Human Resources analysis

Employee Attrition and Performance

Course: Introduction to Business Data Analytics

Section: 1001

Instructor: Dr. Jingzhi ZHANG

Group 1

Ma Mingze 1930006138

Xu Jiayun 1930026140

Lin Xiaotao 1930026081

Shao Yuqing 1930024205

Liang Fengjin 1930024130

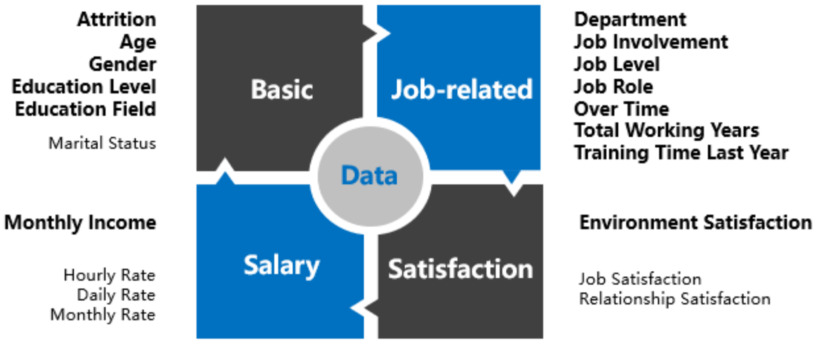
Contents

1. Introduction

The aim of the report is to find out what factors contribute to the attrition of IBM employees. Firstly, we use Chebyshev's inequality to evaluate the distribution of each variable. Then we use decision models to predict whether employees want to leave the company and give partial advice accordingly. Through linear regression and Logistic Regression, we find that over time, the total working year, and monthly income are essential factors in the departure of IBM employees and general advice is given accordingly.

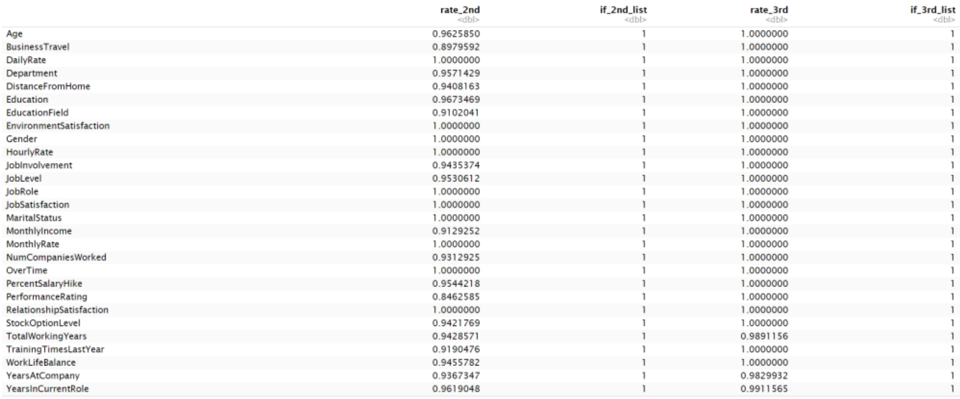
2. Variables description

We find the dataset on Kaggle, which are mainly four categories of data (Details in appendix).



2.1 Preliminary analysis of data - by Chebyshev inequality

In order to have a better understanding of the characteristics of the data and to help us filter out the outliner in the next steps, we use Chebyshev's inequality to evaluate the distribution of each data category, and verify that each quantitative variable satisfies the following condition. Firstly, for a single variable, at least 3/4 (or 75%) of the data lie within 2 standard deviations of the mean. Secondly, at least 8/9 (or 88.9%) of the data lie within 3 standard deviations of the mean. In the table, we display whether the data fit the first and second conditions by 1 for pass and 0 for fail.



	rate_2nd <dbl>	if_2nd_list <dbl>	rate_3rd <dbl>	if_3rd_list <dbl>
Age	0.9625850	1	1.0000000	1
BusinessTravel	0.8979592	1	1.0000000	1
DailyRate	1.0000000	1	1.0000000	1
Department	0.9571429	1	1.0000000	1
DistanceFromHome	0.9408163	1	1.0000000	1
Education	0.9673469	1	1.0000000	1
EducationField	0.9102041	1	1.0000000	1
EnvironmentSatisfaction	1.0000000	1	1.0000000	1
Gender	1.0000000	1	1.0000000	1
HourlyRate	1.0000000	1	1.0000000	1
JobInvolvement	0.9435374	1	1.0000000	1
JobLevel	0.9530612	1	1.0000000	1
JobRole	1.0000000	1	1.0000000	1
JobSatisfaction	1.0000000	1	1.0000000	1
MaritalStatus	1.0000000	1	1.0000000	1
MonthlyIncome	0.9129252	1	1.0000000	1
MonthlyRate	1.0000000	1	1.0000000	1
NumCompaniesWorked	0.9312925	1	1.0000000	1
OverTime	1.0000000	1	1.0000000	1
PercentSalaryHike	0.9544218	1	1.0000000	1
PerformanceRating	0.8462585	1	1.0000000	1
RelationshipSatisfaction	1.0000000	1	1.0000000	1
StockOptionLevel	0.9421769	1	1.0000000	1
TotalWorkingYears	0.9428571	1	0.9891156	1
TrainingTimesLastYear	0.9190476	1	1.0000000	1
WorkLifeBalance	0.9455782	1	1.0000000	1
YearsAtCompany	0.9367347	1	0.9829932	1
YearsInCurrentRole	0.9619048	1	0.9911565	1

Result of testing Chebyshev inequality

From the graph above, we can know that all the data satisfied the first and the second conditions, showing that the distribution of this data set is relatively good, which means outliner has less influence on our analysis.

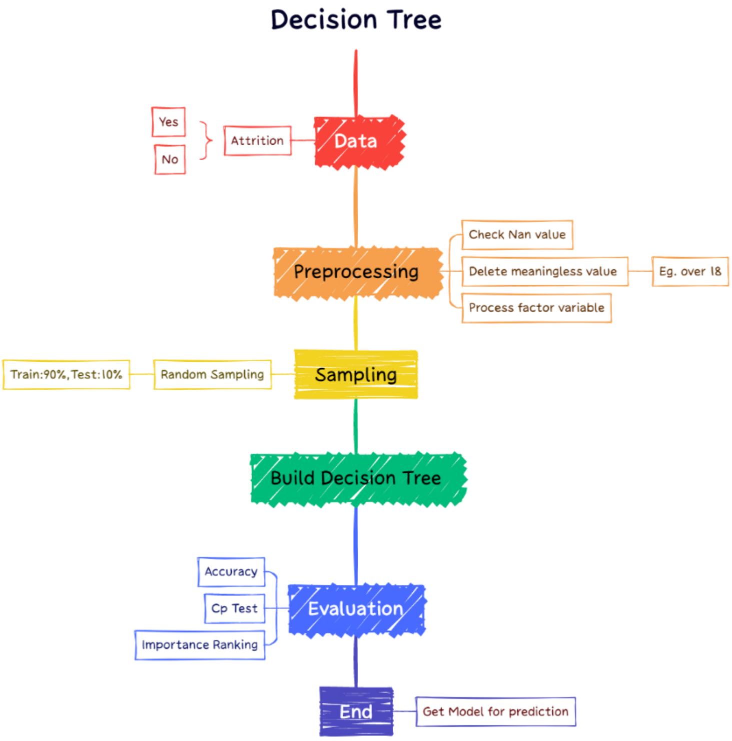
2.2 Employee portrait

We divide the employees into attrition and non-attrition. Of the non-attrition employees, 59% are male, 41% are female, 77% never worked overtime, and most of them are bachelors, studying life science, and married. Besides, most of them are sales executives and research scientists in the sales and R&D departments. Most of the training time last year is 2 or 3 times. Additionally, prominent of them have a high level of environmental satisfaction. Moreover, they are between 24 and 47 years old and have less than ten years of work. Last but not least, most of them have a 2500 to 7500 monthly income, and the job level is 2 or 3.

Of the attrition employees, 63% are female, and 54% have worked overtime before. The distribution of education filed, and level is similar, but most are single. Besides, job roles, departments, and training time are similar, but more attrition employees have low environmental satisfaction. In addition, there is a big difference in the monthly income, and the attrition employees have a lower salary than the non-attrition employees.

3. Attrition analysis

The graph below is the workflow for the attrition analysis.



Our group attempts to use decision models to predict whether employees want to leave the company. The decision tree model is a base model to predict categorical variable. The y label we focus on is “Attrition”. If “Attrition” is yes, it means the staff dismission. Otherwise, the opposite.

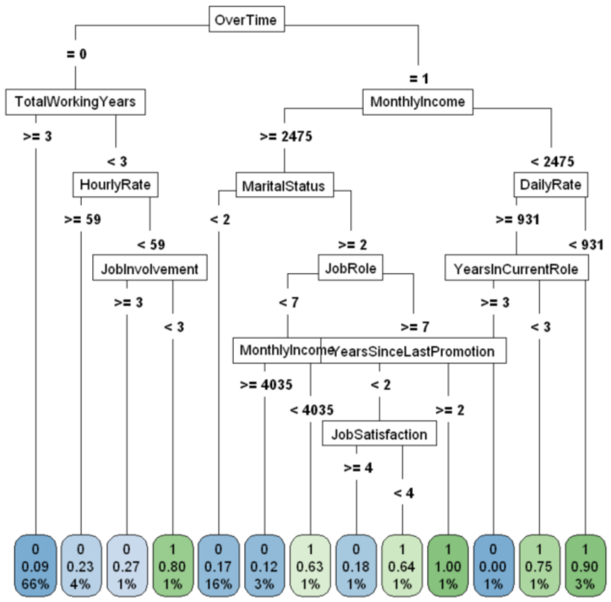
Preprocessing:

For data preprocessing, firstly we check whether the data have null value. After that, we delete the meaningless attributes, which is shown below:

Attributes	Reasons
“Employee Count”	All 1
“Employee Number”	Just like Student ID, no useful information
“Over18”	All “yes”
“Standard Hours”	All 80

For fitting the decision model, we process the categorical value. We transform the string to a factor variable. For example, in “Attrition” label, “Yes” becomes 1. “No” becomes 0.

Building Decision Tree:



Decision Tree Result

Each root represents as a standard attribute for data classification. On the leaf node, the first row represents whether it is classified as attrition or not attrition. The second row represents the proportion of positive classes in the class. The third line represents the proportion of the sample in the class to the total sample. The dark green classes in the figure represent classes with a high probability of leaving, and the dark blue classes represent classes with a high probability of not leaving.

It is obvious that the non-dismission are mainly concentrated on the left side of the figure. Those who leave are concentrated on the right side of the picture. So, for an employee, whether to work overtime becomes a very important criterion for whether an employee leaves the company. But there is an abnormal situation in the 4th leaf node, from the results, this part of the employees' work participation is not high enough, and they do not have a high sense of belonging in the company. Similar anomalies also occur at the 8th and 11th leaf nodes. For the 8th leaf node, we find that most of the employees here are managers with high salaries and high positions. These executives have all experienced job advancement in recent years and enjoy the satisfaction of their jobs. So, job satisfaction also greatly affects how employees feel at work. For the 11th leaf node, the employees here are all low-paid and long-term employees. We speculate that their love for work may have disappeared.

So based on our findings above, we make the following recommendations to the company:

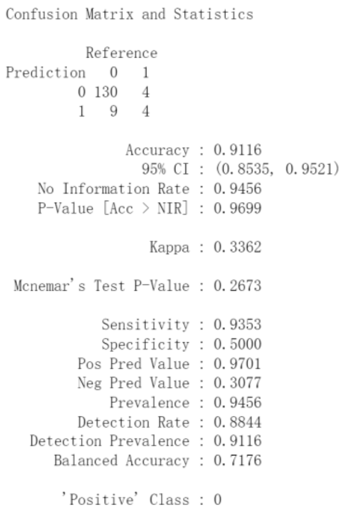
1. For employees with insufficient work participation, an employee incentive work plan can be proposed. Allowing employees to learn on the job can not only complete tasks efficiently, but also enhance employees' trust in their work, thereby greatly increasing employees' participation in work.
2. For employees with high-intensity work and high wages, we suggest that companies should issue relevant decompression policies or activities to ensure that employees

have a normal work attitude. Prevent employees from gradually losing their hearts to the company in a high-pressure work environment.

3. For some employees who do not have passion for work, we suggest that the company can propose some interesting work projects, especially those that are highly related to life and work, which can increase the enthusiasm of employees for work.

Model evaluation:

The graph below shows the confusion matrix and some statistics values.



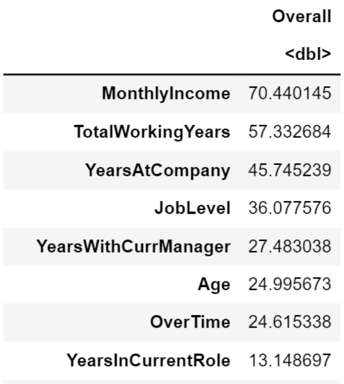
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 130   4
         1   9   4

               Accuracy : 0.9116
                 95% CI : (0.8535, 0.9521)
    No Information Rate : 0.9456
    P-Value [Acc > NIR] : 0.9699

                  Kappa : 0.3362

 Mcnemar's Test P-Value : 0.2673

            Sensitivity : 0.9353
            Specificity : 0.5000
         Pos Pred Value : 0.9701
         Neg Pred Value : 0.3077
             Prevalence : 0.9456
         Detection Rate : 0.8844
   Detection Prevalence : 0.9116
      Balanced Accuracy : 0.7176

       'Positive' Class : 0
```

Performance of decision tree

Since during training, the training set and test set are randomly generated, so every time the decision tree will be different. We ran 10,000 times and found the decision tree model with the highest accuracy.

What’s more, we conduct cp test to prune the tree. We use the $cp = 0.01$ to be the critical value to prune the decision tree. But the new decision tree has lower accuracy. So, we decide to keep the original decision model. The result of CP value detection is shown in appendix.

Finally, we get the importance ranking of each variable:



	Overall
	<dbl>
MonthlyIncome	70.440145
TotalWorkingYears	57.332684
YearsAtCompany	45.745239
JobLevel	36.077576
YearsWithCurrManager	27.483038
Age	24.995673
OverTime	24.615338
YearsInCurrentRole	13.148697

Importance Ranking

From the result we can observe, money, tenure, position, etc. all influence whether an employee leaves. We provide all properties rated above 10, which influence employees' decisions to varying degrees. So companies should mainly work on these issues. Here we take the first three factors to conduct further analysis.

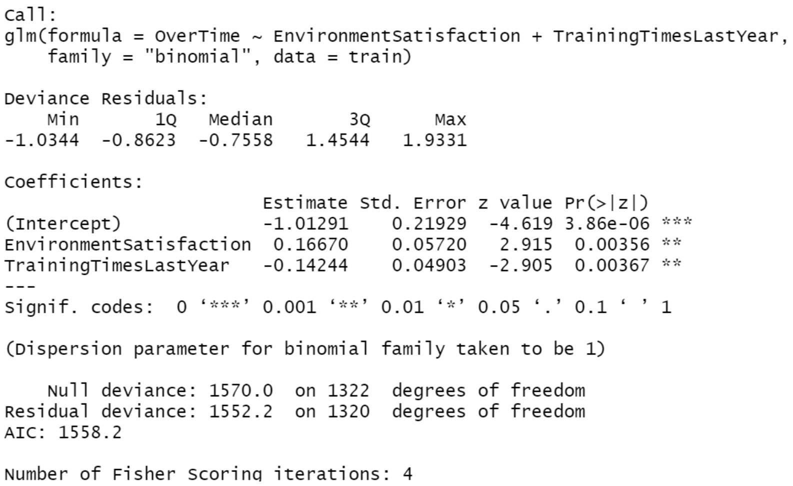
4. Further exploration

4.1 Exploration on Overtime - by Logistic Regression

Through the previous decision tree, employee Overtime has a relatively great effect on whether an employee is likely to leave, so we firstly conduct a study on the qualitative variable Overtime. According to the Tolles and Meurer (2016), Logistic regression can indicate which of the various factors being evaluated has the strongest association with the outcome and provide a measure of the potential degree of influence. So, we apply logistic regression to predict the output of Overtime and analysis the important factors.

Building Logistic Regression

In order to get the variables that have a greater influence, we perform stepwise variable selection by the step function. Next, we fit the model and manually remove the insignificant variables until all of them are significant (smaller than 0.05), and then we obtain a few of the most critical factors, which is our final result.

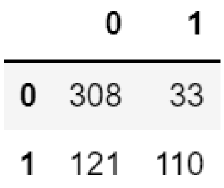


```
Call:
glm(formula = OverTime ~ EnvironmentSatisfaction + TrainingTimesLastYear,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0344  -0.8623  -0.7558   1.4544   1.9331

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.01291    0.21929  -4.619 3.86e-06 ***
EnvironmentSatisfaction  0.16670    0.05720   2.915  0.00356 **
TrainingTimesLastYear   -0.14244    0.04903  -2.905  0.00367 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1570.0  on 1322  degrees of freedom
Residual deviance: 1552.2  on 1320  degrees of freedom
AIC: 1558.2

Number of Fisher Scoring iterations: 4
```

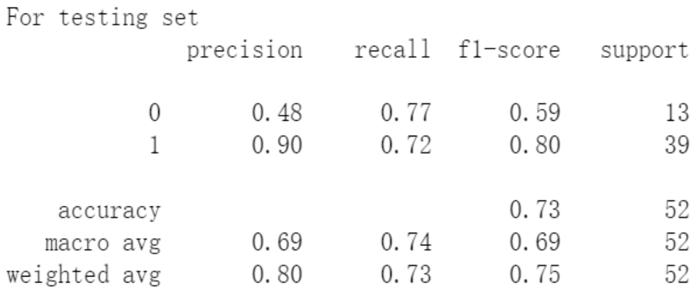
Significant variable after manually selects

Model evaluation

Once the model is built, we need to evaluate it. We split all the data into two parts, one part is used for training, another part is used for testing. For the performance of the model application in the test set, we can make a preliminary analysis of its predicted results by confusion matrix, recall, precision and F1-score. It can be seen that the performance of the model's precision is not so satisfactory. The reason might be the unbalance of the data, which may let the model tend to classify the result into the main group of the result 0. That is the aspect for this model for improvement.



	0	1
0	308	33
1	121	110



```
For testing set
              precision    recall  f1-score   support

           0       0.48      0.77      0.59        13
           1       0.90      0.72      0.80        39

    accuracy                           0.73        52
   macro avg       0.69      0.74      0.69        52
weighted avg       0.80      0.73      0.75        52
```

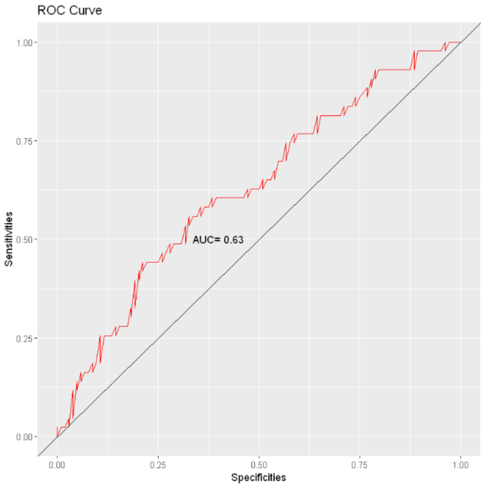
Confusion matrix

Recall, Precision and F1-score

However, the results of the chi-square test and VIF test are ideal. The Chi-square test is used to confirm whether x and y have a strong relationship. And the VIF test is used to describe the multicollinearity among independent variables. Results of chi-square and VIF tests are presented in the appendix.

The p-value of chi-square is small (the number are smaller than 0.1), which means that there is a significant relationship between independent variables and dependent variable, and the VIF for each independent variables are very close to 1, which means that the multicollinearity among independent variables is almost none.

In addition, we also apply the ROC curve to evaluate the model. The ROC curve describes how the performance of the classifier changes with the change of the classifier threshold. From the graph, we can see that it is still possible to make a preliminary estimate of whether there is overtime with these variables.



ROC curve for the model

4.2 Exploration on Total Working Years and Monthly Income - by Linear Regression

From the conclusion of decision tree, we can observe that Total Working Years and

Monthly Income are two important factors affecting Attrition. So, we want to further explore what variables affect these two factors, and use the method of linear regression to explain (Tranmer & Elliot, 2008).

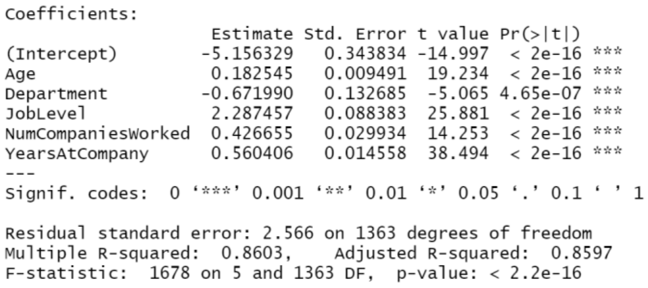
Building Linear Regression

We first analyze the Total Working Years and conduct stepwise regression to get the variables that have the smallest AIC, meaning the most concise and effective model. Then we do the t test on each variable, and manually delete the insignificant variables until all of them are significant. This process improves the R-squared from 0.487 to 0.7092 and Adjusted R-squared from 0.4838 to 0.7894 respectively.

Model evaluation

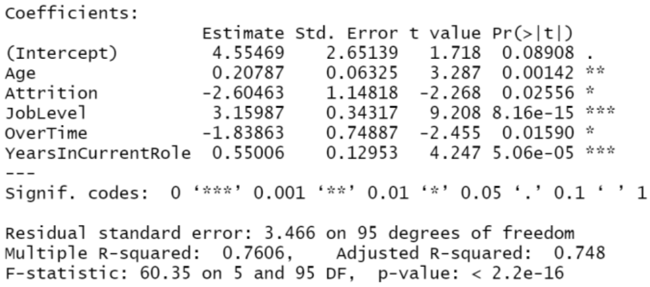
Then we conduct several tests on the model. The VIF test checks whether the model has multicollinearity. From the results we can see, the value of Job Level and Monthly Income is relatively high, meaning they are highly correlated. Thus, we eliminate the Monthly Income and keep the VIF within the threshold. Next, we do the ANOVA test to see if x truly affects y, and the final result is significant. After that, we perform a transformation test to see if y or x needs to be transformed. Since Box-Cox Transformation is not equal to 0 and powerTransform is small, we don’t have significant evidence to proof that y or x needs to be changed.

Finally, we perform the diagnostic plot. After deleting outliers, the model basically holds the assumption of independent residuals, normal distribution, equal variance and has no influential point. By refitting the model, our multiple R-squared and adjusted R-Squared is now 0.8603 and 0.8597, highly improving the performance. We also fit a model for outliers and compare them with the rest of employees to see what are the characteristics of this group of people. Following shows our final model for total working years and its outliers.



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -5.156329   0.343834 -14.997  < 2e-16 ***
Age               0.182545   0.009491  19.234  < 2e-16 ***
Department       -0.671990   0.132685  -5.065 4.65e-07 ***
JobLevel          2.287457   0.088383  25.881  < 2e-16 ***
NumCompaniesWorked 0.426655  0.029934  14.253  < 2e-16 ***
YearsAtCompany    0.560406   0.014558  38.494  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.566 on 1363 degrees of freedom
Multiple R-squared:  0.8603,	Adjusted R-squared:  0.8597 
F-statistic:  1678 on 5 and 1363 DF,  p-value: < 2.2e-16
```

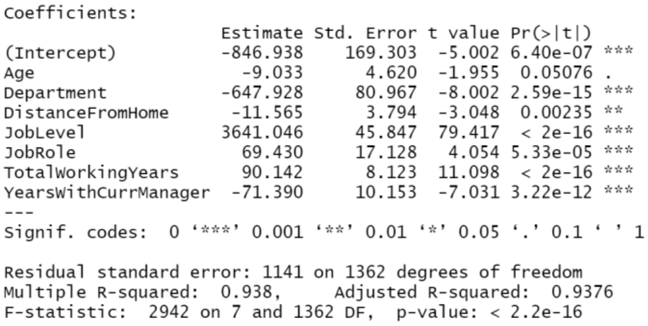
Final model for Total Working Years



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          4.55469    2.65139   1.718  0.08908 .  
Age                  0.20787    0.06325   3.287  0.00142 ** 
Attrition           -2.60463    1.14818  -2.268  0.02556 *  
JobLevel             3.15987    0.34317   9.208 8.16e-15 ***
OverTime            -1.83863    0.74887  -2.455  0.01590 *  
YearsInCurrentRole   0.55006    0.12953   4.247 5.06e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.466 on 95 degrees of freedom
Multiple R-squared:  0.7606,	Adjusted R-squared:  0.748 
F-statistic: 60.35 on 5 and 95 DF,  p-value: < 2.2e-16
```

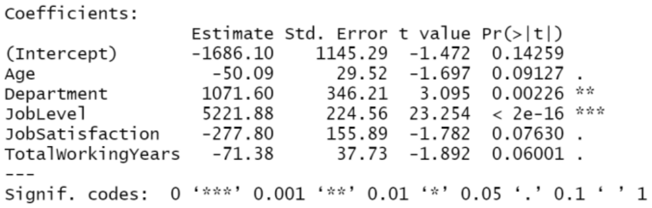
Model for Total Working Years outliers

Same as the previous procedure, we fit the model for monthly income and get the final model which has an R-squared and adjusted R-squared up to 0.938 and 0.9376. A model for outliers is also given as follows.



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -846.938    169.303  -5.002 6.40e-07 ***
Age                  -9.033      4.620  -1.955  0.05076 .  
Department         -647.928     80.967  -8.002 2.59e-15 ***
DistanceFromHome    -11.565      3.794  -3.048  0.00235 ** 
JobLevel           3641.046     45.847  79.417  < 2e-16 ***
JobRole              69.430     17.128   4.054 5.33e-05 ***
TotalWorkingYears    90.142      8.123  11.098  < 2e-16 ***
YearsWithCurrManager -71.390    10.153  -7.031 3.22e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1141 on 1362 degrees of freedom
Multiple R-squared:  0.938,	Adjusted R-squared:  0.9376 
F-statistic:  2942 on 7 and 1362 DF,  p-value: < 2.2e-16
```

Final model for Monthly Income



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -1686.10    1145.29  -1.472  0.14259    
Age                 -50.09      29.52  -1.697  0.09127 .  
Department         1071.60     346.21   3.095  0.00226 ** 
JobLevel           5221.88     224.56  23.254  < 2e-16 ***
JobSatisfaction    -277.80     155.89  -1.782  0.07630 .  
TotalWorkingYears   -71.38      37.73  -1.892  0.06001 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model for Monthly Income outliers

5. Solutions

We can see the significant impact of overtime on attrition from the decision tree. This result is easily explained by the fact that overtime affects the time available for leisure activities, leaving employees with insufficient time to take a break. When employees are trapped in an exhausted state, the higher the probability they want to quit (Anderson & Buchholz, 1988). Therefore, companies should try to avoid employees working overtime. Then what can companies do to reduce the amount of overtime their employees work?

In the logistic regression analysis, we find two variables that were significantly associated with overtime: training times last year and environment satisfaction. The more training times in the previous year, the less likely overtime in the current year. Training can improve employees' productivity and thus reduce the probability of overtime (Konings & Vanormelingen, 2015). Companies should plan to provide employees with training to help them improve their work efficiency and adapt to the dynamic business environment.

There is also a significant relationship between environmental satisfaction and overtime. Employees with higher environment satisfaction for the same amount of training times in the previous year were more likely to work overtime. It might mean that employees are more willing to work overtime when the work environment is friendly. It is a positive phenomenon, for which we do not propose improvements.

Since the total working years and monthly income are the decision tree nodes only second to overtime. We conducted linear regressions for total working years and monthly income, respectively. The factors that have a significant impact on total working years are age, department, job level, number of companies worked, and years of working in IBM. In a regression of the total working years outliers, we found that age, job level, and years in the current role had a significant effect. The factors that have a significant effect on monthly income are department, job level, job role, total working years, and years of working with current managers. For the analysis of monthly income outliers, we find that only department and job level had a significant effect.

Based on the above data we find that there is a significant positive effect of job level on both total years of work and monthly income, regardless of whether the person fits the distribution or is an outlier. Our recommendation to the company is that it should continuously refine the job level system and promptly promote employees with good performance (Guan et al, 2014), which can directly increase the monthly income of employees and indirectly promote their retention.

6. References

Anderson, E. E., & Buchholz, R. A. (1988). Economic instability and occupational injuries: The impact of overtime hours and turnover rates. Labor Studies Journal, 13(4), 33-49.

CFI. (n.d.). Sprout. (n.d.). *Stock Option.* CFI.

Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis*. Elsevier.

Guan, Y., Wen, Y., Chen, S. X., Liu, H., Si, W., Liu, Y., Wang, Y., Fu, R., Zhang, Y., & Dong, Z. (2014). When do salary and job level predict career satisfaction and turnover intention among Chinese managers? the role of perceived organizational career management and career anchor. European Journal of Work and Organizational Psychology, 23(4), 596-607.

Hoffman, M., & Tadelis, S. (2021;2020;). People management skills, employee attrition, and manager rewards: An empirical analysis. *The Journal of Political Economy, 129*(1), 243-285. https://doi.org/10.1086/711409

IBM. (2021). *2021 IBM Annual Report*. IBM. https://www.ibm.com/

Lambooij, M., Flache, A., Sanders, K., & Siegers, J. (2007). Encouraging employees to co-operate: The effects of sponsored training and promotion practices on employees' willingness to work overtime. *International Journal of Human Resource Management, 18*(10), 1748-1767. _

Konings, J., & Vanormelingen, S. (2015). The impact of training on productivity and wages: Firm-level evidence. The Review of Economics and Statistics, 97(2), 485-497.

Myers, R. (2005). Classical and Modern Regression with Applications (Second Edition). Higher Education Press.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *Journal of the Royal*

Statistical Society: Series D (The Statistician), *41*(2), 169-178.

Sprout. (n.d.). *How to Calculate for the Daily Rate from Your Monthly Salary?* Sprout. https://sprout.zendesk.com/hc/en-us/articles/360030922133-How-to-Calculate-for-the-Daily-Rate-from-Your-Monthly-Salary-

Tolles, J., & Meurer, W. J. (2016). Logistic regression: Relating patient characteristics to outcomes. *JAMA: The Journal of the American Medical Association, 316*(5), 533-534. https://doi.org/10.1001/jama.2016.7653

Appendix

1. Data description

Basic information	
Attrition	Employees leave the company or not
Age	Employee’s age
Gender	Employee’s gender
Over 18	Age 18 or over
Distance From Home	Distance from home to company
Education	Education level 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
Salary information	
Hourly Rate	Hourly Rate = Income / Hours
Daily Rate	(Basic Monthly Salary x 12) / (Total Working Days in a Year) = Daily Rate Basic salary refers to the monetary amount paid to an employee, excluding added bonuses or deductions from absences or incurred late

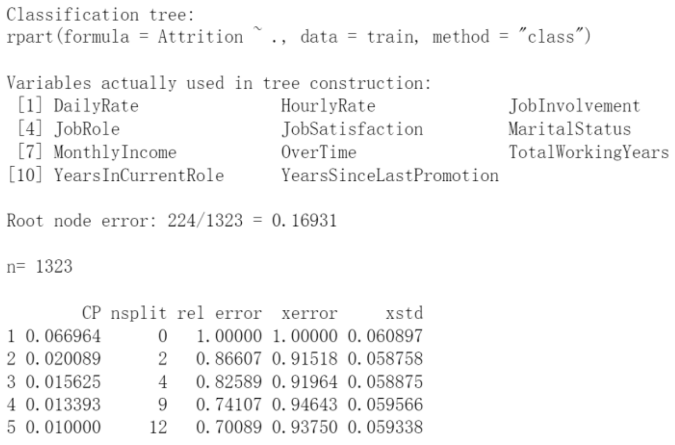
	minutes.
Monthly Rate	(Basic Monthly Salary x 12) / (Total Working Months in a Year) = Monthly Rate
Monthly Income	Salary received each month
Job-related information	
Employee Count	The number of employees (all is one)
Employee Number	Staff No.
Job Involvement	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
Job Level	Staff position levels
Job Role	Staff position (Ordinal variable)
Num Companies Worked	Total number of companies worked for now and in the past
Over Time	Ever worked overtime or not
Percent Salary Hike	Percentage of salary increase
Performance Rating	Staff performance assessment
Standard Hours	Statutory Standard Working Hours

Total Working Years	Total number of years the employee has worked (whether in this company or not)
Training Time Last Year	Training time received in the previous year
Years At Company	Years of working for this company
Years Since Last Promotion	Number of years since last promotion
Years With Current Manager	Years spent working with the current manager
Years In Current Role	Number of years spent in current position
Employee satisfaction	
Environment Satisfaction	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
Job Satisfaction	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
Relationship Satisfaction	Satisfaction with relationships with supervisors or colleagues 1 'Low'

	2 'Medium' 3 'High' 4 'Very High'
Others	
Stock Option Level	A stock option is a contract between two parties that gives the buyer the right to buy or sell underlying stocks at a predetermined price and within a specified time period. The level means the right level of the buyer.
Work Life Balance	1 'Bad' 2 'Good' 3 'Better' 4 'Best'

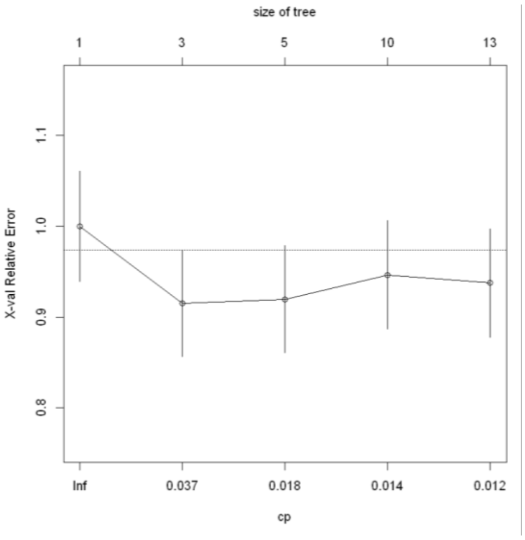
2. Modeling result

Decision tree



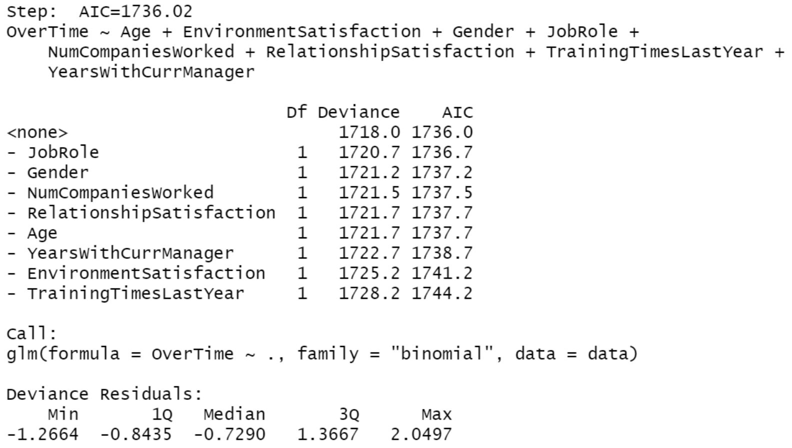
```
Classification tree:
rpart(formula = Attrition ~ ., data = train, method = "class")

Variables actually used in tree construction:
 [1] DailyRate            HourlyRate           JobInvolvement
 [4] JobRole              JobSatisfaction      MaritalStatus
 [7] MonthlyIncome        OverTime             TotalWorkingYears
[10] YearsInCurrentRole   YearsSinceLastPromotion

Root node error: 224/1323 = 0.16931

n= 1323

        CP nsplit rel error  xerror     xstd
1 0.066964      0   1.00000 1.00000 0.060897
2 0.020089      2   0.86607 0.91518 0.058758
3 0.015625      4   0.82589 0.91964 0.058875
4 0.013393      9   0.74107 0.94643 0.059566
5 0.010000     12   0.70089 0.93750 0.059338
```

CP test result



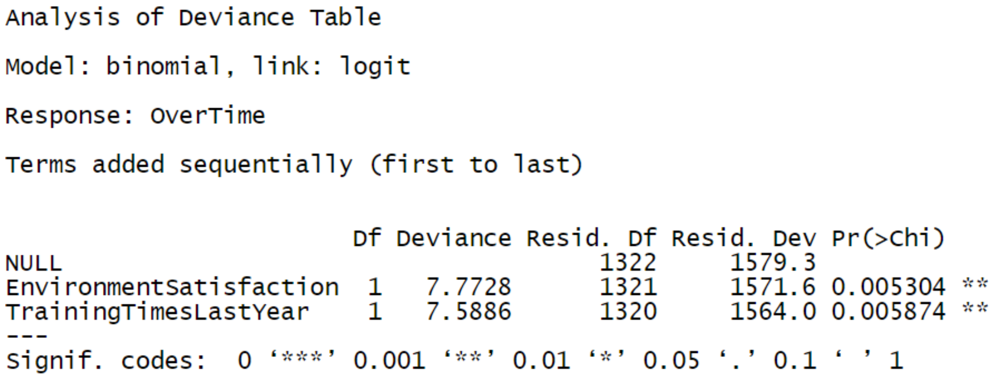
Relative error VS cp statistics

Logistic regression



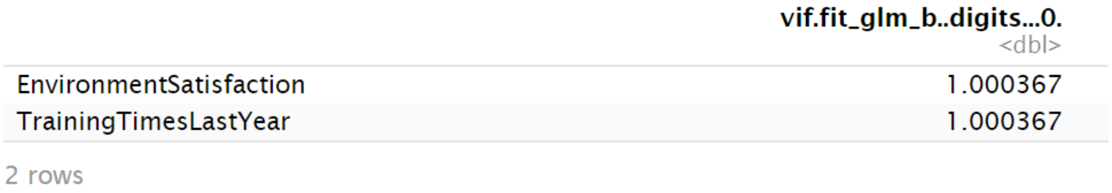
```
Step:  AIC=1736.02
OverTime ~ Age + EnvironmentSatisfaction + Gender + JobRole + 
    NumCompaniesWorked + RelationshipSatisfaction + TrainingTimesLastYear + 
    YearsWithCurrManager

                           Df Deviance    AIC
<none>                          1718.0 1736.0
- JobRole                   1   1720.7 1736.7
- Gender                    1   1721.2 1737.2
- NumCompaniesWorked        1   1721.5 1737.5
- RelationshipSatisfaction  1   1721.7 1737.7
- Age                       1   1721.7 1737.7
- YearsWithCurrManager      1   1722.7 1738.7
- EnvironmentSatisfaction   1   1725.2 1741.2
- TrainingTimesLastYear     1   1728.2 1744.2

Call:
glm(formula = OverTime ~ ., family = "binomial", data = data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.2664  -0.8435  -0.7290   1.3667   2.0497  
```

Stepwise regression for OverTime



```
Analysis of Deviance Table

Model: binomial, link: logit

Response: OverTime

Terms added sequentially (first to last)


                        Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
NULL                                     1322     1579.3            
EnvironmentSatisfaction  1   7.7728      1321     1571.6 0.005304 **
TrainingTimesLastYear    1   7.5886      1320     1564.0 0.005874 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Chi-square test for the model

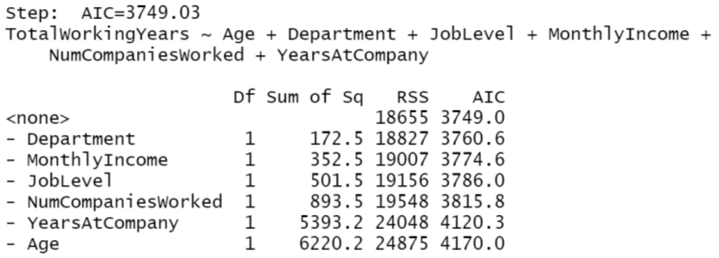


	vif.fit_glm_b..digits...0. <dbl>
EnvironmentSatisfaction	1.000367
TrainingTimesLastYear	1.000367

2 rows

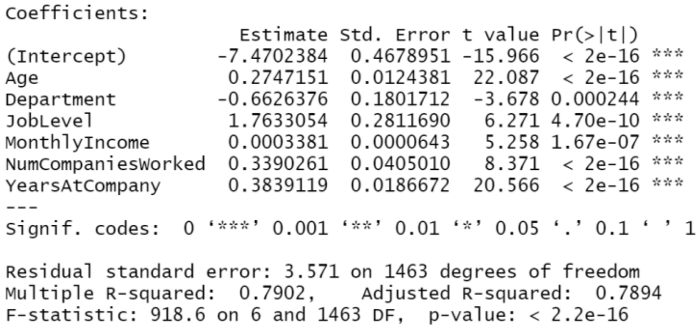
VIF for each independent variable

Linear regression



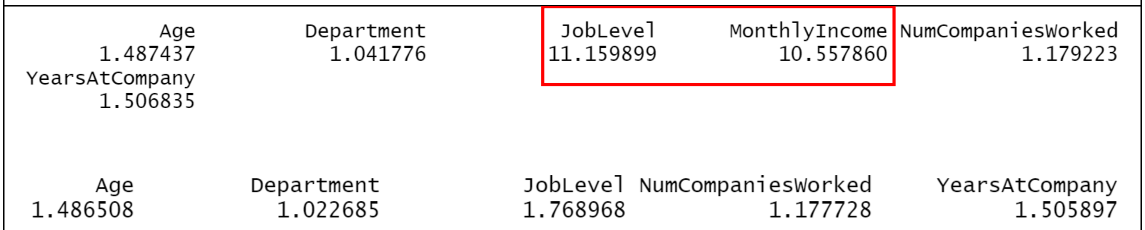
```
Step:  AIC=3749.03
TotalWorkingYears ~ Age + Department + JobLevel + MonthlyIncome + 
    NumCompaniesWorked + YearsAtCompany

                     Df Sum of Sq   RSS    AIC
<none>                            18655 3749.0
- Department          1     172.5 18827 3760.6
- MonthlyIncome       1     352.5 19007 3774.6
- JobLevel            1     501.5 19156 3786.0
- NumCompaniesWorked  1     893.5 19548 3815.8
- YearsAtCompany      1    5393.2 24048 4120.3
- Age                 1    6220.2 24875 4170.0
```

Stepwise regression for Total Working Years



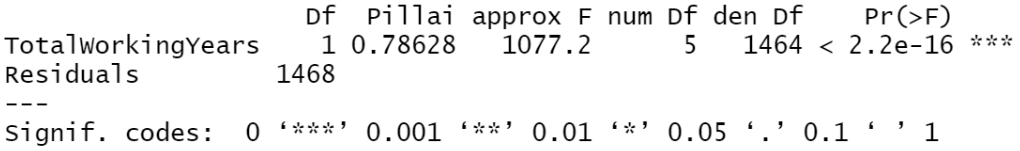
```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -7.4702384  0.4678951 -15.966  < 2e-16 ***
Age                 0.2747151  0.0124381  22.087  < 2e-16 ***
Department         -0.6626376  0.1801712  -3.678 0.000244 ***
JobLevel            1.7633054  0.2811690   6.271 4.70e-10 ***
MonthlyIncome       0.0003381  0.0000643   5.258 1.67e-07 ***
NumCompaniesWorked  0.3390261  0.0405010   8.371  < 2e-16 ***
YearsAtCompany      0.3839119  0.0186672  20.566  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.571 on 1463 degrees of freedom
Multiple R-squared:  0.7902,	Adjusted R-squared:  0.7894 
F-statistic: 918.6 on 6 and 1463 DF,  p-value: < 2.2e-16
```

Total Working Years model with all significant variables



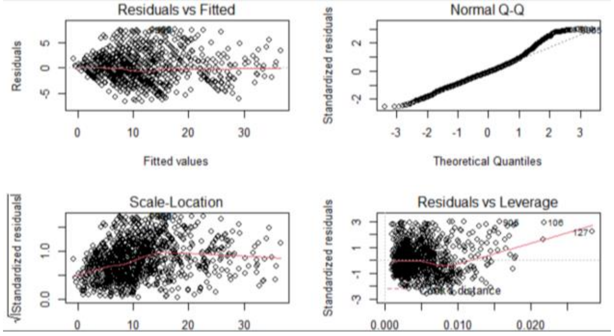
```
               Age         Department           JobLevel      MonthlyIncome NumCompaniesWorked 
          1.487437           1.041776          11.159899          10.557860           1.179223 
    YearsAtCompany 
          1.506835 


               Age         Department           JobLevel NumCompaniesWorked     YearsAtCompany 
          1.486508           1.022685           1.768968           1.177728           1.505897 
```

VIF test for Total Working Years model



```
                    Df  Pillai approx F num Df den Df    Pr(>F)    
TotalWorkingYears    1 0.78628   1077.2      5   1464 < 2.2e-16 ***
Residuals         1468                                             
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA test for Total Working Years model



Diagnostic plot for Total Working Years model