IBM HR Analysis and Prediction

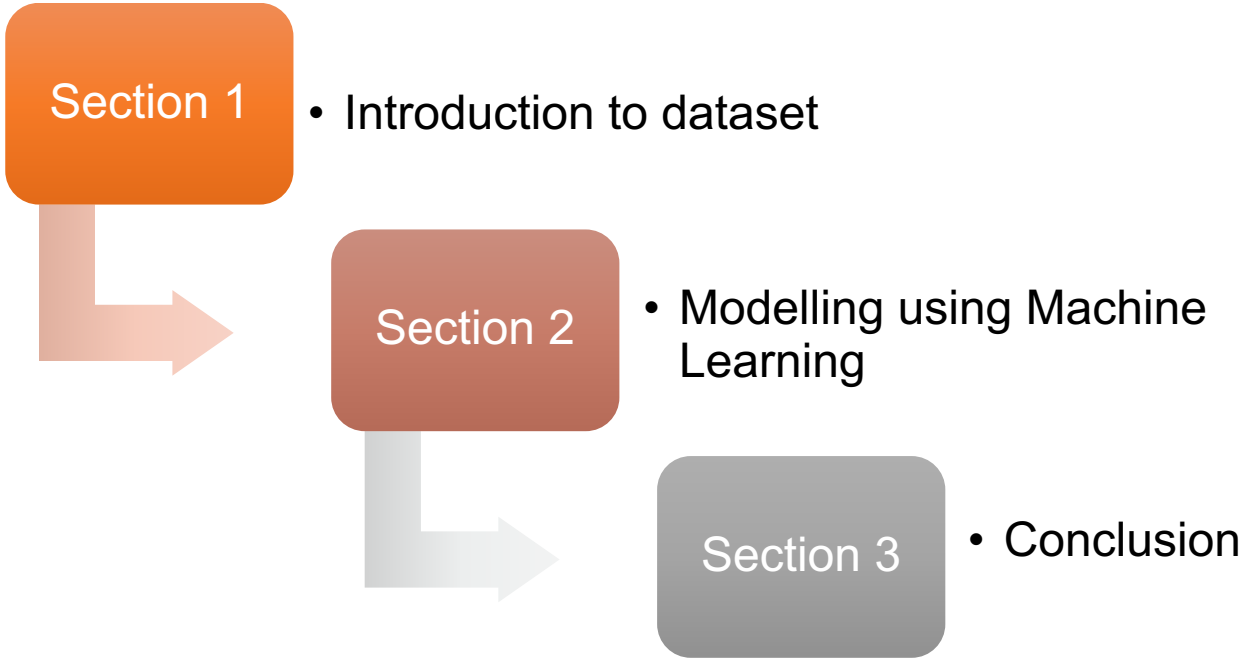
Employee Attrition & Income

Data 201B | Group Project | Dec 2020
Li Ling | Yitong Sun | Ying Yang

Attrition

Salary

IBM HR Analysis and Prediction on Employee Attrition & Income



Introduction

1. Purpose of Study
2. Data Source
3. Data Process
4. Data Exploration

I. Purpose of study

1. Use classification models to predict if an employee is likely to quit or not

- Greatly increase the HR's ability to intervene on time and remedy the situation to prevent attrition

2. Use regression models to predict reasonable monthly income for employees

- Set a reasonable salary level for employees to motivate them

Target:

Significantly help improve the operations of most businesses

II. Data Source

- Source: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset (from Kaggle)
- Content: An employee survey from IBM
- Size: 1,470 observations * 35 features
- Variable Description: See next slide

II. Data Source

Target variables in two models:

- Attrition
- MonthlyIncome

No.	Variable Name	Description	Type
1	Age	Age of employee	Numerical
2	Attrition	Yes if employee leaving, No otherwise	Object
3	BusinessTravel	Frequency of business travel	Object
4	DailyRate	Daily internal charge out rate	Numerical
5	Department	Department which employee works in	Object
6	DistanceFromHome	Distance from work to home	Numerical
7	Education	Education level: 1-5 (lowest to highest level)	Numerical
8	EducationField	Education field	Object
9	EmployeeCount	Number of employee	Numerical
10	EmployeeNumber	Employee ID	Numerical
11	EnvironmentSatisfaction	Satisfaction with working environment: 1-4 (lowest to highest level)	Numerical
12	Gender	Female; Male	Object
13	HourlyRate	Hourly internal charge out rate	Numerical
14	JobInvolvement	Involvement in the job: 1-4 (lowest to highest level)	Numerical
15	JobLevel	Job level: 1-5 (lowest to highest level)	Numerical
16	JobRole	Job position	Object
17	JobSatisfaction	Satisfaction with job: 1-4 (lowest to highest level)	Numerical
18	MaritalStatus	Marital Status of employee	Object
19	MonthlyIncome	Monthly salary of employee	Numerical
20	MonthlyRate	Monthly internal charge out rate	Numerical
21	NumCompaniesWorked	Number of companies worked at	Numerical
22	Over18	Y if employee over 18, N otherwise	Object
23	OverTime	Yes if employee works overtime, No otherwise	Object
24	PercentSalaryHike	Percent increase in salary	Numerical
25	PerformanceRating	Performance rating: 3-4 (lowest to highest level)	Numerical
26	RelationshipSatisfaction	Satisfaction with relationship: 1-4 (lowest to highest level)	Numerical
27	StandardHours	Working standard hours	Numerical
28	StockOptionLevel	Stock option level: 0-3 (lowest to highest level)	Numerical
29	TotalWorkingYears	Total years worked	Numerical
30	TrainingTimesLastYear	Hours spent training	Numerical
31	WorkLifeBalance	Work and life balance level: 1-4 (lowest to highest level)	Numerical
32	YearsAtCompany	Total number of years at the company	Numerical
33	YearsInCurrentRole	Number of years in current role	Numerical
34	YearsSinceLastPromotion	Number of years since last promotion	Numerical
35	YearsWithCurrManager	Number of years spent with current manager	Numerical

III. Data Process

- Remove unrelated or uninformative variables
- Deal with missing data
- Change Boolean features to binari
- Convert the categorical features in
- Size of dataset for modelling:

Variable Name	Description	Type
DailyRate	Daily internal charge out rate	Numerical
EmployeeCount	Number of employee	Numerical
EmployeeNumber	Employee ID	Numerical
HourlyRate	Hourly internal charge out rate	Numerical
Over18	Y if employee over 18, N otherwise	Object
StandardHours	Working standard hours	Numerical

III. Data Process

- Remove unrelated or uninformative variables
- Deal with missing data
- Change Boolean features to binaries
- Convert the categorical features into dummy variables
- Size of dataset for modelling:

```
Age                         0
Attrition                   0
BusinessTravel              0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EnvironmentSatisfaction     0
Gender                      0
JobInvolvement              0
JobLevel                    0
JobRole                     0
JobSatisfaction             0
MaritalStatus               0
MonthlyIncome               0
MonthlyRate                 0
NumCompaniesWorked          0
OverTime                    0
PercentSalaryHike           0
PerformanceRating           0
RelationshipSatisfaction    0
StockOptionLevel            0
TotalWorkingYears           0
TrainingTimesLastYear       0
WorkLifeBalance             0
YearsAtCompany              0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

III. Data Process

- Remove unrelated or uninformative variables
- Deal with missing data
- Change Boolean features to binaries
- Convert the categorical features into dummy variables
- Size of dataset for modelling:

Variable Name	Description	Type
Attrition	Yes if employee leaving, No otherwise	Object

Variable Name	Description	Type
Attrition	1 if employee leaving, 0 otherwise	Number

III. Data Process

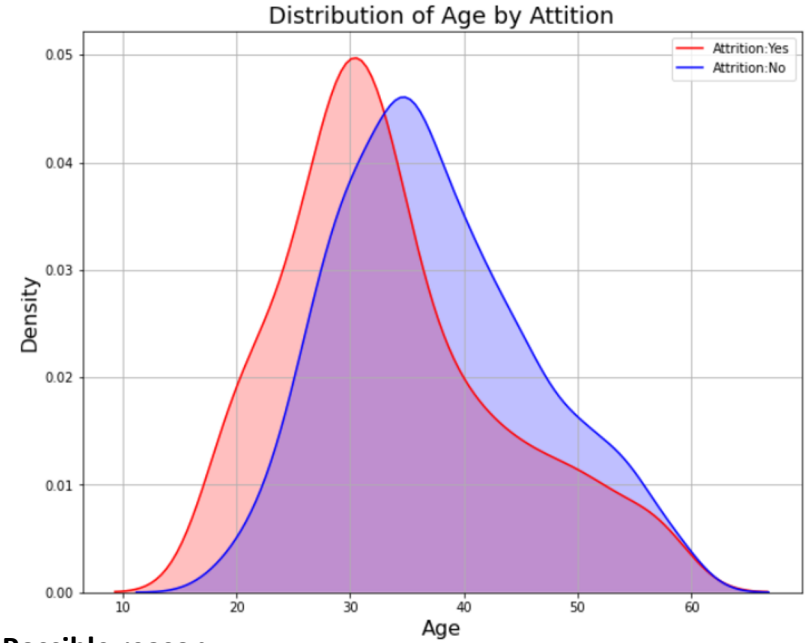
- Remove unrelated or uninformative variables
- Deal with missing data
- Change Boolean features to binaries
- Convert the categorical features into dummy variables
- Size of dataset for modelling:
  1,470 observations * 50 features

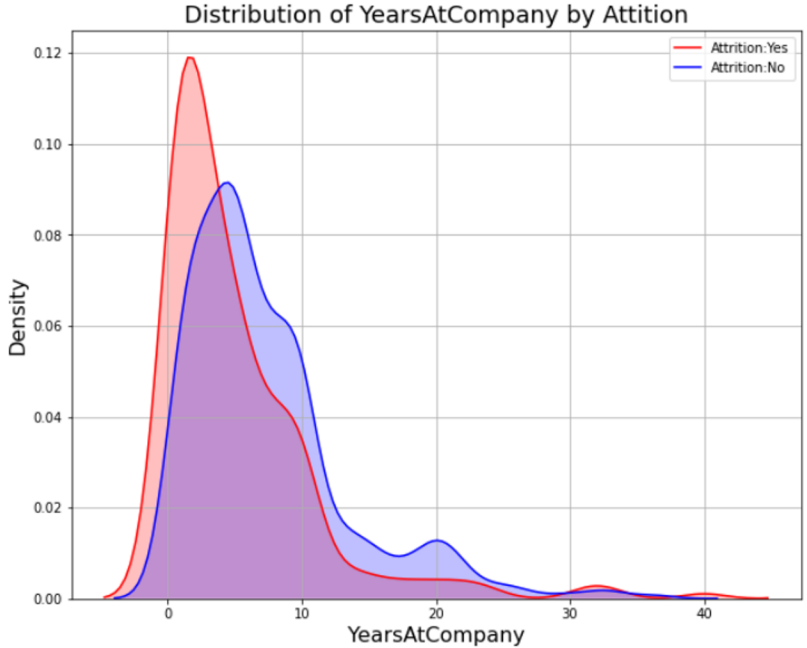
No.	Variable Name	Description	Type
1	Business Travel	Frequency of business travel	Object
2	Department	Department which employee works in	Object
3	Education Field	Education field	Object
4	Gender	Female; Male	Object
5	JobRole	Job position	Object
6	MaritalStatus	Marital Status of employee	Object
7	OverTime	Yes if employee works overtime, No otherwise	Object

```
['Travel_Rarely' 'Travel_Frequently' 'Non-Travel']
['Sales' 'Research & Development' 'Human Resources']
['Life Sciences' 'Other' 'Medical' 'Marketing' 'Technical Degree'
 'Human Resources']
['Female' 'Male']
['Sales Executive' 'Research Scientist' 'Laboratory Technician'
 'Manufacturing Director' 'Healthcare Representative' 'Manager'
 'Sales Representative' 'Research Director' 'Human Resources']
['Single' 'Married' 'Divorced']
['Yes' 'No']
```

IV. Data Exploration-Attrition

A. Experience



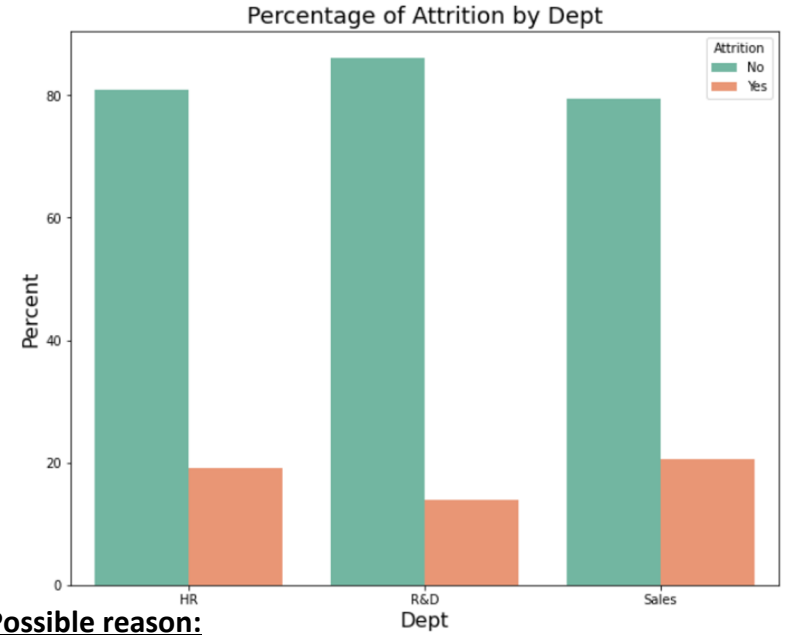


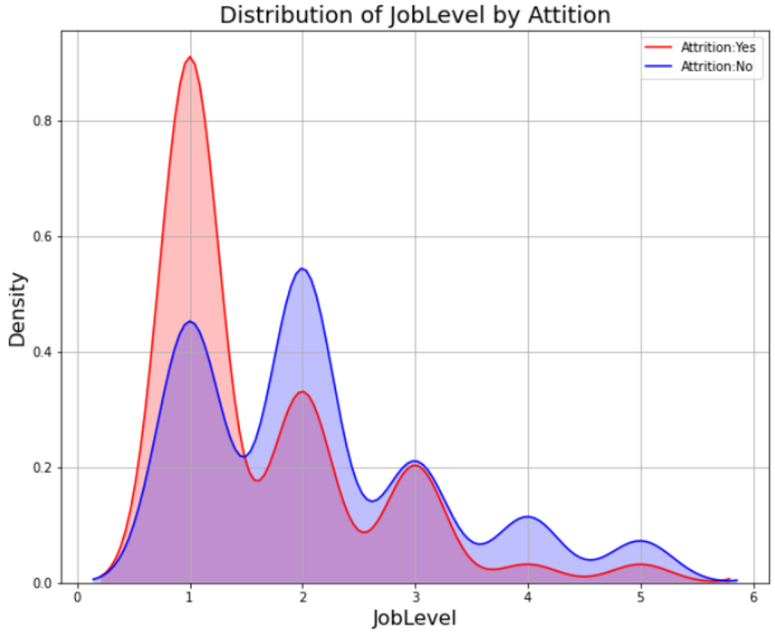
Possible reason:

Young employees tend to try more, and they are relatively confused about their future goals. The high turnover rate also means that it is difficult for such employees to form long-term identification of enterprise values in a short time.

IV. Data Exploration-Attrition

B. Department and Job level





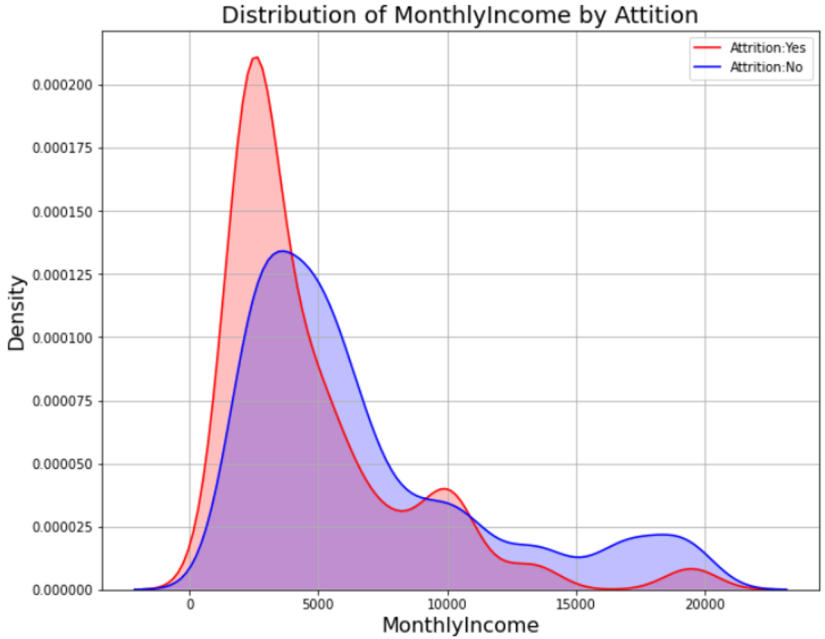
Possible reason:

Employees in Sales Dept may face some higher workload and pressure.

Most employees in lower job level may be not satisfied with their situation, may due to low pay or fewer opportunities to promote.

IV. Data Exploration-Attrition

C. Personal Income



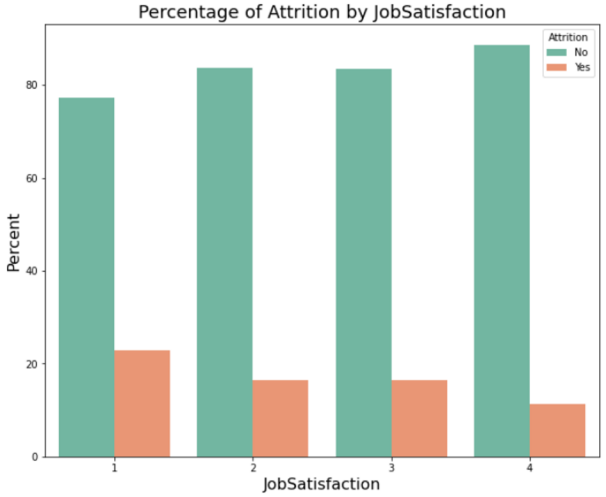
Possible reason:

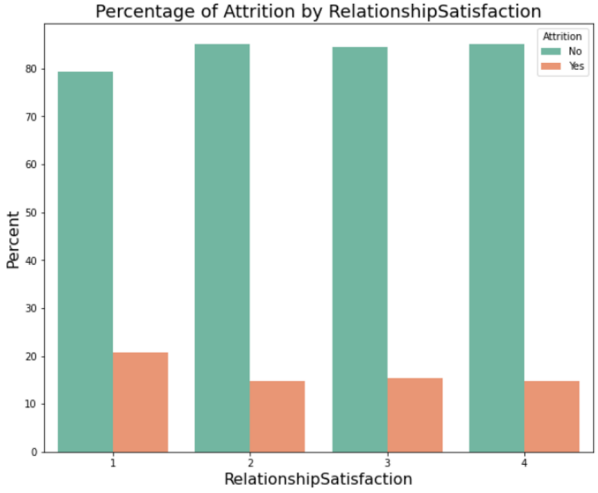
Employees with lower income would like to seek more opportunities outside to improve their salary level.

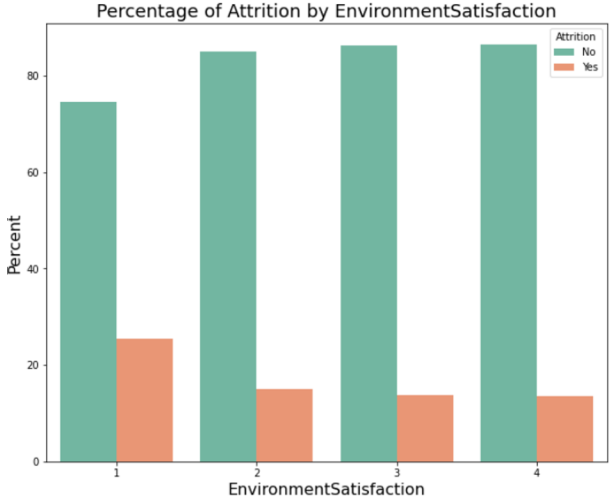
Employees with relatively high pay (about $10,000) are in management level, they may not be satisfied with their current work-life balance or salary level. HR should focus on this phenomenon.

IV. Data Exploration-Attrition

D. Satisfaction

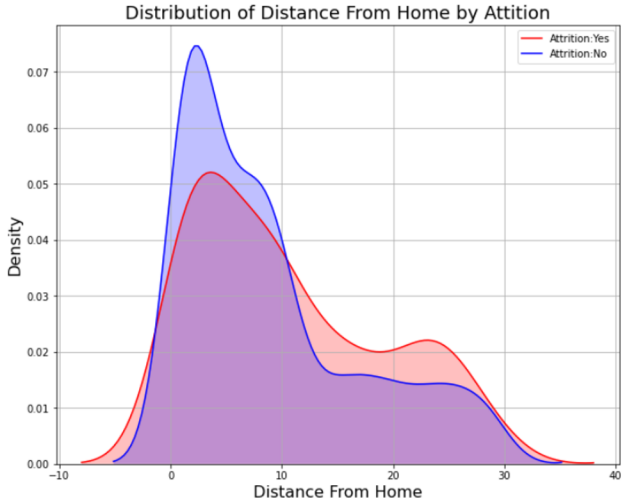


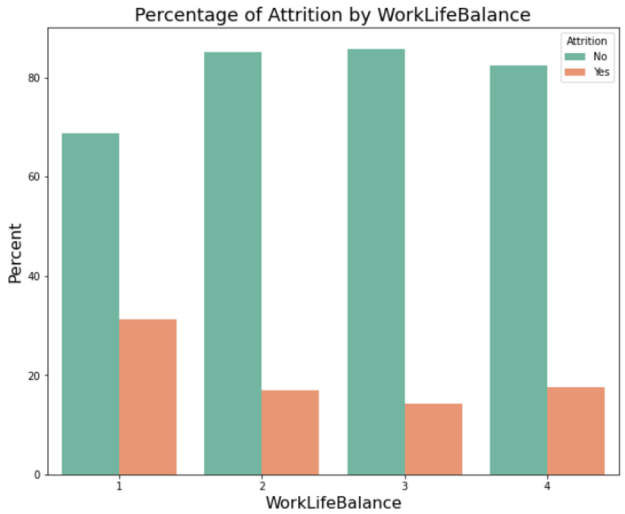


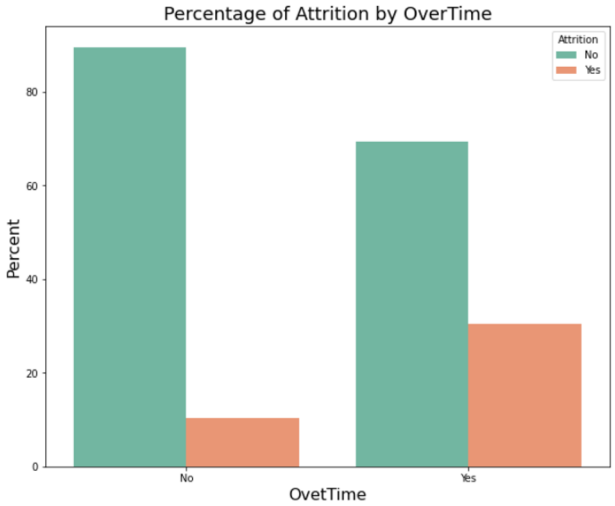


IV. Data Exploration-Attrition

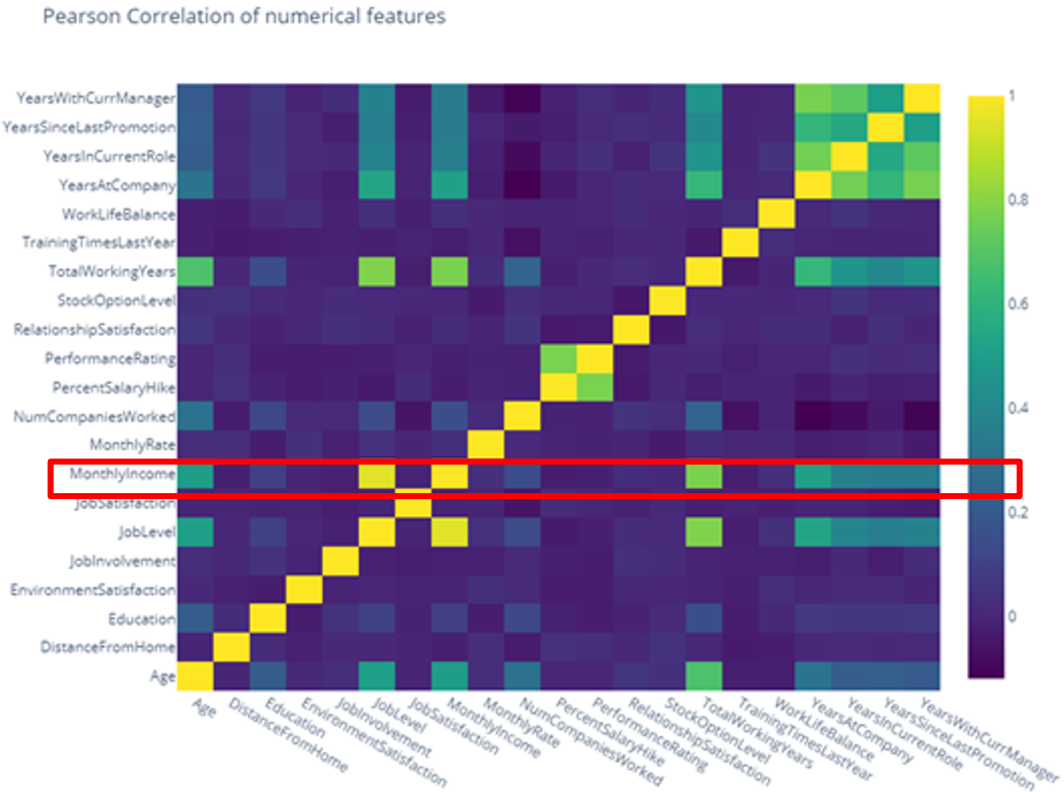
E. Workload







IV. Data Exploration-Monthly Income



Correlated Variables with Monthly Income:

- Age
- Education
- JobLevel
- TotalWorkingYears
- ...

Modelling

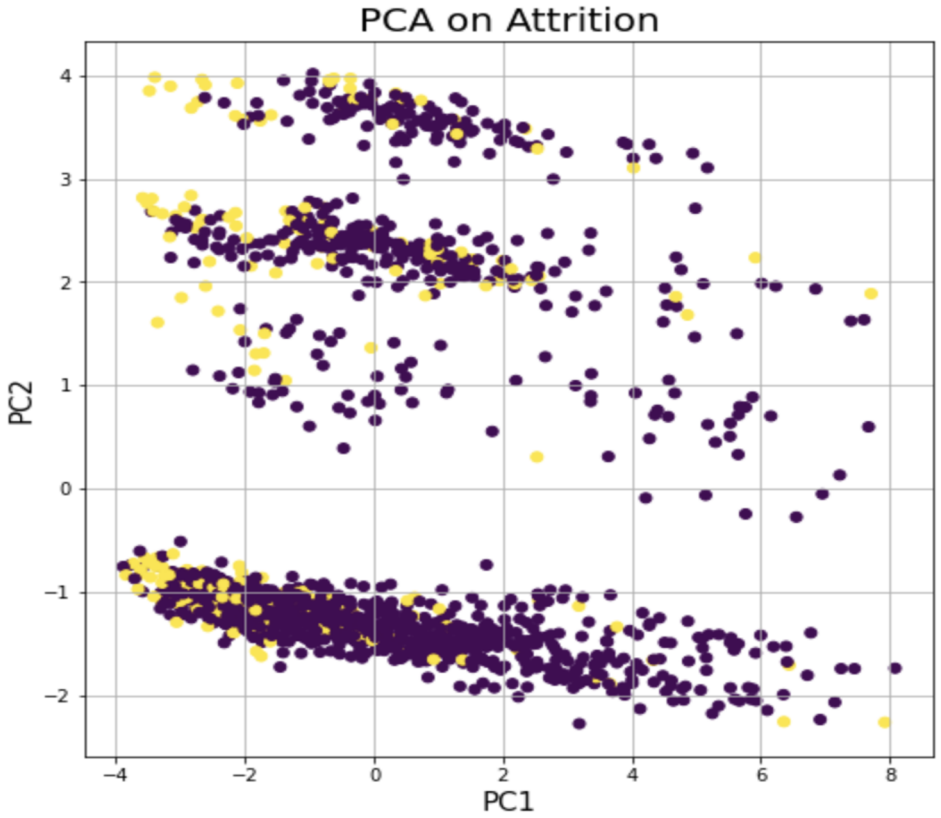
1. Attrition - Classification Model

2. Monthly Income - Regression Model

Classification

Principle Components Analysis

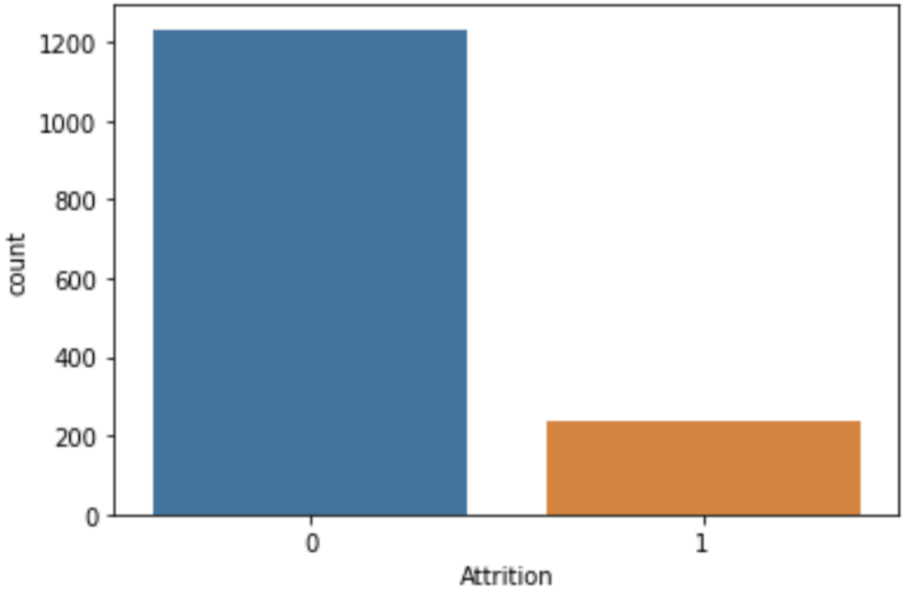
- Distribution
- Outliners
- Balance



Imbalance Problem? - Yes

- SMOTE Method

```
0     1233
1      237
Name: Attrition, dtype: int64
```



K-Nearest Neighbour

```
Test Result:

accuracy score: 0.6394557823129252

Classification Report:
        Precision: 0.1888111888111888
        Recall Score: 0.3857142857142857
        F1 score: 0.2535211267605634

Confusion Matrix:
 [[255 116]
 [ 43  27]]
```

Logistics Regression

```
Test Result:

accuracy score: 0.7505668934240363

Classification Report:
         Precision: 0.3275862068965517
         Recall Score: 0.5428571428571428
         F1 score: 0.4086021505376344

Confusion Matrix:
 [[293  78]
 [ 32  38]]
```

Decision Tree

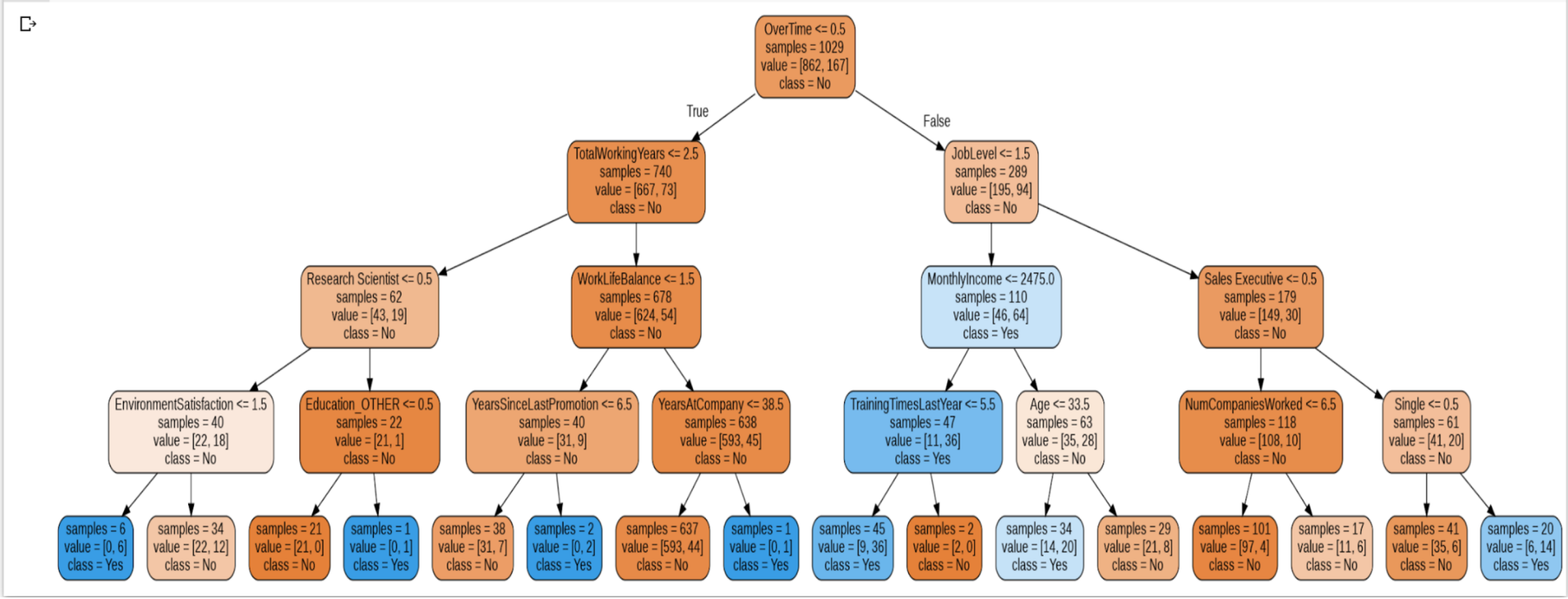
before tunning:

```
Test Result:

accuracy score: 0.7414965986394558

Classification Report:
        Precision: 0.3048780487804878
        Recall Score: 0.3048780487804878
        F1 score: 0.3048780487804878

Confusion Matrix:
 [[302  57]
 [ 57  25]]
```

after tunning:

```
Test Result:

accuracy score: 0.782312925170068

Classification Report:
        Precision: 0.27631578947368424
        Recall Score: 0.3387096774193548
        F1 score: 0.30434782608695654

Confusion Matrix:
 [[324  55]
 [ 41  21]]
```

Decision Tree Visualization



Random Forest

before tuning:

```
Test Result:

accuracy score: 0.8752834467120182

Classification Report:
        Precision: 0.629629629629297
        Recall Score: 0.27419354838709675
        F1 score: 0.3820224719101123

Confusion Matrix:
 [[369  10]
 [ 45  17]]
```

Random Forest hyperparameter tuning

----Randomized Search Cross Validation

```
Best paramters: {'n_estimators': 400, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features':
'sqrt', 'max_depth': None, 'bootstrap': False})
```

```
accuracy score: 0.8866213151927438

Classification Report:
        Precision: 0.6451612903225806
        Recall Score: 0.3389830508474576
        F1 score: 0.4444444444444444

Confusion Matrix:
 [[371  11]
 [ 39  20]]
```

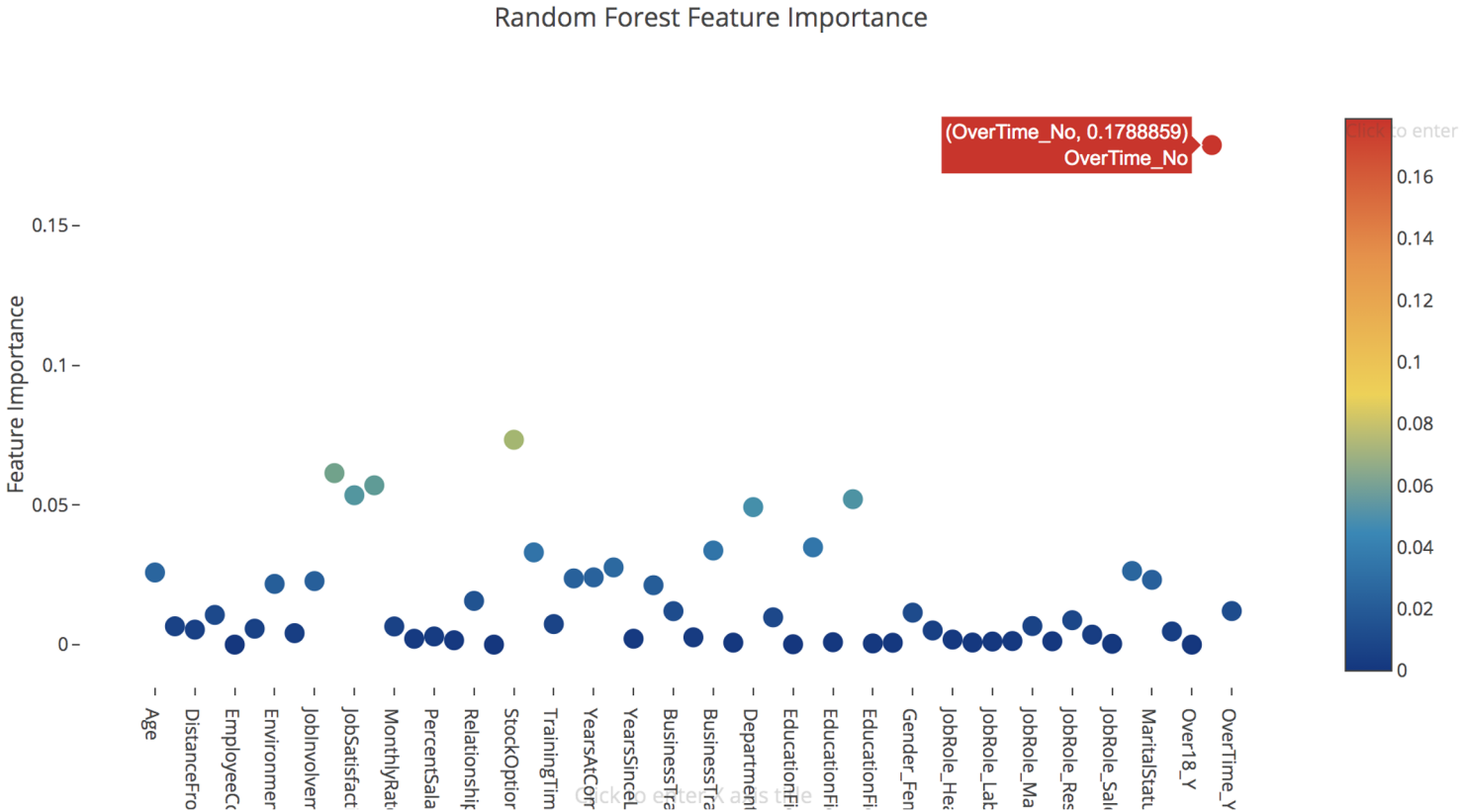
Random Forest hyperparameter tuning

----- Grid Search with Cross Validation

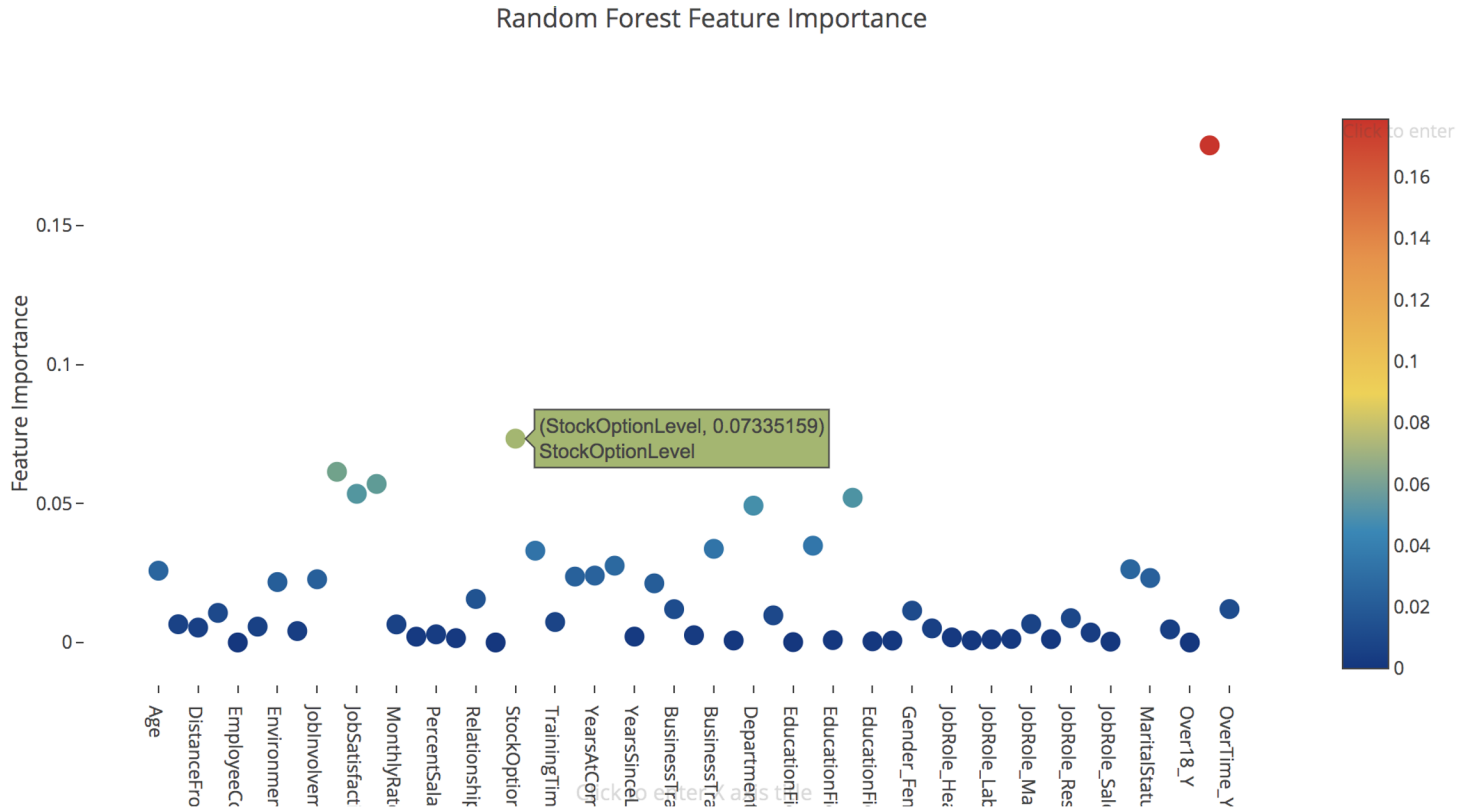
```
Best parameters: {'bootstrap': False, 'max_depth': None, 'max_features': 'auto', 'min_samples_leaf': 1,
'min_samples_split': 2, 'n_estimators': 1500}
```

```
accuracy score: 0.8888888888888888

Classification Report:
        Precision: 0.65625
        Recall Score: 0.3559322033898305
        F1 score: 0.4615384615384616

Confusion Matrix:
 [[371  11]
 [ 38  21]]
```

Feature Importance



Feature Importance



Gradient Boosting

```
Test Result:

accuracy score: 0.8639455782312925

Classification Report:
        Precision: 0.7391304347826086
        Recall Score: 0.4146341463414634
        F1 score: 0.53125

Confusion Matrix:
 [[347  12]
 [ 48  34]]
```

```
Test Result:

accuracy score: 0.8662131519274376

Classification Report:
        Precision: 0.7555555555555555
        Recall Score: 0.4146341463414634
        F1 score: 0.5354330708661418

Confusion Matrix:
 [[348  11]
 [ 48  34]]
```

Regression for Salary

Variables:

Variable Name	Description
MonthlyIncome	Monthly salary of employee
Age	Age of employee
Education	Education level: 1-5 (lowest to highest level)
JobInvolvement	Involvement in the job: 1-4 (lowest to highest level)
JobLevel	Job level: 1-5 (lowest to highest level)
MonthlyRate	Monthly internal charge out rate
NumCompaniesWorked	Number of companies worked at
PerformanceRating	Performance rating: 3-4 (lowest to highest level)
StockOptionLevel	Stock option level: 0-3 (lowest to highest level)
TotalWorkingYears	Total years worked
TrainingTimesLastYear	Hours spent training
WorkLifeBalance	Work and life balance level: 1-4 (lowest to highest level)
YearsAtCompany	Total number of years at the company
YearsInCurrentRole	Number of years in current role
YearsSinceLastPromotion	Number of years since last promotion
YearsWithCurrManager	Number of years spent with current manager
Department	Department which employee works in
EducationField	Education field
Gender	Female; Male
JobRole	Job position
MaritalStatus	Marital Status of employee
OverTime	Yes if employee works overtime, No otherwise

Reference group:

Department-HR, Education field-HR, Job role- HR, Marital status-single

Method: linear regression, ridge regression, lasso regression, elastic net, random forest

*Note: data are scaled

Linear Regression

$$\min_{w} ||Xw - y||_2^2$$

Holding all other regressors fixed, an unit increase in job level could lead to 3201 dollars increase in monthly income on average.

A manager earns 1005 dollars more than an HR monthly on average and holding other regressors constant.

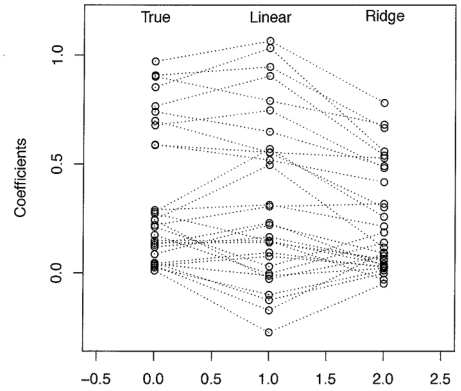
An employee in R & D department earns 282 dollars more than an employee in HR department on average and holding other variables constant.

etc...

Age	-10.296672
Education	8.326474
JobInvolvement	-50.252396
JobLevel	3201.006227
MonthlyRate	-28.921052
NumCompaniesWorked	40.794858
PerformanceRating	-25.922504
StockOptionLevel	-2.748921
TotalWorkingYears	251.697478
TrainingTimesLastYear	-0.032968
WorkLifeBalance	-20.859778
YearsAtCompany	64.285717
YearsInCurrentRole	-6.293626
YearsSinceLastPromotion	69.311207
YearsWithCurrManager	-98.714999
Research & Development	282.318608
Sales	129.056564
Life Sciences	-148.281251
Marketing	-56.449361
Medical	-105.383649
Education_OTHER	-103.151592
Technical Degree	-55.799615
Healthcare Representative	-138.137068
Laboratory Technician	-365.662100
Manager	1005.024487
Manufacturing Director	-130.020161
Research Director	827.399748
Research Scientist	-325.039730
Sales Executive	-63.097193
Sales Representative	-184.728449
Female	-27.779905
Divorced	-14.117592
Married	-2.689789
OverTime_YES	26.512325

Ridge Regression

$$\min_w ||Xw - y||_2^2 + \alpha||w||_2^2$$



Age	52.023960
Education	-10.951790
JobInvolvement	-51.333914
JobLevel	2385.608492
MonthlyRate	-10.300375
NumCompaniesWorked	32.856293
PerformanceRating	-34.004723
StockOptionLevel	-11.059505
TotalWorkingYears	558.610978
TrainingTimesLastYear	-2.134151
WorkLifeBalance	-1.200961
YearsAtCompany	149.872666
YearsInCurrentRole	-5.887883
YearsSinceLastPromotion	87.425235
YearsWithCurrManager	-104.569199
Research & Development	146.930742
Sales	54.805962
Life Sciences	-13.049549
Marketing	20.064324
Medical	13.608003
Education_OTHER	-30.453012
Technical Degree	12.592699
Healthcare Representative	-48.355952
Laboratory Technician	-429.975673
Manager	1151.456257
Manufacturing Director	-29.827198
Research Director	982.133430
Research Scientist	-421.969274
Sales Executive	25.576258
Sales Representative	-240.901567
Female	-18.699556
Divorced	-13.036250
Married	1.401758
OverTime_YES	23.310644

Lasso Regression

$$\min_w \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha||w||_1$$

Several variables whose coefficients are significant:

job level,
total working years,
department,
job role

Age	1.578464
Education	5.496498
JobInvolvement	0.000000
JobLevel	3231.912983
MonthlyRate	0.000000
NumCompaniesWorked	49.099602
PerformanceRating	0.000000
StockOptionLevel	0.000000
TotalWorkingYears	224.479003
TrainingTimesLastYear	0.000000
WorkLifeBalance	0.000000
YearsAtCompany	0.000000
YearsInCurrentRole	0.000000
YearsSinceLastPromotion	64.793396
YearsWithCurrManager	0.000000
Research & Development	0.000000
Sales	0.000000
Life Sciences	0.000000
Marketing	23.139097
Medical	44.718949
Education_OTHER	0.000000
Technical Degree	25.700620
Healthcare Representative	114.387153
Laboratory Technician	0.000000
Manager	1188.093238
Manufacturing Director	121.607280
Research Director	1027.064908
Research Scientist	48.445095
Sales Executive	174.818764
Sales Representative	0.000000
Female	0.000000
Divorced	0.000000
Married	6.427241
OverTime_YES	29.089828

Elastic Net

$$\min_{w} \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$$

Variables significant in 4 models::

job level,
total working years,
department,
job role

Age	-6.316050
Education	4.761486
JobInvolvement	-51.064703
JobLevel	3055.641088
MonthlyRate	-25.988380
NumCompaniesWorked	37.603849
PerformanceRating	-27.840989
StockOptionLevel	-3.240261
TotalWorkingYears	310.608189
TrainingTimesLastYear	-0.689665
WorkLifeBalance	-18.030834
YearsAtCompany	75.881314
YearsInCurrentRole	-6.813314
YearsSinceLastPromotion	71.832552
YearsWithCurrManager	-102.664887
Research & Development	212.046446
Sales	63.258229
Life Sciences	-75.266254
Marketing	-11.143870

Medical	-38.285350
Education_OTHER	-65.925430
Technical Degree	-15.528571
Healthcare Representative	-96.487913
Laboratory Technician	-347.763813
Manager	1064.914868
Manufacturing Director	-85.587931
Research Director	885.392858
Research Scientist	-310.591981
Sales Executive	-3.494058
Sales Representative	-173.051465
Female	-27.700522
Divorced	-14.145645
Married	-2.058166
OverTime_YES	26.111346

Model Comparison

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Linear Regression	880.726445	1.317020e+06	1147.614953	0.934883	0.939593
1	Ridge Regression	914.063591	1.356797e+06	1164.816503	0.932916	0.939586
2	Lasso Regression	877.946050	1.320198e+06	1148.998481	0.934726	0.939794
3	Elastic Net Regression	877.313396	1.298775e+06	1139.638181	0.935785	0.831882
4	Random Forest Regressor	817.826522	1.233829e+06	1110.778668	0.938996	0.000000

n estimator=1000

```
MAE: 836.0955102040816
MSE: 1275935.7541968254
RMSE: 1129.5732619873868
R2 Square 0.9369139566176896
```

n estimator=50

```
MAE: 821.3025850340138
MSE: 1245844.4766988663
RMSE: 1116.1740351302149
R2 Square 0.9384017585163521
```

n estimator=100

```
MAE: 816.6204336734694
MSE: 1234311.5845236005
RMSE: 1110.9957626038006
R2 Square 0.9389719788692964
```

n estimator=800

Conclusion

1. Importance of the Analysis
2. Features of employees with higher attrition probablity
3. Models & Results
4. Takeaways
5. Future Work

Importance of the analysis

Features of employees with higher attrition probability

- Employees who are younger(18-30 years)
- Employees who have worked in the company for a shorter time(0-4 years)
- Employees in lower job level
- Employees with lower pay (under $4000 per month) & relatively high pay(around $10000 per month)
- Employees with lower satisfaction of job, relationship and environment
- Employees who work farther away from home
- Employees with lower level of work life balance
- Employees who work over time
- Employees in sales department

Models and Results

Classification part:

logistic regression, KNN, decision tree, gradient boosting and random forest

Regression part:

linear regression, ridge, lasso, elastic net and random forest regressor

Takeaways

- Pay more attention to young employees, employees who have just joined and senior managers, and try to develop a more reasonable benefit system for each type.
- Hold one-on-one meetings regularly, in order to know employees better and solve potential problems that would lead to attrition.
- Set optimum workload for employees, and improve the treatment for employees who have to work over time occasionally/usually.
- Consider a more flexible working time or regulation to facilitate employees to balance work and life.
- Several factors should be taken into considerations when devise the salary system:
  job level, numbers of working year, job role, and whether work over time

Future Work

- Tune parameters to make regression model better for interpretation.
- Try out some advanced models to predict, such as neural network.
- Take industry difference into account and look into different types of companies.
- Find out whether area/region would hugely affect the results.
- etc...

Thank you!