Attrition Analysis

A Case Study of IBM’s Attrition Prediction

Xiao Li
Xuan Wang
Yi Nian
Jingrou Wei

Introduction

Dataset

IBM HR Analytics Employee Attrition & Performance

Predict attrition of your valuable employees

Motivation

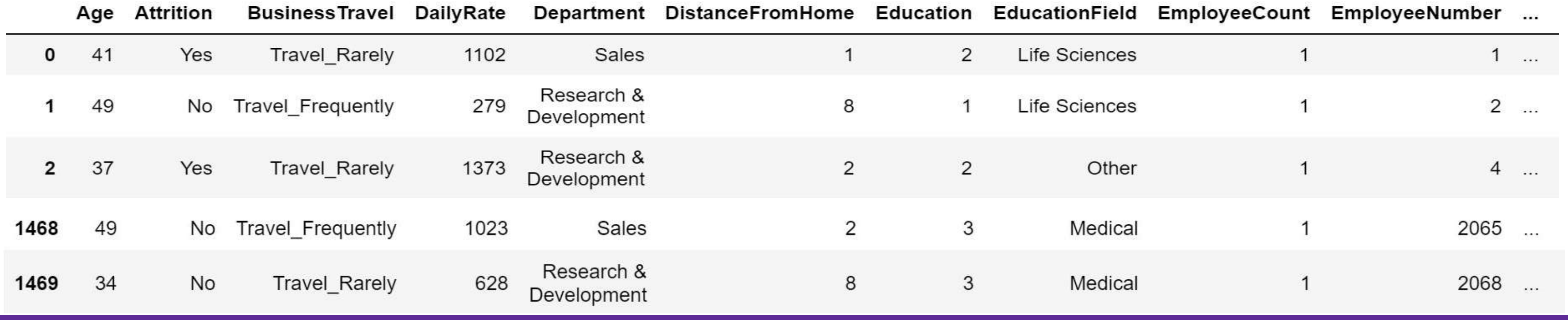
Construct correlations between factors

Analyze attrition distribution of features

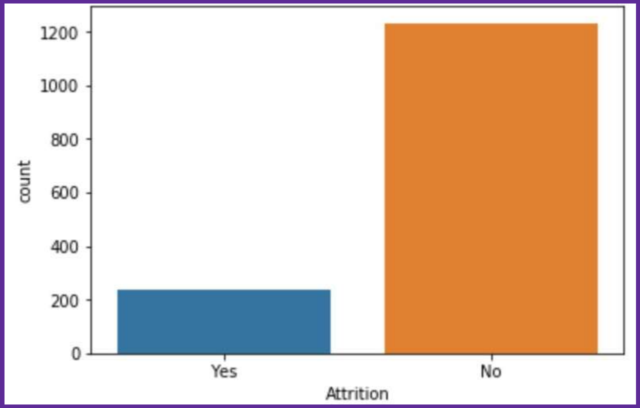
Implement and evaluate models to predict the attrition rate

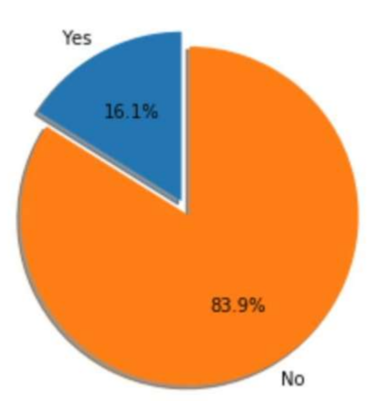
Find best plan to minimize attrition rate

Data Visualization



	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...
1468	49	No	Travel_Frequently	1023	Sales	2	3	Medical	1	2065	...
1469	34	No	Travel_Rarely	628	Research & Development	8	3	Medical	1	2068	...

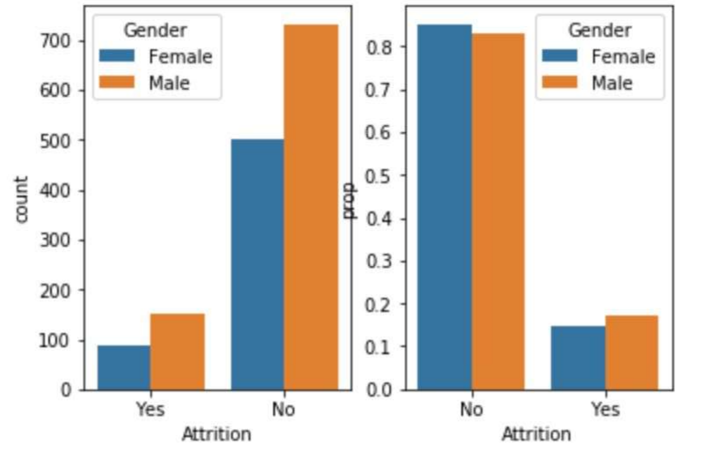




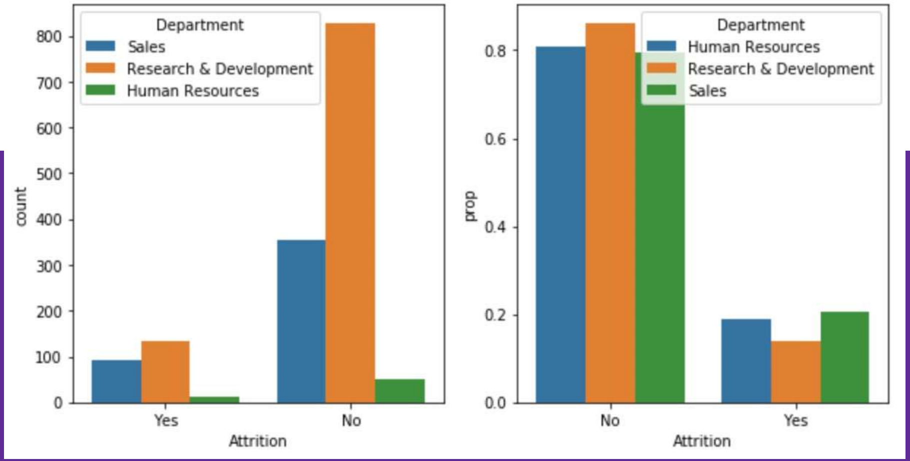
Attrition Total: 237
Attrition Rate: 16.1% (high)

Attrition Distribution Analysis

Comparisons of attrition amount and attrition rate in ‘Gender’ and ‘Department’

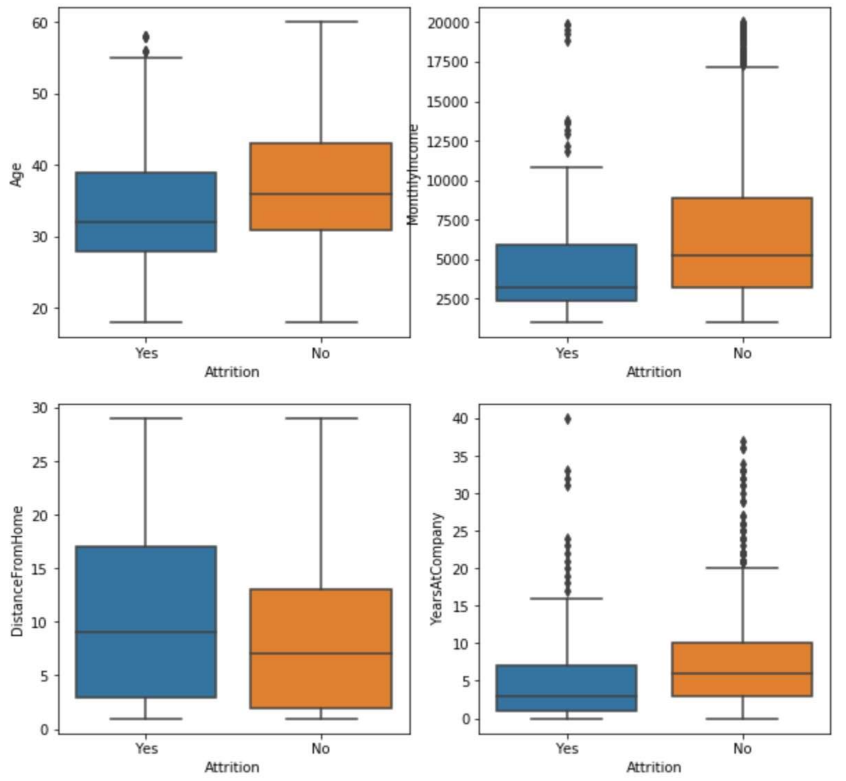


Higher proportion of males are likely for attrition as compared to females.



Employee in Sales department are more likely for attrition as compared to the other two.

Attrition Distribution Analysis

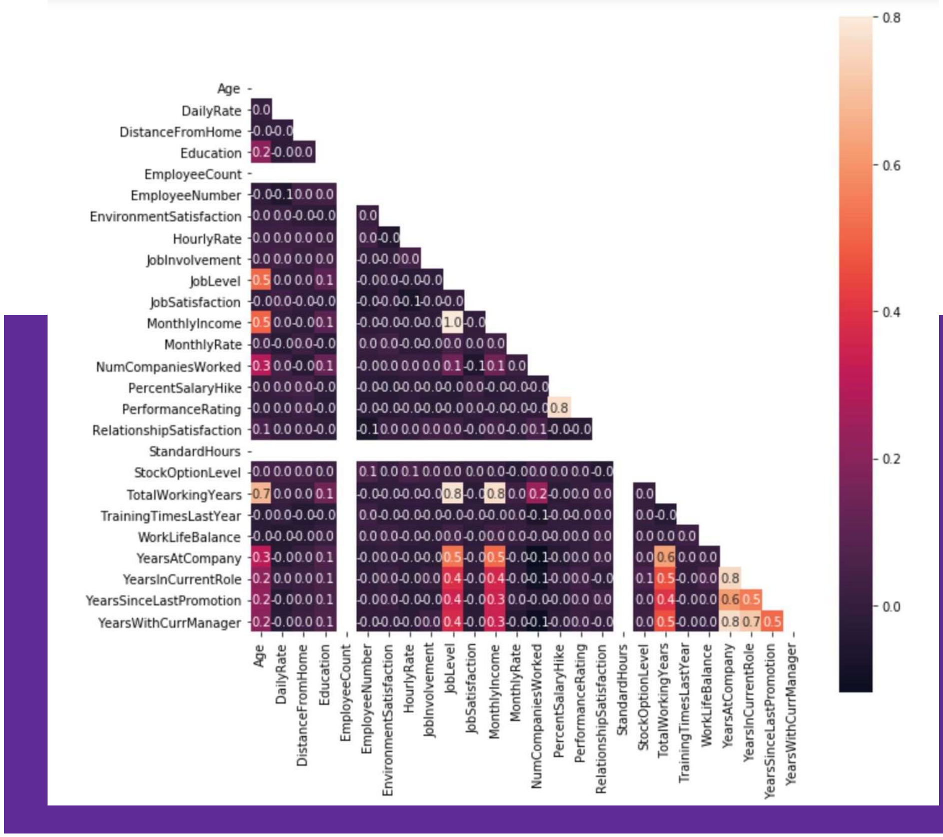


Comparisons of attrition in

- ‘Age’ ,
- ‘Monthly Income’,
- ‘Distance From Home’,
- ‘Years at Company’

with Box Plot.

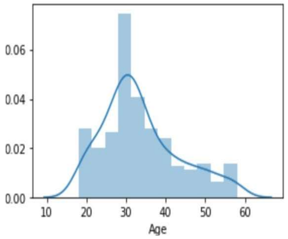
Correlations between Features

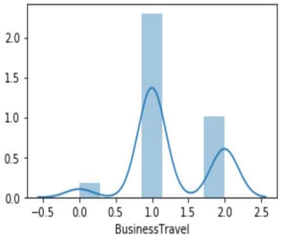


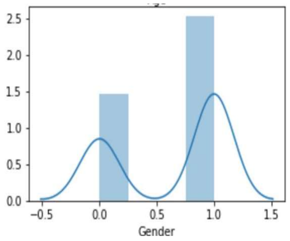
Most of the features are uncorrelated

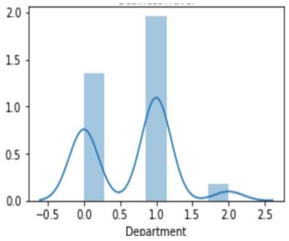
But, there is a high correlation between Monthly Income and Job Level(1.0), Monthly Income and Total Working Years(0.8), Total Working Years and Job Level(0.8)

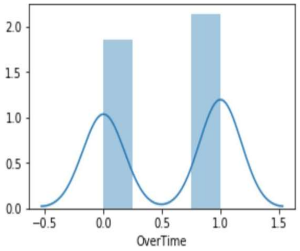
Implement Models

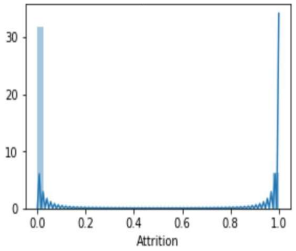




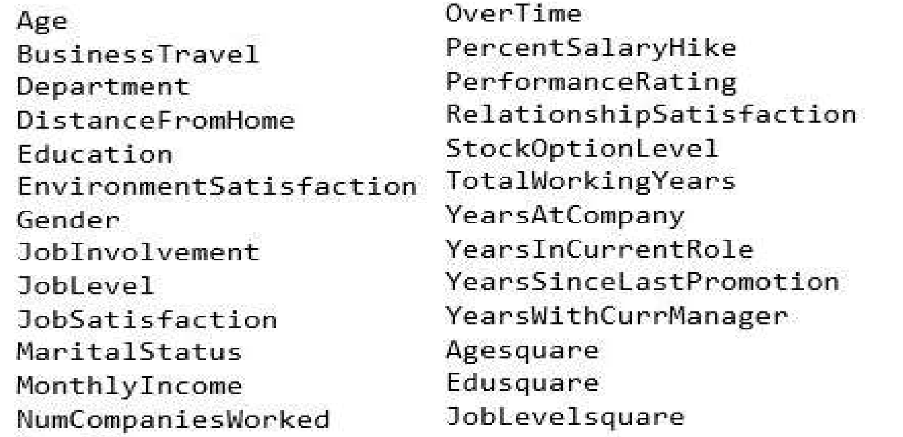






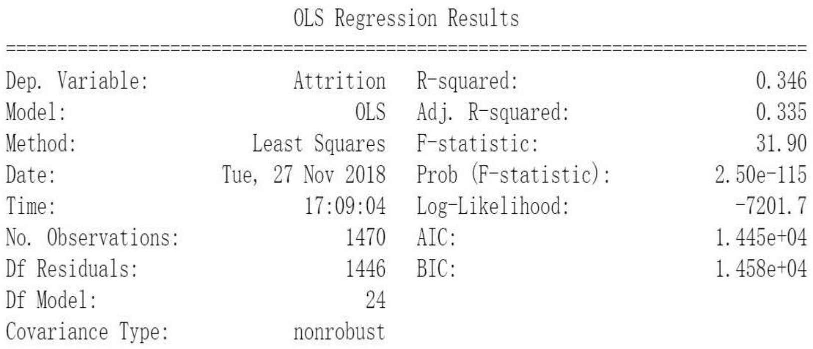


Try to fit a linear model. Add second order or third order terms to features appeared non-linear.



```
Age                       OverTime
BusinessTravel            PercentSalaryHike
Department                PerformanceRating
DistanceFromHome          RelationshipSatisfaction
Education                 StockOptionLevel
EnvironmentSatisfaction   TotalWorkingYears
Gender                    YearsAtCompany
JobInvolvement            YearsInCurrentRole
JobLevel                  YearsSinceLastPromotion
JobSatisfaction           YearsWithCurrManager
MaritalStatus             Agesquare
MonthlyIncome             Edusquare
NumCompaniesWorked        JobLevelsquare
```

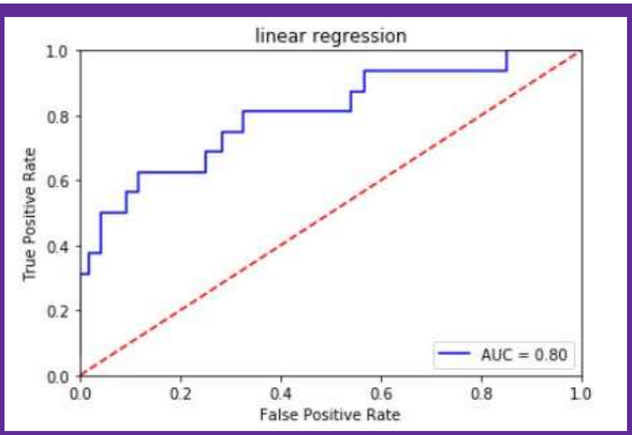
Implement Models



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Attrition   R-squared:                       0.346
Model:                            OLS   Adj. R-squared:                  0.335
Method:                 Least Squares   F-statistic:                     31.90
Date:                Tue, 27 Nov 2018   Prob (F-statistic):          2.50e-115
Time:                        17:09:04   Log-Likelihood:                -7201.7
No. Observations:                1470   AIC:                         1.445e+04
Df Residuals:                    1446   BIC:                         1.458e+04
Df Model:                          24
Covariance Type:            nonrobust
```

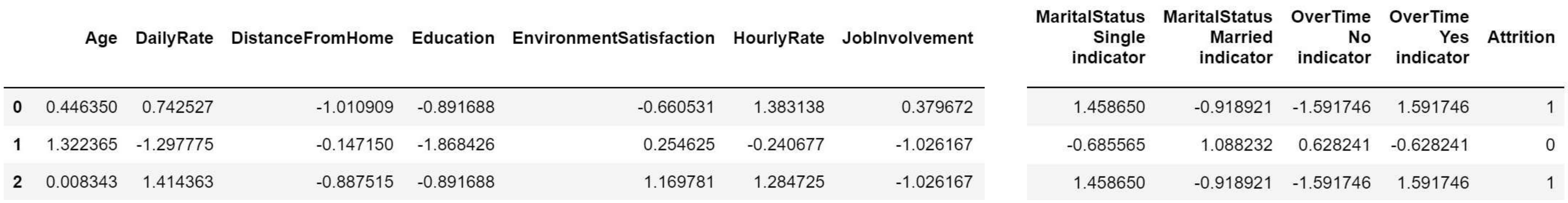
After comparing AIC and significance test, a linear model with terms shows on the left with 34.6% R square level.

ROC = 0.80
Looks good but relations between features and attrition may be more complicated...



Implement Models

Data Cleaning

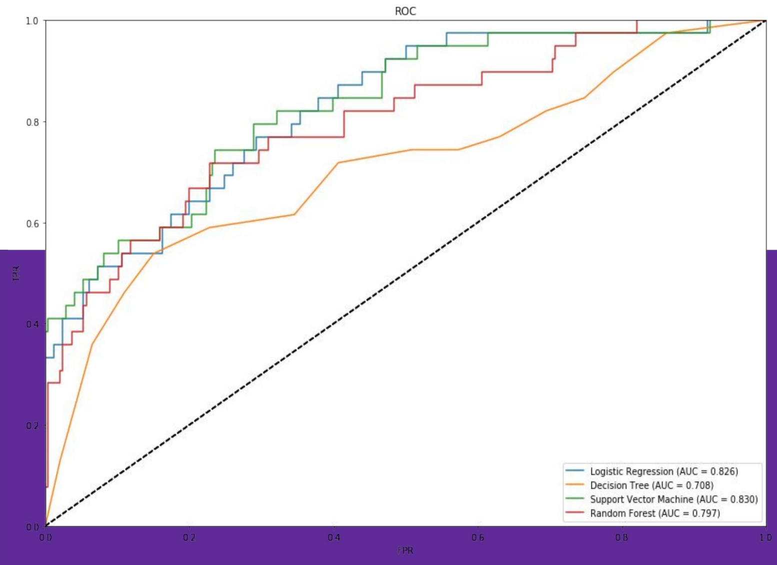


	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	MaritalStatus Single indicator	MaritalStatus Married indicator	OverTime No indicator	OverTime Yes indicator	Attrition
0	0.446350	0.742527	-1.010909	-0.891688	-0.660531	1.383138	0.379672	1.458650	-0.918921	-1.591746	1.591746	1
1	1.322365	-1.297775	-0.147150	-1.868426	0.254625	-0.240677	-1.026167	-0.685565	1.088232	0.628241	-0.628241	0
2	0.008343	1.414363	-0.887515	-0.891688	1.169781	1.284725	-1.026167	1.458650	-0.918921	-1.591746	1.591746	1

- Check if there exists missing data
- Make the attrition column numeric, 1 for 'yes' and 0 for 'no'
- Remove variables that are the same all the time and not meaningful
- Assign an unique indicator for each categorical variables in the dataset
- Make all the features in our dataset have mean 0 and std 1 for future convenience

Implement Models

Logistic Regression vs. Decision Tree vs. Support Vector Machine vs. Random Forest



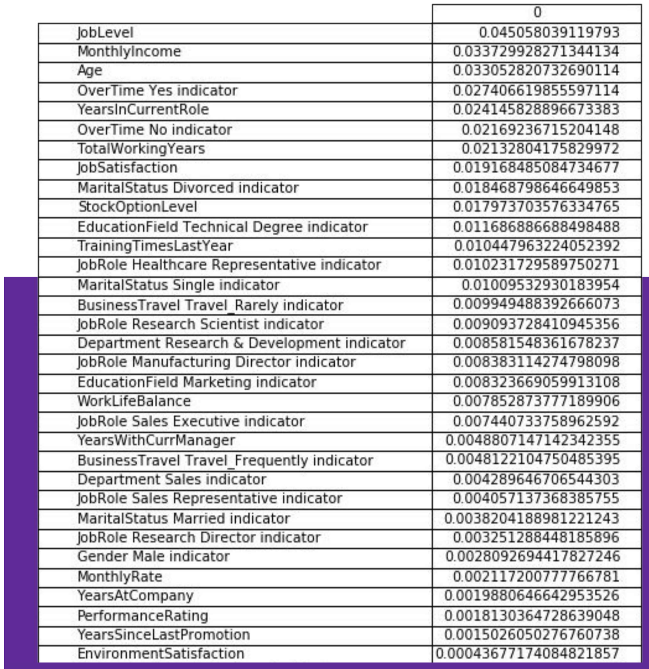
Models Comparison

Compared by ROC:

Logistic Regression: 0.826
Decision Tree: 0.708
SVM: 0.830
Random Forest: 0.797

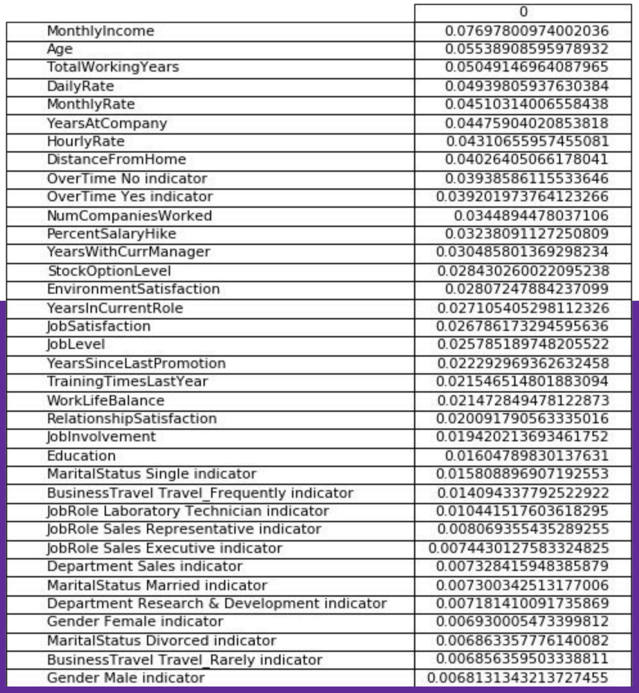
Feature Importance

Mutual Information



	0
JobLevel	0.045058039119793
MonthlyIncome	0.033729928271344134
Age	0.033052820732690114
OverTime Yes indicator	0.027406619855597114
YearsInCurrentRole	0.024145828896673383
OverTime No indicator	0.02169236715204148
TotalWorkingYears	0.02132804175829972
JobSatisfaction	0.019168485084734677
MaritalStatus Divorced indicator	0.018468798646649853
StockOptionLevel	0.017973703576334765
EducationField Technical Degree indicator	0.011686886688498488
TrainingTimesLastYear	0.010447963224052392
JobRole Healthcare Representative indicator	0.010231729589750271
MaritalStatus Single indicator	0.01009532930183954
BusinessTravel Travel_Rarely indicator	0.009949488392666073
JobRole Research Scientist indicator	0.009093728410945356
Department Research & Development indicator	0.008581548361678237
JobRole Manufacturing Director indicator	0.008383114274798098
EducationField Marketing indicator	0.008323669059913108
WorkLifeBalance	0.007852873777189906
JobRole Sales Executive indicator	0.007440733758962592
YearsWithCurrManager	0.0048807147142342355
BusinessTravel Travel_Frequently indicator	0.0048122104750485395
Department Sales indicator	0.004289646706544303
JobRole Sales Representative indicator	0.004057137368385755
MaritalStatus Married indicator	0.0038204188981221243
JobRole Research Director indicator	0.003251288448185896
Gender Male indicator	0.0028092694417827246
MonthlyRate	0.002117200777766781
YearsAtCompany	0.0019880646642953526
PerformanceRating	0.0018130364728639048
YearsSinceLastPromotion	0.0015026050276760738
EnvironmentSatisfaction	0.00043677174084821857

Feature Importance from Random Forest

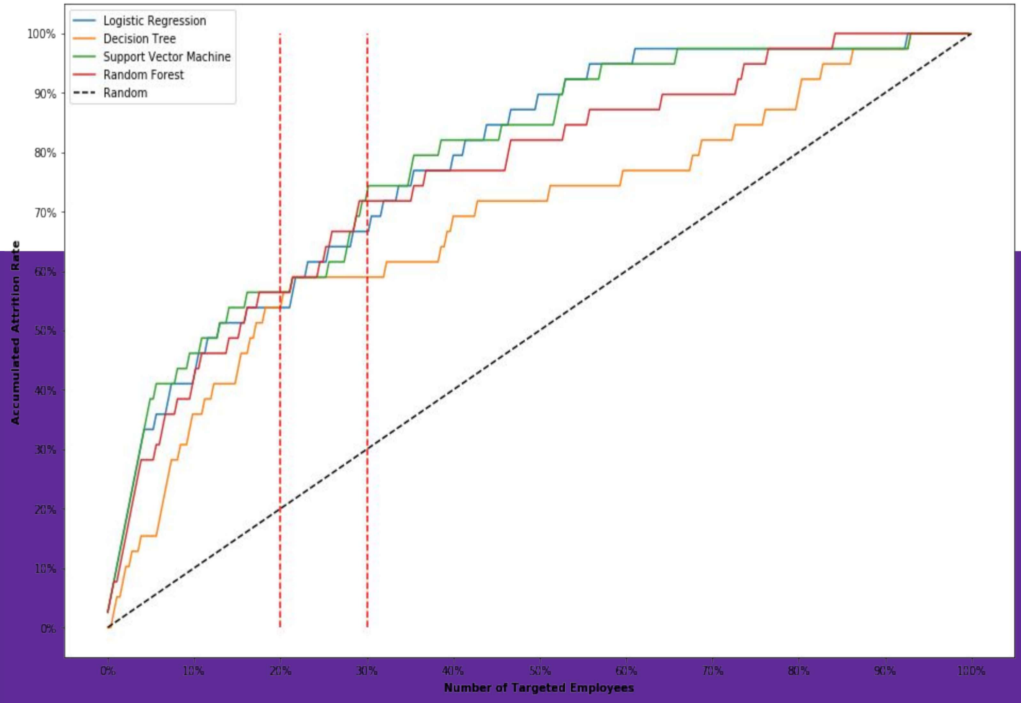


	0
MonthlyIncome	0.07697800974002036
Age	0.05538908595978932
TotalWorkingYears	0.05049146964087965
DailyRate	0.04939805937630384
MonthlyRate	0.04510314006558438
YearsAtCompany	0.04475904020853818
HourlyRate	0.04310655957455081
DistanceFromHome	0.04026405066178041
OverTime No indicator	0.03938586115533646
OverTime Yes indicator	0.039201973764123266
NumCompaniesWorked	0.0344894478037106
PercentSalaryHike	0.03238091127250809
YearsWithCurrManager	0.030485801369298234
StockOptionLevel	0.028430260022095238
EnvironmentSatisfaction	0.02807247884237099
YearsInCurrentRole	0.027105405298112326
JobSatisfaction	0.026786173294595636
JobLevel	0.025785189748205522
YearsSinceLastPromotion	0.022292969362632458
TrainingTimesLastYear	0.021546514801883094
WorkLifeBalance	0.021472849478122873
RelationshipSatisfaction	0.020091790563335016
JobInvolvement	0.019420213693461752
Education	0.01604789830137631
MaritalStatus Single indicator	0.015808896907192553
BusinessTravel Travel_Frequently indicator	0.014094337792522922
JobRole Laboratory Technician indicator	0.010441517603618295
JobRole Sales Representative indicator	0.008069355435289255
JobRole Sales Executive indicator	0.0074430127583324825
Department Sales indicator	0.007328415948385879
MaritalStatus Married indicator	0.007300342513177006
Department Research & Development indicator	0.007181410091735869
Gender Female indicator	0.006930005473399812
MaritalStatus Divorced indicator	0.006863357776140082
BusinessTravel Travel_Rarely indicator	0.006856359503338811
Gender Male indicator	0.0068131343213727455

Features that significantly affect the attrition:

- Age
- Income
- Working Overtime
- Working Years
- Job Satisfaction

Cumulative Gain Curve



Target top 30% most possible leaving employees

EQUALS

Target over 70% of employees who actually would leave!

Solve the Problem

Main Idea: change part of an employee’s status to make the employee more likely to stay.

Approach:

- **Naive approach:** change some features of a employee and predict the turn off probability again and check if the attrition rate drops.
- **Interactive approach**: design a questionnaire for the employees who are detected to be most likely to leave and change the employee’s status accordingly
- Features that
  cannot be changed: Age, Total Working Years, Gender…
  can be changed: Monthly Income, Job Level, Years at Current Role…
  can be changed gradually: Job Satisfaction…

THANKS FOR LISTENING!