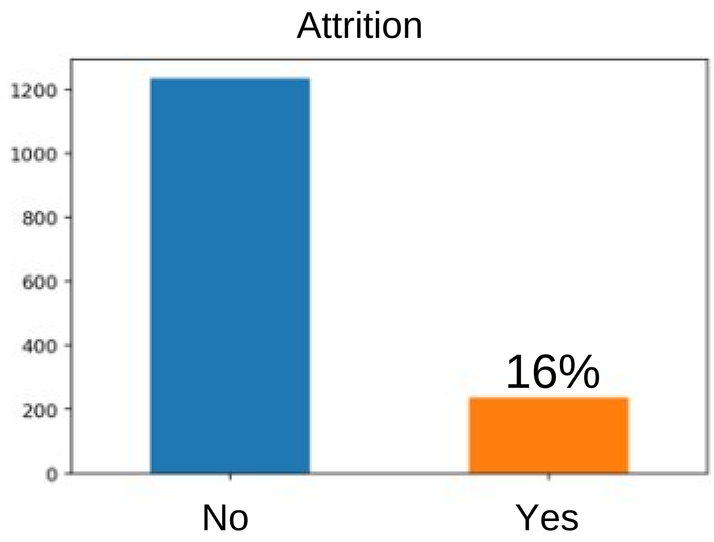
EMPLOYEE ATTRITION AND HUMAN RESOURCE (HR) ANALYTICS

BY SHIVANSH TRIPATHI

Dataset

There are 1,470 observations with 34 characteristics of employees (Features) and 1 target variable (attrition 0 for ”no” or 1 for ”yes”)

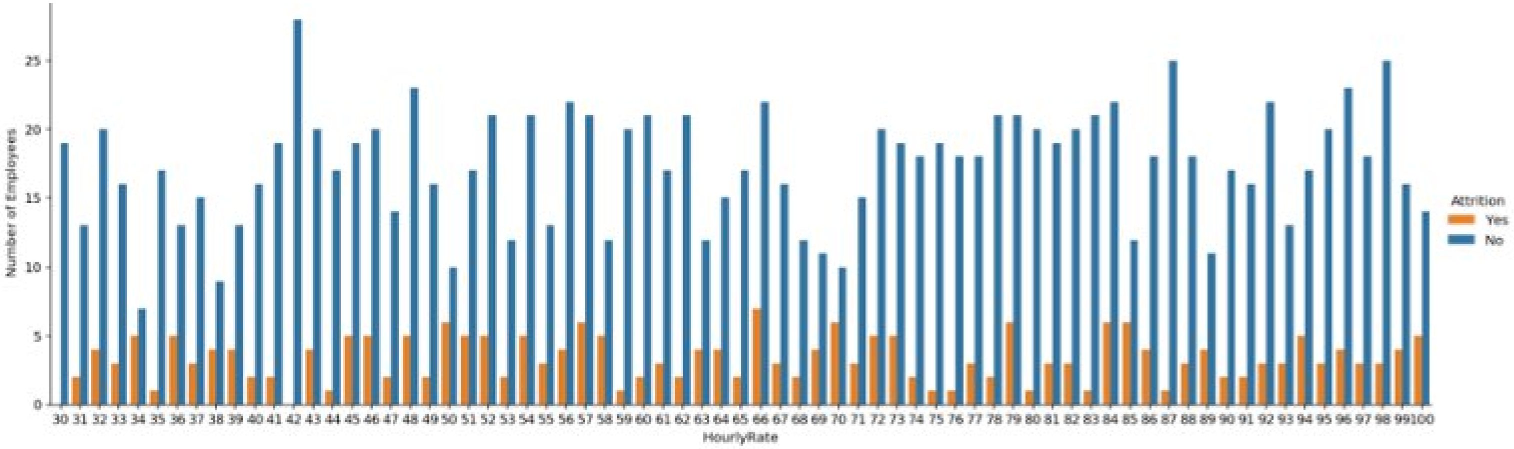


Features: ['Age', 'BusinessTravel', 'DailyRate', 'Department', 'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount', 'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager']

Exploratory Data Analysis (EDA)

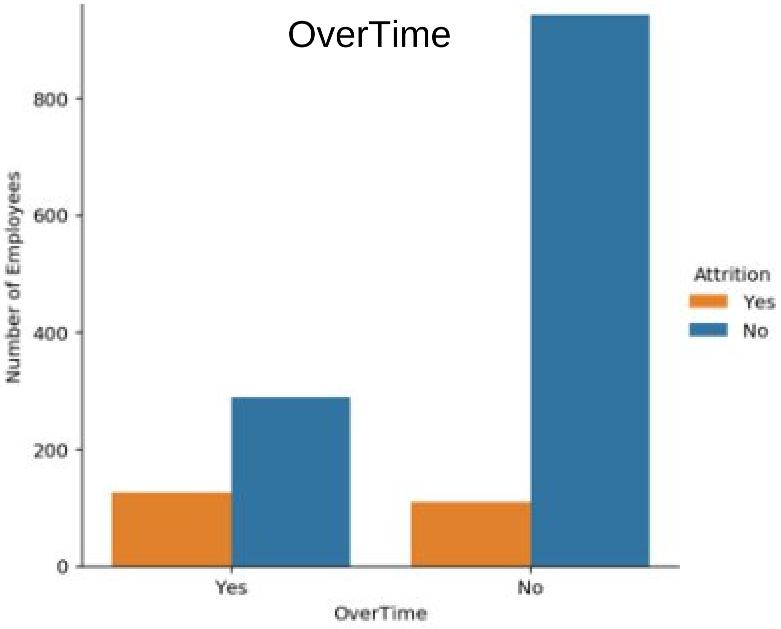
Attrition seems to happen at every level regardless of employee hourly rate

HourlyRate



Overtime could be a key factor leading to attrition

OverTime



Iteration 1: Baseline Model Performance

Below 10 models have been evaluated:

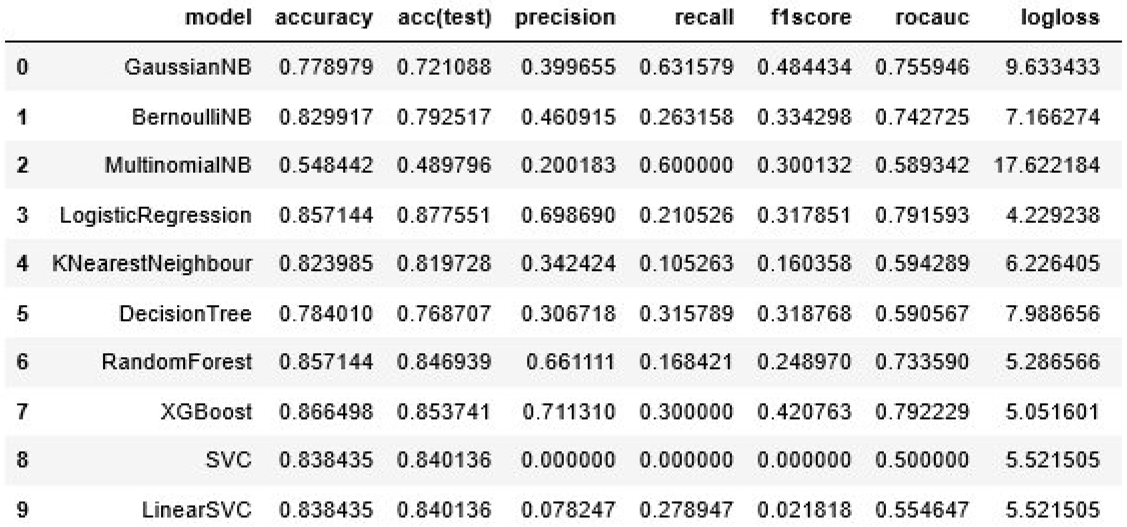
· Gaussian Naive Bayes
· Bernoulli Naive Bayes
· Multinomial Naive Bayes
· Logistic Regression
· K Nearest Neighbour
· Decision Tree Classifier
· Random Forest Classifier
· Extreme Gradient Boost (XGBoost)
· Support Vector Classification (SVC)
· Linear SVC

The performance metrics used in the evaluation are:

· Accuracy Score: proportion of correct predictions out of the whole dataset. Be careful when the target class is imbalance, for example, if a model predicts all flight passengers as non-terrorist, then the useless model would be 99.99% accurate.
· Precision Score: proportion of correct predictions out of all predicted attrition cases.
· Recall Score: proportion of correct predictions out of all actual attrition cases.
· F1 Score: optimised balance between Precision and Recall for the selected relevant target.

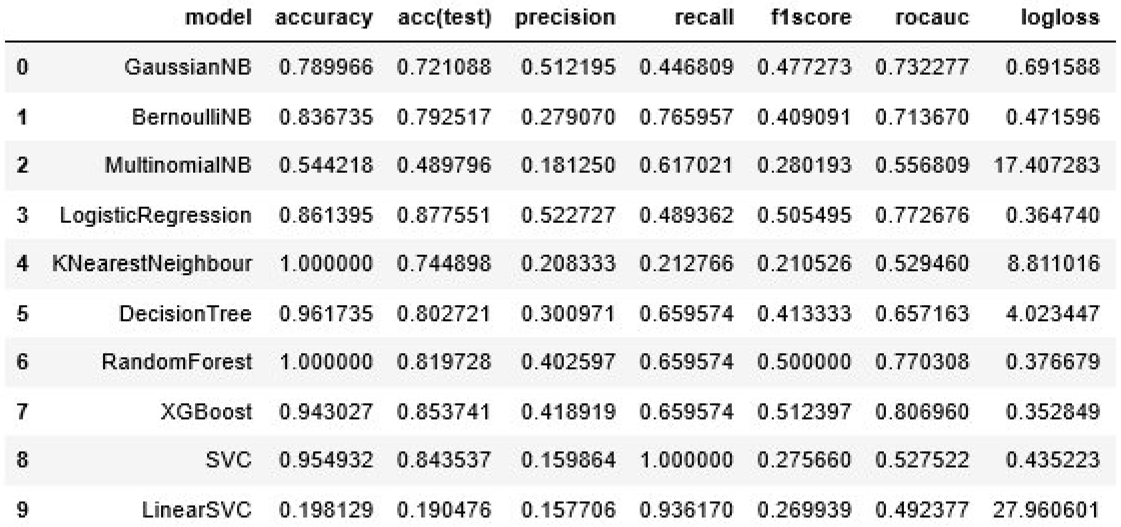
Iteration 1: Baseline Model Performance

Baseline model performance



	model	accuracy	acc(test)	precision	recall	f1score	rocauc	logloss
0	GaussianNB	0.778979	0.721088	0.399655	0.631579	0.484434	0.755946	9.633433
1	BernoulliNB	0.829917	0.792517	0.460915	0.263158	0.334298	0.742725	7.166274
2	MultinomialNB	0.548442	0.489796	0.200183	0.600000	0.300132	0.589342	17.622184
3	LogisticRegression	0.857144	0.877551	0.698690	0.210526	0.317851	0.791593	4.229238
4	KNearestNeighbour	0.823985	0.819728	0.342424	0.105263	0.160358	0.594289	6.226405
5	DecisionTree	0.784010	0.768707	0.306718	0.315789	0.318768	0.590567	7.988656
6	RandomForest	0.857144	0.846939	0.661111	0.168421	0.248970	0.733590	5.286566
7	XGBoost	0.866498	0.853741	0.711310	0.300000	0.420763	0.792229	5.051601
8	SVC	0.838435	0.840136	0.000000	0.000000	0.000000	0.500000	5.521505
9	LinearSVC	0.838435	0.840136	0.078247	0.278947	0.021818	0.554647	5.521505

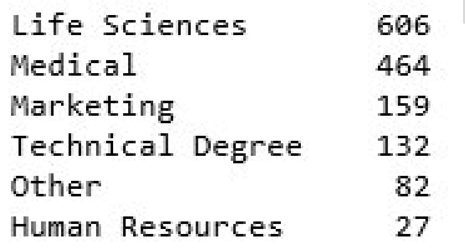
After tuning hyperparameters and threshold



	model	accuracy	acc(test)	precision	recall	f1score	rocauc	logloss
0	GaussianNB	0.789966	0.721088	0.512195	0.446809	0.477273	0.732277	0.691588
1	BernoulliNB	0.836735	0.792517	0.279070	0.765957	0.409091	0.713670	0.471596
2	MultinomialNB	0.544218	0.489796	0.181250	0.617021	0.280193	0.556809	17.407283
3	LogisticRegression	0.861395	0.877551	0.522727	0.489362	0.505495	0.772676	0.364740
4	KNearestNeighbour	1.000000	0.744898	0.208333	0.212766	0.210526	0.529460	8.811016
5	DecisionTree	0.961735	0.802721	0.300971	0.659574	0.413333	0.657163	4.023447
6	RandomForest	1.000000	0.819728	0.402597	0.659574	0.500000	0.770308	0.376679
7	XGBoost	0.943027	0.853741	0.418919	0.659574	0.512397	0.806960	0.352849
8	SVC	0.954932	0.843537	0.159864	1.000000	0.275660	0.527522	0.435223
9	LinearSVC	0.198129	0.190476	0.157706	0.936170	0.269939	0.492377	27.960601

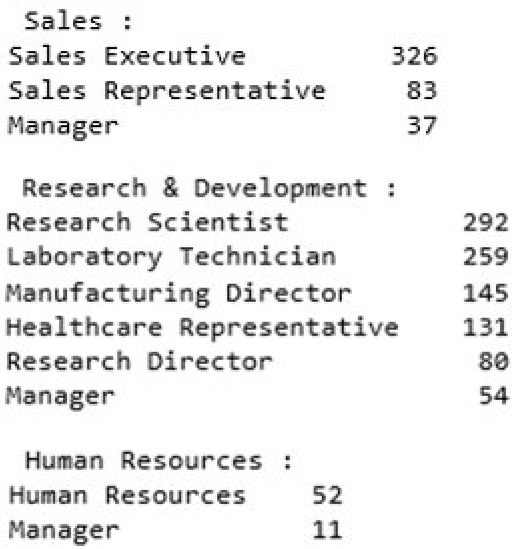
Iteration 2: Feature Engineering, Feature Selection

'EducationField'



```
Life Sciences        606
Medical              464
Marketing            159
Technical Degree     132
Other                 82
Human Resources       27
```

‘JobRole’



```
 Sales :
Sales Executive           326
Sales Representative       83
Manager                    37

 Research & Development :
Research Scientist                 292
Laboratory Technician              259
Manufacturing Director             145
Healthcare Representative          131
Research Director                   80
Manager                             54

 Human Resources :
Human Resources     52
Manager             11
```

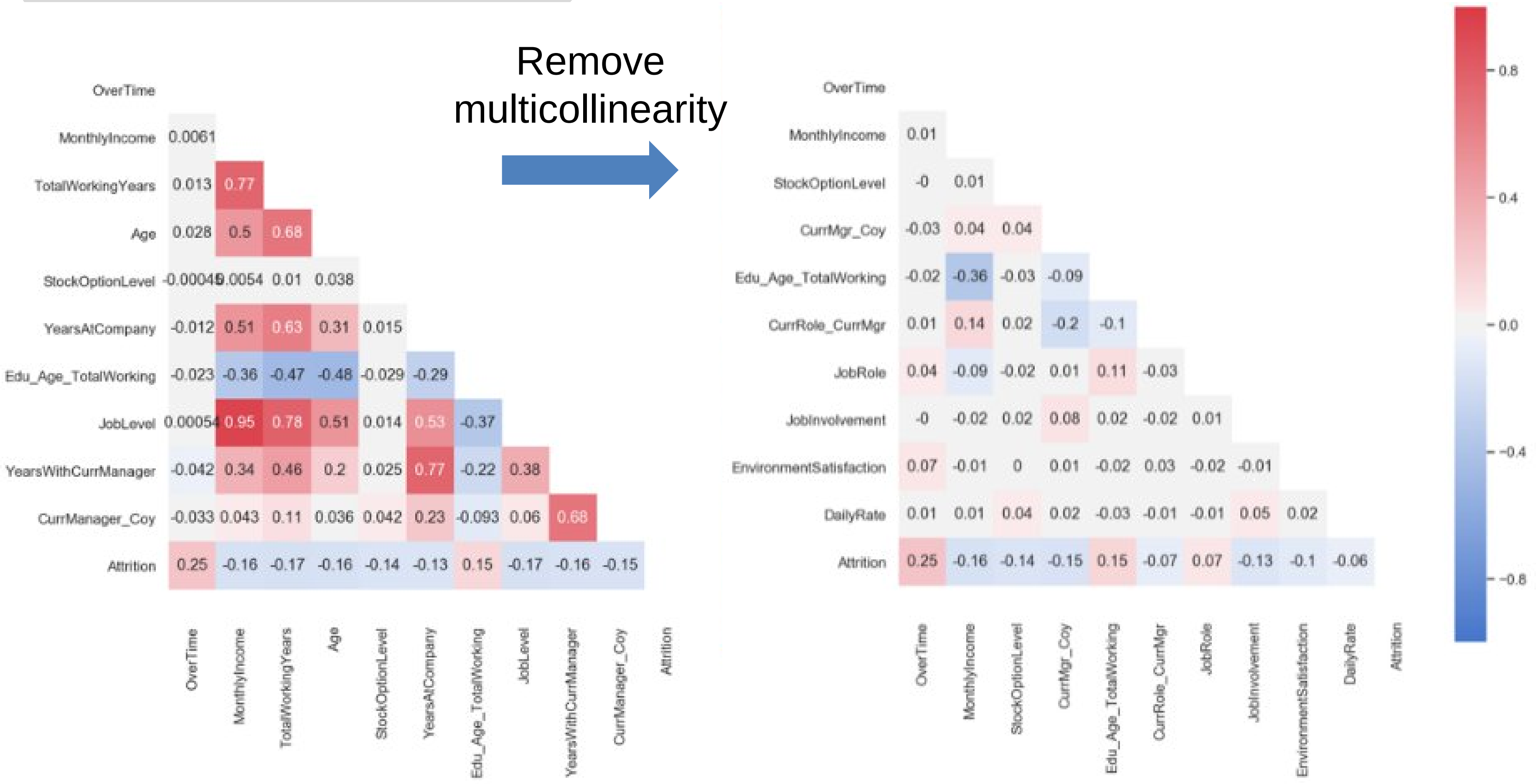
New feature ‘**EduField_Dept**’

whether JobRole is related to EducationField:

- 0 = not related,
- 1 = related,
- 2 = somewhat related

Other new features created:
· **Job_Coy** = JobLevel / (YearsAtCompany + 1)
· **Edu_Age_TotalWorking** = Education / (Age + TotalWorkingYears)
· **CurrMgr_Coy** = YearsWithCurrManager / (YearsAtCompany + 1)
· **CurrRole_CurrMgr** = YearsInCurrentRole / (YearsWithCurrManager + 1)

Iteration 2: Feature Engineering, Feature Selection

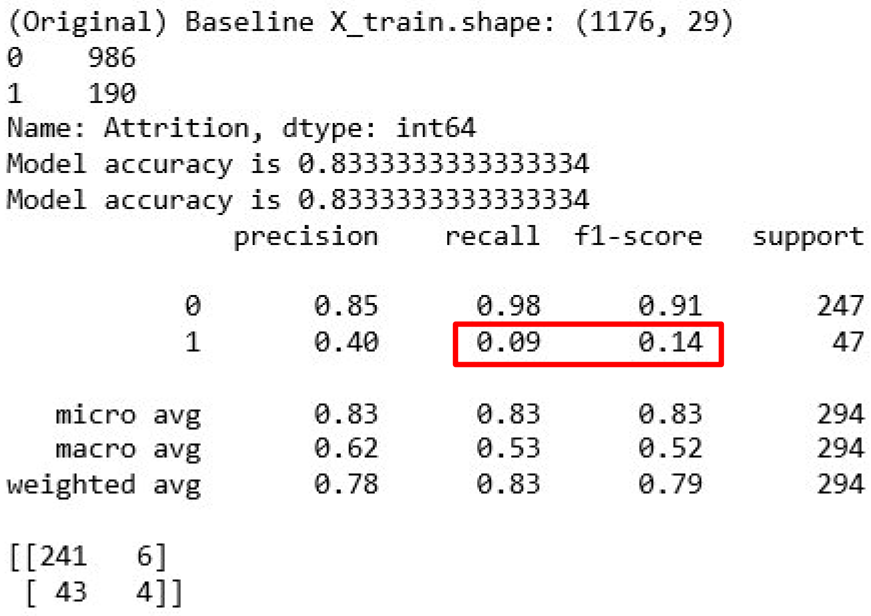


	OverTime	MonthlyIncome	TotalWorkingYears	Age	StockOptionLevel	YearsAtCompany	Edu_Age_TotalWorking	JobLevel	YearsWithCurrManager	CurrManager_Coy	Attrition
OverTime											
MonthlyIncome	0.0061										
TotalWorkingYears	0.013	0.77									
Age	0.028	0.5	0.68								
StockOptionLevel	-0.00045	0.0054	0.01	0.038							
YearsAtCompany	-0.012	0.51	0.63	0.31	0.015						
Edu_Age_TotalWorking	-0.023	-0.36	-0.47	-0.48	-0.029	-0.29					
JobLevel	0.00054	0.95	0.78	0.51	0.014	0.53	-0.37				
YearsWithCurrManager	-0.042	0.34	0.46	0.2	0.025	0.77	-0.22	0.38			
CurrManager_Coy	-0.033	0.043	0.11	0.036	0.042	0.23	-0.093	0.06	0.68		
Attrition	0.25	-0.16	-0.17	-0.16	-0.14	-0.13	0.15	-0.17	-0.16	-0.15	

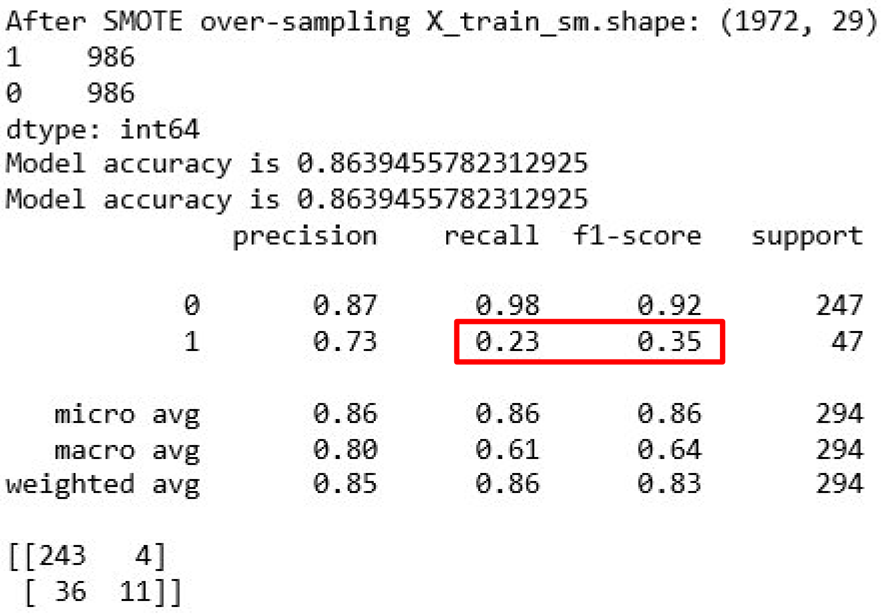
Remove multicollinearity

	OverTime	MonthlyIncome	StockOptionLevel	CurrMgr_Coy	Edu_Age_TotalWorking	CurrRole_CurrMgr	JobRole	JobInvolvement	EnvironmentSatisfaction	DailyRate	Attrition
OverTime											
MonthlyIncome	0.01										
StockOptionLevel	-0	0.01									
CurrMgr_Coy	-0.03	0.04	0.04								
Edu_Age_TotalWorking	-0.02	-0.36	-0.03	-0.09							
CurrRole_CurrMgr	0.01	0.14	0.02	-0.2	-0.1						
JobRole	0.04	-0.09	-0.02	0.01	0.11	-0.03					
JobInvolvement	-0	-0.02	0.02	0.08	0.02	-0.02	0.01				
EnvironmentSatisfaction	0.07	-0.01	0	0.01	-0.02	0.03	-0.02	-0.01			
DailyRate	0.01	0.01	0.04	0.02	-0.03	-0.01	-0.01	0.05	0.02		
Attrition	0.25	-0.16	-0.14	-0.15	0.15	-0.07	0.07	-0.13	-0.1	-0.06	

Iteration 3: Over-sampling with SMOTE

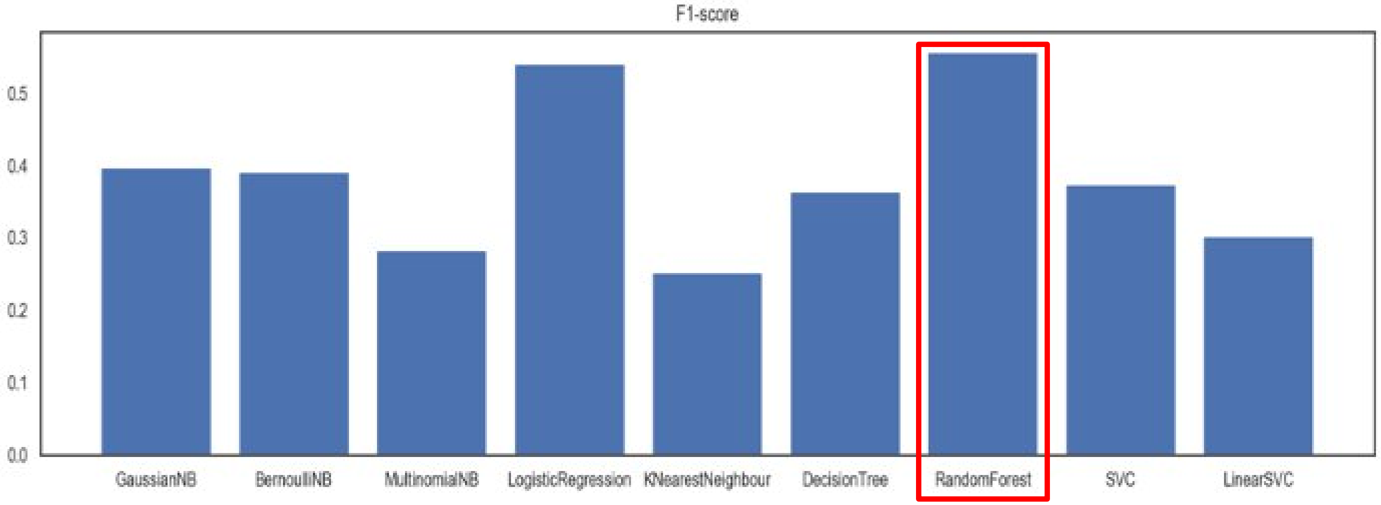


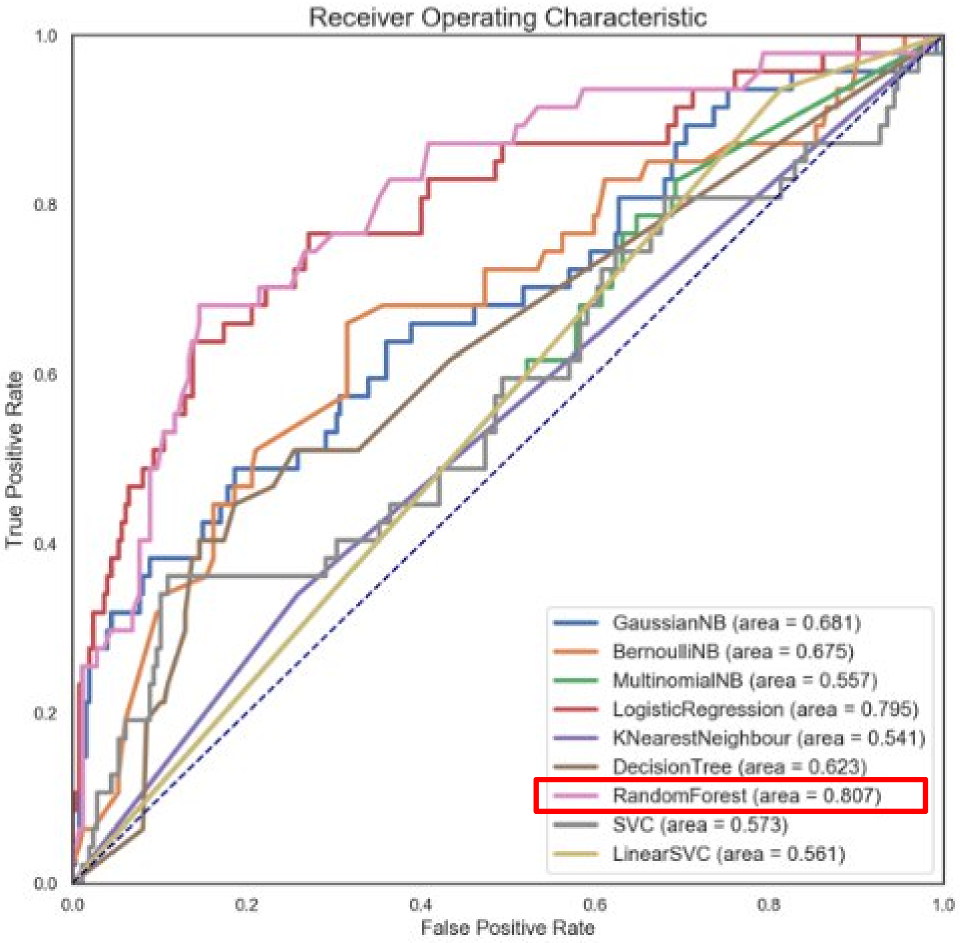
```
(Original) Baseline X_train.shape: (1176, 29)
0    986
1    190
Name: Attrition, dtype: int64
Model accuracy is 0.8333333333333334
Model accuracy is 0.8333333333333334
              precision    recall  f1-score   support

           0       0.85      0.98      0.91       247
           1       0.40      0.09      0.14        47

   micro avg       0.83      0.83      0.83       294
   macro avg       0.62      0.53      0.52       294
weighted avg       0.78      0.83      0.79       294

[[241   6]
 [ 43   4]]
```



```
After SMOTE over-sampling X_train_sm.shape: (1972, 29)
1    986
0    986
dtype: int64
Model accuracy is 0.8639455782312925
Model accuracy is 0.8639455782312925
              precision    recall  f1-score   support

           0       0.87      0.98      0.92       247
           1       0.73      0.23      0.35        47

   micro avg       0.86      0.86      0.86       294
   macro avg       0.80      0.61      0.64       294
weighted avg       0.85      0.86      0.83       294

[[243   4]
 [ 36  11]]
```

Final Model Performance (including tuning)





Best performing model:
Random Forest Classifier

Key Findings – Employee Attrition Factors

Factor 1: **monetary**

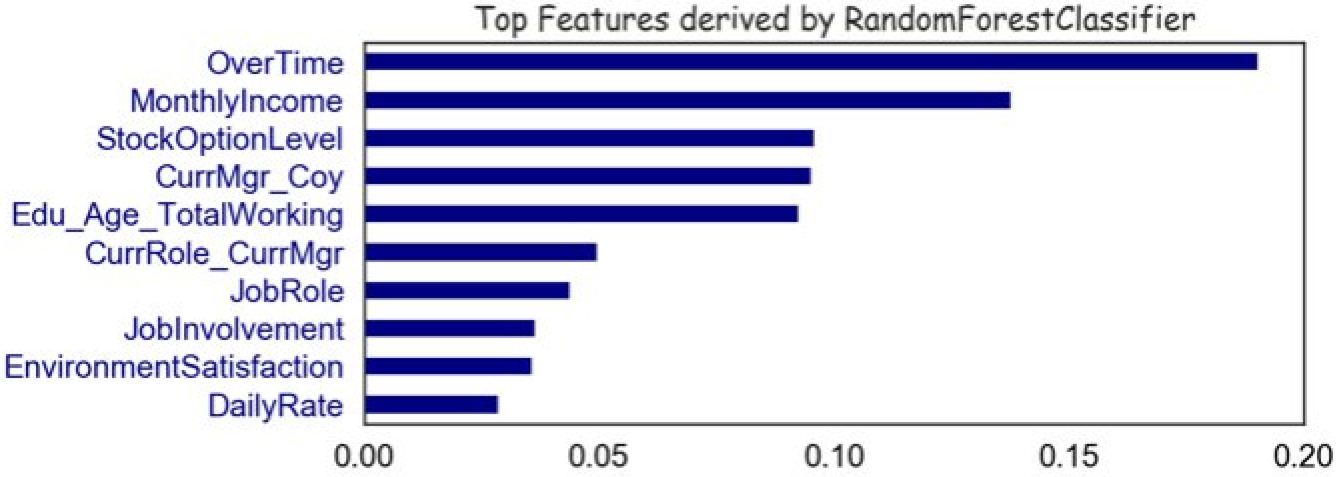
- ‘OverTime’
- ‘MonthlyIncome’
- ‘StockOptionLevel’

Factor 2: **personal relationships**

- ‘CurrMgr_Coy’
- ‘CurrRole_CurrMgr’
- ‘JobRole’

Factor 3: **employee engagement**

- ‘JobInvolvement’,
- ‘EnvironmentSatisfaction’



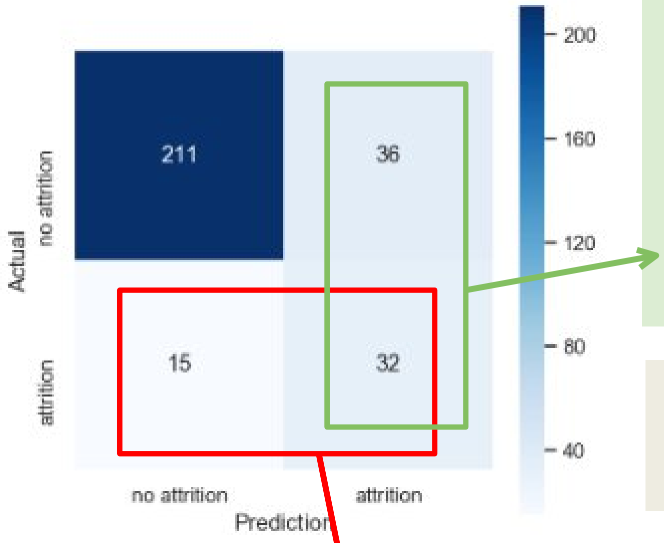
Note: Results are only based on the hypothetical dataset

Conclusion

In this project, the **Random Forest Classifier** model has achieved prediction (Recall) score of **68.1%**

Out of all 47 attrition employees, 68.1% of them will be classified correctly using their background attributes and characteristics

Optimal threshold 0.288
F1 Score = 55.7%



proportion of correct predictions out of all predicted attrition cases

Precision = 32/68 =

Accuracy = (211+32)/294 = 82.7%

Recall = 32/47 = 68.1%

proportion of correct predictions out of all actual attrition cases

THANK YOU