

Data 501 Project

Analysis on Employees’ Attrition by Statistical Learning Methods

HaoQiu Wu

93166163

# Background and Problem

Machine learning is a choice for HR to better understand what is happening among employees. The machine learning method has been applied in human resource management for a long time. For example, Glint is not an AI company, but they use the machine learning method to speed up the process of selecting new employees. They record where the applicants find job ads; they split all resume information into different categories; they record the applicants’ social media activities; they record every detail of the applicants. Finally, they feed all the information into machine learning models to get the analysis results. Obviously, algorithmic-based assessment can more effectively help a company choose the ideal members from the candidates.

In terms of the employee attrition problem, some companies even build a corresponding engagement platform for advanced emotional pattern recognition through investigated survey and comments. This is because experienced employees’ leaving is a big loss to companies. To be more specific, employees are the basic unit of an organization and it is the employees who build the functionality of the organization by working day and night.

Therefore, the high turnover rate of employees can dramatically influence a company’s overall competitiveness. Such effects can be mainly divided into several aspects. Firstly, an organization suffering from high rate of turnover is more likely to incur higher cost, because replacing employees means companies have to spend more expenditure on recruiting advertisement and training. Secondly, productivity of a company will decrease. With losing experienced employees, companies have to spend extra time and energy on helping new employees instead of dedicating to designing products and developing clients. Thirdly, employees who leave companies will join in the vast job market again, which means there is a high probability for them to be hired by competitors, so they will apply the experience gained from the previous company towards serving the new company. Finally, frequently replacing employees can destroy the positive company culture, because some employees may have to spend a long time to develop a close relationship with new colleagues. Continuous cooperation among employees improves work efficiency where no matter who is in trouble, others are always willing to help. Such positive culture makes employees feel like a company is their family so they can make more efforts to working. Hence, by successfully sustaining employees, organizations can cut cost, develop better work culture and increase competitiveness.

Within such an environment, it does make sense for HR professionals to pay more attention to sustaining hardworking and high-evaluated employees. HR should establish a better employee retention system, which can help detect unhappy employees in advance so that HR can have enough time to help these employees become motivated, happy and personally engaged again. Therefore, it is necessary for HR to know that why previous experienced employees leave and which factors mainly result in leaving with the assistance from the machine learning method.

In this report, PCA analysis, unsupervised learning, logistic regression, classification tree, random forest, KNN and support vector machine will be applied. After using cross-validation and ROC curve to assess models, I can get both better prediction of who will leave and explanations of which factor mainly results in leaving. In summary, I have two goals to help HR: firstly, I want to understand why valuable employees leave, and secondly, I want to predict who will leave next.

# Data Set Introduction

The data I used comes from Kaggle. It is provided by an HR department. This dataset has nine predictors: satisfaction level, latest evaluation (yearly), the number of projects worked on, average monthly hours, time spent in the company (in years), work accidents (within the past 2 years), promotions within the past 5 years, department and salary.

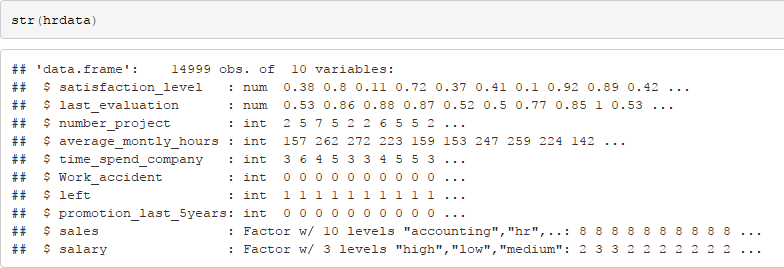


Fig.1. Data type of each variable

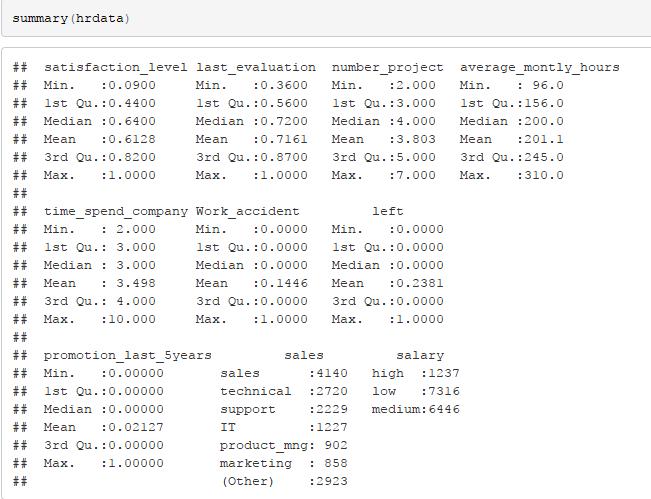


Fig.2. Summary of data

# Overview of Software

Microsoft Excel will be used for global data exploration and data cleaning. It has a simple user interface and allows people to easily understand operations and perform basic operations in a short time.

R is good at visualizing data. In data science, people usually choose plotting data to show analysis results to clients. Therefore, whether a software can generate outstanding plots is a key selecting criteria here. Thanks to Hadley Wickham for an incredible ggplot2 package, I will use R here for data visualization and generating supervised machine learning models.

# Interpretation

## Data Overview and Data Clean

### Data Clean

In this step, I will use excel to clean the data and show the overall distribution of the data. Firstly, I will check whether there exists NULL values and transform them if they do exist. Secondly, I will rename variables to make them more readable. Finally, through plotting each variable’s distribution, I will try to find and describe their statistical features.

In the excel, “COUNTIF” function can be used to count the number of cells that satisfy some particular conditions. Therefore, I firstly use “IF” function to replace blank cells with “NA”. Then I use “COUNTIF” function to check whether there exist blank cells. The result shows that there are not blank cells in this data set.

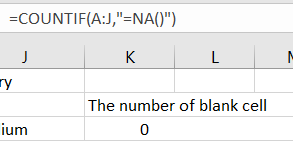
 C:\Users\WHQ672~1\AppData\Local\Temp\1520471263(1).png

Fig.3. Result of Data Clean

In terms of some confusing variable names like “time spent in company” and “sales”, I rename them as “years spent in company” and “department”.

### Data Overview

Again, “COUNTIF” function is used to count the number of employees who left so overall leave rate can be easily calculated, which is 0.238.

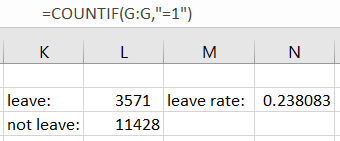


Fig.4. Overall Calculate Rate

Because here I only have nine variables, it is possible to plot the distribution of each variable. In terms of numeric ones, statistical methods will be used to test whether employees who left have different properties from those who stay.

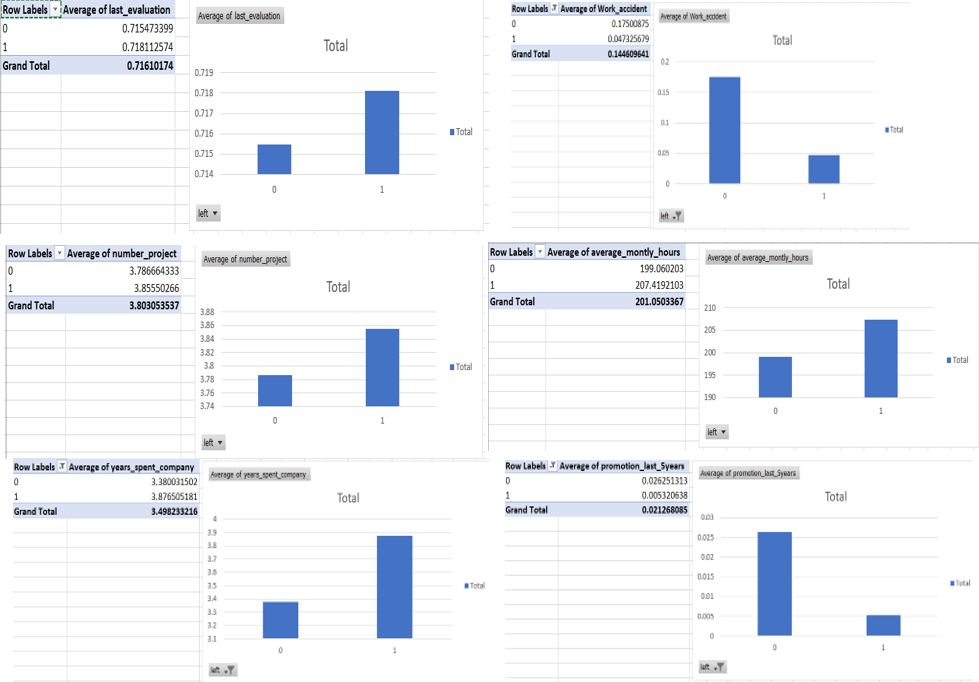


Fig.4. Global Exploration of Variable

From above plots, it is clear that people who left and people who stay do not have same mean values. In order to verify whether these mean values are truly different from each other, two-sample T-Test is used. Also, F-Test is used to check whether two samples have different variance values, which is necessary for conducting the two-sample T-Test.

In the hypothesis test, whether two samples have different means is checked under the significant level 0.05. Null hypothesis(H0) is: the difference between two sample means is zero. Alternative hypothesis(H1) is: the difference between two sample means is not zero.

According to the result, the p-value for “last evaluation” is 0.4683, the p-value for “number of projects” is 0.03034, so there are enough evidences to say people who left and who stay do not have different means of “last evaluation” and “number of projects”. By contrast, the p-value for else other variables are all far less than 0.05, so we can say that people who left and people who stay have different means.

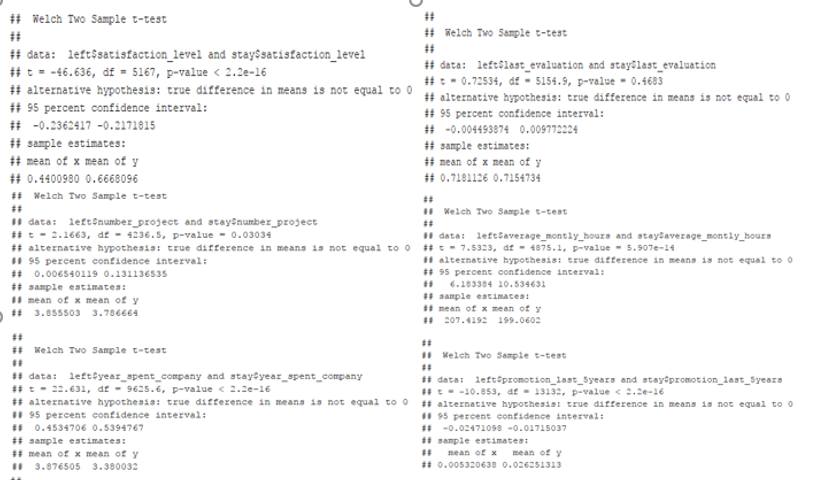


Fig.5. Results of Two Sample T-Test

## Correlation Check and Pairwise Variable Distribution Plot

In this step, I will use R to check correlated variables and plot their relationships. In general, “correlation variable” means that two variables are connected so that one changes with a specific trend if you change the other. The closer this coefficient is to zero the weaker the correlation is. Both 1 and -1 are the cases of strongest correlation and anti-correlation.

From the figure of correlation matrix, variable pairs that have positively correlated relationship are “number\_project” and “last\_evaluation,month\_hours”, “last\_evaluation” and “number projects”, “month\_hours” and “number\_projects”. Variable pairs that have negatively correlated relationship are “left” and “satisfaction\_level”.

For pairs that have the positively correlated relationship, reasons may be that people who work on more projects will always have to work for longer hours and therefore have higher evaluations from the company. Also, it is reasonable that employees who are not satisfied with the company (low satisfaction level) will eventually leave. Finally, in terms of the response variable – “left”, there exists a slight correlated relationship with “average month hours”, “year spent in company” and “work accident”.

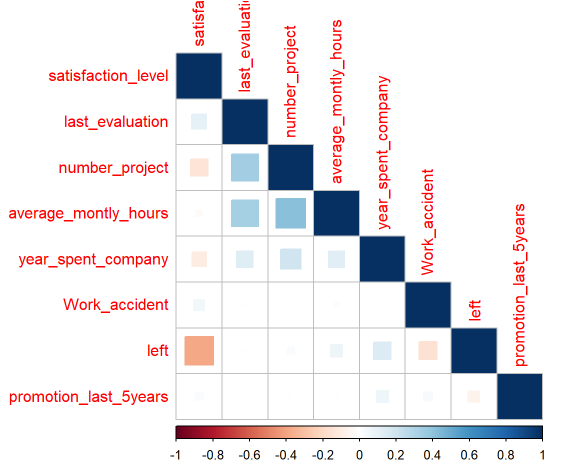


Fig.6. Correlation Matrix

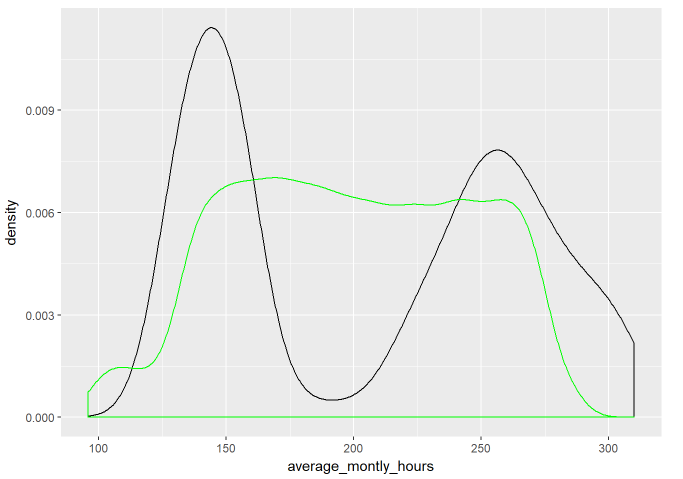
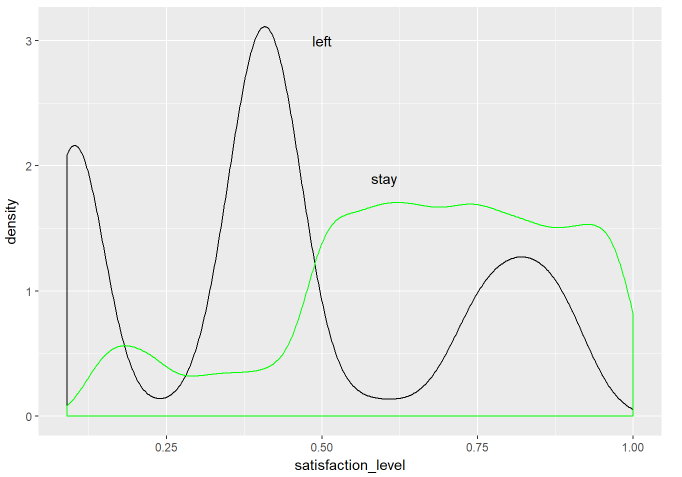
Therefore, according to the result of previous T-Test and the correlation matrix plot, I will check detailed distributions between those variable pairs.

For “left” and “satisfaction” plot, it is obvious that there are three huge spikes for people who left. They are around 0.1,0.4 and 0.8 respectively. For people who stay, their satisfaction levels are usually greater than 0.5.

For “left” and “month hours” plot, there are two spikes that the value of “month hours” of people who left is usually around 150 or 260, which means that people who left are either overworked or underworked. By contrast, people who stay usually have stable and medium working hours.

For “left” and “number of projects” plot, all of people who have seven projects and the majority of people who have two or six projects leave the company

For “left” and “year spent in company” plot, the value of “year spent in company” of people who left focuses on 3,4 and 5. In other words, people who works over 6 years or less than 2 years seldom leave the company.



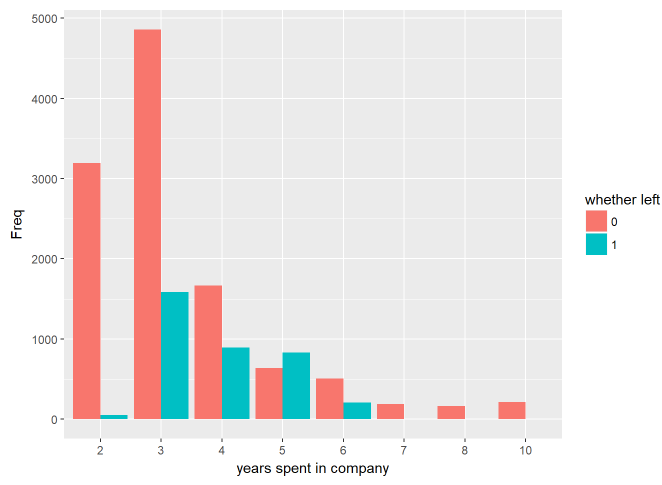
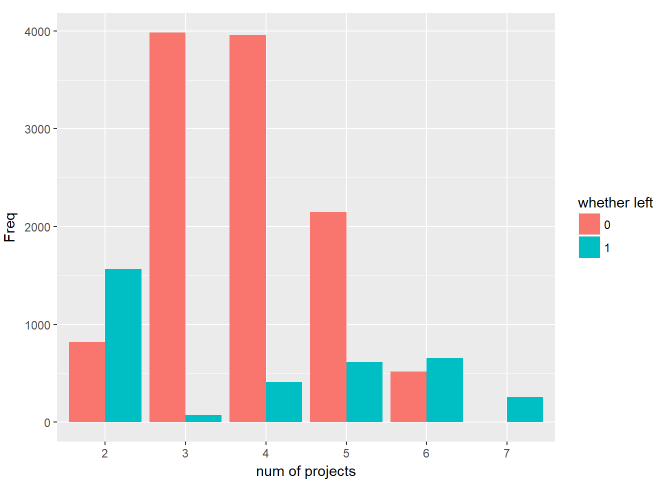
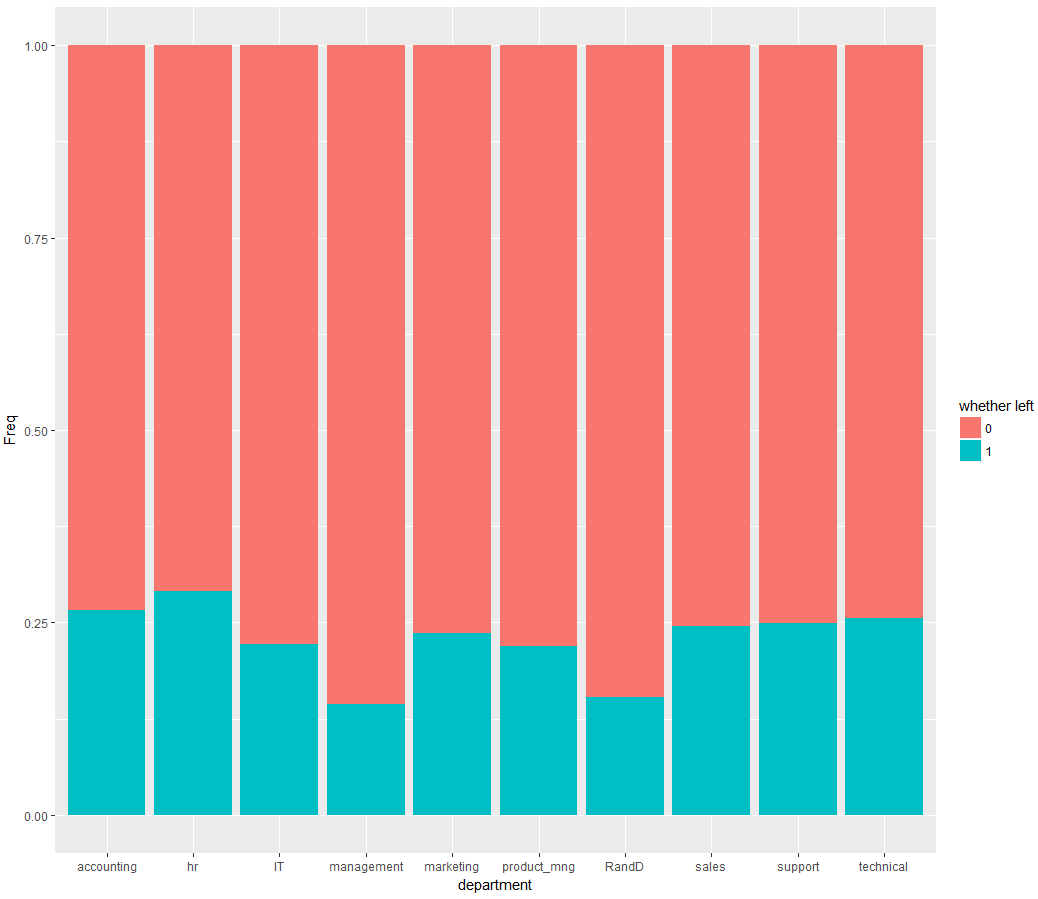
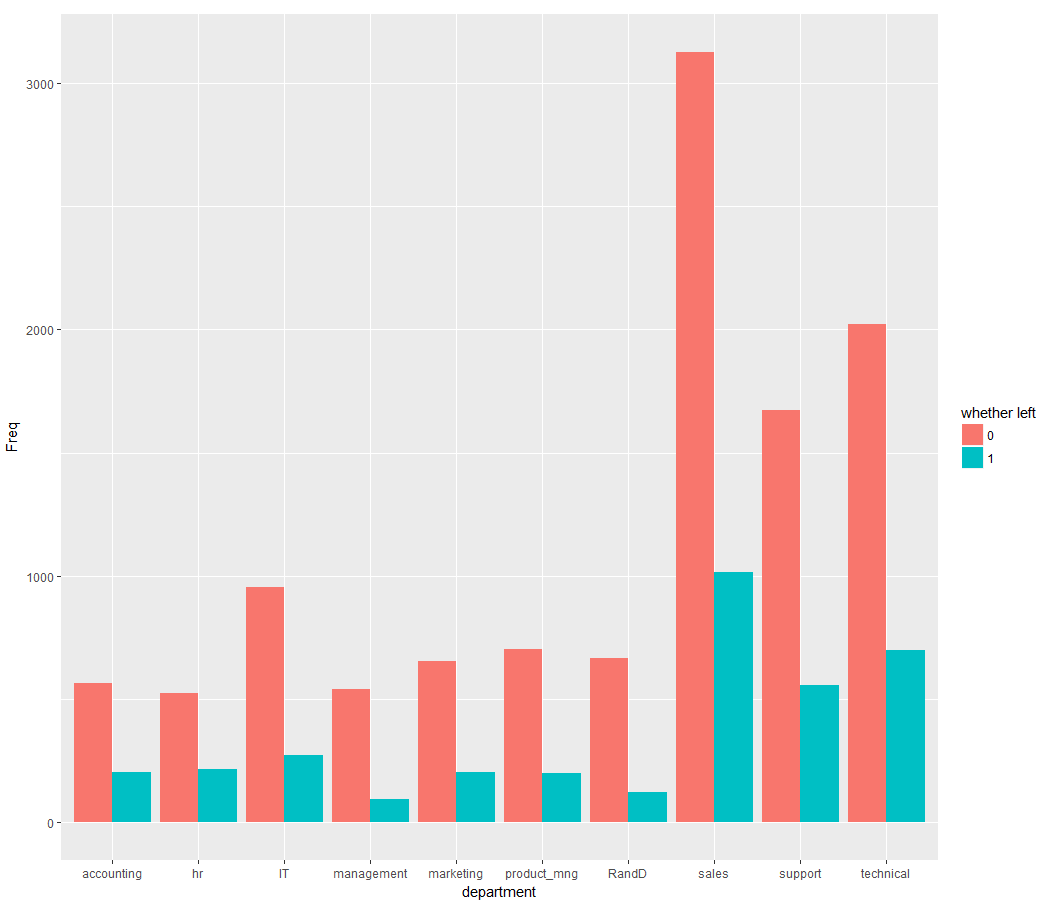


Fig.8. Plots of Some Numeric Variables

For “left” and “department” plot, sales department ranks first in terms of the number of leaving people and management department has the lowest number of leaving people.

For “left” and “salary” plots, there are not any interesting phenomenon because it is reasonable that people has low salary are more likely to leave compared with people who have the high salary.



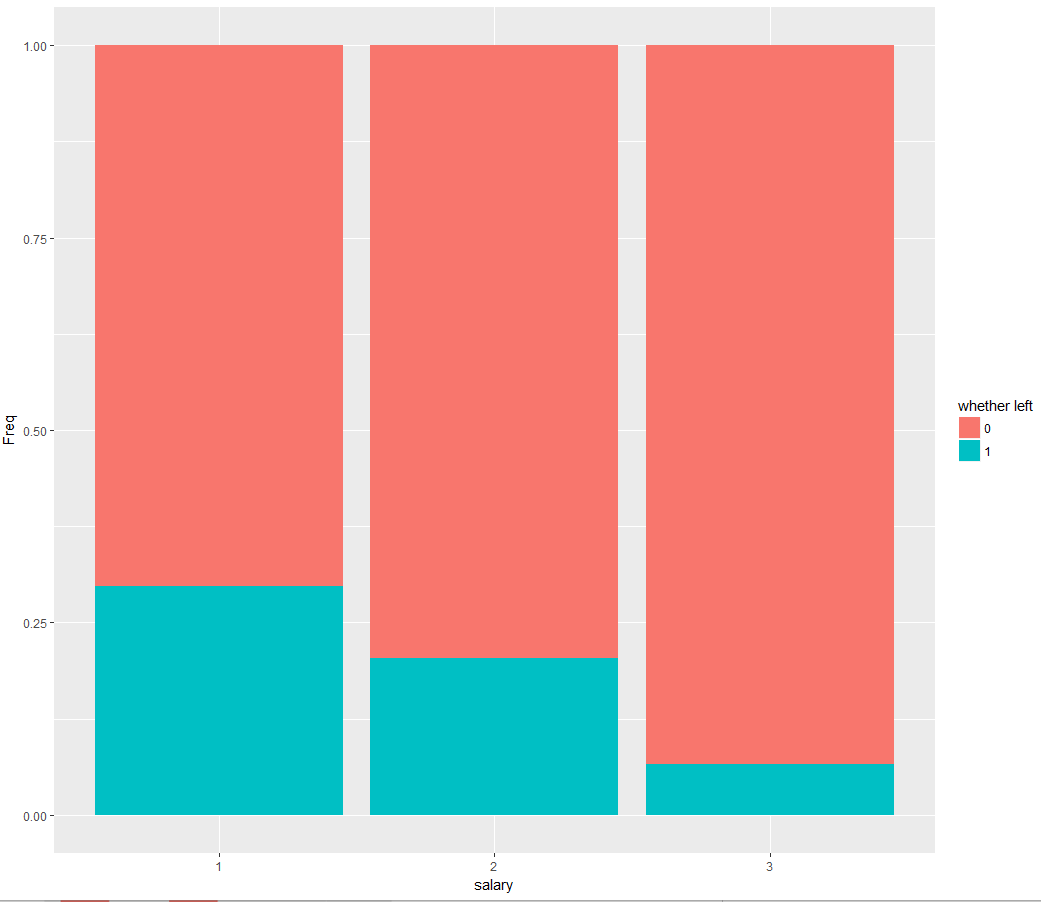
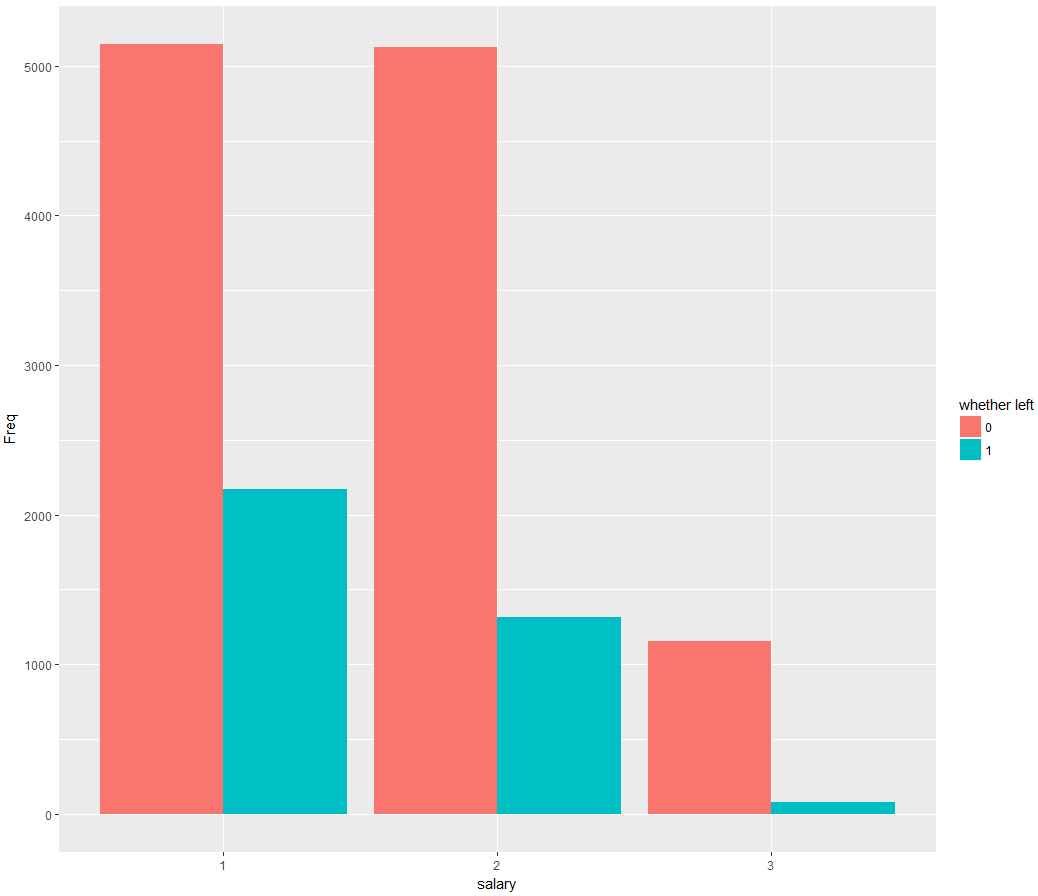


Fig.9. Plots of Categorical Variables

## Standardization, Principle Component Analysis and Hierarchical Clustering

In this step, I will firstly standardize variables to make them have mean zero and standard deviation one before using the dimension reduction method because variables with different variances can invalidate models. In other words, without scaling variables, the PCA method may not return the right result. After standardization, I will use PCA to try to show data in lower dimensions and use hierarchical clustering methods to see whether there exist potential clusters.

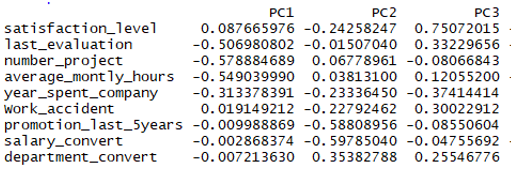
PCA is often used as a tool in visualizing unsupervised data. This method aims at representing data features in a lower dimension through removing redundant features from data sets, especially ones containing more than three variables, without losing much information.

The hierarchical clustering method looks forward to finding potential subgroups among observations. This is an easy and natural approach to split data into distinct and non-overlapping groups.

The following is the formula of first principle component, which is the liner combination of original *p* features. The order of the principle components is made according to how much variance of observations each component can explain. Constant coefficients in the formula are called loadings that can be considered as the weight of a particular predictor.

C:\Users\whq672437089\AppData\Roaming\Tencent\Users\672437089\TIM\WinTemp\RichOle\H20$9Q3U]45}]TWW_0V[L]G.png (1)

Therefore, after performing the PCA analysis, the result shows that the first principle component explains 20% of the total variance, and both the second principle component and the third principle component explain 12% of the total variance. In detail, for the first principle component, it assigns high weights to “last evaluation”, “number projects” and “average month hours”, so it can be concluded that it focuses on how hardworking the employees are. For the second principle component, it assigns high weights to “promotion last year” and “salary”, so it can be concluded that it focuses on how well the employees are treated in the company. Finally, for the third principle component, it assigns the highest value to “satisfaction level”, therefore it focuses on the how employees feel about the company.



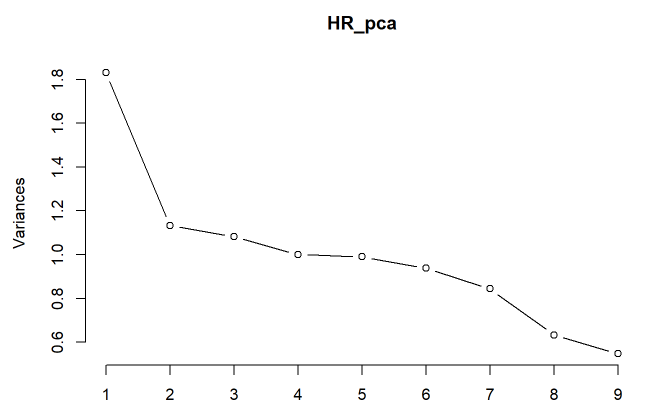
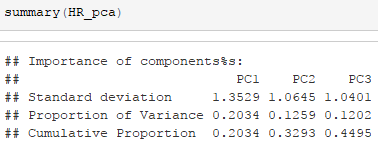


Fig.10. Results of PCA Analysis

The idea behind the PCA analysis is that this method tries to project data from high dimensions to low dimensions. Plots of PC1 and PC2, PC1 and PC3 will be then generated and results are presented in fig11. Geometrically speaking, such plots equal to the result of projecting original observations on the subspace spanned by PC1 and PC2 or PC1 and PC3.

From the plotting of PC1 and PC2, it can be seen that there exist clusters of people who left the company. For the subspace spanned by PC1 and PC2, there are three clusters. These three clusters have much more variance along the horizontal axes than vertical axes, which means that people who left are either overworked or underworked.

From the plotting of PC1 and PC3, again there are two clusters. Evidently, both groups have low values of PC3, which means that people who left always have a very low “satisfaction level”.

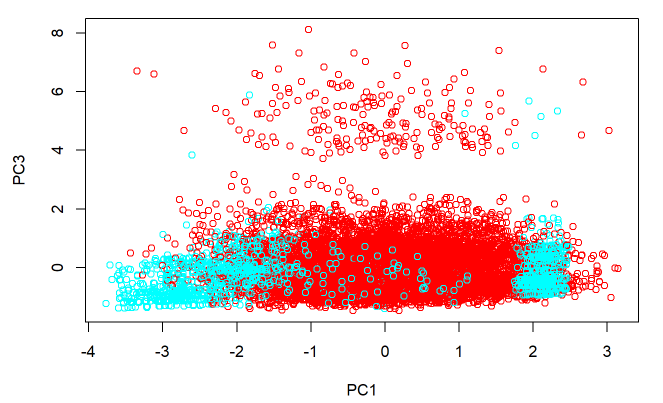
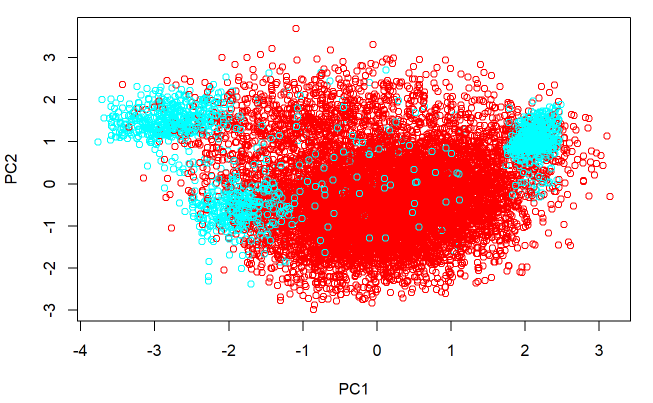


Fig.11. Plot of PCA Subspace

To further explore reasons behind the phenomenon observed from PCA plots, hierarchical clustering was performed. This method can generate a tree-based representation of observations and those observations are combined to the trunk according to the distance. From bottom to top, the earlier the fusion occurs, the more similar the groups of observations are to each other. The height of the fusion is measured on the vertical axes, indicating how different two groups are.

From the result of clustering, the observations are naturally divided into two or three groups according to different calculating methods of linkage between two groups, which agrees with the result of PCA analysis again.

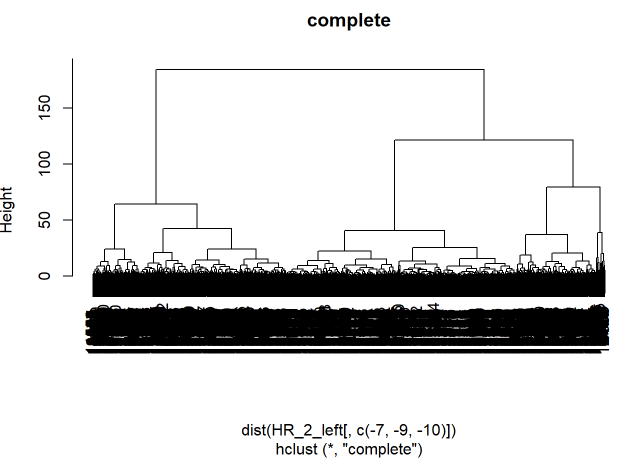
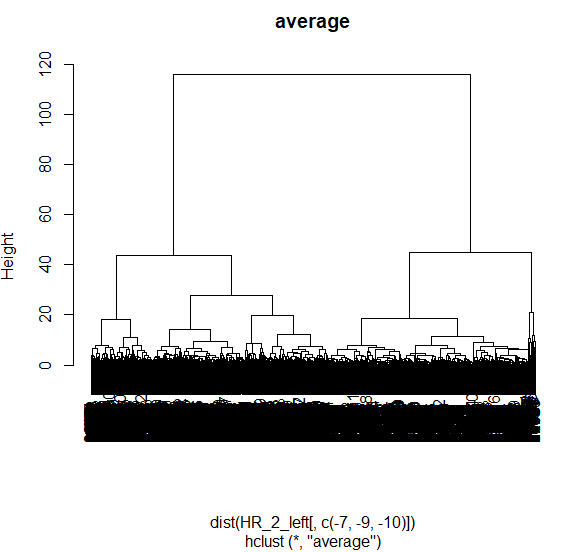


Fig.12. Dendrogram of Clustering

Finally, the pair-plot of some important variables in PC1, PC2 and PC3 will be performed.

In “satisfaction level” against “last evaluation” plot, and “satisfaction level” against “average month hours” plot, there are three different clusters for employees who left.

Here is the detailed description of these clusters.

Cluster1: extremely unhappy (satisfaction level is around 0.1) and hardworking or successful workers. Those workers are hardworking and highly evaluated but feel bad about company.

Cluster2: unhappy (satisfaction level is around 0.35) and unsuccessful or underworked workers. Those workers are badly evaluated and feel bad about company as well.

Cluster3: happy (satisfaction level is over 0.75) and successful or hardworking workers. Those workers are highly evaluated and feel good about the company.

In “last evaluation” and “average month hours” plot, there are two distinct groups of people who left.

Cluster1: successful (evaluation is around 0.9) and hardworking workers (month working hour is over 250). Those workers are hardworking and highly evaluated.

Cluster2: unsuccessful (evaluation is around 0.5) and underworked workers (month working hours is less than 160). Those workers work less and are badly evaluated.

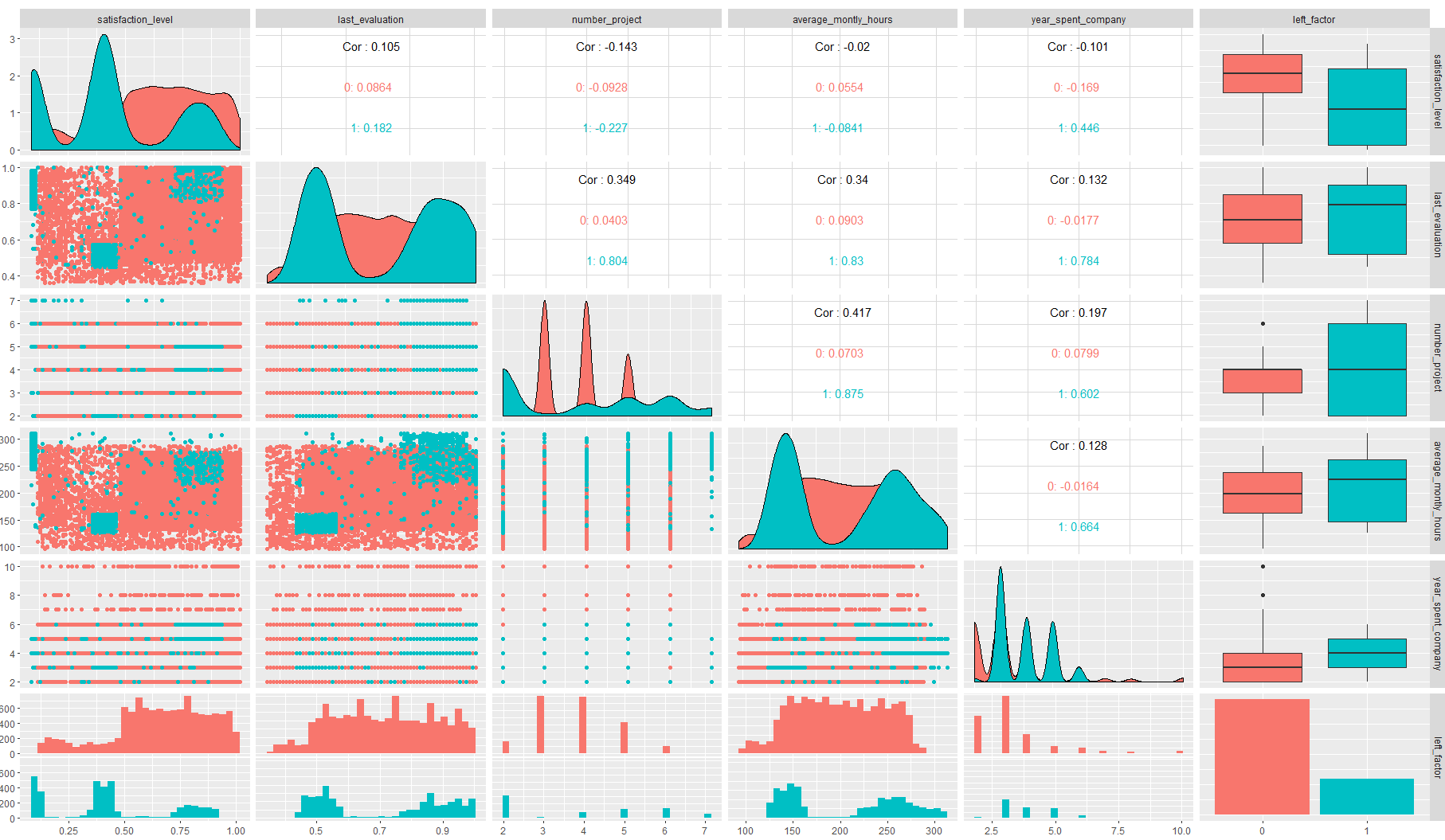


Fig.13. Pair Plot of some Variables

From the above pair-plot, apart from clusters, I also found that there is an increasing trend on “average month working hour” for people who left when they have more projects. By contrast, people who stay have a consistent value of “average month hours”, despite the increase of projects. In order to better visualize such trend, another boxplot is drawn.

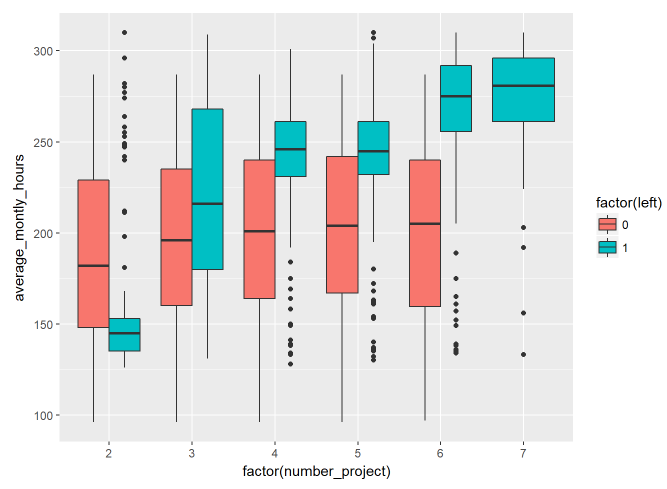
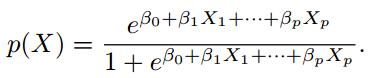
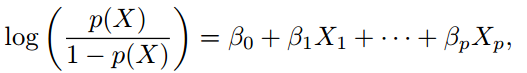


Fig.14. Box Plot of “Average Monthly Hours” against “Number Projects”

## Logistic Regression, Shrinkage Methods and Cross-validation

Logistic regression is always the first choice to get the preliminary analysis of variable importance and inference. It is a very convenient and straightforward way to generate classification probabilities associated with each observation. You can apply standard or custom threshold values on this probability score to get a cutoff and in turn classify output in a way which best fits your problem. In addition, it is robust to small noise and can be implemented quickly. The following formula is the multiple logistic function. After the transforming, we can see that the logistic regression model is linear in X from formula (3). The left side of formula (3) is called logit or log-odds.

 (2)

 (3)

Therefore, in logistic regression background, the following coefficient means the average change in logit when its corresponding predictor change one unit. According to each predictor’s *p*-value, some preliminary variable selections can be made. Therefore, if the significance level is set to equal to 0.05, then the variable-“department” will be abandoned.

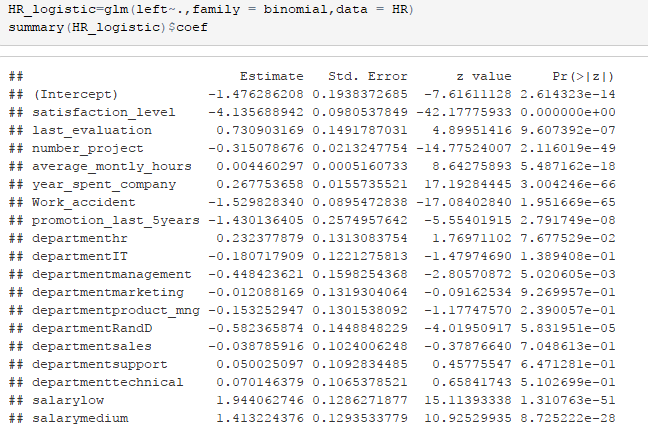


Fig.15. Coefficients of the logistic model

After fitting the logistic model, CV (cross-validation) is performed. In terms of CV, it is always used to evaluate the test error in a given statistic model, which is also known as model assessment. The result of CV on this model shows that the misclassification rate is 0.1399.

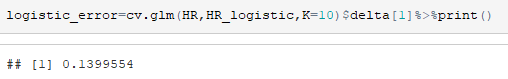
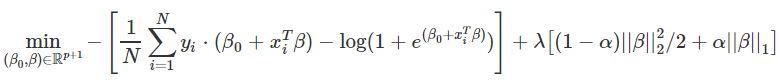


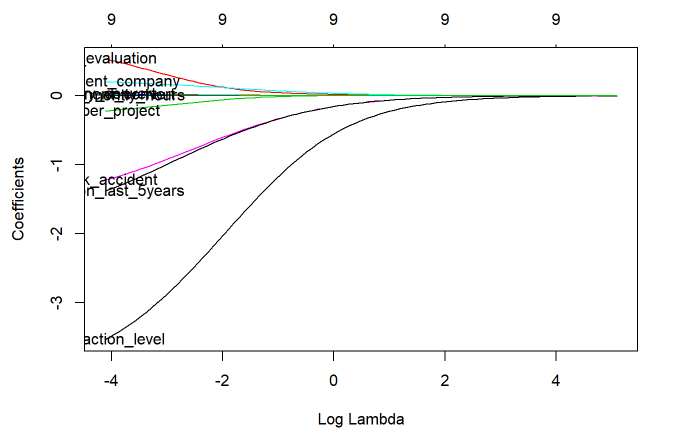
Fig.16. The CV misclassification rate

Then some shrinkage methods like lasso and ridge are performed. These methods aim at constraining estimated coefficients to improve the fitted model and significantly reduce the variance of the coefficient estimate. These two methods allow people to use more complex models and avoid over-fitting at the same time by forcing estimated coefficients towards zero. A tuning parameter λ is used to restrict the coefficient. If this tuning parameter equals to zero, then the normal logistic regression is performed. Otherwise, if the parameter equals to infinity, then all coefficients will be assigned to zero.

Also, although the ridge regression is usually more stable, lasso can perform the variable selection. Therefore, in terms of interpretation, lasso is better because it can generate simpler models.

 (4)

The ridge logistic regression is fitted by training data set, then through cross-validation, the best tuning parameter can be selected. Here the best tuning parameter is 0.087, resulting in the minimum training misclassification rate - 0.087 and the minimum CV misclassification rate - 0.206. Then this parameter is used to predict the test data and the test misclassification rate is 0.207.



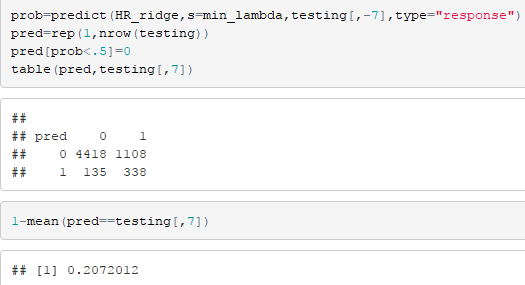
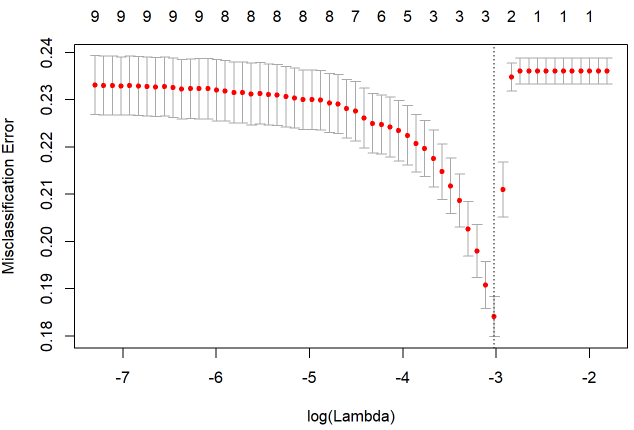
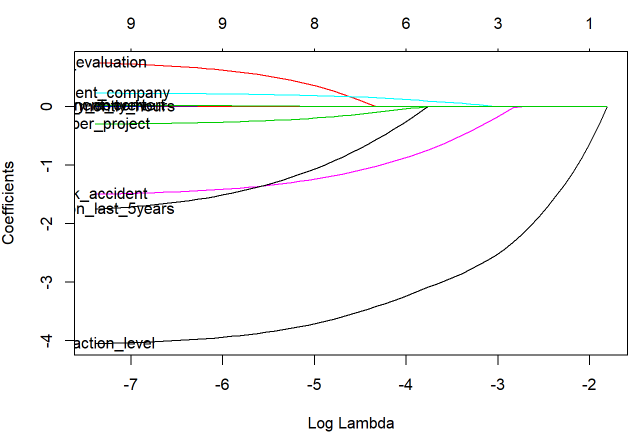


Fig.17. Results of the ridge logistic model

Then the lasso logistic regression is fitted. This time the optimum tuning parameter is 0.048. The CV misclassification rate is 0.184. Then this parameter is used to predict the test data set and the misclassification rate is 0.183.

The selected variables are “satisfaction level”, “year spent in company” and “work accident”. Here the reason why the misclassification rate dramatically decreases and increases after and before a particular λ - value is probably that the variable, “work accident”, has been downsized to zero at that point. Therefore, the variable “work accident” is very likely to be a valuable predictor.



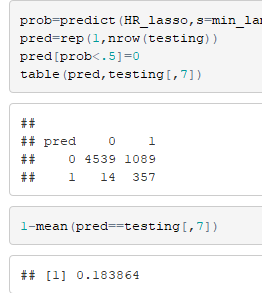
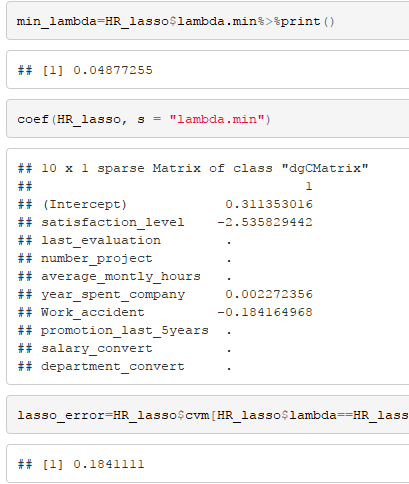


Fig.17. Results of the lasso logistic model

Overall, important variables are “satisfaction level”, “years spent in company” and “work accidents” according to results of the lasso logistic model.

However, what should be noticed here is that if shrinkage methods are applied on the logistic model, the misclassification rate increases instead of decreasing, which means that the model fitted by shrinkage methods performs worse. The reason may lie on the bias-variance tradeoff. Shrinkage methods are always used in situations where the model has high variance or large coefficients. Then through shrinking coefficients and thereby dramatically decreasing the variance at the little cost of increasing the bias, models are improved to be more predictable and interpretable. However, in this logistic model, which is used for classification at first in this report, it can be seen that its coefficients and the corresponding standard errors are already small enough, which means that this model may not have a high variance, so applying these two methods may only increase bias without decreasing the variance.

## Classification tree

In this step, a classification tree is built and it can also be used to tell the importance of variables. This is a simple and straightforward model for interpretation. Each time this model chooses a predictor X and the corresponding cut-point to split the original predictor space into two spaces

C:\Users\whq672437089\AppData\Roaming\Tencent\Users\672437089\TIM\WinTemp\RichOle\`HV`3FW}O%4PJ{PXXLPIE$C.png (formula 5)

Then formula 5 tries to seek a particular *j* and *s* to minimize the following RSS equation.

C:\Users\whq672437089\AppData\Roaming\Tencent\Users\672437089\TIM\WinTemp\RichOle\MRG532`3}38)R82_YH7~_BV.png (formula 6)

Therefore, the earlier the predictor occur, the more important it is. According the output of the following classification tree model, it can be seen that the first four appearing variables are “satisfaction level”, “year spent in company”, “number of projects” and “last evaluation”. Although this result is a bit different from the result generated from the lasso logistic regression, these two variable selection methods both select “satisfaction level” and “years spent in company” as important variables.

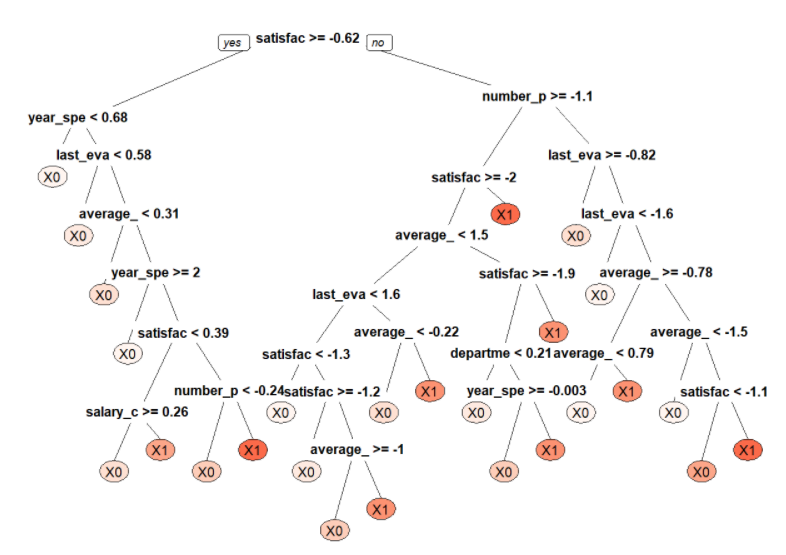
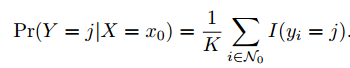


Fig.18. Pot of the classification tree

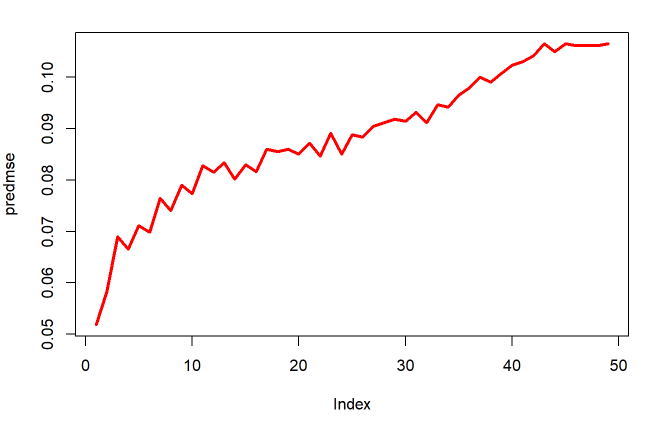
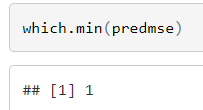
# Prediction

## KNN

KNN (K-nearest neighbors) is a kind of non-parametric classifier, which means that this method is less interpretable. The idea behind this method is trying to estimate the conditional distribution of Y given X, which is also the key factor of Bayes classifier. To be specific, the KNN method at first find the K points closest to the targeted predicting point from training observations and then estimate the conditional probability by calculation the fraction of those K nearest points whose is belong to class *j*.

 (formula 7)

In order to select the value of K that can generate the most accurate model, cross validation is used and then the relationship between the value of K and its corresponding misclassification rate is plotted. According to the output, when the value of K is one, the misclassification rate is lowest, so the model with K=1 is used to predict the testing data set and the test misclassification rate is 0.05184. Therefore, in terms of the prediction, the KNN model is better than the logistic regression model.

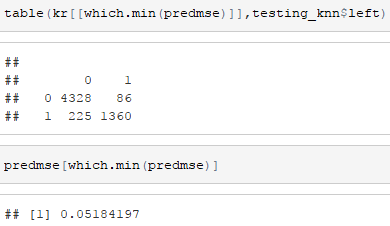
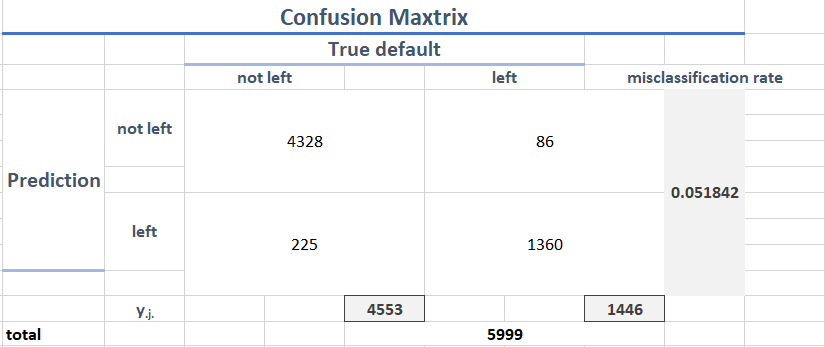


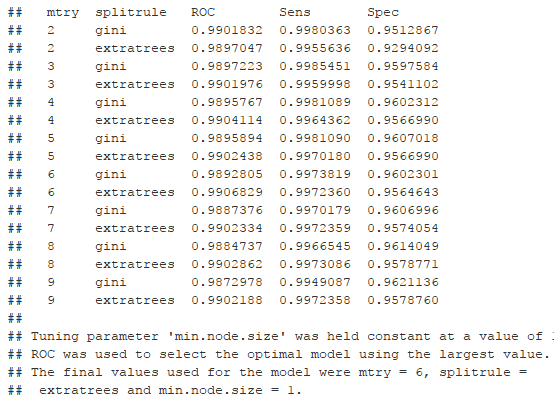
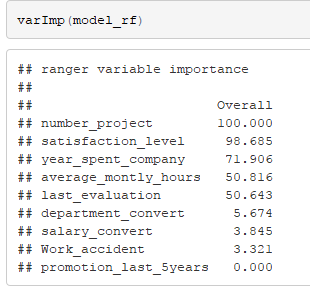
Fig.19. Results of KNN

## Random Forest

Random forest is a generalized model of bagging trees. It improves the bagged small tweak that decorrelates the trees. More specifically, the random forest can do better at minimizing the variance between decision tree branches and allow weak predictors to have a greater chance to contribute to the prediction. This is because for normal bagging decision trees, each time when splitting the tree, the random chosen predictor is selected from the full set of p predictors. By contrast, in random forest, each time when splitting the tree, a particular splitting variable is selected from a subset (size = *m* < *p*) of full variables. Then the cross validation is used decide how large this subset should be so that the test misclassification can be minimized.

According to the output of the R code, the random forest can generate the most accurate predicting result when *m* = 6. Then the model with *m* = 6 is used to predict the testing data set and the misclassification rate is only 0.013169.

In addition, the random forest model can also be used to generate the importance of variables. Here the result shows that the first five important variables are “number project”, “satisfaction level”, “year spent company”, “average monthly hours” and “last evaluation”.

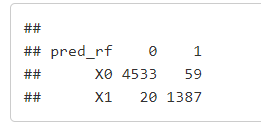
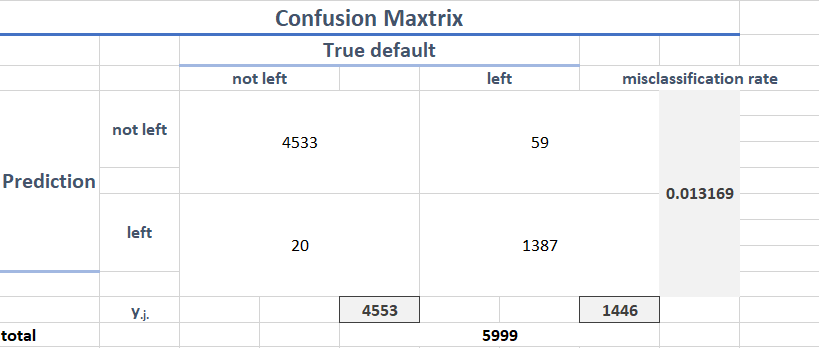
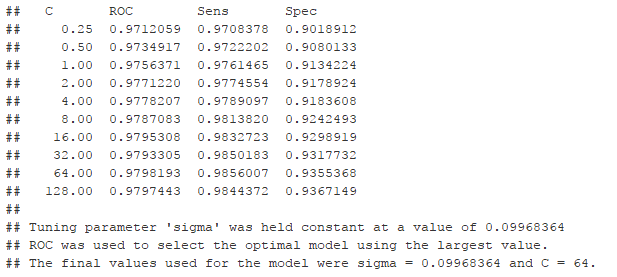


Fig.20. Results of Random Forest

## Support Vector Machine

In this step I will build the last classifier model – the support vector machine, which is good at fitting non-linear boundaries and designed particularly for a classification in the binary setting. SVM was developed as a classification tool in the 1990s and it has become popular since then. It is a general extension of another simple and straightforward classifier – the maximal margin model.

Specifically, here the model I built is called support vector machine with a Gaussian radial basis function kernel. Here the C is the parameter to control how soft the SVM allows the margin to be, which means that the larger C can make SVM treat margin more strictly. The sigma is the parameter of the Gaussian radial basis function and the larger this value is, the more linear the Gaussian kernel of radial based function is going to be. Here after the cross validation, output shows that when sigma=0.09968364 and C=64, the model is the most accurate. The misclassification rate is 0.032505.



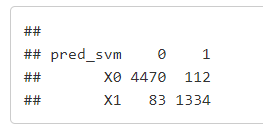
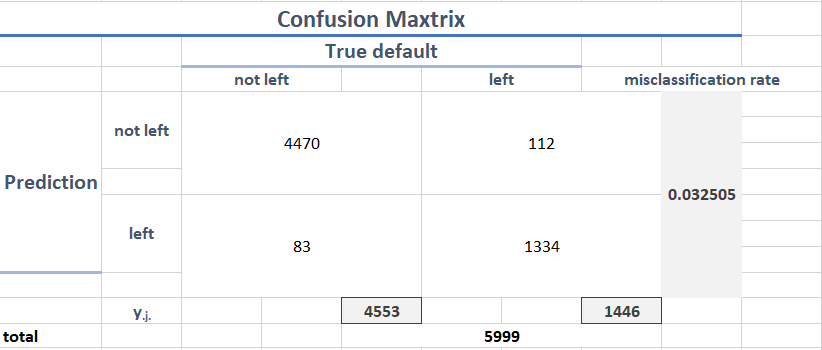


Fig.21. Results of SVM

## Comparing Model by Receiver Operating Characteristic Curve

The ROC curve is always used to visualize the tradeoff between sensitivity and specificity in two-class settings. Its vertical proportion unit represents sensitivity; its horizontal proportion unit represents specificity.

It is a useful plot to give the overview of two types of errors of a binary classifier. Therefore, usually people can compute the AUC (area under the curve) to represent the overall performance of a binary classifier. The larger the AUC is, the better the classifier is. In this step, I will plot different models’ ROC Curves and compute their AUC to select the best model.

In this setting, the sensitivity means the rate of correctly classifying people who not left to “not left”, and the specificity means the rate of correctly classifying people who left to “left”. Green line represents “Random Forest Model”; Blue line represents “Support Vector Machine Model”; Red line represents “KNN Model”. The output shows that the best model is “Random Forest Model”. This result agrees with the previous analysis where the “Random Forest Model” also has the lowest misclassification rate.

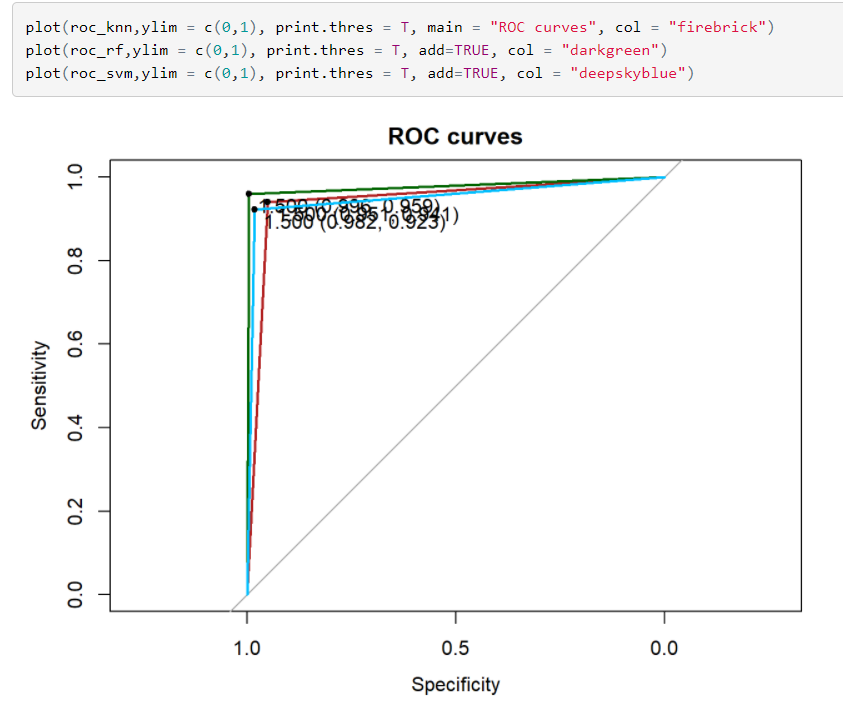


Fig.22. ROC Curves of Models

# Conclusion

In terms of interpretation, what is interesting is that “salary” and “promotion” that seems like key determining factor are not identified as an important predictor by any one of variable selection methods. Instead, all of three variable selection methods identify “satisfaction level” and “year spent in company” as important predictors. Therefore, HRs should pay more attention on employees who have a low satisfaction level or who have spent several years in the company. In addition, people who left can be clustered into two groups, “high last evaluation and hardworking” and “low last evaluation and underworked” or three groups, “happy and hardworking”, “unhappy and hardworking” and “unhappy and underworked” according to the previous analysis.

In terms of the prediction, the random forest model seems like the most accurate one according to the previous ROC plot. Also, the logistic regression may be another option to predict because it can not only give us the probability of employee’s leaving but also the threshold of the probability to classify can be set according to different settings.

# References

Kim, H. C.-M. (2002). Support Vector Machine Ensemble with Bagging. *Proceedings of the First International Workshop on Pattern Recognition with Support Vector Machines*, 397-407.

E. schapire Robert, "The boosting approach to machine learning: An overview", *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

Buciu, "Demonstrating the stability of support vector machines for classification", *Signal Processing*, vol. 86, pp. 2364-2380, 2006. (G. Casella, 2013)

In, Z.; Li, B.; Kotagiri, R.; Chen, J.; Chen, Y.; and Zhang, R. 2016. Improving efficiency of SVM kfold cross-validation by alpha seeding. *Technical Report arXiv*:1611.07659 [cs.LG], arXiv

B.P. Roe et. al., “Boosted decision trees as an alternative to artificial neural networks for particle identification”, NIM A543, pp. 577-584 (2005).

G. Casella, S. F. (2013). *An Introduction to Statistical Learning with Appliccation in R.* New York: Springer New York Heidelberg Dordrecht London.

Ally Edwards.”5 Reansons to Retain Employees.” *Blog.DATIS*. 22 April, 2015, <http://blog.datis.com/2015/04/22/5-reasons-to-retain-employees/>