.Intuitively speaking, the farther the classification boundary is from the nearest training data point, the better, because this can reduce the generalization error of the classifier.



```
TRAINIG RESULTS:
================================
CONFUSION MATRIX:
[[863   0]
 [ 61 105]]
ACCURACY SCORE:
0.9407
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision   0.93   1.00      0.94       0.97          0.94
recall      1.00   0.63      0.94       0.82          0.94
f1-score    0.97   0.77      0.94       0.87          0.94
support   863.00 166.00      0.94    1029.00       1029.00
TESTING RESULTS:
================================
CONFUSION MATRIX:
[[358  12]
 [ 52  19]]
ACCURACY SCORE:
0.8549
CLASSIFICATION REPORT:
               0     1  accuracy  macro avg  weighted avg
precision   0.87  0.61      0.85       0.74          0.83
recall      0.97  0.27      0.85       0.62          0.85
f1-score    0.92  0.37      0.85       0.65          0.83
support   370.00 71.00      0.85     441.00        441.00
```

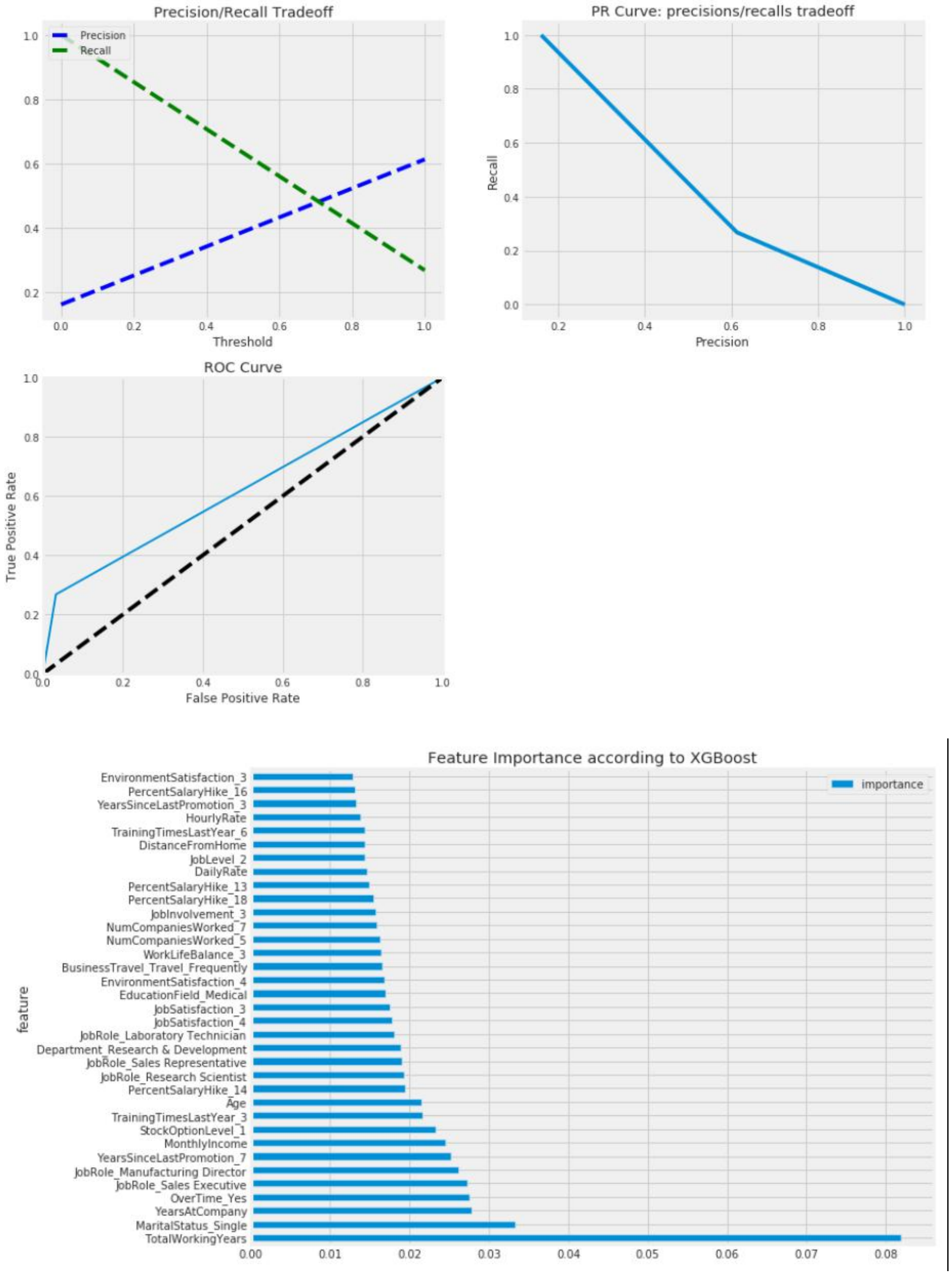
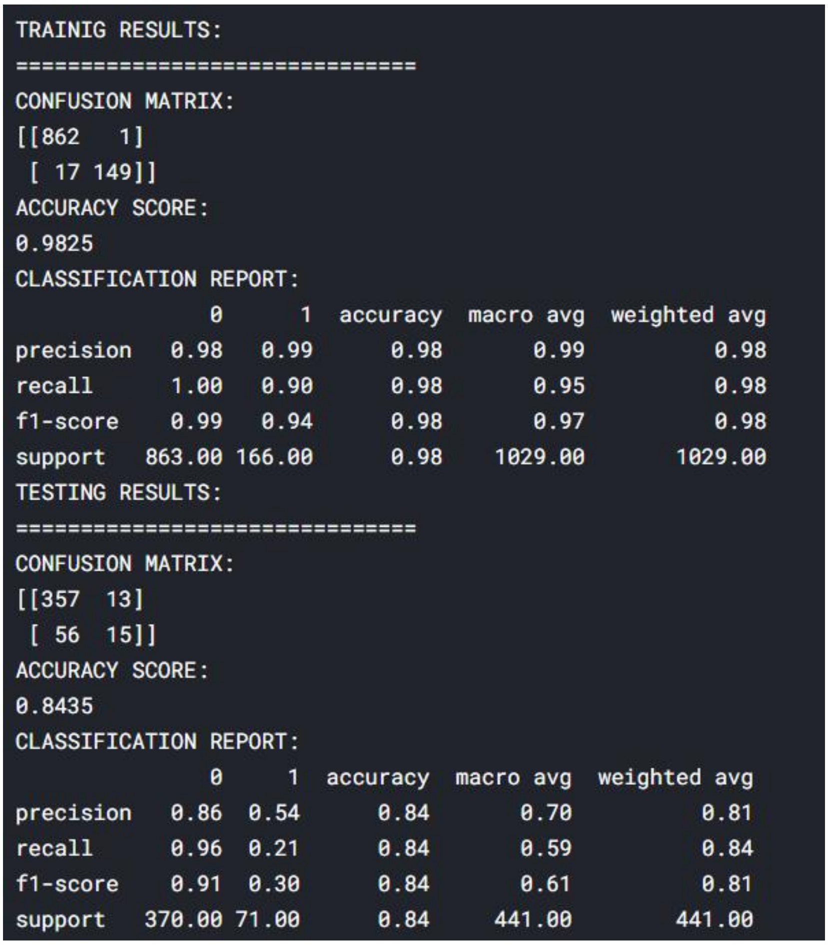


Fig.13 The XGBOOST

XGBOOST CLASSIFIER

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.



```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[862   1]
 [ 17 149]]
ACCURACY SCORE:
0.9825
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision   0.98   0.99      0.98       0.99          0.98
recall      1.00   0.90      0.98       0.95          0.98
f1-score    0.99   0.94      0.98       0.97          0.98
support   863.00 166.00      0.98    1029.00       1029.00
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[357  13]
 [ 56  15]]
ACCURACY SCORE:
0.8435
CLASSIFICATION REPORT:
               0     1  accuracy  macro avg  weighted avg
precision   0.86  0.54      0.84       0.70          0.81
recall      0.96  0.21      0.84       0.59          0.84
f1-score    0.91  0.30      0.84       0.61          0.81
support   370.00 71.00      0.84     441.00        441.00
```

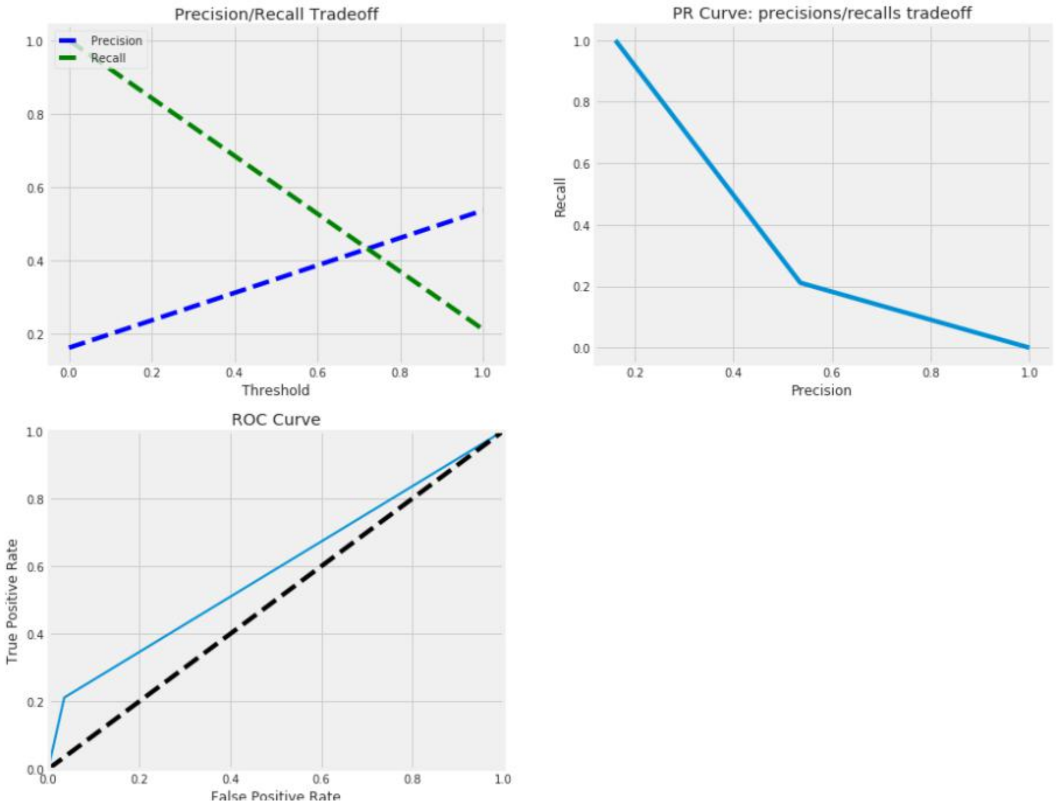
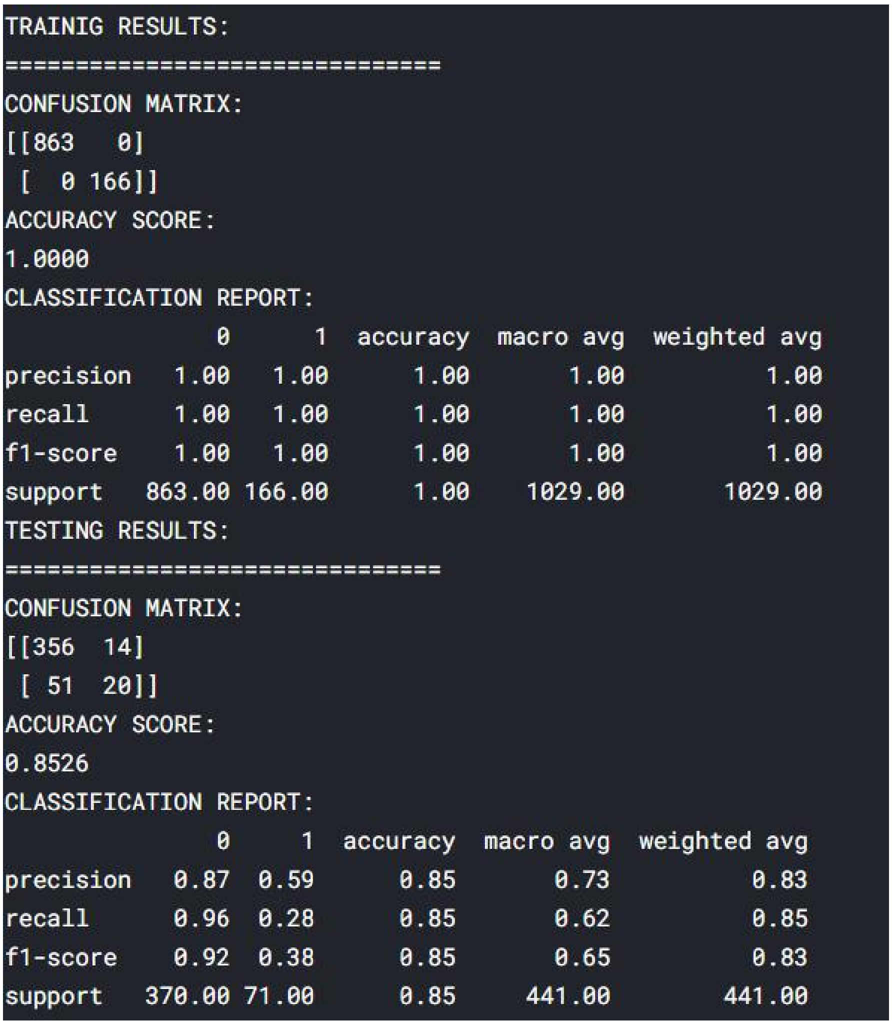


Fig.14 The CATBOOST

CATBOOST

CatBoost is an algorithm for gradient boosting on decision trees. Developed by Yandex researchers and engineers, it is the successor of the MatrixNet algorithm that is widely used within the company for ranking tasks, forecasting and making recommendations.



```
TRAINIG RESULTS:
================================
CONFUSION MATRIX:
[[863   0]
 [  0 166]]
ACCURACY SCORE:
1.0000
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision   1.00   1.00      1.00       1.00          1.00
recall      1.00   1.00      1.00       1.00          1.00
f1-score    1.00   1.00      1.00       1.00          1.00
support   863.00 166.00      1.00    1029.00       1029.00
TESTING RESULTS:
================================
CONFUSION MATRIX:
[[356  14]
 [ 51  20]]
ACCURACY SCORE:
0.8526
CLASSIFICATION REPORT:
               0     1  accuracy  macro avg  weighted avg
precision   0.87  0.59      0.85       0.73          0.83
recall      0.96  0.28      0.85       0.62          0.85
f1-score    0.92  0.38      0.85       0.65          0.83
support   370.00 71.00      0.85     441.00        441.00
```

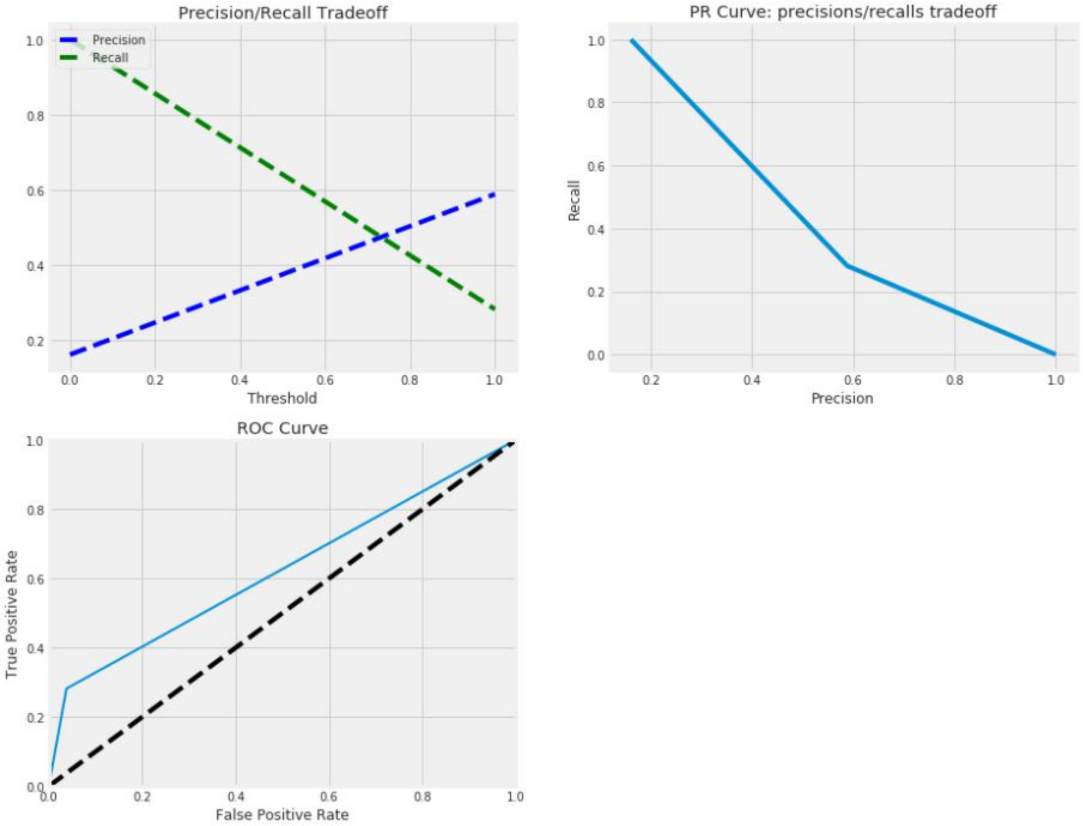
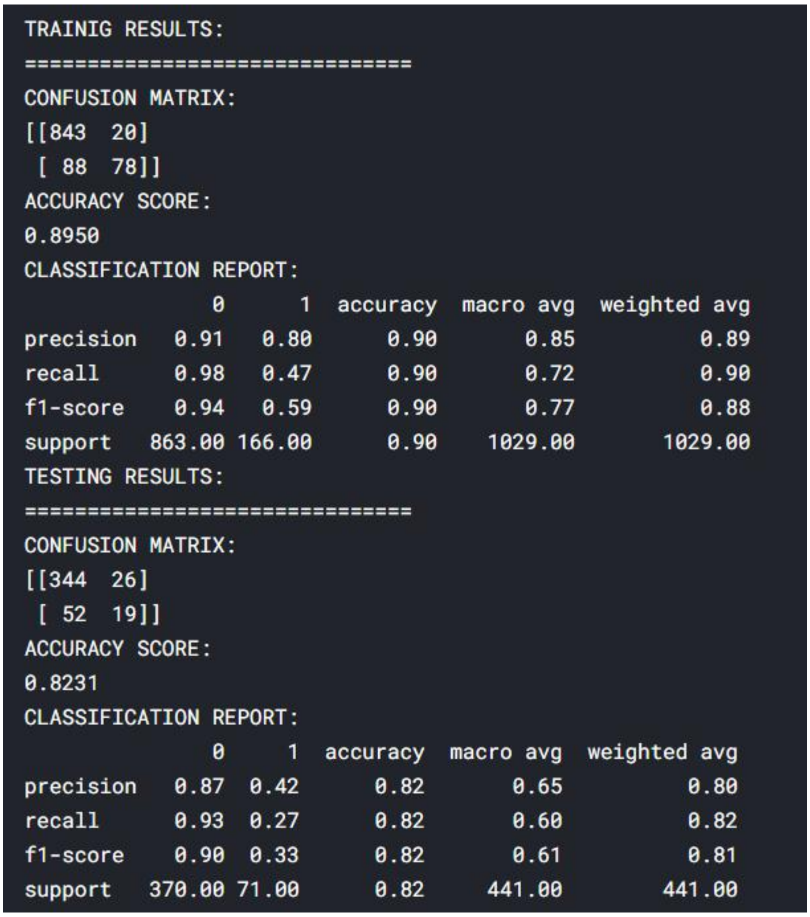


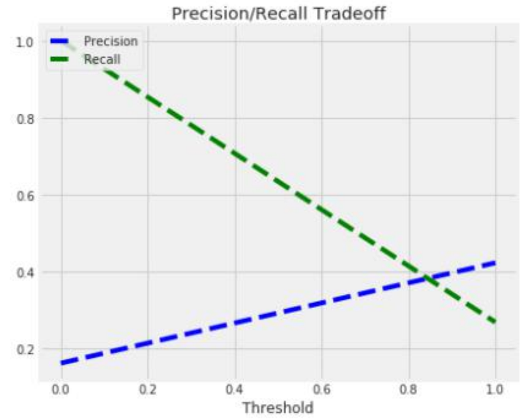
Fig.15 The LIGHTGBM

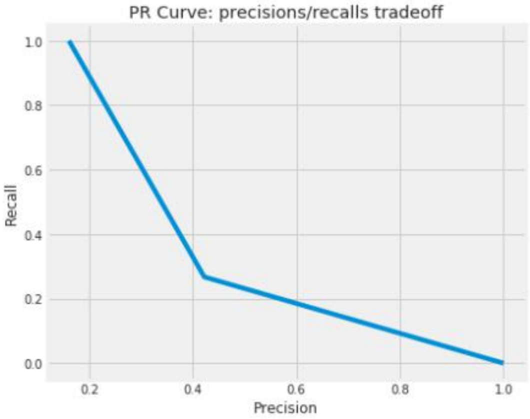
LIGHTGBM

The main idea is to use the weak classifier (decision tree) iterative training to get the optimal model, which has the advantages of good training effect and not easy to overfit. It is commonly used for multi-classification, click-through rate prediction, search sorting and other tasks; It supports efficient parallel training, and has the advantages of faster training speed, lower memory consumption, better accuracy, distributed support and rapid processing of massive data.



```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[843  20]
 [ 88  78]]
ACCURACY SCORE:
0.8950
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    0.91   0.80      0.90       0.85          0.89
recall       0.98   0.47      0.90       0.72          0.90
f1-score     0.94   0.59      0.90       0.77          0.88
support    863.00 166.00      0.90    1029.00       1029.00
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[344  26]
 [ 52  19]]
ACCURACY SCORE:
0.8231
CLASSIFICATION REPORT:
                0     1  accuracy  macro avg  weighted avg
precision    0.87  0.42      0.82       0.65          0.80
recall       0.93  0.27      0.82       0.60          0.82
f1-score     0.90  0.33      0.82       0.61          0.81
support    370.00 71.00      0.82     441.00        441.00
```





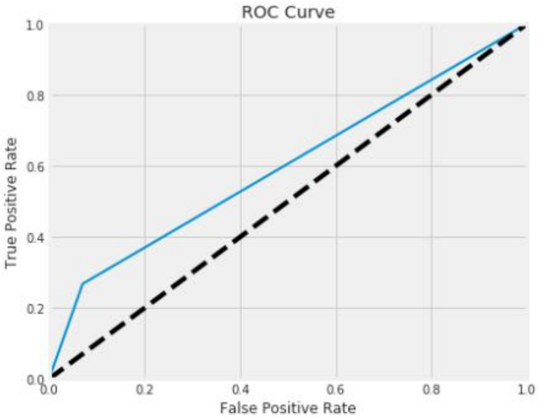
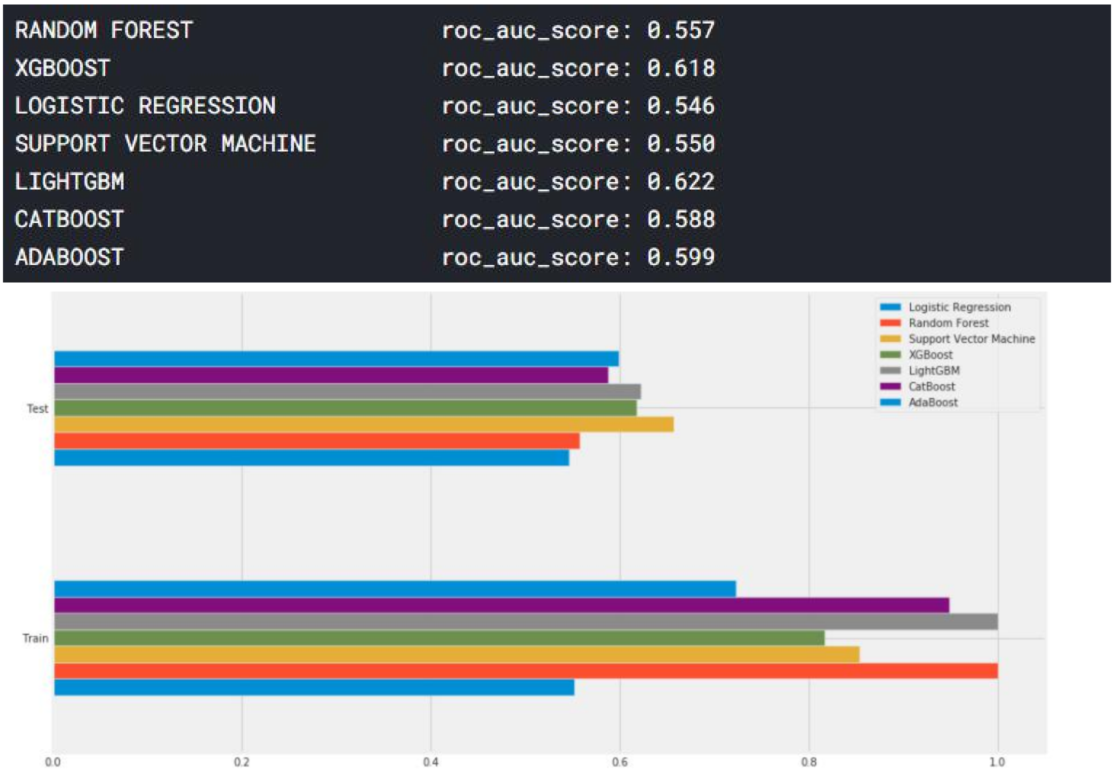


Fig.16 The ADABOOST

ADABOOST

AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

Comparison of the final results



```
RANDOM FOREST                  roc_auc_score: 0.557
XGBOOST                        roc_auc_score: 0.618
LOGISTIC REGRESSION            roc_auc_score: 0.546
SUPPORT VECTOR MACHINE         roc_auc_score: 0.550
LIGHTGBM                       roc_auc_score: 0.622
CATBOOST                       roc_auc_score: 0.588
ADABOOST                       roc_auc_score: 0.599
```

Fig.17 The Comparison of the Final Results

Comparing models' performances, we made a comprehensive comparison between the train and test processes. We can still observe that *LIGHTGBM* model has the best score.

Summary

Models: GBM model

Recommendations: Companies should give young employees more chances to show their ability. The company should give appropriate raises or promotions to the employees who perform well.

In addition, company management should focus on improving work efficiency rather than making employees work longer hours.