Employee Turnover Analysis

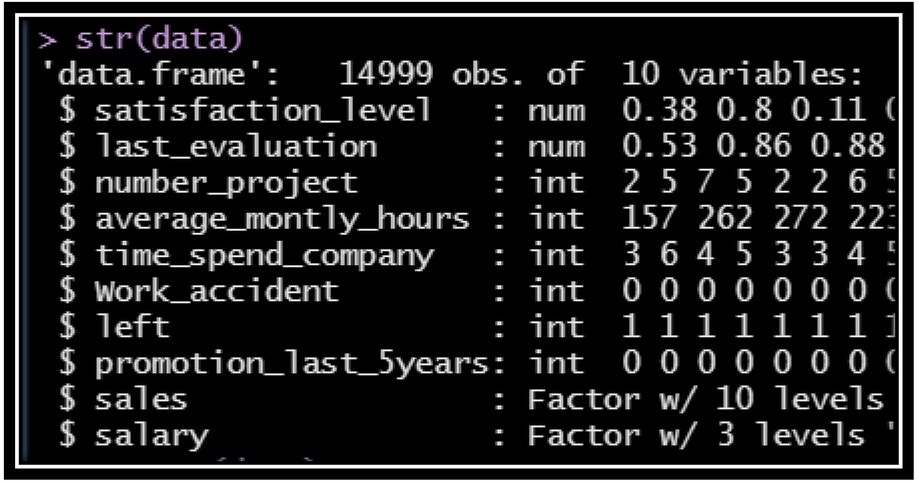
Jovan Trajceski

Agenda

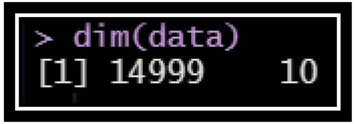
1. Exploratory Data Analysis
2. Modelling and Evaluation
3. Discussion
4. Conclusion

Exploratory Data Analysis

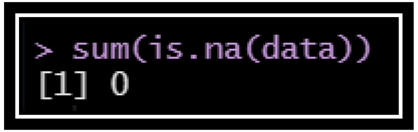
EDA – data overview



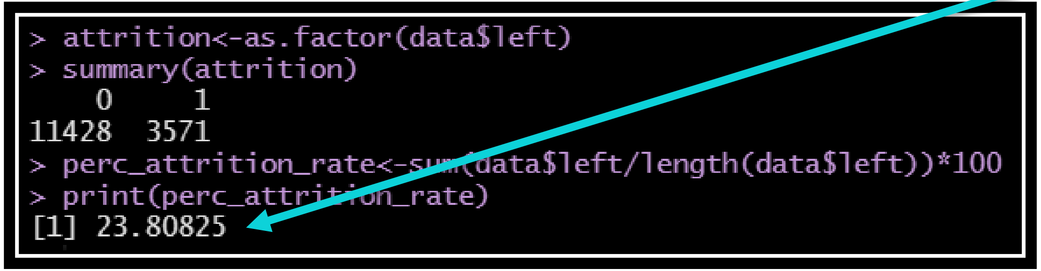
```
> str(data)
'data.frame':   14999 obs. of  10 variables:
 $ satisfaction_level   : num  0.38 0.8 0.11
 $ last_evaluation      : num  0.53 0.86 0.88
 $ number_project       : int  2 5 7 5 2 2 6
 $ average_montly_hours : int  157 262 272 22
 $ time_spend_company   : int  3 6 4 5 3 3 4
 $ Work_accident        : int  0 0 0 0 0 0 0
 $ left                 : int  1 1 1 1 1 1 1
 $ promotion_last_5years: int  0 0 0 0 0 0 0
 $ sales                : Factor w/ 10 levels
 $ salary               : Factor w/ 3 levels
```



```
> dim(data)
[1] 14999    10
```



```
> sum(is.na(data))
[1] 0
```



```
> attrition<-as.factor(data$left)
> summary(attrition)
    0     1
11428  3571
> perc_attrition_rate<-sum(data$left/length(data$left))*100
> print(perc_attrition_rate)
[1] 23.80825
```

❑ 15,000 employees

❑ 10 variables (features)

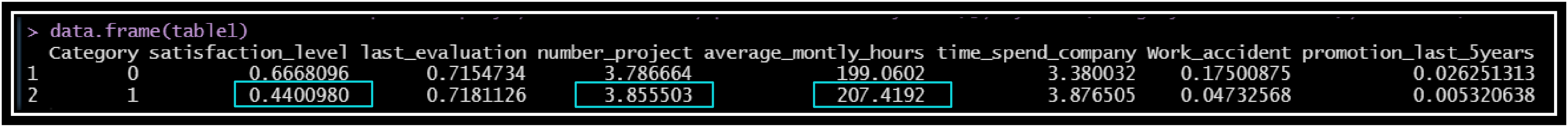
❑ No NaN values

❑ Turnover rate of 23.81%

EDA – transformation

Metrics for the employee population that left the company:

❑ Lower Satisfaction level, Higher # of Projects, and Higher # of Hours

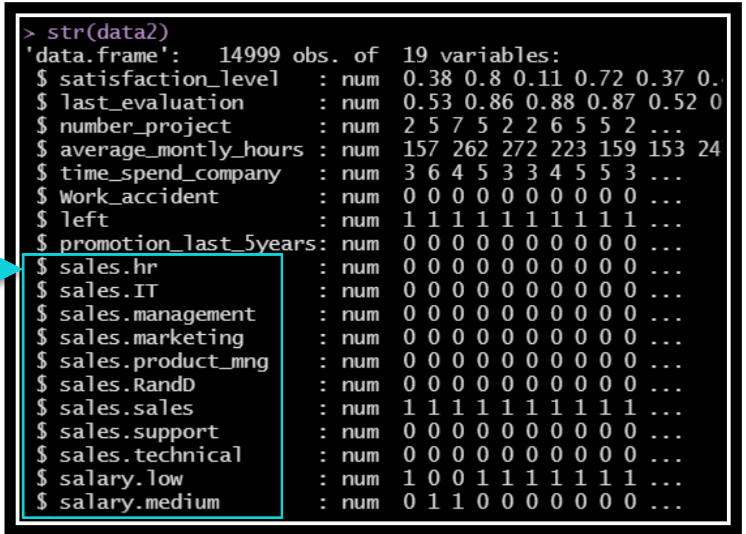


```
> data.frame(table1)
  Category satisfaction_level last_evaluation number_project average_montly_hours time_spend_company Work_accident promotion_last_5years
1        0          0.6668096       0.7154734       3.786664             199.0602           3.380032    0.17500875           0.026251313
2        1          0.4400980       0.7181126       3.855503             207.4192           3.876505    0.04732568           0.005320638
```

Created dummy variables for:

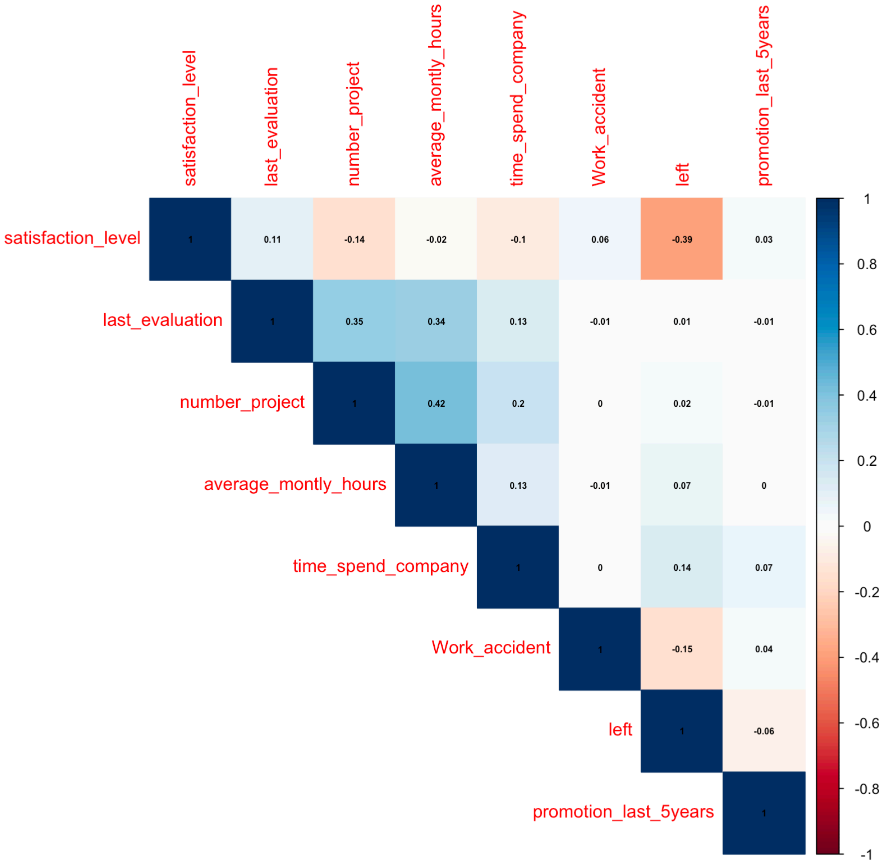
❑ Departments, and,

❑ Salary



```
> str(data2)
'data.frame':	14999 obs. of  19 variables:
 $ satisfaction_level   : num  0.38 0.8 0.11 0.72 0.37 0.
 $ last_evaluation      : num  0.53 0.86 0.88 0.87 0.52 0
 $ number_project       : num  2 5 7 5 2 2 6 5 5 2 ...
 $ average_montly_hours : num  157 262 272 223 159 153 24
 $ time_spend_company   : num  3 6 4 5 3 3 4 5 5 3 ...
 $ Work_accident        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ left                 : num  1 1 1 1 1 1 1 1 1 1 ...
 $ promotion_last_5years: num  0 0 0 0 0 0 0 0 0 0 ...
 $ sales.hr             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sales.IT             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sales.management     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sales.marketing      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sales.product_mng    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sales.RandD          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sales.sales          : num  1 1 1 1 1 1 1 1 1 1 ...
 $ sales.support        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sales.technical      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ salary.low           : num  1 0 0 1 1 1 1 1 1 1 ...
 $ salary.medium        : num  0 1 1 0 0 0 0 0 0 0 ...
```

EDA – correlation map



	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years
satisfaction_level	1	0.11	-0.14	-0.02	-0.1	0.06	-0.39	0.03
last_evaluation		1	0.35	0.34	0.13	-0.01	0.01	-0.01
number_project			1	0.42	0.2	0	0.02	-0.01
average_montly_hours				1	0.13	-0.01	0.07	0
time_spend_company					1	0	0.14	0.07
Work_accident						1	-0.15	0.04
left							1	-0.06
promotion_last_5years								1

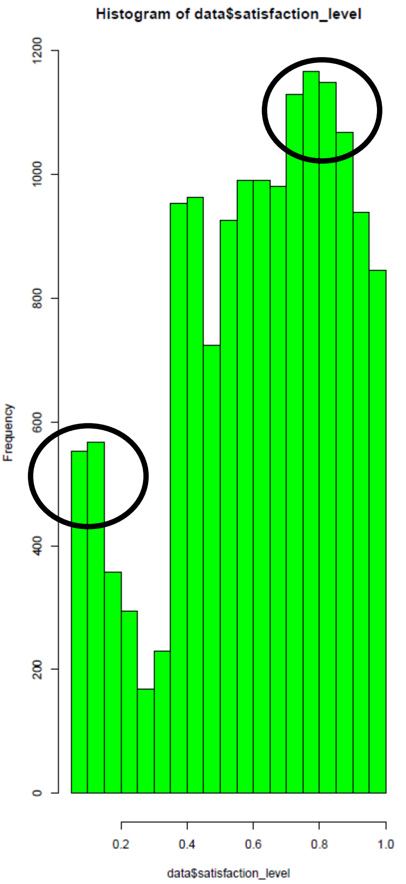
Positive correlation:

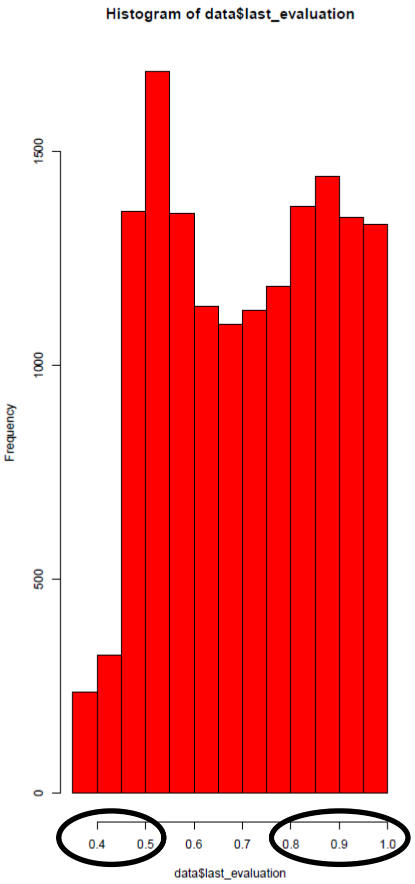
- ❑ Hours and Projects (0.42)
- ❑ Hours and Evaluation (0.34)
- ❑ Projects and Evaluation (0.35)

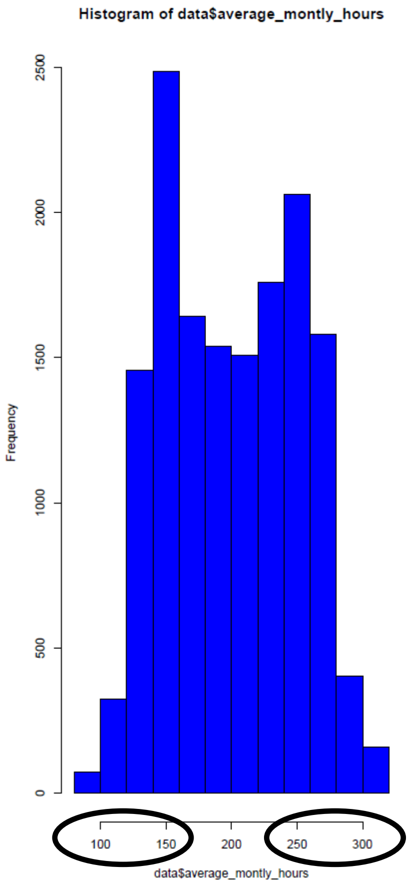
Negative correlation:

- ❑ Left and Satisfaction (-0.39)

EDA – Distribution







Satisfaction:

- □ Low spike
- □ High spike

Evaluation:

- □ Bimodal

  <0.6

  >0.8

Monthly hours:

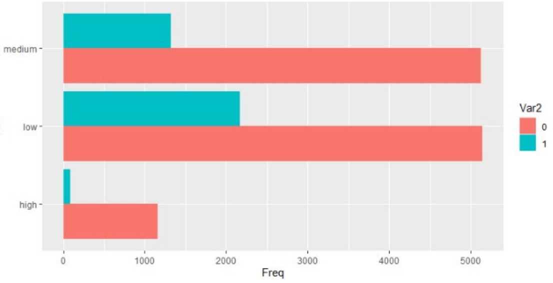
- □ Bimodal

  <150

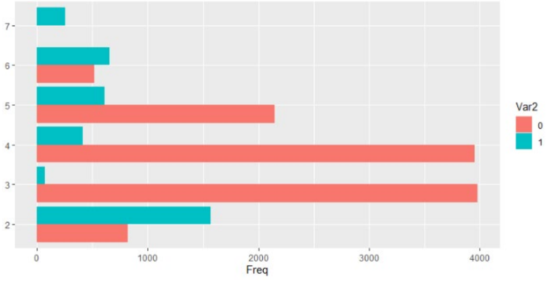
  >250

EDA – variables and turnover

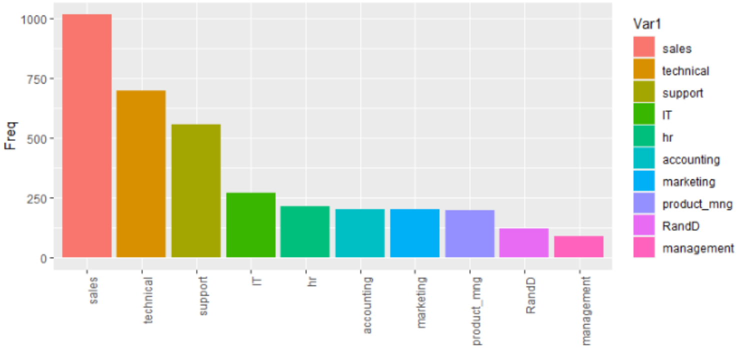
Salary vs. Turnover



Turnover vs. Project



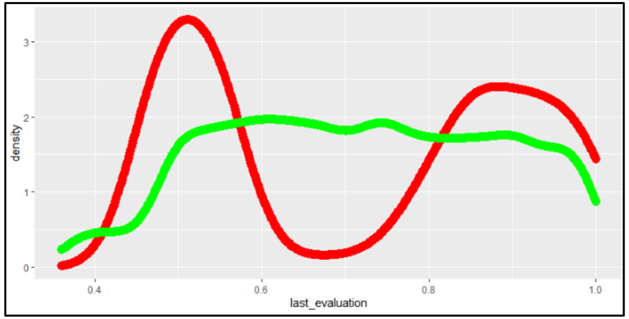
Turnover by Department



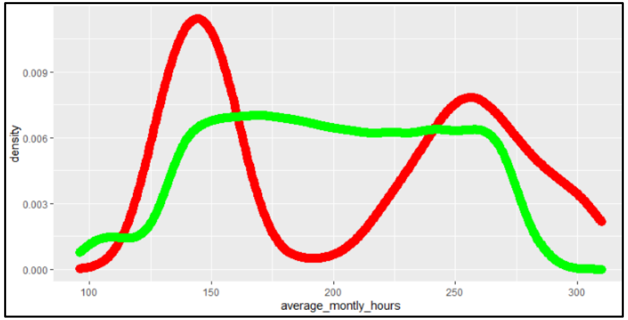
- ❑ Employees with low/avg salary leave
- ❑ Almost no one left with high salary
- ❑ All employees with 7 projects left
- ❑ Increase in turnover as project count increases
- ❑ Sales, technical, and support department have highest turnover
- ❑ Management has lowest turnover

EDA – Turnover Density

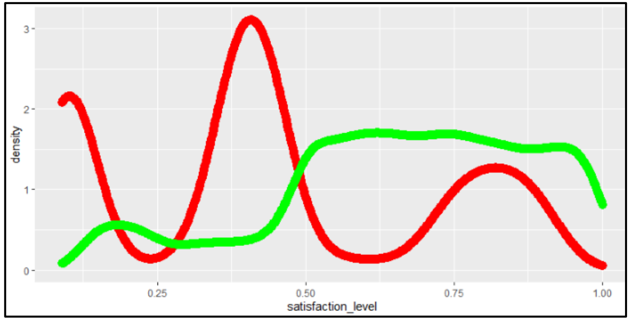
Turnover & Evaluation



Turnover & Hours



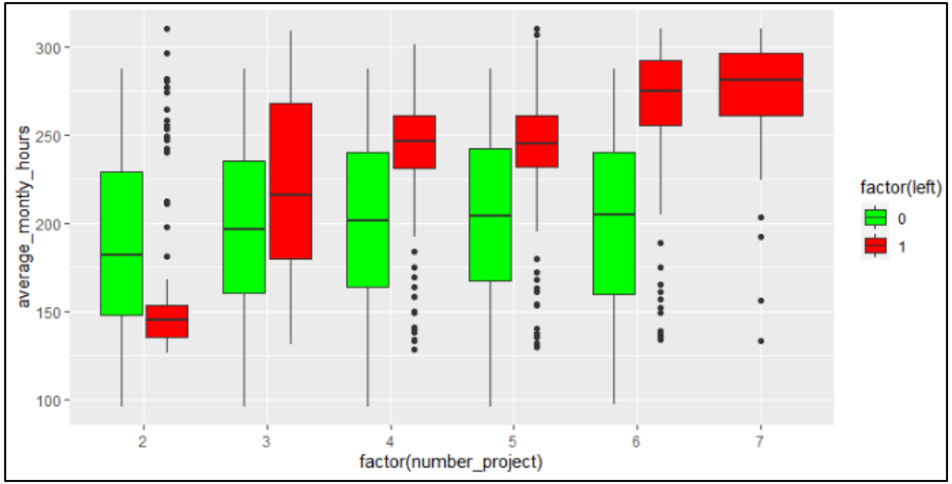
Turnover & Satisfaction



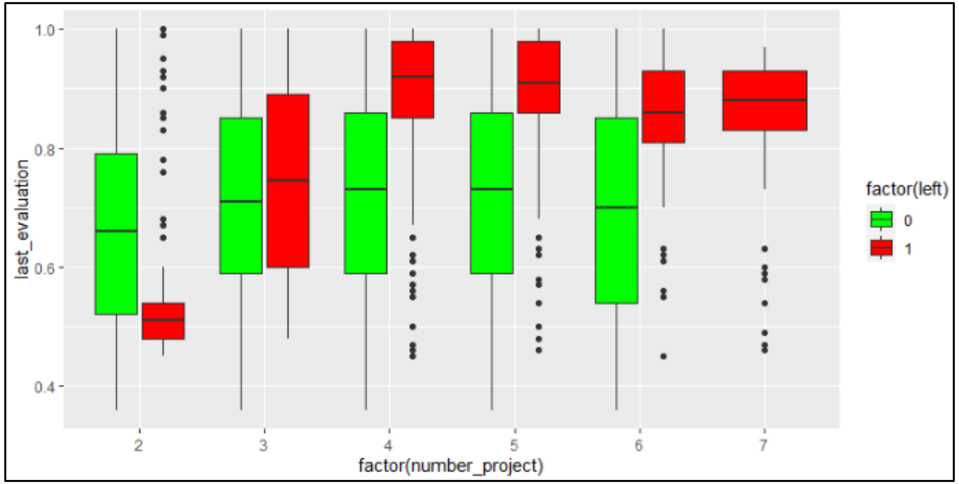
- ❑ Employees with low and high evaluation leave
- ❑ Employees with 0.6-0.8 stay
- ❑ Employees with hours<150 (underworked) and hours>250 (overworked) leave
- ❑ Employees who had 150-250 hours stay
- ❑ Employees with low satisfaction <0.2 and 0.3-0.5 leave
- ❑ Employees with high satisfaction (over 0.75) leave more than stay

EDA – Number of Projects

Projects & Hours



Projects & Evaluation

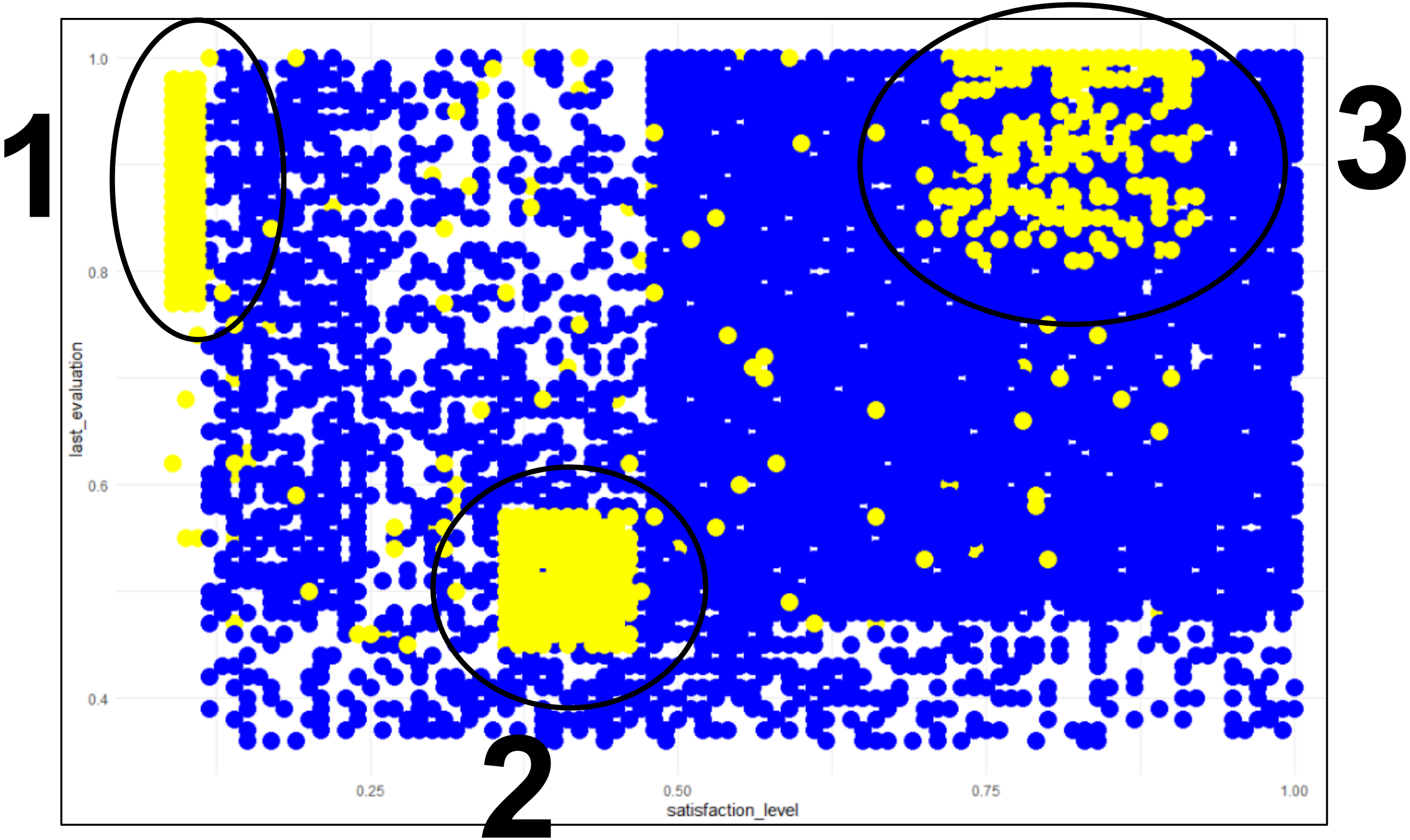


- ❑ Employees that stayed had 200 average hours, regardless of projects number
- ❑ Employees that left had increased hours as projects increased in count

- ❑ Employees that left with high project count had better evaluation (0.9)
- ❑ Employees that stayed had consistent evaluation (0.7) even when project count increased

EDA – Clusters

1. **Overworked:** Good workers (0.8-1), not satisfied (<0.2)
2. **Low Performers**: Poor workers(<0.6), not satisfied (0.3-0.5)
3. **Found new jobs**: Good workers (0.8-1), satisfied (0.7-1)



Modelling and Evaluation

Handle skewed data

Use random sampling
to reduce problem of skewed data

Data transformation

Logistic regression & decision tree
based on 4 datasets

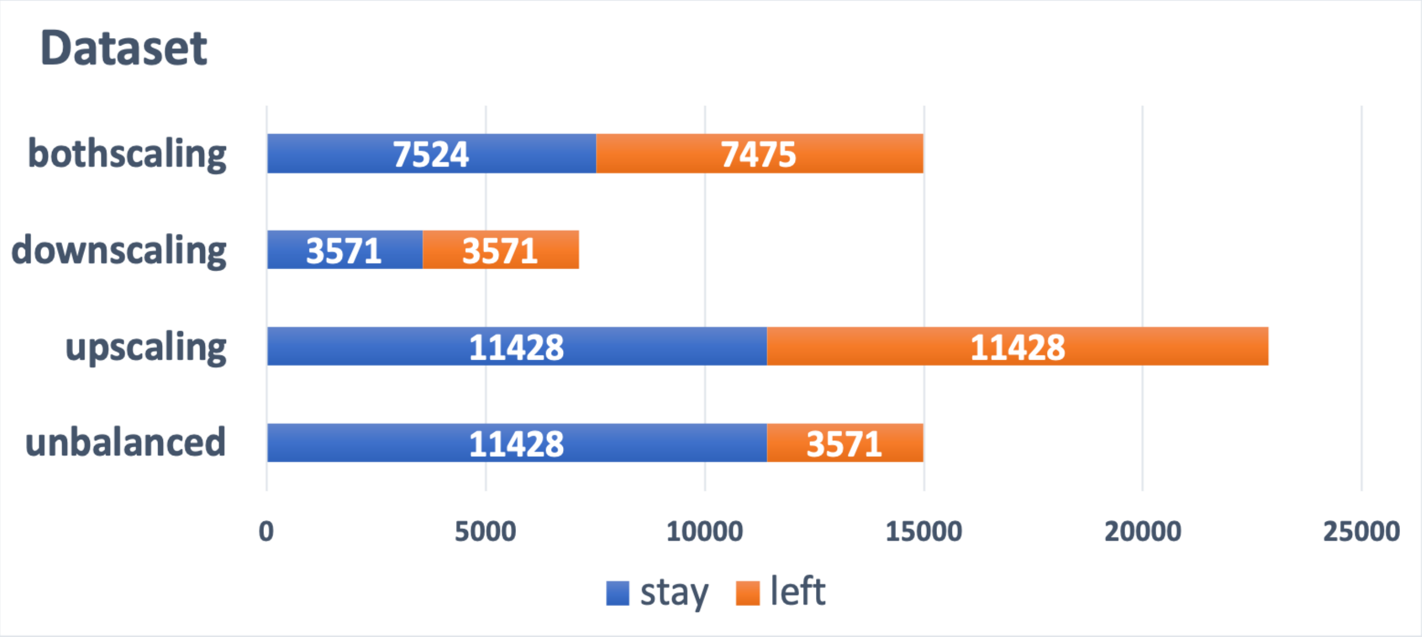
1 **Unbalanced**
target variable: **stay 76%** / **left 24%**
prediction maybe biased

2 **Upscaling**
repetitive sampling minority
no loss of information
possible overfitting because of repetition

3 **Downscaling**
decrease observations of majority
loss of information because of deletion

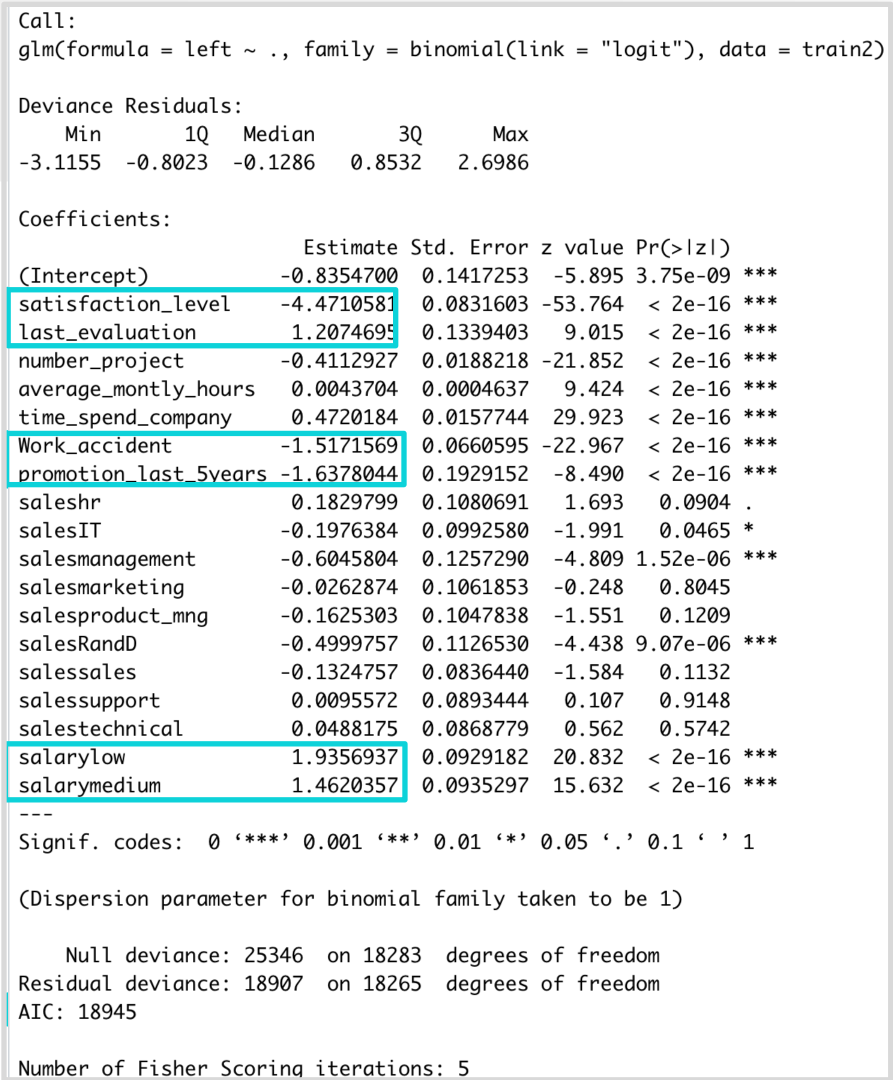
4 **Combine upscaling and downscaling**
upscaling minority
downscaling majority

Data structure and transformation



dataset	∆ left	∆ stay
both scaling	+4042	-3592
downscaling	0	-7857
upscaling	+7857	0
unbalanced	0	0

Logistic regression



```
Call:
glm(formula = left ~ ., family = binomial(link = "logit"), data = train2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1155  -0.8023  -0.1286   0.8532   2.6986

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.8354700  0.1417253  -5.895 3.75e-09 ***
satisfaction_level    -4.4710581  0.0831603 -53.764  < 2e-16 ***
last_evaluation        1.2074695  0.1339403   9.015  < 2e-16 ***
number_project        -0.4112927  0.0188218 -21.852  < 2e-16 ***
average_montly_hours   0.0043704  0.0004637   9.424  < 2e-16 ***
time_spend_company     0.4720184  0.0157744  29.923  < 2e-16 ***
Work_accident         -1.5171569  0.0660595 -22.967  < 2e-16 ***
promotion_last_5years -1.6378044  0.1929152  -8.490  < 2e-16 ***
saleshr                0.1829799  0.1080691   1.693   0.0904 .
salesIT               -0.1976384  0.0992580  -1.991   0.0465 *
salesmanagement       -0.6045804  0.1257290  -4.809 1.52e-06 ***
salesmarketing        -0.0262874  0.1061853  -0.248   0.8045
salesproduct_mng      -0.1625303  0.1047838  -1.551   0.1209
salesRandD            -0.4999757  0.1126530  -4.438 9.07e-06 ***
salessales            -0.1324757  0.0836440  -1.584   0.1132
salessupport           0.0095572  0.0893444   0.107   0.9148
salestechnical         0.0488175  0.0868779   0.562   0.5742
salarylow              1.9356937  0.0929182  20.832  < 2e-16 ***
salarymedium           1.4620357  0.0935297  15.632  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25346  on 18283  degrees of freedom
Residual deviance: 18907  on 18265  degrees of freedom
AIC: 18945

Number of Fisher Scoring iterations: 5
```

Upscaling Example

Data Structure
total: 22856
left: 11428 (50%)
stay: 11428 (50%)
train/test = 4:1

Package & Function
ROSE package
ovun.sample()

Classification threshold
probability: 0.5

Coefficients by descending order
satisfaction level
low salary
promotion in last 5 years
work accident
medium salary
last evaluation

Decision Tree

3 X 3 Trees

- Build 9 decision trees in total, for each of the sampling methods,
- Complete 3 scenarios,
- Select the most appropriate scenario

Under-Sampling

Combined

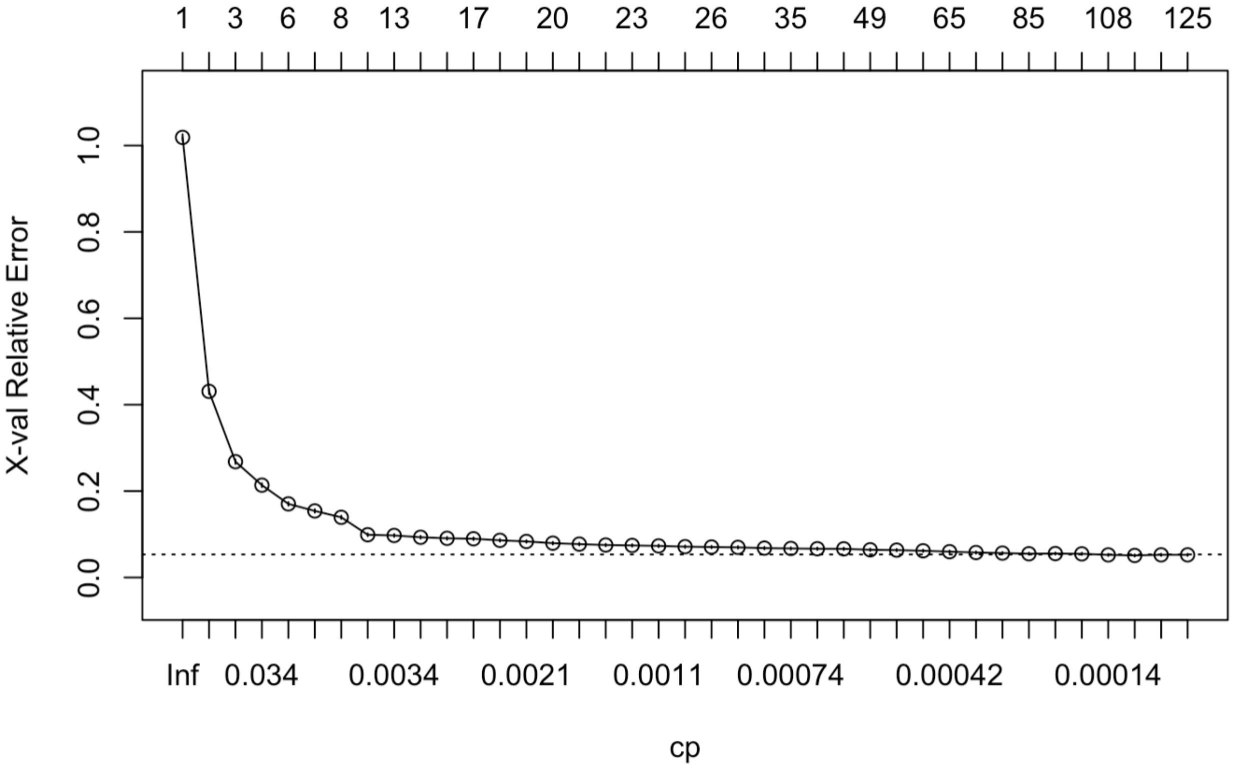
Over-Sampling

Pre-Pruning

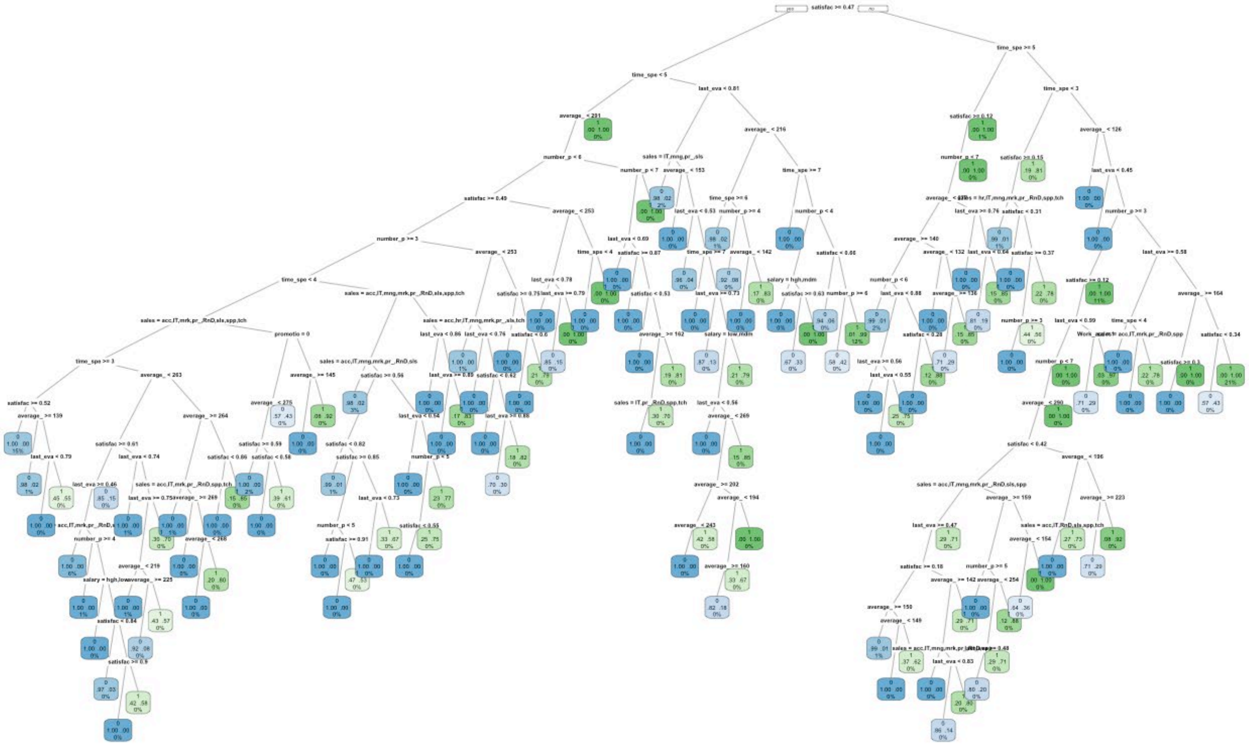
Base

Post-Pruning

CP Plot


1
3
6
8
13
17
20
23
26
35
49
65
85
108
125
X-val Relative Error
0.0
0.2
0.4
0.6
0.8
1.0
Inf
0.034
0.0034
0.0021
0.0011
0.00074
0.00042
0.00014
cp

Tree Without Pruning

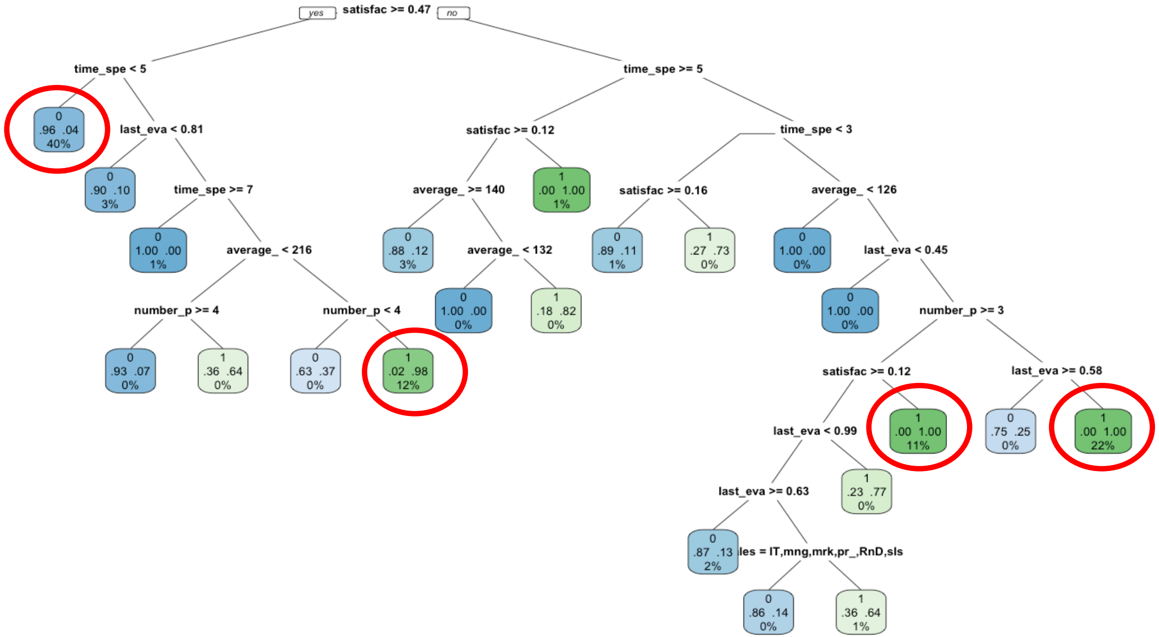


Why we need pruning?

↓

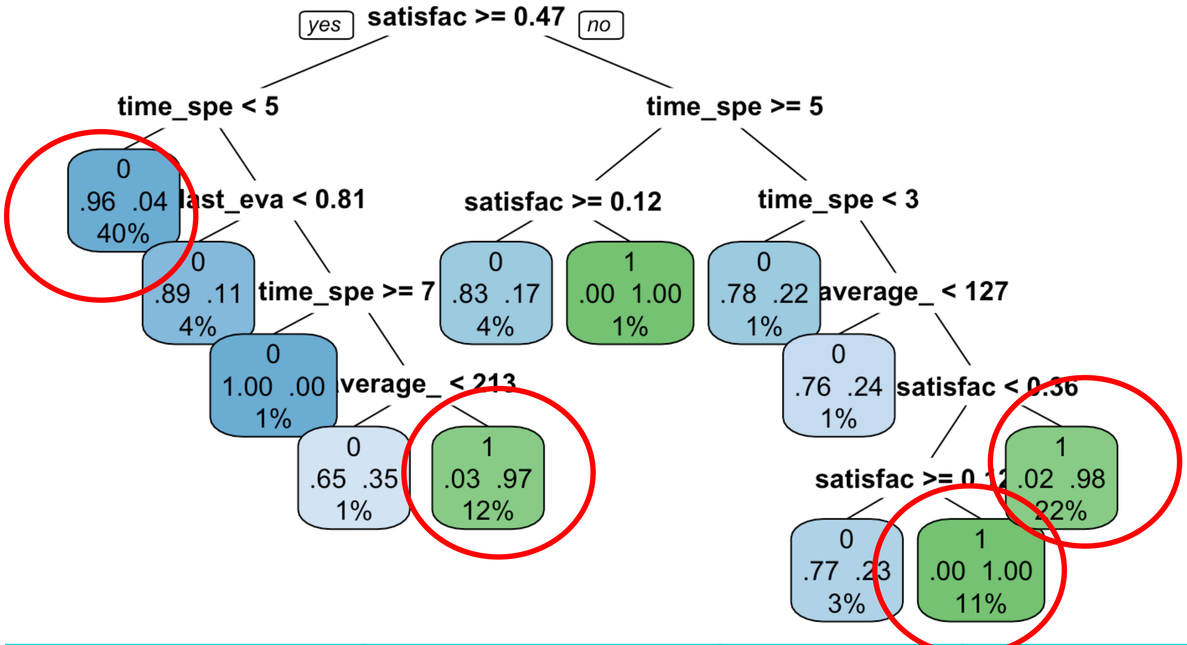
Over fitting

Over-Sampling Post-Pruning



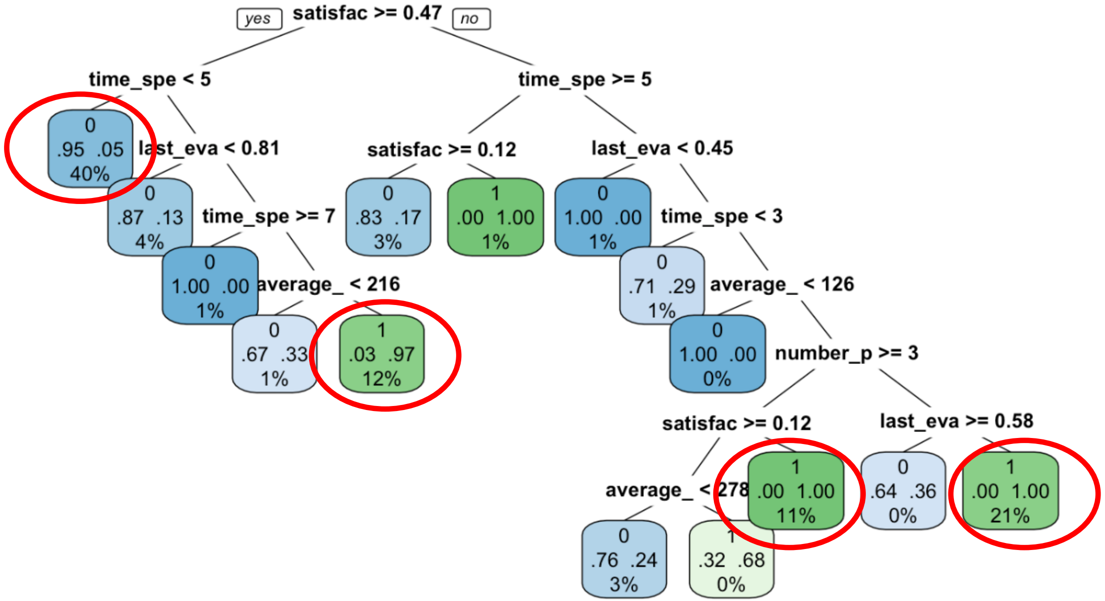
satisfaction_level	time_spend_company	last_evaluation	number_project
3918.83508	3342.14371	2483.91322	2046.43165
average_montly_hours	salary	sales	promotion_last_5years
1950.82373	70.21370	44.21282	13.39175

Under-Sampling Pre-Pruning



satisfaction_level	time_spend_company	last_evaluation	number_project
1277.054544	1019.755047	713.394546	656.744017
average_montly_hours	promotion_last_5years	sales	Work_acciden
625.559809	20.798779	8.946267	6.652014

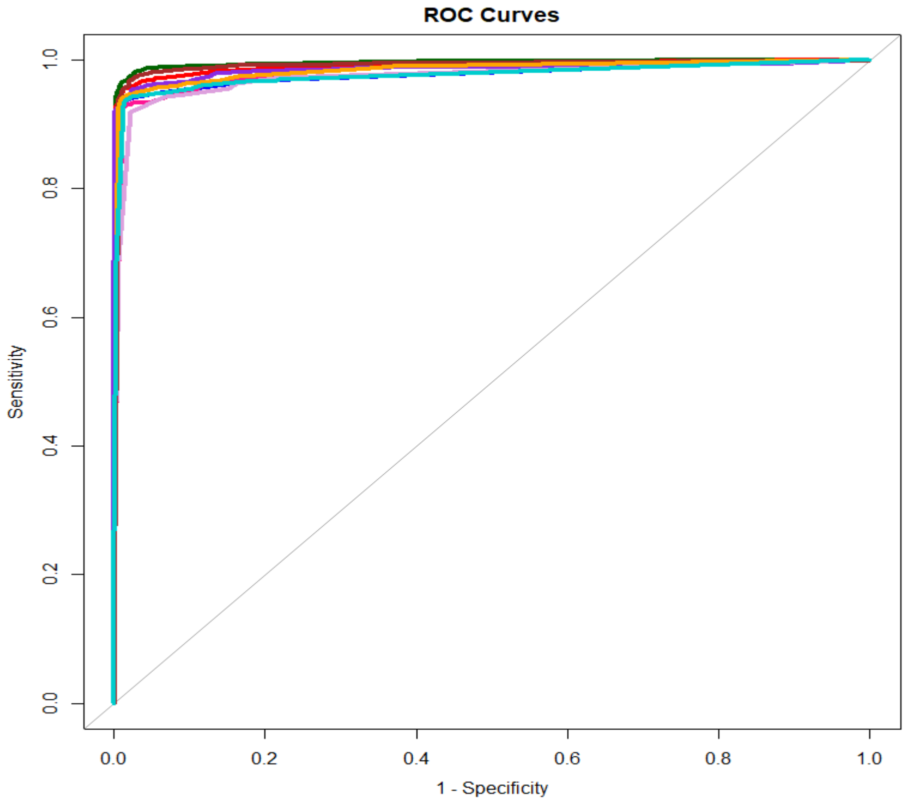
Both-Sampling Post-Pruning



satisfaction_level	time_spend_company	last_evaluation	number_project
2491.88129	2138.28538	1587.84219	1273.81460
average_montly_hours	salary	promotion_last_5years	sales
1187.82363	34.76908	34.60430	10.02167

Model Evaluation: Decision Tree

◆ ROC and AUC for Decision Tree Models



	ROC	AUC
—	Over Base	0.995
—	Over Pre	0.991
—	Over Post	0.977
—	Under Base	0.986
—	Under Pre	0.976
—	Under Post	0.974
—	Both Base	0.991
—	Both Pre	0.985
—	Both Post	0.976

Highest AUC

We choose
Decision Tree Over Sampling Post

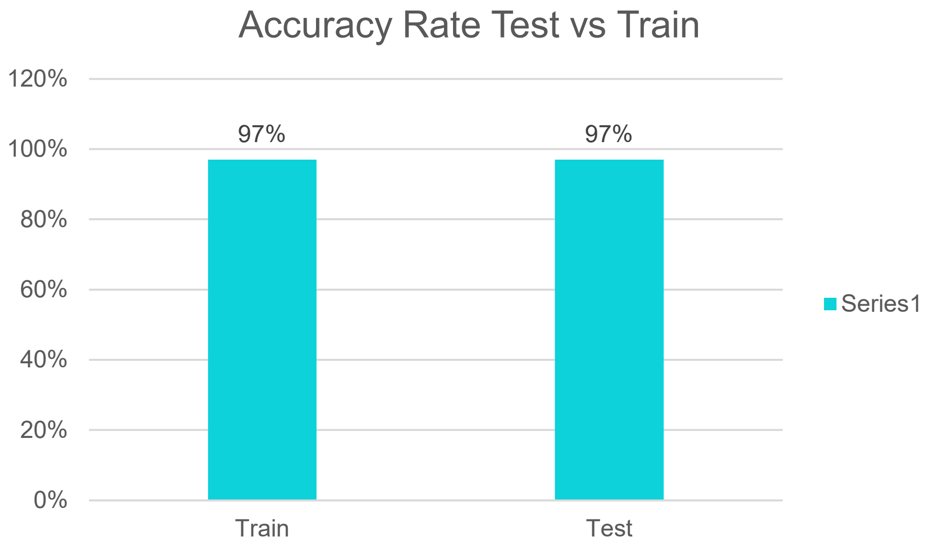
Decision Tree Over Sampling Post: Confusion Matrix Test vs Train

train data

actual predict	left	stay
left	8851	215
stay	330	8888

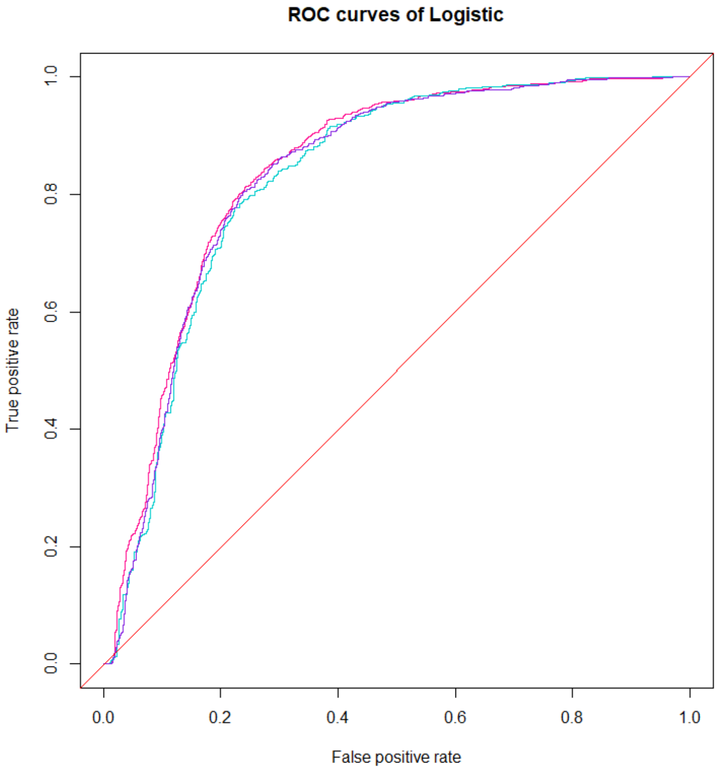
test data

actual predict	left	stay
left	2154	39
stay	93	2286



Model Evaluation: Logistic

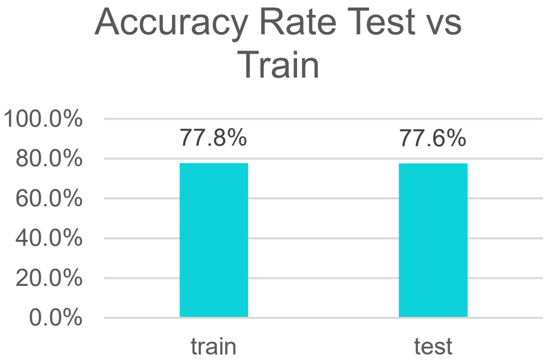
◆ **ROC and AUC for Logistic Models**



ROC		AUC
🟥	Over	0.840
🟦	Under	0.827
🟪	Both	0.832

Highest AUC

We choose
Logistic Over Sampling



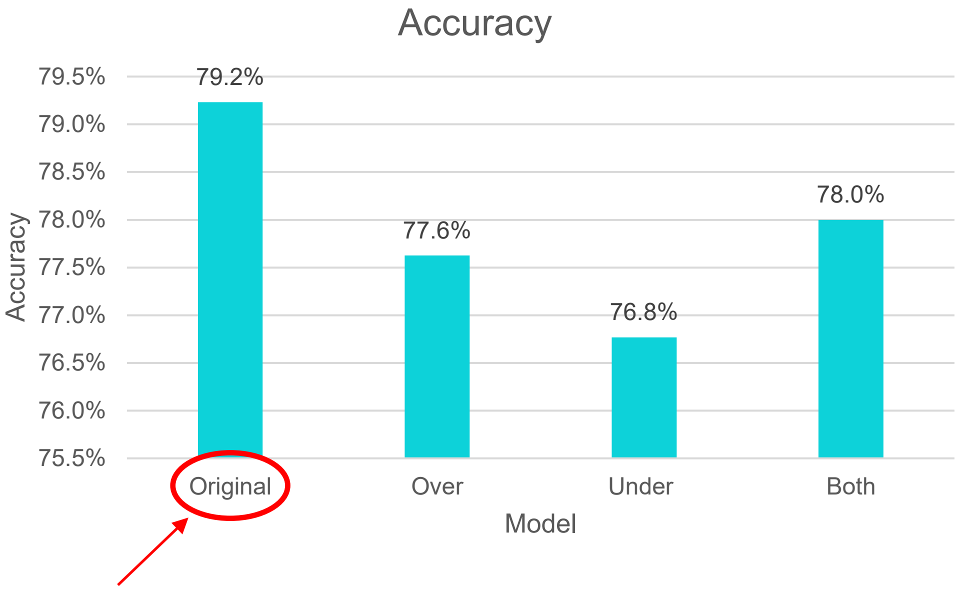
test data

actual predict	Left	Stay
Left	1509	793
Stay	363	1907

train data

actual predict	left	stay
left	6121	1468
stay	3021	7674

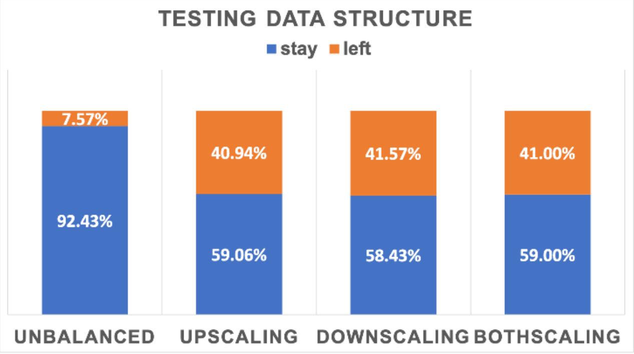
Problem of plain accuracy



Original Dataset

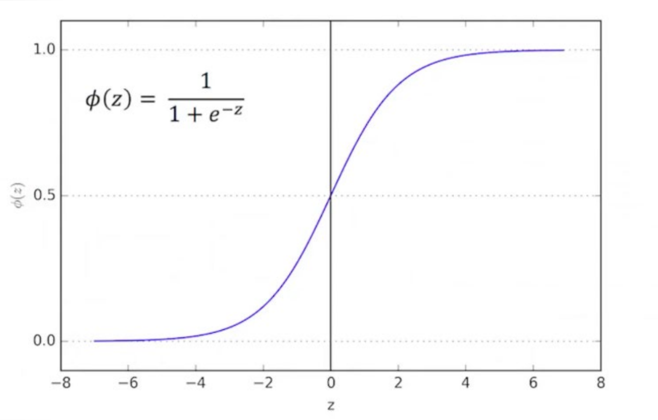
Why unbalanced dataset has highest accuracy rate?

- When the dataset is imbalanced, plain accuracy as metrics is unreliable
- In this scenario, majority of target variables are ”stay”

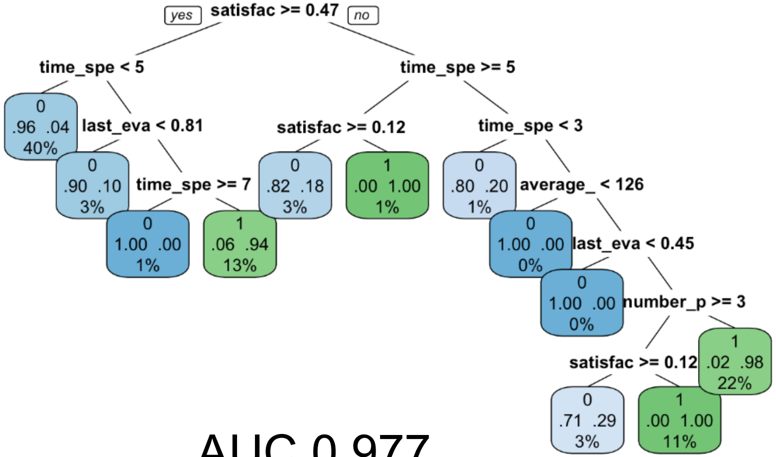


Best Model Selection

Logistic Over Sampling ***vs*** ***Decision Tree Over Sampling Post***



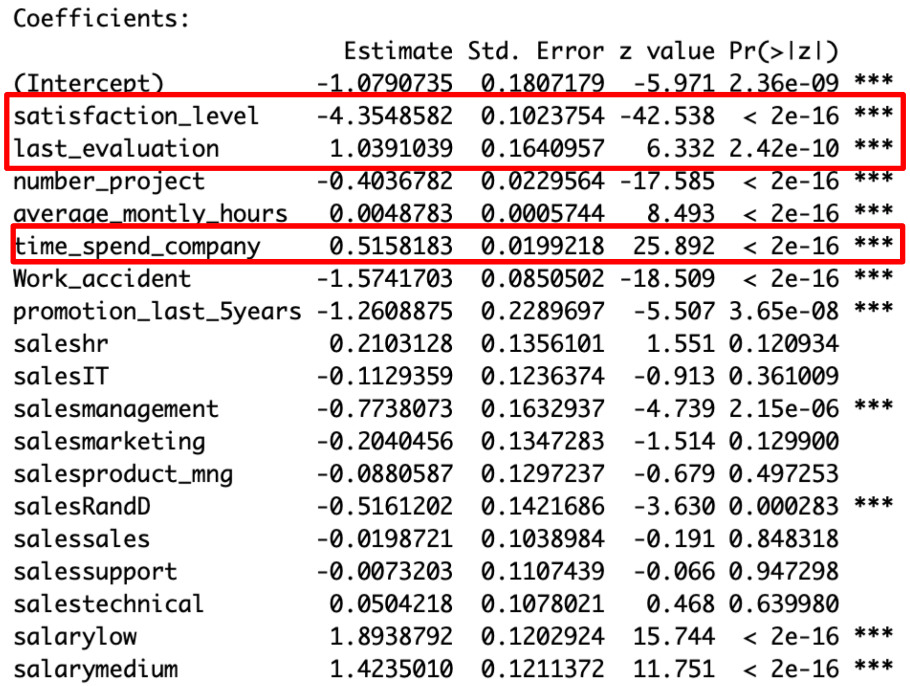
AUC 0.84



AUC 0.977

Discussion

Significance and Variable Importance



```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -1.0790735  0.1807179  -5.971 2.36e-09 ***
satisfaction_level   -4.3548582  0.1023754 -42.538  < 2e-16 ***
last_evaluation       1.0391039  0.1640957   6.332 2.42e-10 ***
number_project       -0.4036782  0.0229564 -17.585  < 2e-16 ***
average_montly_hours  0.0048783  0.0005744   8.493  < 2e-16 ***
time_spend_company    0.5158183  0.0199218  25.892  < 2e-16 ***
Work_accident        -1.5741703  0.0850502 -18.509  < 2e-16 ***
promotion_last_5years -1.2608875  0.2289697  -5.507 3.65e-08 ***
saleshr               0.2103128  0.1356101   1.551 0.120934    
salesIT              -0.1129359  0.1236374  -0.913 0.361009    
salesmanagement      -0.7738073  0.1632937  -4.739 2.15e-06 ***
salesmarketing       -0.2040456  0.1347283  -1.514 0.129900    
salesproduct_mng     -0.0880587  0.1297237  -0.679 0.497253    
salesRandD           -0.5161202  0.1421686  -3.630 0.000283 ***
salessales           -0.0198721  0.1038984  -0.191 0.848318    
salessupport         -0.0073203  0.1107439  -0.066 0.947298    
salestechnical        0.0504218  0.1078021   0.468 0.639980    
salarylow             1.8938792  0.1202924  15.744  < 2e-16 ***
salarymedium          1.4235010  0.1211372  11.751  < 2e-16 ***
```

	Satisfaction_level	**Last_evaluation**	**Time_spend_company**
Over_Post	3873	3311	2410

Satisfaction Level

- Berties, etc.(2019):

  •People with greater work autonomy exhibit critical-thinking skills and lower propensity to leave

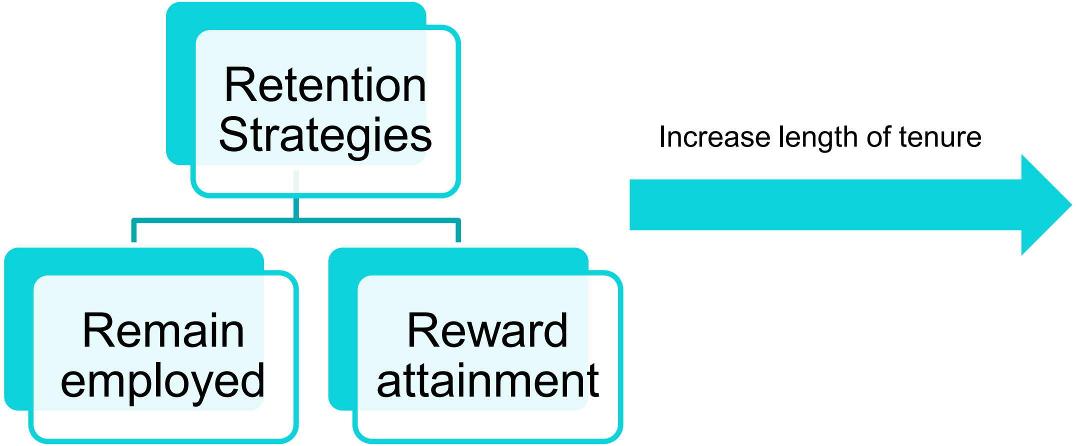
  •Therefore, having lower work autonomy leads to low job satisfaction that can contribute to further employee turnover

Recommendations

- Prioritise employee well-being
  - Motivating employees towards achieving a fitness milestone
  - Encouraging employees to disconnect when they showcase early signs of burnout

Time_Spend_Company (Tenure)

- HR professionals associate high employee tenure with employee’s job and organizational satisfaction (Ng & Feldman, 2013)



Long-tenured employees are **reluctant to leave** because of the accumulation of organizational investment

Recommendations

- ❑ Celebrate Milestones
  - ❑ E.g. Organizations should reward/recognize employees who have stayed with the company for certain years (5 years, 10 years, etc.)
- ❑ Celebrate positive experiences
  - ❑ Enable front line management to conduct pulse-check with their team

Last Evaluation (Job Performance)

- Performance directly affects the employee's motivation and intent to search for other jobs (Jackofsky,1986)
  - High-performance employees leave if they are not challenged

Recommendations

- Align Task to Employee`s Skillset and provide Market Based Salary
  - Knowing employee's skills and behavioural styles
  - For example, a creative thinker is probably a good fit for pitching ideas to clients.
  - However, they might struggle if they are given a more rule-intensive or detail-oriented task

Reference

- Lévy-Garboua, L. et al., 2007. Job satisfaction and quits. Labour Economics Volume 14, Issue 2, April 2007, Pages 251-268.

- D'Ambrosio, Conchita. et al., 2018. Unfairness at work: Well-being and quits. Labour Economics Volume 51, April 2018, Pages 307-316.

- Freeman, R.B., 1978. Job Satisfaction as an Economic Variable. American economic association 68, 135 – 141. Available at: https://www.nber.org/system/files/working_papers/w0225/w0225.pdf

- Martin, T., Price, J. , Mueller, C.(1981). Job performance and turnover. [J]. Journal of Applied Psychology, 66, 116-119