Employee Attrition

Statistical Learning MOD B, 2022/2023

Anna Cerbaro, Erica Marras, Eleni Papadopulos

Problem Presentation

Dataset	Employee Attrition and Factors

Why	• Employee turnover is expensive • Disruption of the workplace stability and productivity

Goal	Predicting if an employee leaves or not

Dataset

Sythetically generated by **IBM**

1470 instances

33 features

Target variable

Nominal

1. **Attrition**
2. Department
3. EducationField
4. Gender
5. JobRole
6. MaritalStatus
7. ~~Over18~~
8. OverTime
9. BusinessTravel

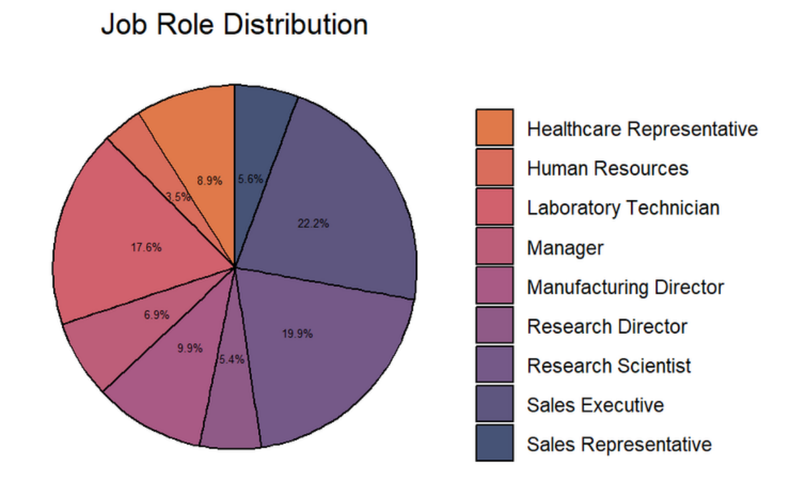
Ordinal

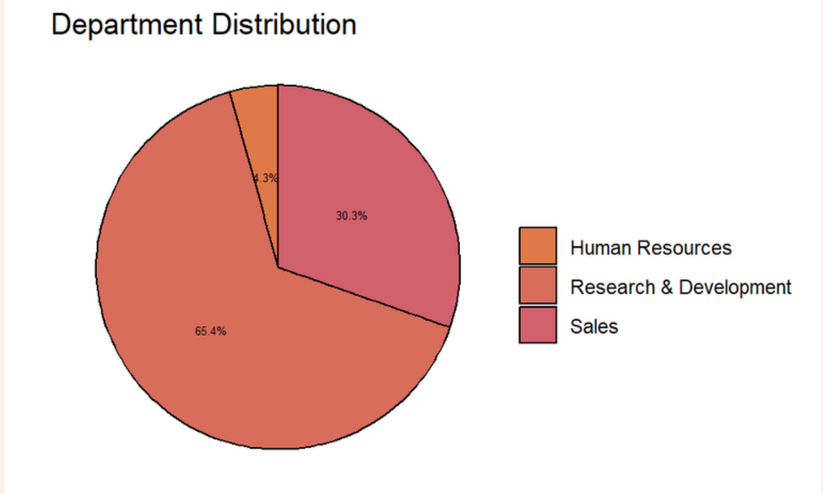
10. Education
11. EnvironmentSatisfaction
12. JobInvolvement
13. JobLevel
14. JobSatisfaction
15. PerformanceRating
16. RelationshipSatisfaction
17. StockOptionLevel
18. WorkLifeBalance

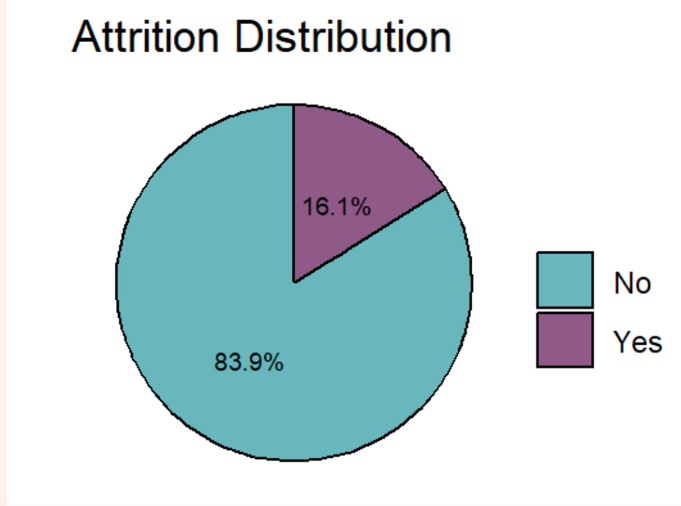
Numerical

19. MonthlyIncome
20. DistanceFromHome
21. ~~EmployeeCount~~
22. ~~EmployeeNumber~~
23. EnvironmentSatisfaction
24. *NumCompaniesWorked*
25. PercentSalaryHike
26. ~~StandardHours~~
27. TotalWorkingYears
28. TrainingTimesLastYear
29. WorkLifeBalance
30. YearsAtCompany
31. YearsInCurrentRole
32. YearsSinceLastPromotion
33. YearsWithCurrManager

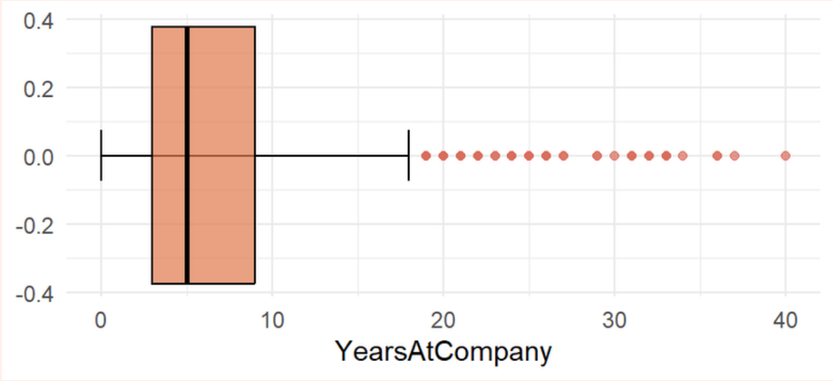
More about the company





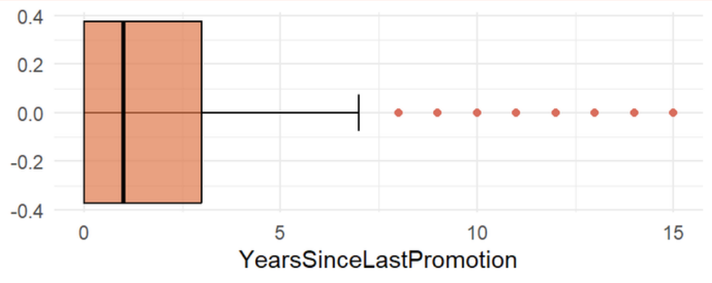


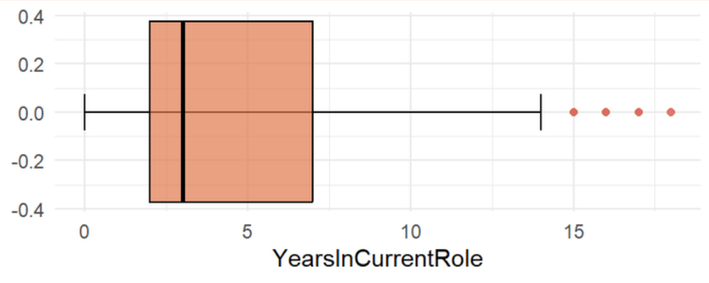
The company appears to be in the pharmaceutical, healthcare, or biotechnology industry. It is not a newly established company since some employees have been working there for almost 40 years.

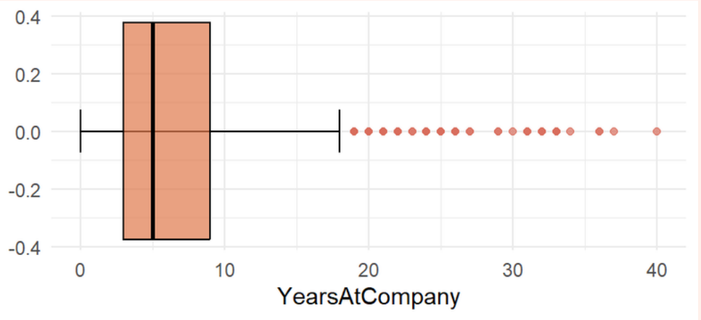


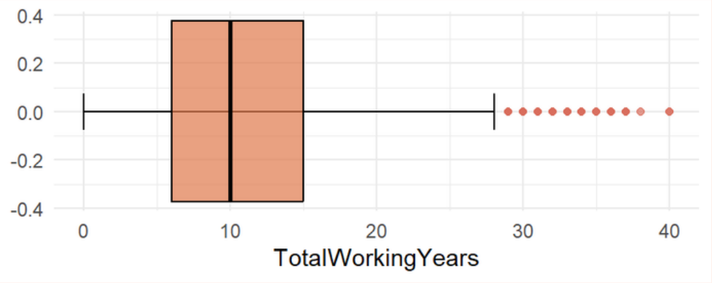
In the company at hand, this means that 16% of employees have left the company at some point during the time covered by the dataset

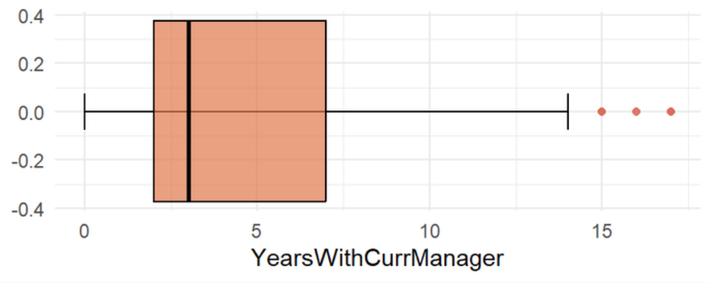
Feature Outliers

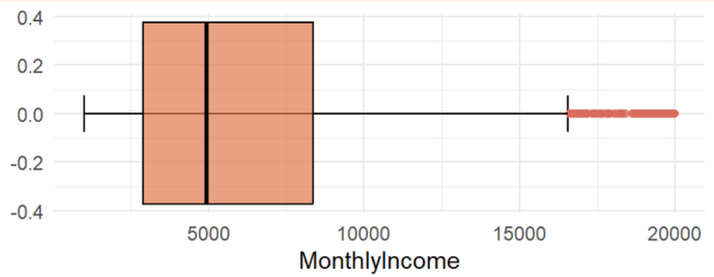


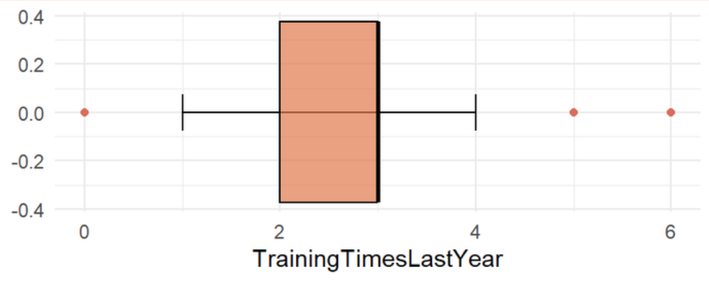


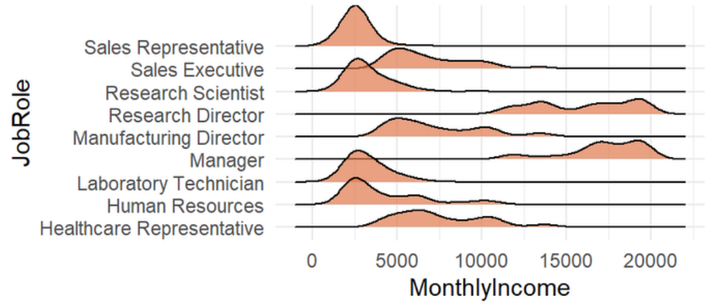




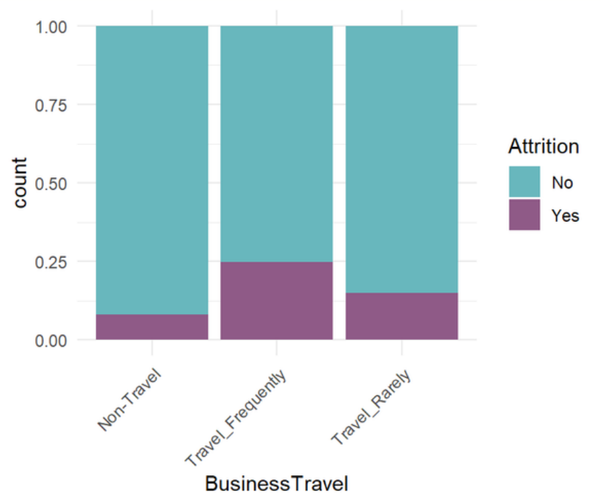


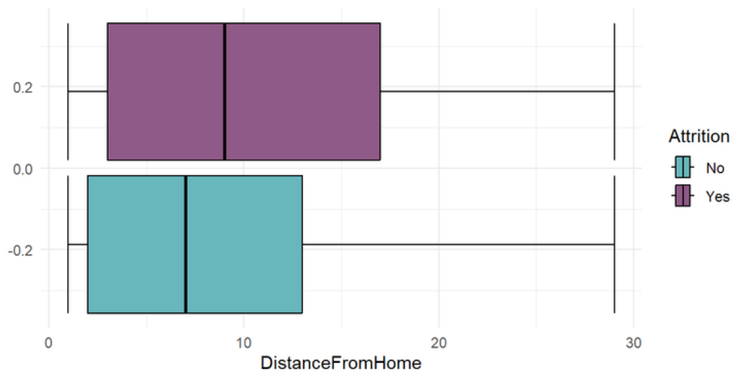


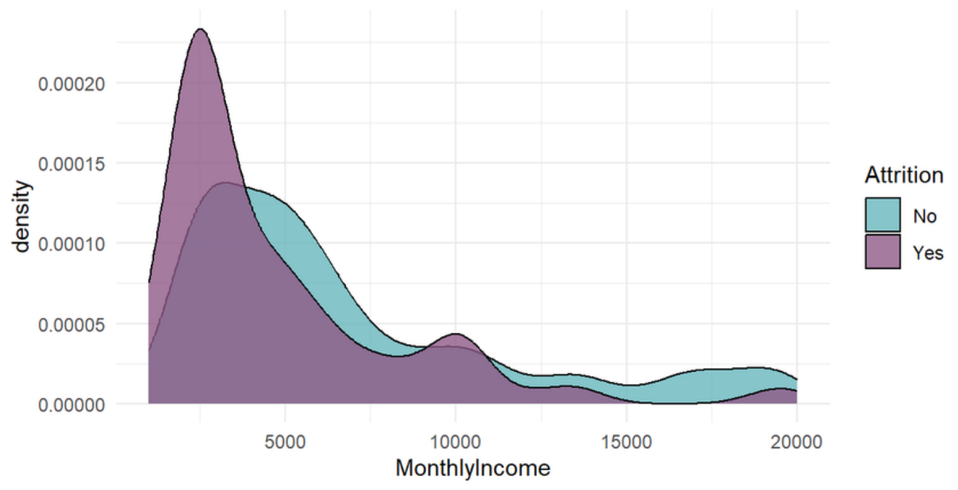


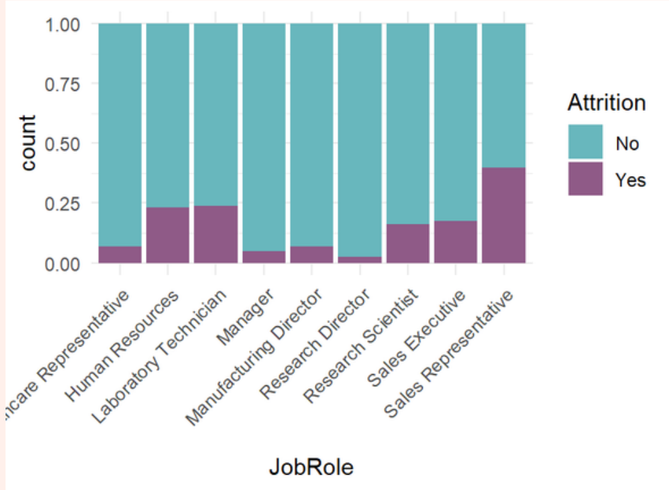


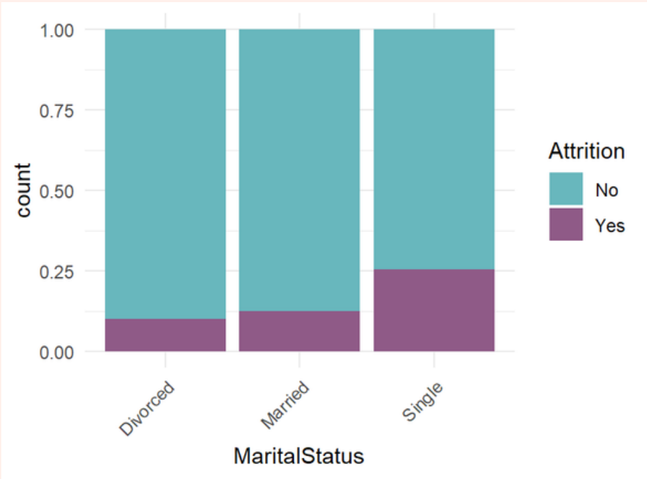
Some feature distributions wrt Attrition

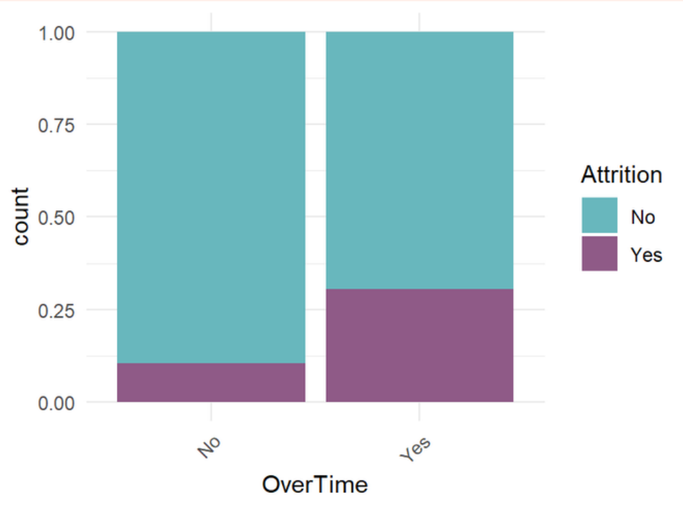




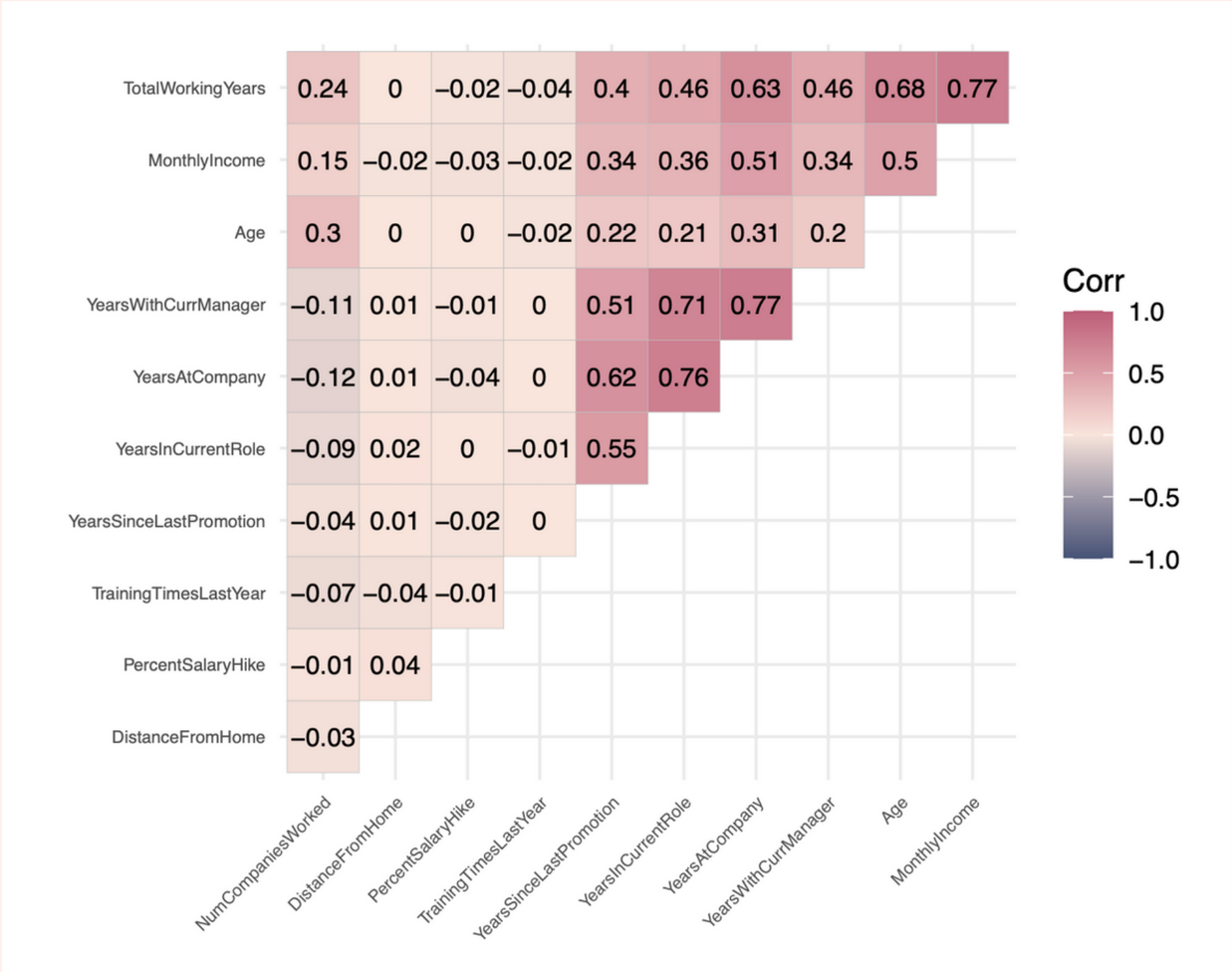








Correlation matrix



Correlation among categorical

Cramer's V

```
## [1] "Attrition-BusinessTravel           0.1283"
## [1] "Attrition-Department               0.0857"
## [1] "Attrition-EducationField           0.1044"
## [1] "Attrition-Gender                   0.02945"
## [1] "Attrition-MaritalStatus            0.1772"
```

```
## [1] "Attrition-OverTime                 0.2461"
## [1] "Attrition-JobRole                  0.2421"
## [1] "Attrition-Education                0.04573"
## [1] "Attrition-JobLevel                 0.2221"
## [1] "Attrition-StockOptionLevel         0.203"
```

```
## [1] "Attrition-PerformanceRating        0.002889"
## [1] "Attrition-JobInvolvement           0.1392"
## [1] "Attrition-JobSatisfaction          0.1091"
## [1] "Attrition-EnvironmentSatisfaction  0.1237"
## [1] "Attrition-RelationshipSatisfaction 0.05971"
```

Questions

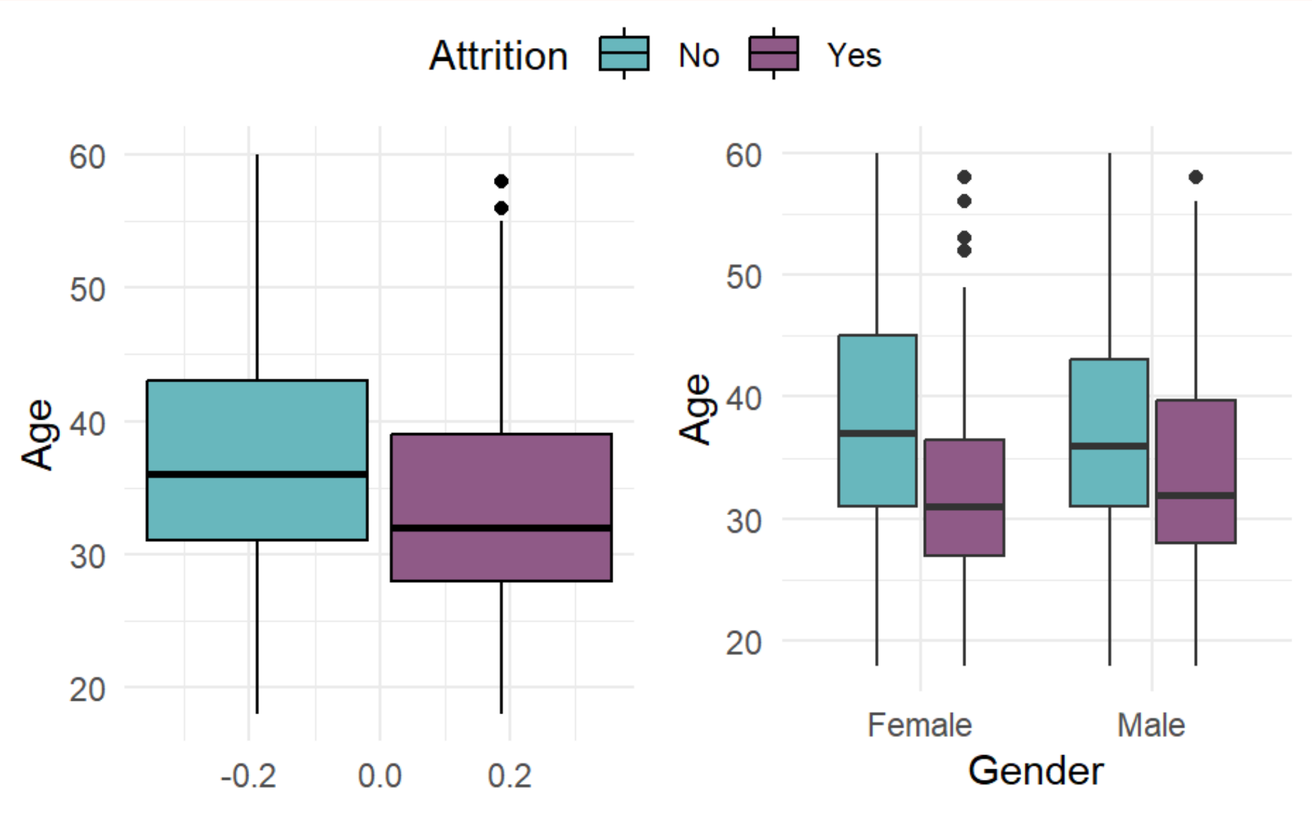
1. What is the profile of a leaving employee?

2. Does employee well-being impact the choice to leave?

3. Does the decision to work overtime affect the choice to leave the company?

4. Which are the most critical roles?

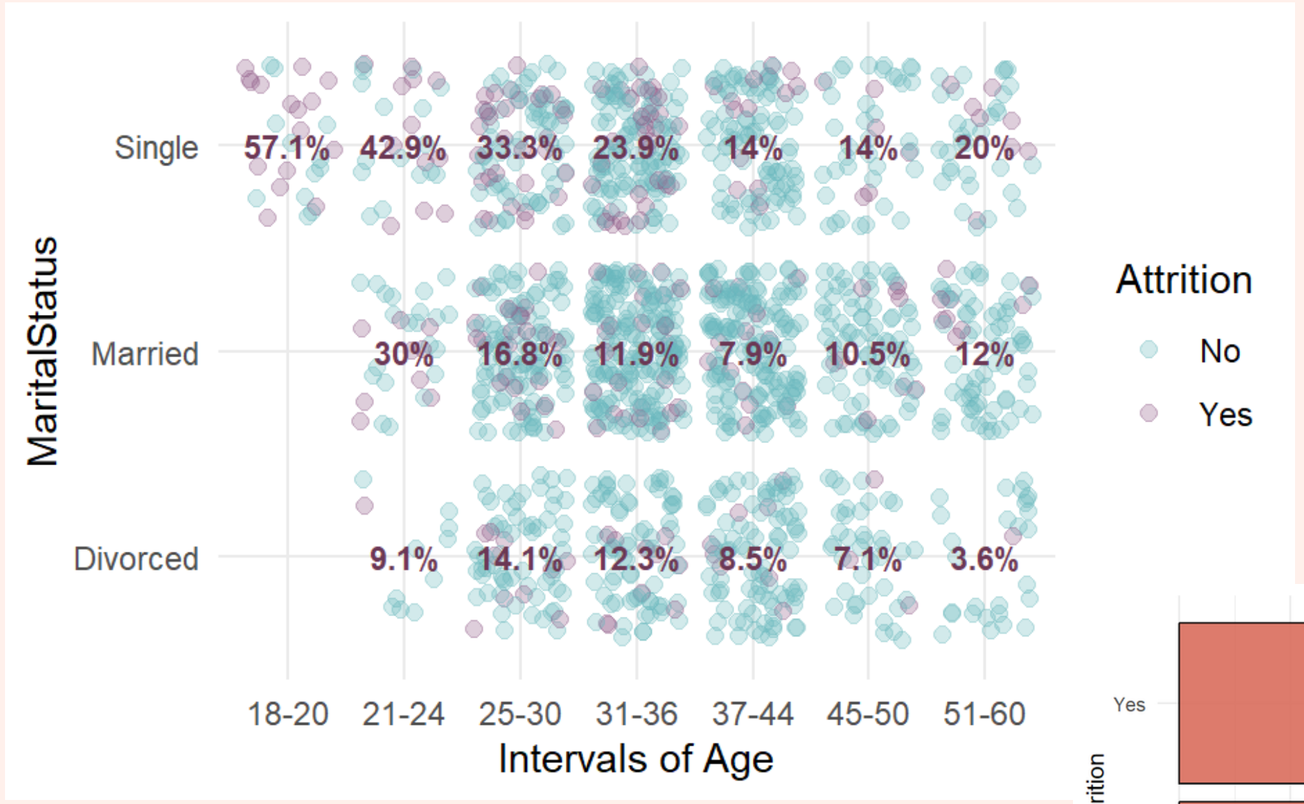
1. What is the profile of a leaving employee?



Who leaves is mostly:

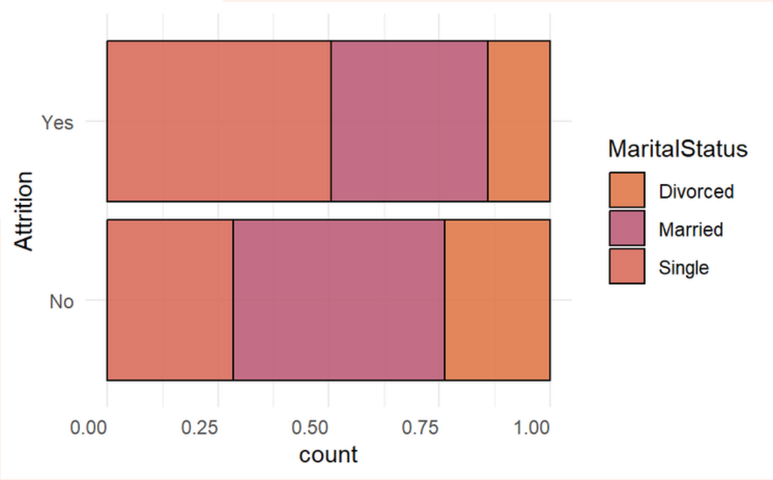
- **younger**

1. What is the profile of a leaving employee?



Who leaves is mostly:

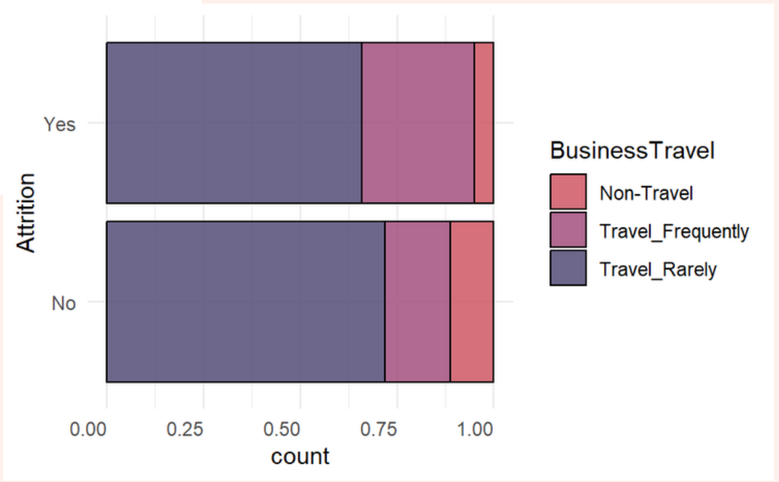
- **younger**
- **single**



1. What is the profile of a leaving employee?

Who leaves is mostly:

- **younger**
- **single**
- a **frequent traveler**

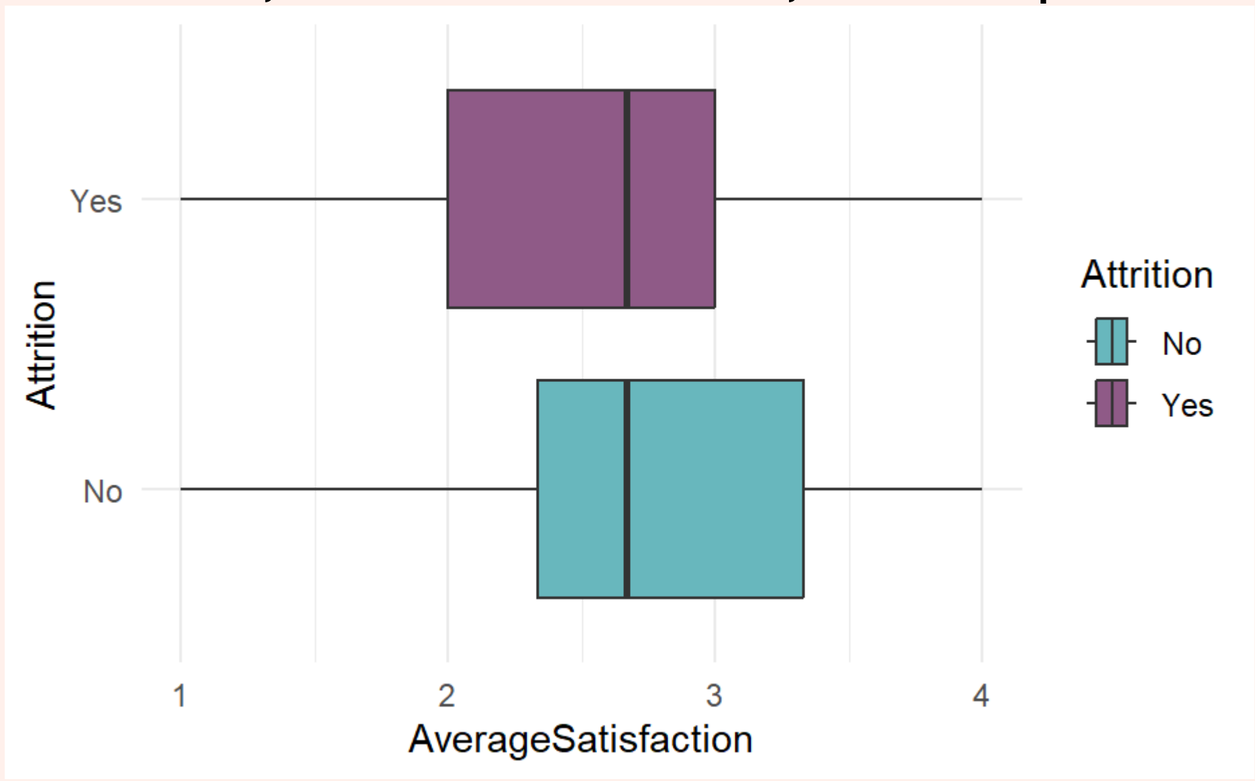


2. Does employee well-being impact the choice to leave?

Features that can express employee well-being

- Job Satisfaction
- Environment Satisfaction
- Relationship Satisfaction
- Work-Life Balance
- Job Involvement

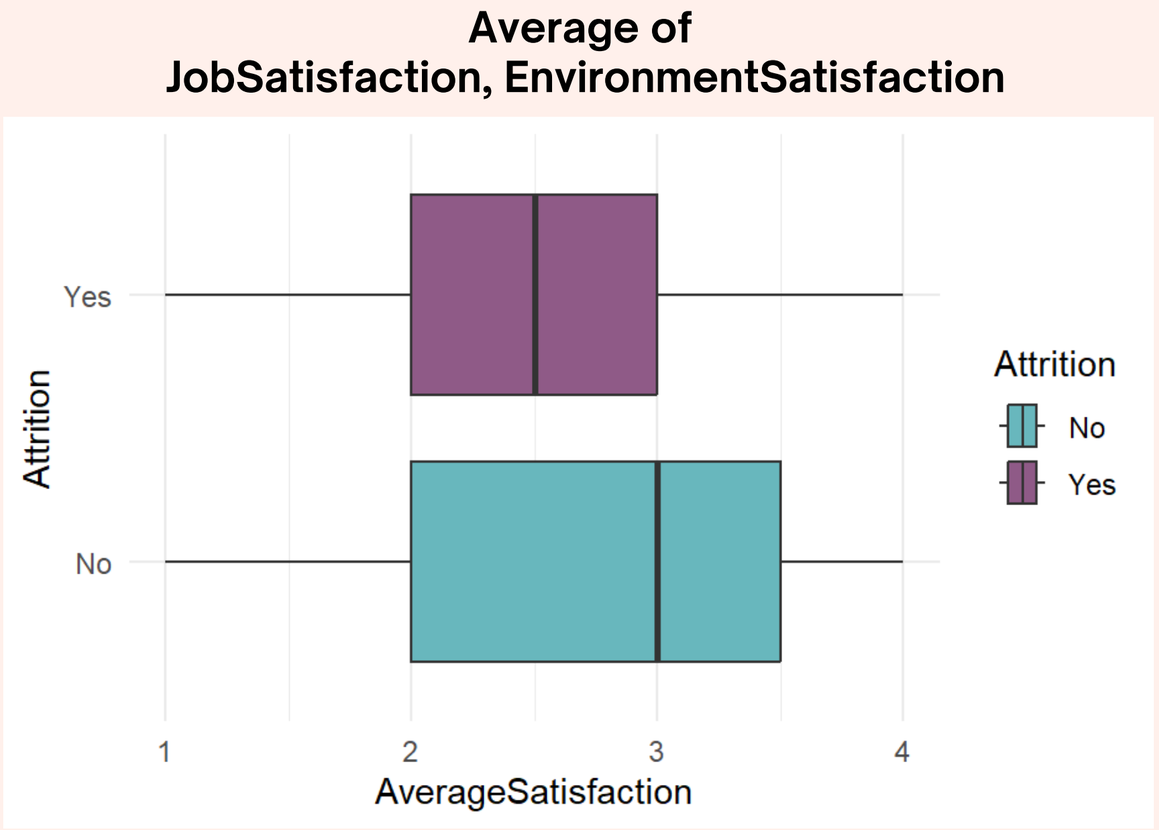
Average of JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction



2. Does employee well-being impact the choice to leave?

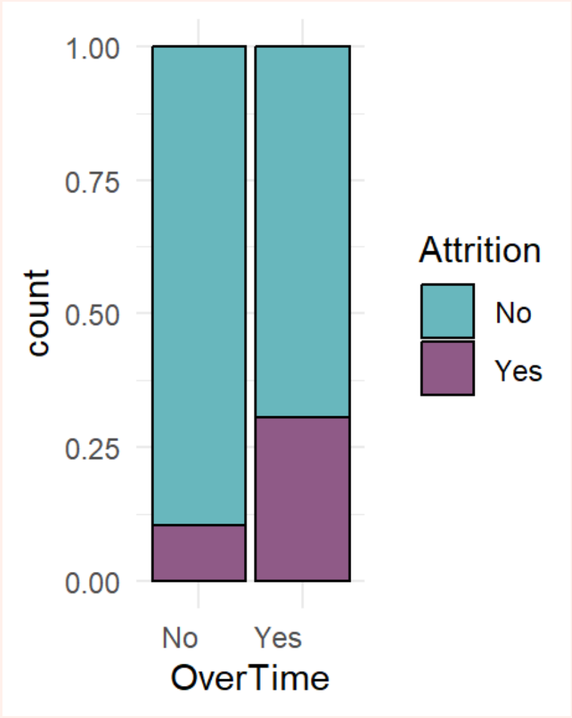
Results of chi-square test with Attrition

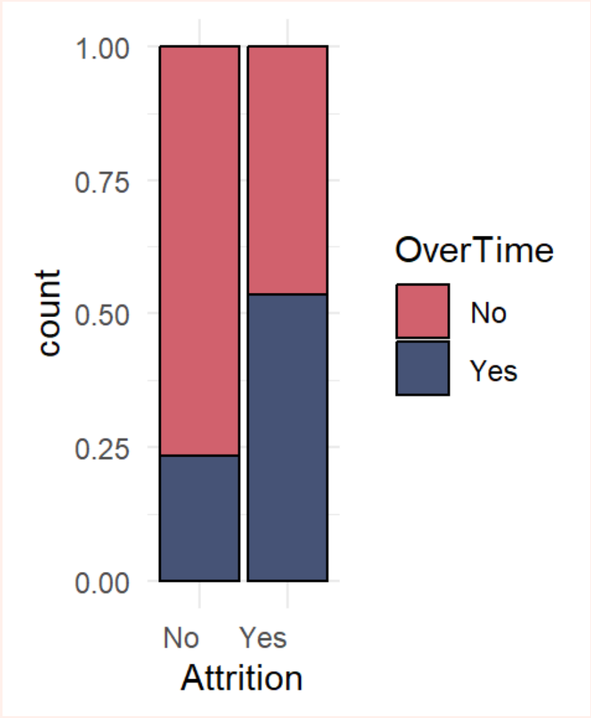
Job Satisfaction	p-value < **0.000556**
Environment Satisfaction	p-value < **0.000051**
Relationship Satisfaction	p-value = 0.154972
Work-Life Balance	p-value < **0.00097**
Job Involvement	p-value < **0.000002**



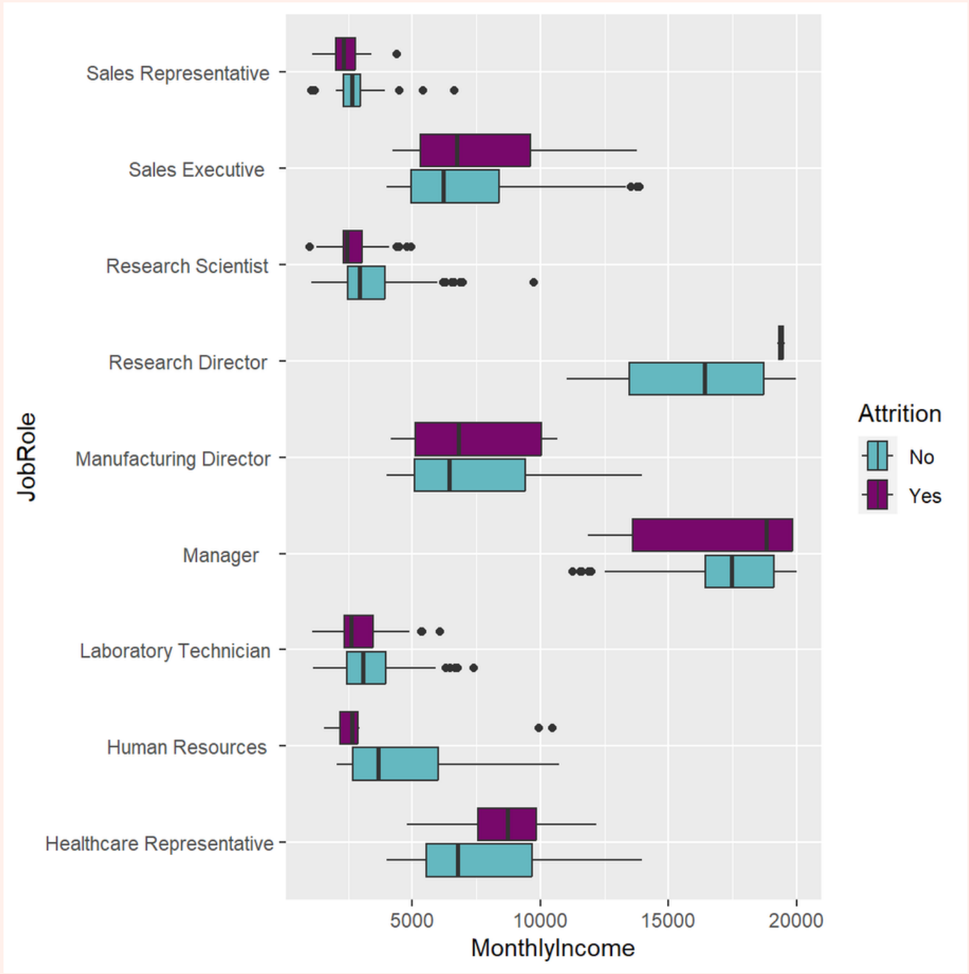
3. Does the decision to work overtime affect on the choice to leave the company?

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(HR$Attrition, HR$OverTime)
## X-squared = 87.564, df = 1, p-value < 0.00000000000000022
```

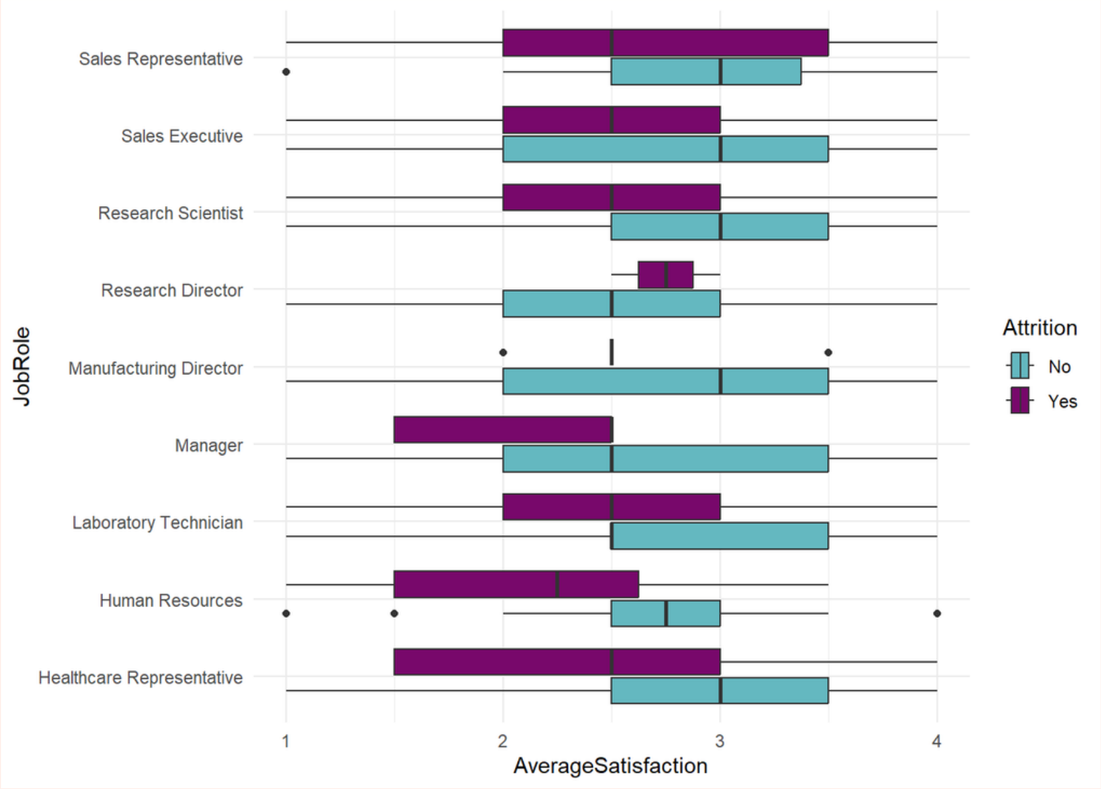




4. Which are the most critical roles?



4. Which are the most critical roles?



Classification

The classification models proposed are:

- Logistic Regression
- Ridge Regularization
- Lasso Regularization
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naive Bayes
- KNN

FOCUS ON Recall

Data preparation

Training set	Test set
75	25

Since our dataset is unbalanced, we need to apply some sampling techniques to achieve a 50/50 class ratio in the training set.

We employed:

- sampling keeping the original training set size and p = 0.5
- **oversampling**
- SMOTE

Logistic Regression

We compute the VIF (Variance Inflation Factor) to observe the presence of any collinearity among the variables, removing any predictor having value over 5.

Age	BusinessTravel	Department	DistanceFromHome	Education
1.981978	1.114924	1.286799	1.084124	1.084458
EducationField	EnvironmentSatisfaction	Gender	JobInvolvement	JobLevel
1.061957	1.089613	1.061261	1.059996	8.943373
JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome	NumCompaniesWorked
1.287902	1.083157	1.857679	8.121151	1.369443
OverTime	PercentSalaryHike	PerformanceRating	RelationshipSatisfaction	StockOptionLevel
1.202241	2.562508	2.552855	1.108647	1.822890
TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole
4.477195	1.061533	1.106284	5.368819	2.756052
YearsSinceLastPromotion	YearsWithCurrManager			
2.327484	2.749695			

We iteratively delete the variable with the highest VIF until any value is under 5. We ended up eliminating *JobLevel* and *YearsAtCompany*.

Logistic Regression

Feature selection → **Backward selection**

Forward selection

The algorithm removed:

- Education
- PerformanceRating
- StockOptionLevel

Logistic Regression

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	5.478e+00	1.038e+00	5.279	1.30e-07	***
Age	-3.142e-02	1.588e-02	-1.979	0.047829	*
BusinessTravel	9.336e-01	2.089e-01	4.469	7.86e-06	***
Department	-6.857e-01	2.200e-01	-3.118	0.001823	**
DistanceFromHome	3.344e-02	1.289e-02	2.593	0.009503	**
EducationField	1.357e-01	7.513e-02	1.806	0.070864	.
EnvironmentSatisfaction	-4.177e-01	9.907e-02	-4.217	2.48e-05	***
Gender	-5.659e-01	2.248e-01	-2.517	0.011827	*
JobInvolvement	-5.484e-01	1.450e-01	-3.782	0.000155	***
JobRole	8.456e-02	5.316e-02	1.591	0.111694	
JobSatisfaction	-3.232e-01	9.842e-02	-3.284	0.001025	**
MaritalStatus	-6.942e-01	1.526e-01	-4.549	5.39e-06	***
MonthlyIncome	-1.019e-04	4.299e-05	-2.371	0.017757	*
NumCompaniesWorked	1.557e-01	4.675e-02	3.330	0.000869	***
OverTime	2.057e+00	2.292e-01	8.974	< 2e-16	***
PercentSalaryHike	-4.826e-02	3.029e-02	-1.593	0.111073	
RelationshipSatisfaction	-1.613e-01	1.010e-01	-1.597	0.110188	
TotalWorkingYears	-4.941e-02	3.003e-02	-1.645	0.099898	.
TrainingTimesLastYear	-1.549e-01	8.986e-02	-1.724	0.084669	.
WorkLifeBalance	-5.425e-01	1.462e-01	-3.711	0.000206	***
YearsInCurrentRole	-9.314e-02	5.285e-02	-1.762	0.078009	.
YearsSinceLastPromotion	1.978e-01	4.784e-02	4.134	3.57e-05	***
YearsWithCurrManager	-9.085e-02	5.571e-02	-1.631	0.102930	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 879.09 on 954 degrees of freedom
Residual deviance: 590.36 on 932 degrees of freedom
AIC: 636.36

Accuracy	Precision	Recall	F1 score
0.8459119	0.5000000	0.4285714	0.4615385

predicted.classes	0	1
0	248	28
1	21	21

Threshold: 0.45

Logistic Regression (oversampled)

Feature selection → **Backward selection**

Forward selection

The algorithm removed:

- PerformanceRating
- StockOptionLevel
- TotalWorkingYears

Logistic Regression (oversampled)

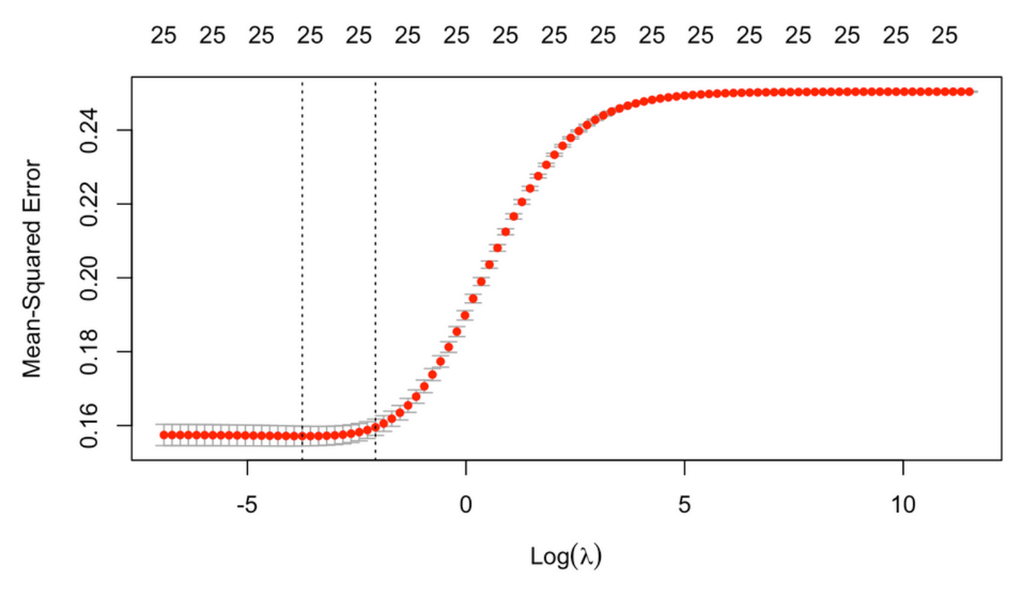
```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              7.136e+00  6.654e-01  10.724  < 2e-16 ***
Age                     -2.808e-02  8.462e-03  -3.319 0.000903 ***
BusinessTravel           9.670e-01  1.335e-01   7.244 4.35e-13 ***
Department              -5.680e-01  1.398e-01  -4.063 4.85e-05 ***
DistanceFromHome         3.241e-02  8.456e-03   3.832 0.000127 ***
Education               -1.829e-01  6.732e-02  -2.717 0.006580 **
EducationField           1.530e-01  4.613e-02   3.317 0.000911 ***
EnvironmentSatisfaction -3.611e-01  6.050e-02  -5.969 2.39e-09 ***
Gender                  -4.963e-01  1.417e-01  -3.503 0.000460 ***
JobInvolvement          -5.496e-01  9.363e-02  -5.869 4.37e-09 ***
JobRole                  8.467e-02  3.461e-02   2.446 0.014438 *
JobSatisfaction         -3.468e-01  6.168e-02  -5.622 1.89e-08 ***
MaritalStatus           -7.728e-01  9.438e-02  -8.188 2.66e-16 ***
MonthlyIncome           -1.341e-04  2.134e-05  -6.285 3.28e-10 ***
NumCompaniesWorked       1.346e-01  2.907e-02   4.629 3.68e-06 ***
OverTime                 1.955e+00  1.434e-01  13.630  < 2e-16 ***
PercentSalaryHike       -4.718e-02  1.869e-02  -2.524 0.011594 *
RelationshipSatisfaction -1.224e-01  6.242e-02  -1.961 0.049841 *
TrainingTimesLastYear   -2.113e-01  5.750e-02  -3.675 0.000238 ***
WorkLifeBalance         -5.509e-01  8.949e-02  -6.156 7.48e-10 ***
YearsInCurrentRole      -8.600e-02  3.332e-02  -2.581 0.009862 **
YearsSinceLastPromotion  1.998e-01  2.877e-02   6.946 3.76e-12 ***
YearsWithCurrManager    -1.117e-01  3.243e-02  -3.446 0.000569 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2190.3  on 1579  degrees of freedom
Residual deviance: 1424.5  on 1557  degrees of freedom
AIC: 1470.5
```

Accuracy	Precision	Recall	F1 score
0.7044025	0.3109244	0.7551020	0.4404762

predicted.classes.step.over	0	1
0	187	12
1	82	37

Threshold: 0.45

Ridge Regularization (Oversampled)



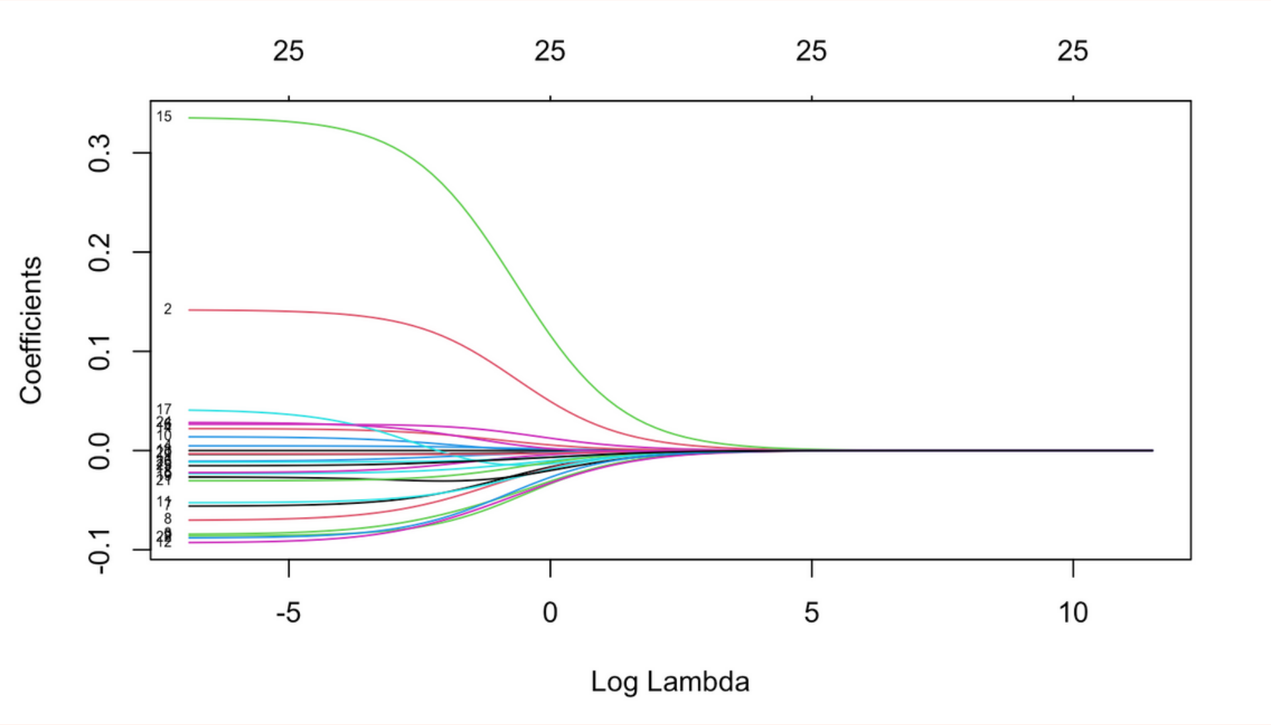
Accuracy	Precision	Recall	F1 score
0.6949686	0.3095238	0.7959184	0.4457143

ridge.pred_class	0	1
0	182	10
1	87	39

```
## [1] 0.02364489
```

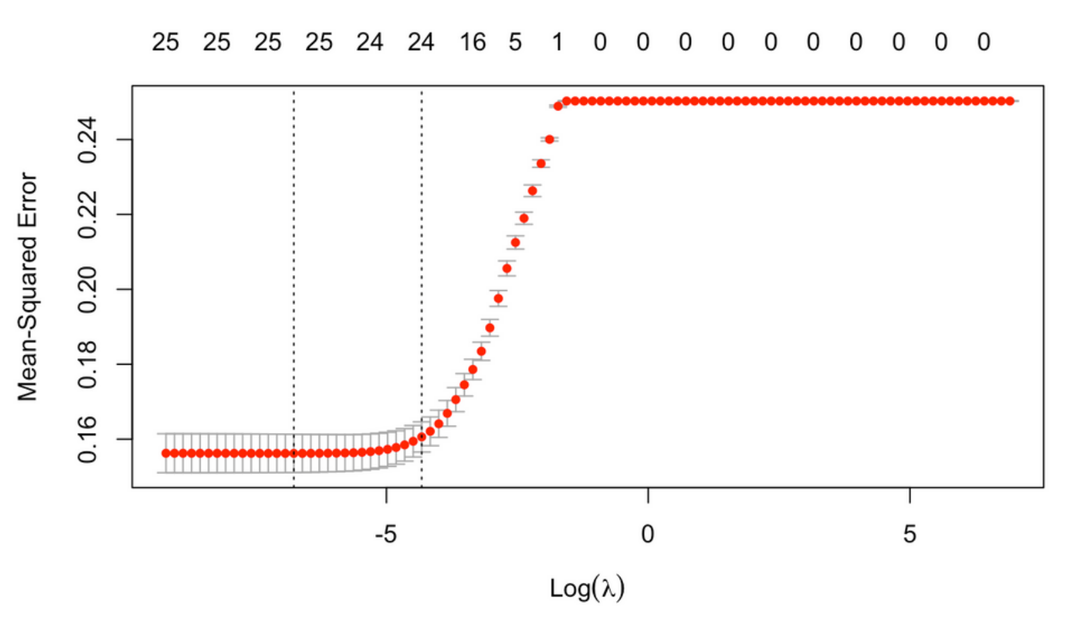
Threshold: 0.45

Ridge Regularization (Oversampled)



	s1
(Intercept)	1.470407e+00
Age	-3.544739e-03
BusinessTravel	1.323367e-01
Department	-7.554738e-02
DistanceFromHome	4.311224e-03
Education	-2.177334e-02
EducationField	2.591622e-02
EnvironmentSatisfaction	-5.089871e-02
Gender	-6.182340e-02
JobInvolvement	-8.133138e-02
JobRole	1.052541e-02
JobSatisfaction	-4.893302e-02
MaritalStatus	-8.333648e-02
MonthlyIncome	-1.489378e-05
NumCompaniesWorked	1.973251e-02
OverTime	3.104466e-01
PercentSalaryHike	-8.360841e-03
PerformanceRating	1.806123e-02
RelationshipSatisfaction	-1.889100e-02
StockOptionLevel	-2.928155e-02
TotalWorkingYears	-3.658546e-03
TrainingTimesLastYear	-2.892727e-02
WorkLifeBalance	-8.036310e-02
YearsInCurrentRole	-1.131325e-02
YearsSinceLastPromotion	2.357923e-02
YearsWithCurrManager	-1.378238e-02

Lasso Regularization (Oversampled)



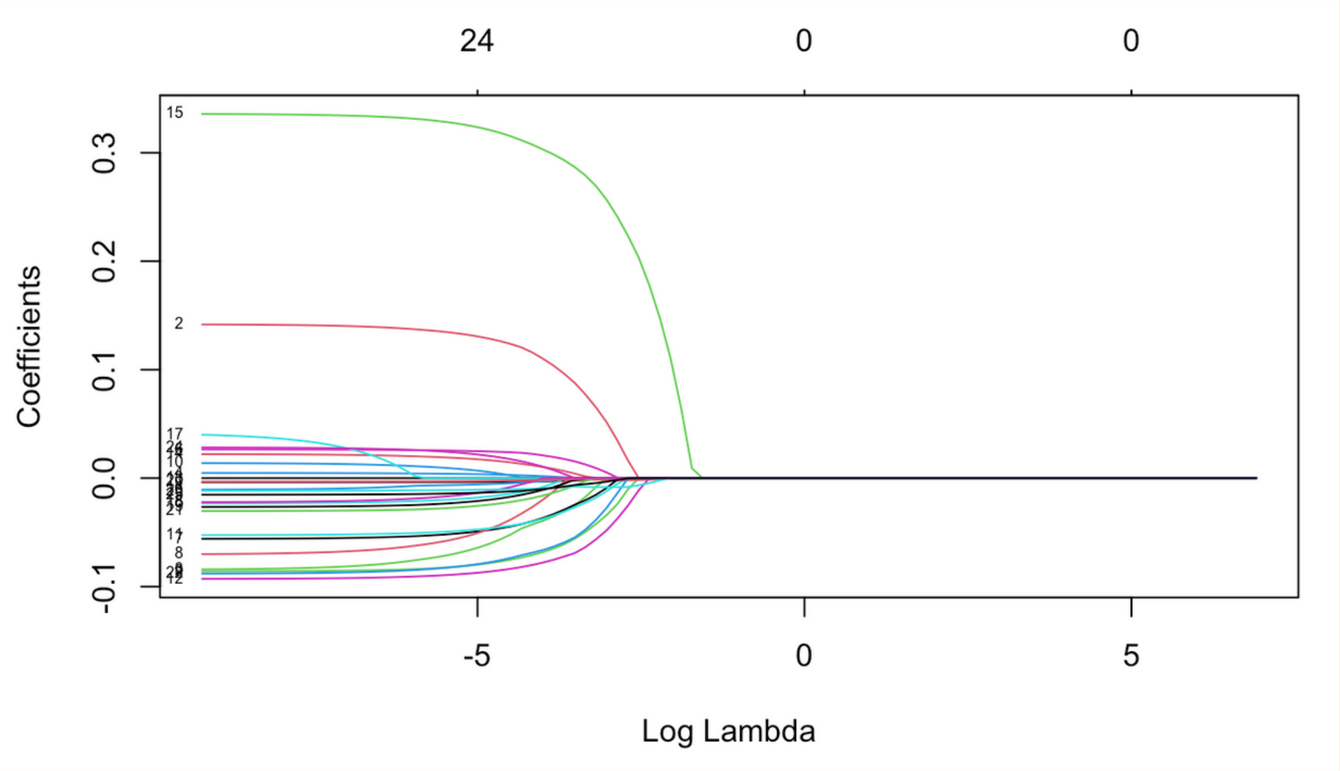
Accuracy	Precision	Recall	F1 score
0.6855346	0.3053435	0.8163265	0.4444444

```
lasso.over   0   1
         0 178   9
         1  91  40
```

```
[1] 0.001149757
```

Threshold: 0.45

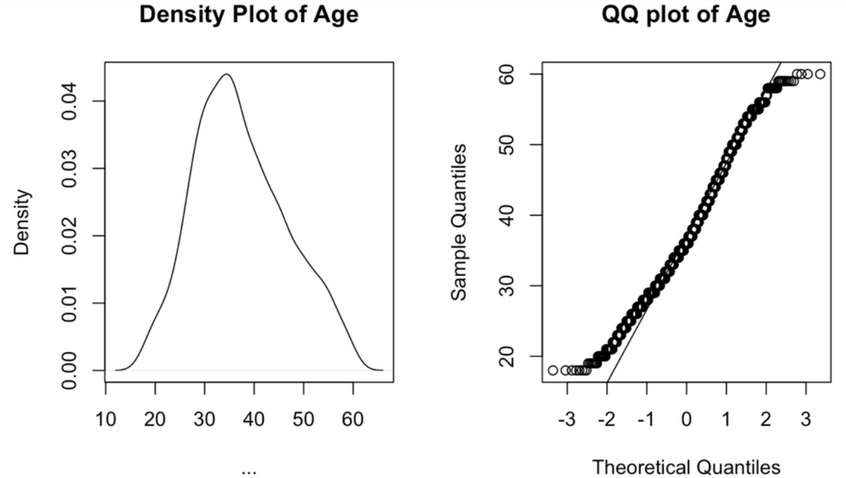
Lasso Regularization (Oversampled)

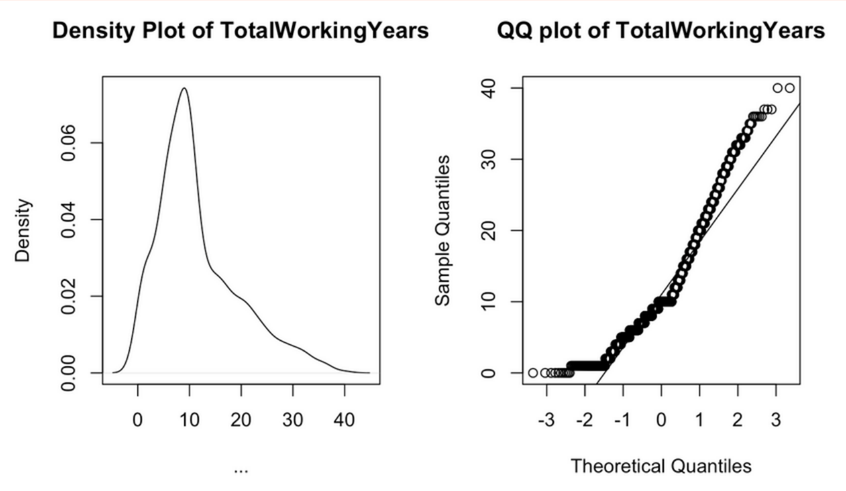


	s1
(Intercept)	1.514824e+00
Age	-3.751509e-03
BusinessTravel	1.390242e-01
Department	-7.911511e-02
DistanceFromHome	4.445228e-03
Education	-2.197093e-02
EducationField	2.605209e-02
EnvironmentSatisfaction	-5.452582e-02
Gender	-6.541360e-02
JobInvolvement	-8.475287e-02
JobRole	1.212538e-02
JobSatisfaction	-5.141131e-02
MaritalStatus	-9.160532e-02
MonthlyIncome	-1.690407e-05
NumCompaniesWorked	2.090381e-02
OverTime	3.330952e-01
PercentSalaryHike	-8.451265e-03
PerformanceRating	1.686658e-02
RelationshipSatisfaction	-2.016337e-02
StockOptionLevel	-2.547156e-02
TotalWorkingYears	-3.060546e-03
TrainingTimesLastYear	-2.954096e-02
WorkLifeBalance	-8.605519e-02
YearsInCurrentRole	-1.142782e-02
YearsSinceLastPromotion	2.681537e-02
YearsWithCurrManager	-1.494266e-02

Normality assumption

```
[1] "Shapiro-Wilk Test Age"

        Shapiro-Wilk normality test

data:  column
W = 0.98063, p-value = 4.959e-12

[1] "Shapiro-Wilk Test TotalWorkingYears"

        Shapiro-Wilk normality test

data:  column
W = 0.91676, p-value < 2.2e-16
```





LDA

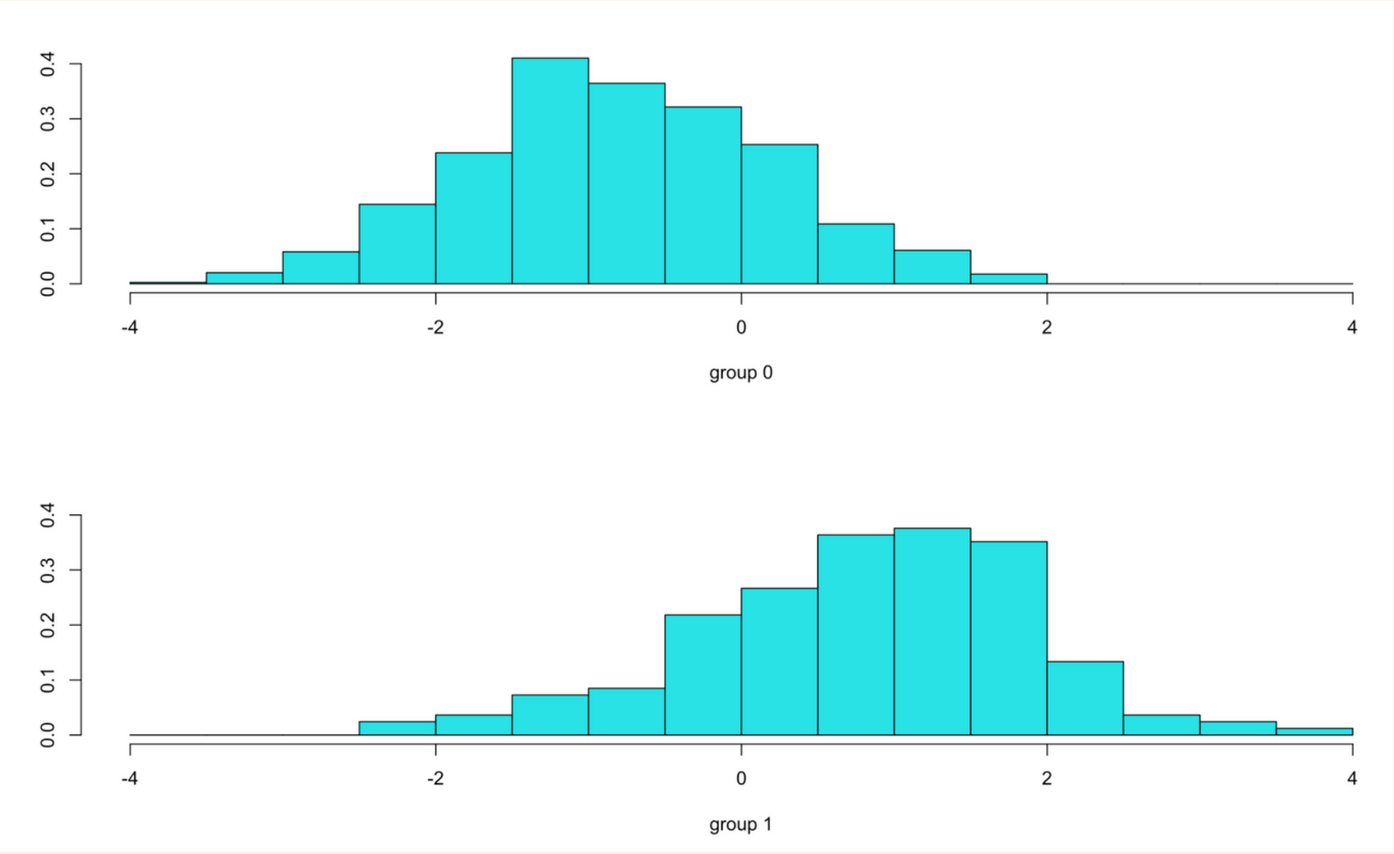
Assumptions:

- data within each class follows a normal distribution
- covariance matrices of different classes are equal

Accuracy	Precision	Recall	F1 score
0.6949686	0.3095238	0.7959184	0.4457143

Threshold: 0.4

LDA



lda.pred.best	0	1
0	182	10
1	87	39

QDA

Assumptions:

- each class is drawn from a normal distribution
- each class has a different covariance matrix

Accuracy	Precision	Recall	F1 score
0.7893082	0.3392857	0.3877551	0.3619048

Threshold: 0.4

qda.pred.best	0	1
0	197	29
1	72	20

Naive Bayes

naivebayes.pred_over	0	1
0	132	11
1	137	38

Accuracy	Precision	Recall	F1 score
0.5566038	0.2195122	0.7346939	0.3380282

Threshold: 0.4

KNN

We scale our whole dataset before splitting it in training and test set.

Accuracy	0.7421384
Precision	0.3225806
Recall	0.6122449
F1 score	0.4225352

knn.pred	0	1
0	206	19
1	63	30

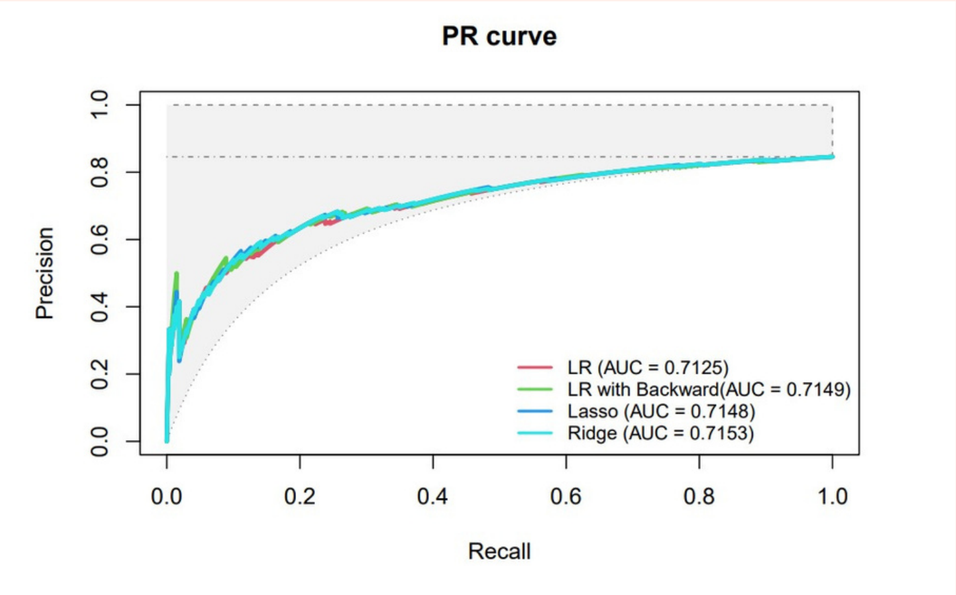
Best K: 7

Comparison

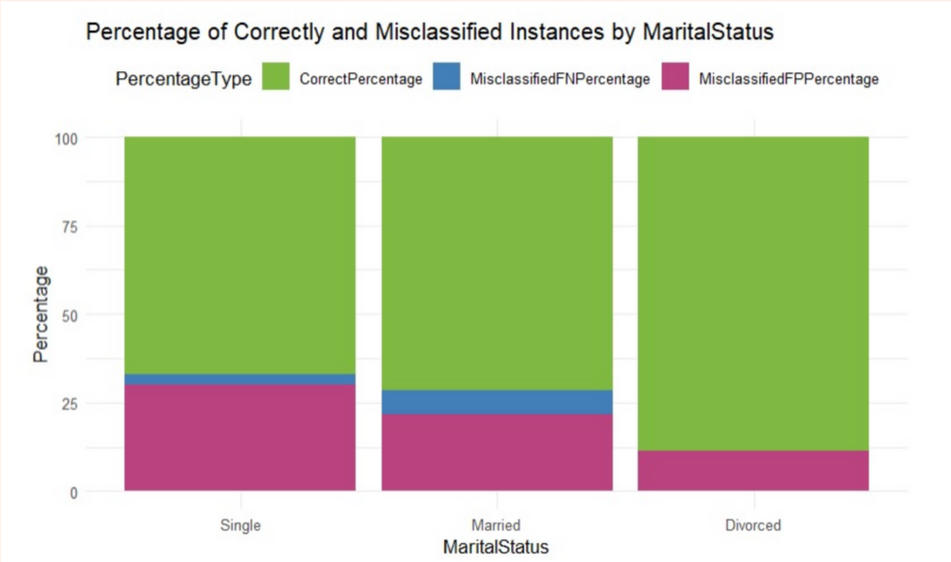
Model_name	Accuracy	Precision	Recall	F1_score
Lasso Regularization	0.6855346	0.3053435	0.8163265	0.4444444
Ridge Regularization	0.6949686	0.3095238	0.7959184	0.4457143
LDA	0.6949686	0.3095238	0.7959184	0.4457143
Naive Bayes	0.5345912	0.2171429	0.7755102	0.3392857
Logistic regression	0.7138365	0.3189655	0.7551020	0.4484848
Logistic regression with Backward Selection	0.7044025	0.3109244	0.7551020	0.4404762
KNN	0.7421384	0.3225806	0.6122449	0.4225352
Logistic regression Unbalanced	0.8459119	0.5000000	0.4285714	0.4615385
QDA	0.6823899	0.2173913	0.4081633	0.2836879

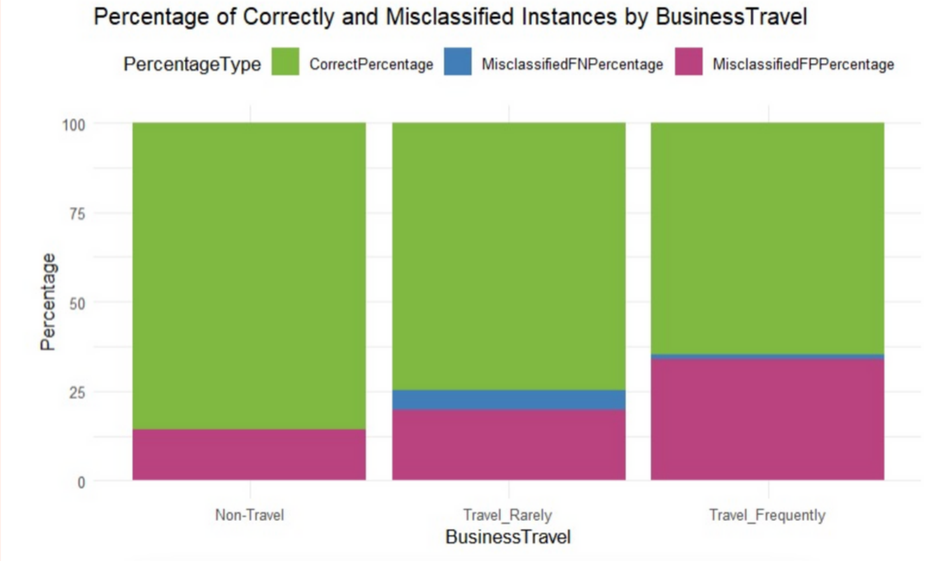
Best model:

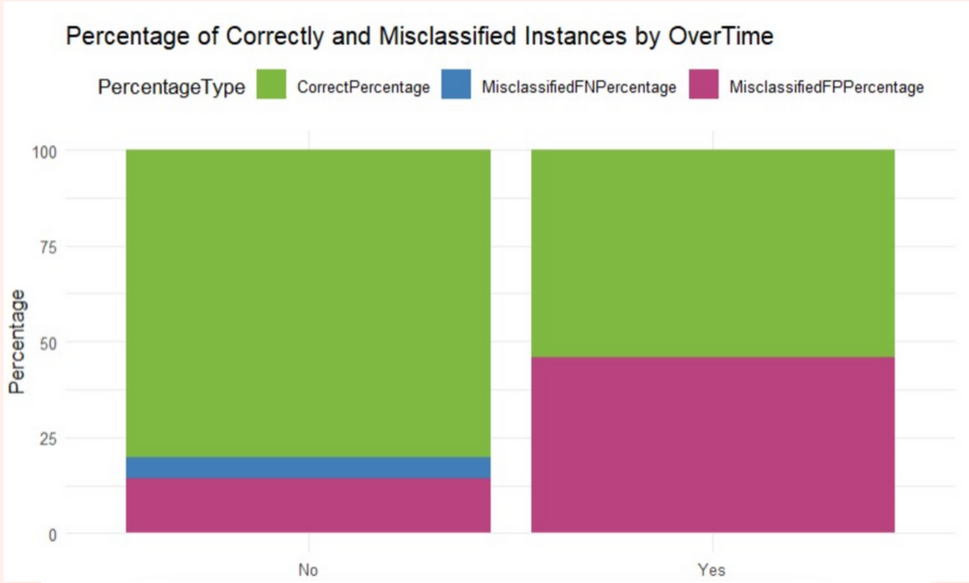
LASSO

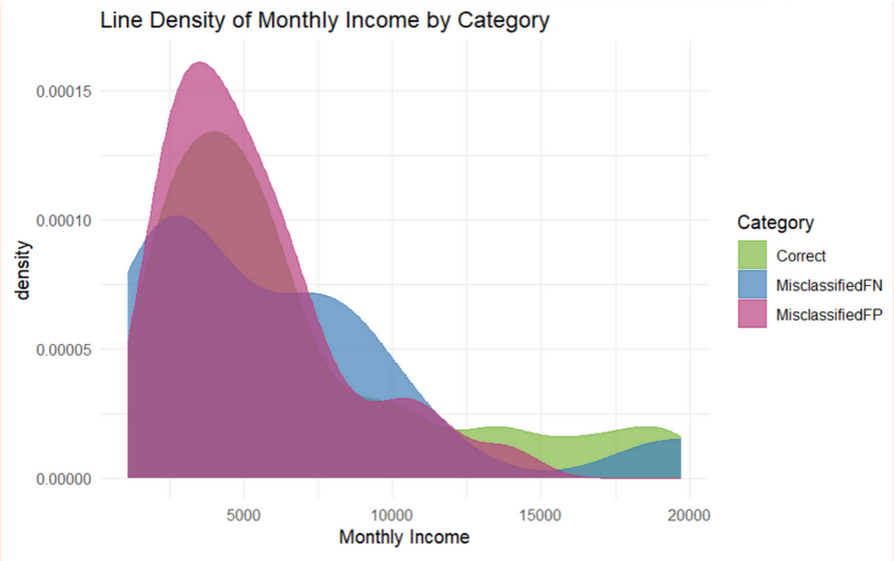


Misclassification Analysis









Conclusions and suggestions

OVERTIME

- higher retribution for employees working overtime
- setting a limit to overtime hours weekly
- analysis of workflow and reasons behind overtime

BUSINESS TRAVEL

- better compensation for travel expenses

Conclusions and suggestions

KEEPING UPDATE

- periodic surveys to get an updated idea of the employees' conditions
- check model predictions annually so that timely actions can be taken

PROPOSED ADDITIONAL QUESTIONS

- ManagerSatisfaction: evaluation of your direct employer
- TeamWork: evaluation of the quality of team work collaboration