Group 3

**Why do Workers Quit?**

Yuan Deng, Linh Diep, Mukil Raj Kannan,

Malav Patel, Shivam Srivastava

Date: 04-17-2019

# 1. Executive summary

Based on a data set of 1470 records and 16 attributes, we have conducted the classification analysis to find out the relationship between Attrition and other 15 attributes. Two classification methods, Decision Tree and Logistic Regression, are used based on three scenarios, in which the data set is divided into training and validation sets with varying percentage ratios. After multiple tests, we conclude the four most important factors in the attrition decision of employees:

1. Overtime. Working overtime can affect the life quality of employees. Especially now, people are more concerned about the balance between life and work.
2. Job Role. Obviously, some employees don't like their jobs and want to change.
3. Job Level. Some employees who believe they deserve a higher job level are prone to leaving the organization.
4. Total Working Years. A portion of employees who have worked for many years might be retiring or looking for more challenges.

Suggestions are included in the final section to provide guidance for employers during the decision making in the future.

2. Background

## 2.1 Defining Attrition

Attrition is generally defined as a gradual reduction in the size of a workforce. As opposed to turnover, it usually refers specifically to voluntary resignations rather than layoffs or other terminations. For companies intending to downsize over a long period or expecting certain jobs to become obsolete after a time, attrition can be a useful occurrence. Companies can induce attrition by leaving positions unfilled as employees resign, reducing their workforce size in a less disruptive manner than terminating workers.

There are two types of attrition: negative and positive. Negative attrition refers to the resignation of high performing workers, whereas positive attrition refers to the resignation of low performing workers. One is, obviously, desirable to a company and the other is not. For corporations seeking maximum worker productivity with minimum costs, employee attrition is one factor that is often tracked as a measure of health. Much like turnover, high attrition can be indicative of a hostile or otherwise undesirable workplace, but considering the nature of voluntary resignations, can also signal that a given position lacks an effective incentives structure or opportunities for advancement, or any number of factors that might induce an employee to leave the company.

## 2.2 Costs of Attrition

In recent years, numerous studies have been conducted to determine the primary causes of employee attrition, as well as quantifiable costs. Although it’s difficult to determine the causative factors of an employee’s resignation, what reasons can be tracked are just as often within a company’s control as not. For example, in any given case of employee attrition, it may be that the employee is unsatisfied with their position, or simply must change jobs due to circumstances external to their work. Knowing this, it is still generally advisable for companies to track attrition rates and keep them to desirable levels. The costs of attrition are well known, and even though attrition refers to something different than turnover, it incurs much of the same effects.

A study conducted by Tom Marsden for the *Strategic HR Review* (2016) looked into the monetary effects of employee resignation. Thanks to the expense of hiring, training, and maintaining employees, it is estimated that each leaver costs an organization 1 to 1.2 times the leaver’s annual salary as it loses the investment made for the departing employee, and must pay to hire someone to replace them. He goes further in estimating the overall costs of employee attrition for the US economy in a given year, a figure which totals to about 27 billion USD even using conservative numbers.

In Mahesh and Holtom’s long-term study of employee attrition and service-oriented companies (2012), they found that there are measurable effects on employees leaving and the profitability of a given company. They found that when employees - especially front-end employees who actually interface with clients - resign, it can damage the overall customer experience.

Clients become accustomed to working with certain employees, and tend to rate their service as lower. In fact, companies with a higher rate of attrition were rated at an average of 3.94 out of a 5 point scale, whereas companies with the lowest rates received an average score of 4.19. As a point of comparison, for companies undergoing severe periods of layoffs or terminations, the average score was 3.93 as opposed to 4.38 for those companies with the lowest. In general, clients associated employees departing from the company as an indicator of poor performance.

This association goes further. The authors found that such offices with low scores saw on average $486,000 in profit, whereas those with the highest scores earned more than $1.9 million, essentially four times higher than low performing counterparts. As such, at least for the service sector, employee attrition can directly affect a company’s bottom line - for better or worse.

Why do employees quit? When companies can confidently answer this question, they can effectively address human resources issues within their organization. It is therefore crucial for successful companies to track attrition rates and control them where possible. Tracking rates will allow the company to identify positions with high rates of attrition and help to assess the factors that might contribute to it. Determining the causative factors may help a company to ameliorate them and prevent employee attrition, preventing certain consequences from being incurred.

# 3. Data preprocessing

## 3.1 Data description

We use a fictional data set created by IBM scientists from Kaggle website: <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>. The data set contains the employee attrition information, which consists of 1470 records of 35 attributes. It has 26 numerical attributes and 9 category attributes in the original data set. The structure information of this data set is listed as follows:

'Data.frame': 1470 obs. of 35 variables:

$ Age : int 41 49 37 33 27 32 59 30 38 36 ...

$ Attrition : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...

$ BusinessTravel : Factor w/ 3 levels "Non-Travel","Travel\_Frequently",..: 3 2 ...

$ DailyRate : int 1102 279 1373 1392 591 1005 1324 1358 216 1299 ...

$ Department : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 ...

$ DistanceFromHome : int 1 8 2 3 2 2 3 24 23 27 ...

$ Education : int 2 1 2 4 1 2 3 1 3 3 ...

$ EducationField : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 ...

$ EmployeeCount : int 1 1 1 1 1 1 1 1 1 1 ...

$ EmployeeNumber : int 1 2 4 5 7 8 10 11 12 13 ...

$ EnvironmentSatisfaction : int 2 3 4 4 1 4 3 4 4 3 ...

$ Gender : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...

$ HourlyRate : int 94 61 92 56 40 79 81 67 44 94 ...

$ JobInvolvement : int 3 2 2 3 3 3 4 3 2 3 ...

$ JobLevel : int 2 2 1 1 1 1 1 1 3 2 ...

$ JobRole : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 ...

$ JobSatisfaction : int 4 2 3 3 2 4 1 3 3 3 ...

$ MaritalStatus : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 ...

$ MonthlyIncome : int 5993 5130 2090 2909 3468 3068 2670 2693 9526 ...

$ MonthlyRate : int 19479 24907 2396 23159 16632 11864 9964 13335 ...

$ NumCompaniesWorked : int 8 1 6 1 9 0 4 1 0 6 ...

$ Over18 : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...

$ OverTime : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...

$ PercentSalaryHike : int 11 23 15 11 12 13 20 22 21 13 ...

$ PerformanceRating : int 3 4 3 3 3 3 4 4 4 3 ...

$ RelationshipSatisfaction : int 1 4 2 3 4 3 1 2 2 2 ...

$ StandardHours : int 80 80 80 80 80 80 80 80 80 80 ...

$ StockOptionLevel : int 0 1 0 0 1 0 3 1 0 2 ...

$ TotalWorkingYears : int 8 10 7 8 6 8 12 1 10 17 ...

$ TrainingTimesLastYear : int 0 3 3 3 3 2 3 2 2 3 ...

$ WorkLifeBalance : int 1 3 3 3 3 2 2 3 3 2 ...

$ YearsAtCompany : int 6 10 0 8 2 7 1 1 9 7 ...

$ YearsInCurrentRole : int 4 7 0 7 2 7 0 0 7 7 ...

$ YearsSinceLastPromotion : int 0 1 0 3 2 3 0 0 1 7 ...

$ YearsWithCurrManager : int 5 7 0 0 2 6 0 0 8 7 …

For a better visualization, some variable names are shortened as listed below.

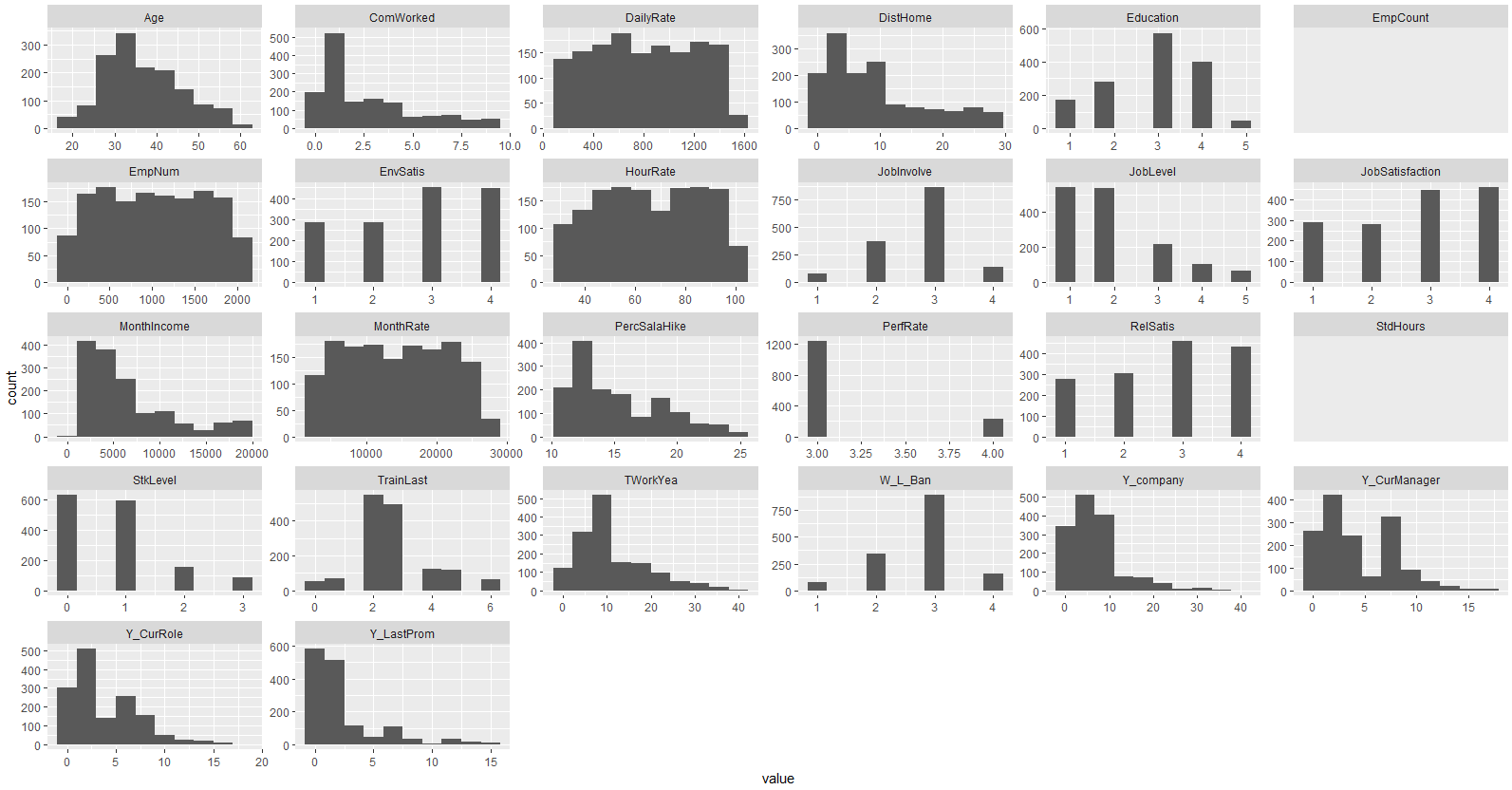
|  |  |
| --- | --- |
| Original name | Short name |
| BusinessTravel | ***Travel*** |
| DistanceFromHome | ***DistHome*** |
| EducationField | ***EduField*** |
| EmployeeCount | ***EmpCount*** |
| EmployeeNumber | ***EmpNum*** |
| EnvironmentSatisfaction | ***EnvSatis*** |
| HourlyRate | ***HourRate*** |
| JobInvolvement | ***JobInvolve*** |
| MonthlyIncome | ***MonthIncome*** |
| MonthlyRate | ***MonthRate*** |
| NumCompaniesWorked | ***ComWorked*** |
| PercentSalaryHike | ***SalHike%*** |
| PerformanceRating | ***PerfRate*** |
| RelationshipSatisfaction | ***RelSatis*** |
| StandardHours | ***StdHours*** |
| StockOptionLevel | ***StkLevel*** |
| TotalWorkingYears | ***TWorkYea*** |
| TrainingTimesLastYear | ***TrainLast*** |
| WorkLifeBalance | ***W\_L\_Ban*** |
| YearsAtCompany | ***Y\_company*** |
| YearsInCurrentRole | ***Y\_CurRole*** |
| YearsSinceLastPromotion | ***Y\_LastProm*** |
| YearsWithCurrManager | ***Y\_CurManager*** |

Here, we have some explanation from the original data set to clarify the category meaning.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | 1 | 2 | 3 | 4 | 5 |
| *Education* | Below College | College | Bachelor | Master | Doctor |
| *EnvSatis* | Low | Medium | High | Very High |  |
| *JobInvolve* | Low | Medium | High | Very High |  |
| *JobSatisfaction* | Low | Medium | High | Very High |  |
| *PerfRate* | Low | Good | Excellent | Outstanding |  |
| *RelSatis* | Low | Medium | High | Very High |  |
| *W\_L\_Ban* | Bad | Good | Better | Best |  |

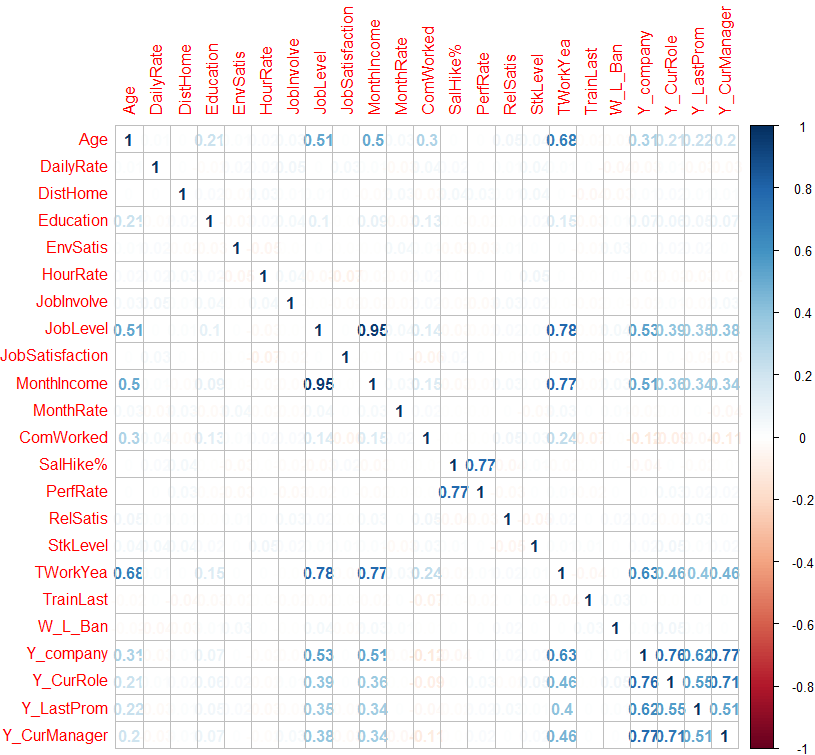
## 3.2 Data cleaning

For the 26 numerical attributes, the histogram distribution is shown below.

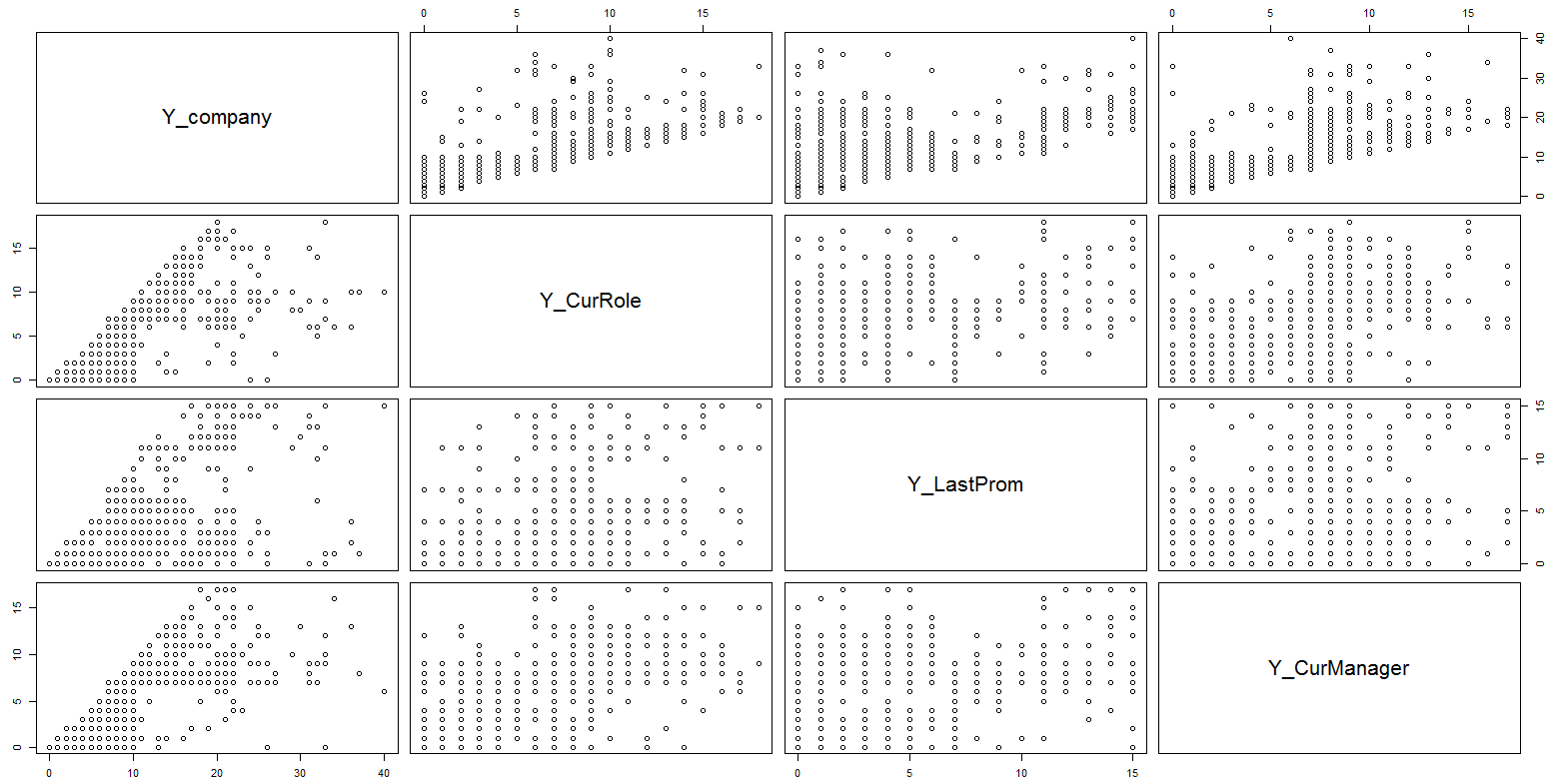


1. It is clear that two attributes (***EmpCount*** and ***StdHours***) are not useful because all records have the same value in these attributes.
2. There are four attributes related to employee income: ***MonthIncome***, ***MonthRate***, ***DailyRate***, ***HourRate***. The first attribute correctly reflects the employee income distribution that the percentage of employee decreases with the increasing income. However, the latter three attributes fail to represent this reasonable trend. The original data set does not provide an explanation of the “rate” used in these three attributes. And the latter three attributes cannot convert to each other considering the time equivalent. For example, in this data set. Thus, only the first attribute ***MonthIncome*** is used in the following analyses and other three are ignored.
3. ***EmpNum*** is the employee ID number, which is not useful and should be ignored.
4. For these nine attributes which are numeric variables, (***Education***, ***EnvSatis***, ***JobInvolve***, ***JobLevel***, ***JobSatisfaction***, ***PerfRate***, ***RelSatis***, ***StkLevel***, ***W\_L\_Ban),*** they should be converted to category variable in the logistic regression analysis.

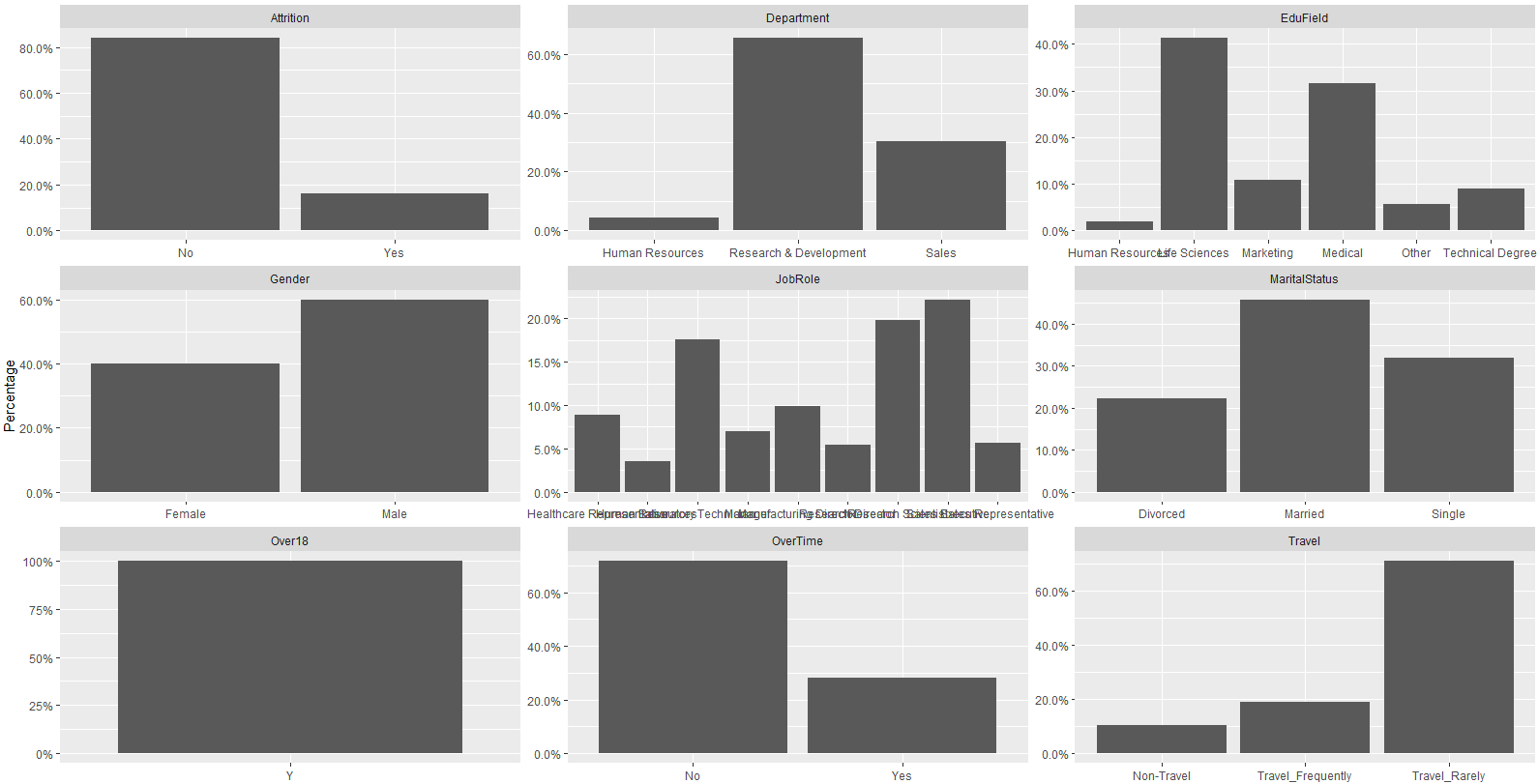
In addition, to have a better understanding of this data set, the correlation is calculated for these numerical attributes except these three (***EmpCount***, ***EmpNum***, ***StdHours***). Note that ***JobLevel*** has a high correlation coefficient with ***MonthIncome*** (0.95) and ***TWorkYear*** (0.78). This agrees with our common sense that employee with higher job level has a higher monthly income. And the longer you work, the higher job level you can get. On the other hand, the three ignored attributes (***MonthRate***, ***HourRate***, ***DailyRate***) are independent of all other variables, which is not a correct relationship in real life.



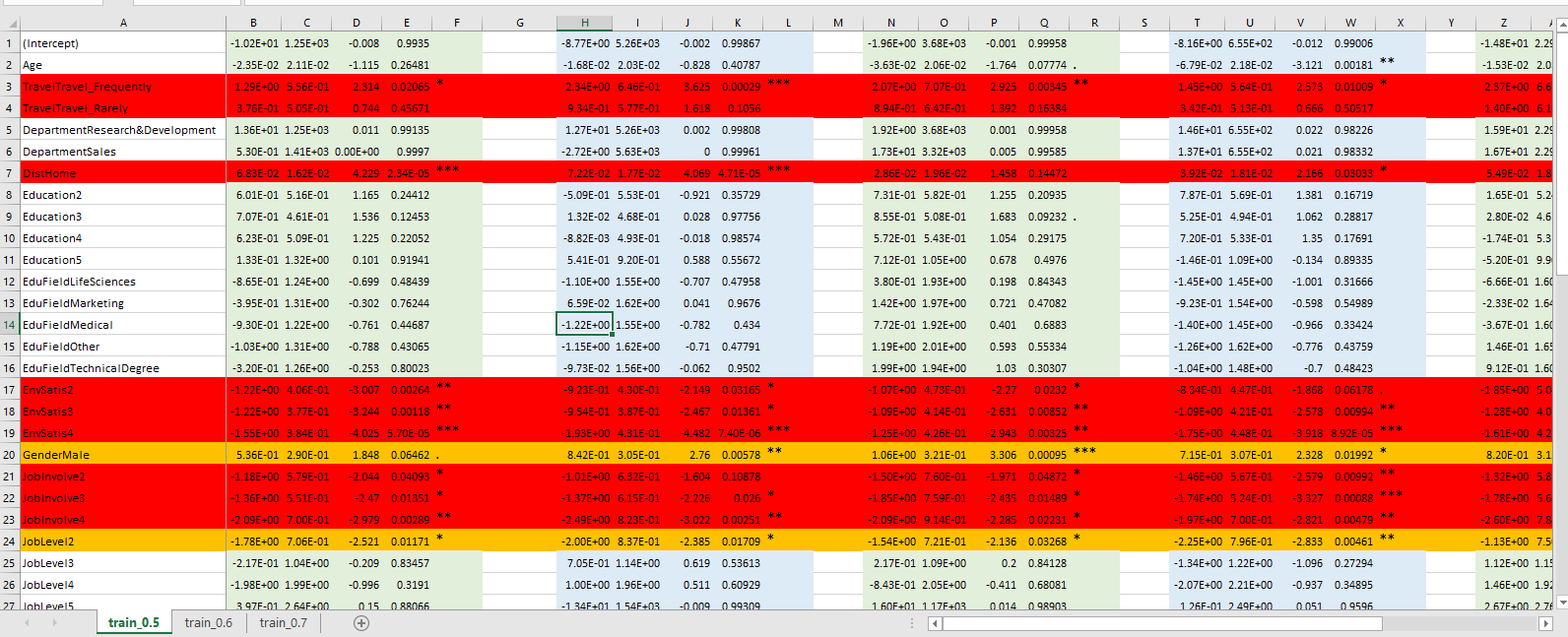
It should be noted that the correlation coefficients are relatively high between these four attributes (***Y\_company***, ***Y\_CurRole***, ***Y\_LastProm***, ***Y\_CurManager***). We use the scatter plot to visualize their relationships as follow. One causal relationship is obvious that an employee can take the current position, or get a promotion, or work with the current manager only after joining the company. However, there is not a clear relationship between ***Y\_CurRole***, ***Y\_LastProm***, and ***Y\_CurManager***.



For the 9 category attributes from the original data set, the bar plot is shown below. Note that the ***Over18*** attribute is not useful since all employee is adult.



After deleting these useless or redundant variables (***StdHours, EmpCount, EmpNum, DailyRate, MonthRate, HourRate, Over18***), the data set is reduced to 28 variables. However, the number of parameters is still too large for classification analysis. To figure out the important variables from the parameter space, we plan to use the Logistic Regression analysis. The data set is divided into two subsets: training and validation sets. Three scenarios (50%/50%, 60%/40%, 70%/30%) are generated by varying the percentage between these two groups. For each scenario, the Logistic Regression classification has been performed six times. One screenshot of the logistic regression analysis results is shown below as an example. By comparing the significant coefficient of each variable between these six tests, the important variables are selected in each scenario and combined into the final conclusion that only the 15 variables are considered as impacting parameters and will be used in the following sections. These 15 variables are: ***Travel, DistHome, EnvSatis, JobInvolve, JobLevel, JobSatisfaction, ComWorked, OverTime, RelSatis, StkLevel, W\_L\_Ban, Y\_LastProm, Y\_company, Y\_CurManager, Gender***. Please note in these variables only the ***Gender*** is selected based on our common sense and its significant coefficient is not as large as others.



# 4. BI model

## 4.1 Research plan

Based on the processed data set of 1470 records, this report uses two classification method, Decision Tree and Logistic Regression, to study the relationship between the target attribute (***Attrition***) and 15 other attributes. Of course, there are other methods which could perform the classification analysis, such as Random Forest and Bootstrap Aggregation, which are more complicated than the chosen methods. However, in this report, our target is to investigate the reasons behind the attrition and provide management suggestion, not finding the most accurate model. We prefer to use simple and straightforward methods to get reasonable conclusions.

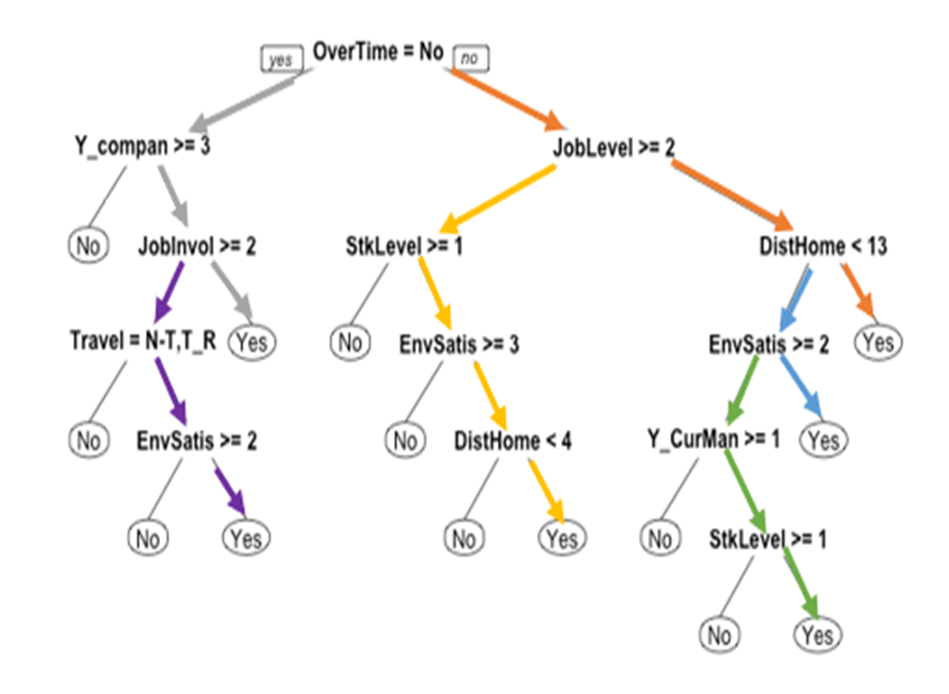
As mentioned above, the data set is divided into two subsets: training and validation subsets. Their percentage ratio varies between the three scenarios (50%/50%, 60%/40%, 70%/30%). For each scenario, the two classification methods will be conducted for analysis. Based on the Bootstrap resampling theory, the analysis in each scenario will be repeated multiple times by resampling the data set to get a robust solution. Of all the classification results in three scenarios, only the results with high accuracy-coefficients are considered as potential solutions. Taking into account our common sense and the similarities between these solutions, the best solution is chosen as the ultimate solution for this report and discussed below.

## 4.2 Results and Interpretation

### 4.2.1 Decision Tree

The output figure of the decision tree is shown as follows. Obviously, there are six patterns where employees are likely to leave the company.

1. Orange: employees with lower job level working overtime whose home distance is greater than 13 miles.
2. Blue: employees with lower job level working overtime with lower environment satisfaction.
3. Green: employees with lower job level working overtime with lower stock option.
4. Yellow: employees with higher job level with lower stock option and lower environment satisfaction working overtime whose home distance is greater than 4 miles.
5. Purple: employees with less than 3 years in a company and lower environment satisfaction with frequent travels.
6. Grey: employees with less than 3 years in a company and low job satisfaction.



From the confusion matrix of the training set in the list below, the accuracy of this model is calculated to be 0.881. In addition, the accuracy of the validation set is 0.8554, which represents a high percentage of correct predictions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Training | | Validation | |
| Predicted | No | Yes | No | Yes |
| No | 718 | 91 | 478 | 62 |
| Yes | 14 | 49 | 23 | 25 |

And the Receiver Operating Characteristic curve is shown below with an AUC value of 0.7334.



### 4.2.2 Logistic Regression

From the summary of the logistic regression model, we get the most significant variables as follows: ***Travel, DistHome, EnvSatis, JobInvolve, JobLevel, JobSatisfaction, ComWorked, OverTime, RelSatis, StkLevel, W\_L\_Ban, Y\_LastProm, Y\_Company, Y\_CurManager***, which are consistent with the chosen variables above.

On the other hand, we can focus on the coefficient of each variable and get some conclusions:

1. Employees who travel frequently are 1.96 times to leave the organization than those who don't travel
2. Employees with High Environment Satisfaction are -1.44 times to leave the organization than those with low Environment Satisfaction
3. Employees with High Job Involvement are -2.03 times to leave the organization than those with low Job Involvement
4. Employees with high Job Level are -2.31 times to leave the organization than those with low Job level.
5. Employees with high Job Satisfaction are -1.01 times to leave the organization than those with low Job Satisfaction.
6. Employees who work Overtime are 1.78 times to leave the organization than those who don't work overtime.
7. Employees with Medium Stock option are -1.24 times to leave the organization than those with low Stock option.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |  |
| (Intercept) | 1.56416 | 0.81803 | 1.912 | 0.055863 | . |
| TravelTravel\_Frequently | 1.96141 | 0.52143 | 3.762 | 0.000169 | \*\*\* |
| TravelTravel\_Rarely | 1.00348 | 0.49189 | 2.04 | 0.041347 | \* |
| DistHome | 0.04674 | 0.01312 | 3.562 | 0.000368 | \*\*\* |
| EnvSatis2 | -0.89552 | 0.33015 | -2.713 | 0.006678 | \*\* |
| EnvSatis3 | -0.86768 | 0.30064 | -2.886 | 0.003901 | \*\* |
| EnvSatis4 | -1.4435 | 0.32506 | -4.441 | 8.97E-06 | \*\*\* |
| GenderMale | 0.33124 | 0.22467 | 1.474 | 0.140388 |  |
| JobInvolve2 | -1.44734 | 0.43012 | -3.365 | 0.000766 | \*\*\* |
| JobInvolve3 | -1.70559 | 0.40435 | -4.218 | 2.46E-05 | \*\*\* |
| JobInvolve4 | -2.03541 | 0.57382 | -3.547 | 0.000389 | \*\*\* |
| JobLevel2 | -1.30662 | 0.28212 | -4.631 | 3.63E-06 | \*\*\* |
| JobLevel3 | -0.81643 | 0.35733 | -2.285 | 0.022324 | \* |
| JobLevel4 | -2.31779 | 0.68117 | -3.403 | 0.000667 | \*\*\* |
| JobLevel5 | -1.89994 | 0.84414 | -2.251 | 0.024402 | \* |
| JobSatisfaction2 | -0.3806 | 0.3362 | -1.132 | 0.257603 |  |
| JobSatisfaction3 | -0.4701 | 0.29758 | -1.58 | 0.114157 |  |
| JobSatisfaction4 | -1.01046 | 0.32405 | -3.118 | 0.00182 | \*\* |
| ComWorked | 0.09373 | 0.04376 | 2.142 | 0.032215 | \* |
| OverTimeYes | 1.78554 | 0.23522 | 7.591 | 3.17E-14 | \*\*\* |
| RelSatis2 | -1.05977 | 0.34918 | -3.035 | 0.002405 | \*\* |
| RelSatis3 | -0.93412 | 0.30585 | -3.054 | 0.002257 | \*\* |
| RelSatis4 | -0.7542 | 0.30936 | -2.438 | 0.014771 | \* |
| StkLevel1 | -1.24399 | 0.25746 | -4.832 | 1.35E-06 | \*\*\* |
| StkLevel2 | -1.03406 | 0.41255 | -2.506 | 0.012194 | \* |
| StkLevel3 | -1.04079 | 0.47437 | -2.194 | 0.028231 | \* |
| W\_L\_Ban2 | -0.7581 | 0.44286 | -1.712 | 0.086925 | . |
| W\_L\_Ban3 | -1.03936 | 0.4144 | -2.508 | 0.012137 | \* |
| W\_L\_Ban4 | -0.74223 | 0.51814 | -1.432 | 0.152004 |  |
| Y\_company | 0.0406 | 0.04029 | 1.008 | 0.313546 |  |
| Y\_LastProm | 0.13208 | 0.05077 | 2.602 | 0.009282 | \*\* |
| Y\_CurManager | -0.21761 | 0.05864 | -3.711 | 0.000207 | \*\*\* |

## 4.3 Conclusion

By comparing the analysis results from both models, we can conclude the four most important variable in the attrition decision of employees.

1. Overtime. Working overtime can affect the life quality of employees. Especially now, people are more concerned about the balance between life and work.
2. Job Role. Obviously, some employees don't like their jobs and want to change.
3. Job Level. Some employees who believe they deserve a higher job level are prone to leaving the organization.
4. Total Working Years. A portion of employees who have worked for many years might be retiring or looking for more challenges.

Other Factors include: frequent travel, a large distance from home, the employee has worked in a lot of companies before, and work-life balance.

# 5. Application and concern

We conclude that the most critical factors that play a role in attrition decisions as: frequent travel, years of experience of employees in the organization, overtime working, and dissatisfaction with roles and responsibilities. HR leaders must pay close attention to employee dissatisfaction with career development to prevent work loss and employee turnover.

Global Talent Monitor’s report on workforce activity in 2Q18 shows that the lack of future career development remains a key driver of employee attrition — cited by 40% of departing employees as a dissatisfying factor in their job. At the same time, 28% of employees are actively seeking a job and 42% are passively open to new opportunities.

Attrition has always been costly for companies, and in many industries, the cost of losing employees is rising, due to tight labor markets and the increasingly collaborative nature of jobs. If employees cannot see you investing in their future with you, they are going to look somewhere else. Gartner research shows that since the financial crisis of 2008, many organizations have removed several layers of middle management, so there are fewer opportunities for internal promotion. These things discourage employees, making them think their efforts are not reflecting as per expected as there is no climbing up the management hierarchy and no hike in salary.

Aforementioned similar issues and the ones that our analysis brought up together are the most concerned and important factors to look into when talking about the strategies to curb attrition within any organization. And as the employed spends and invests on every leaving employee, each attrition comes up with the heavy cost to the organization. Hence, to curb attrition in one’s organization, one needs to have efficient resource planning for the selection, training, placement, performance management, and decent compensations.

# 6. References

Marsden, T. (2016). What is the true cost of attrition? *Strategic HR Review*, *15*(4), 189–190. <https://doi.org/10.1108/SHR-05-2016-0039>

Subramony, M., & Holtom, B. (2012). The Long-Term Influence of Service Employee Attrition on Customer Outcomes and Profits. *Journal of Service Research*, 15(4), 460–473. <https://doi.org/10.1177/1094670512452792>

Steve D. Mullins, Janis L. Prewitt. (2018). The Application of Workplace Attrition in a Multiple-Claimant Gender Discrimination Case: The EEOC v. New Prime Trucking. *Journal of Logistics Management*. <http://article.sapub.org/10.5923.j.logistics.20180701.01.html>

Sarah M. (2018). Lack of Career Development Drives Employee Attrition. Sep. 25, HUMAN RESOURCES, Smarter with Garther.

<https://www.gartner.com/smarterwithgartner/lack-of-career-development-drives-employee-attrition/>

# 7. Source code

(Only important section is included here.)

# ----------------------------- 0 load data ---------------------------

Attrition = read.csv("BI.03.Data.Why\_Workers\_Quit.csv")

class(Attrition)

str(Attrition)

# ----------------------------- 1 Data cleaning ---------------------------

# following names are changed to a shorter length for better visualization in DCT

colnames(Attrition)[colnames(Attrition)=="ï..Age"] = "Age" # change to normal name

colnames(Attrition)[colnames(Attrition)=="BusinessTravel"] = "Travel"

colnames(Attrition)[colnames(Attrition)=="DistanceFromHome"] = "DistHome"

colnames(Attrition)[colnames(Attrition)=="EducationField"] = "EduField"

colnames(Attrition)[colnames(Attrition)=="EmployeeCount"] = "EmpCount"

colnames(Attrition)[colnames(Attrition)=="EmployeeNumber"] = "EmpNum"

colnames(Attrition)[colnames(Attrition)=="EnvironmentSatisfaction"] = "EnvSatis"

colnames(Attrition)[colnames(Attrition)=="HourlyRate"] = "HourRate"

colnames(Attrition)[colnames(Attrition)=="JobInvolvement"] = "JobInvolve"

colnames(Attrition)[colnames(Attrition)=="MonthlyIncome"] = "MonthIncome"

colnames(Attrition)[colnames(Attrition)=="MonthlyRate"] = "MonthRate"

colnames(Attrition)[colnames(Attrition)=="NumCompaniesWorked"] = "ComWorked"

colnames(Attrition)[colnames(Attrition)=="PercentSalaryHike"] = "PercSalaHike"

colnames(Attrition)[colnames(Attrition)=="PerformanceRating"] = "PerfRate"

colnames(Attrition)[colnames(Attrition)=="RelationshipSatisfaction"] = "RelSatis"

colnames(Attrition)[colnames(Attrition)=="StandardHours"] = "StdHours"

colnames(Attrition)[colnames(Attrition)=="StockOptionLevel"] = "StkLevel"

colnames(Attrition)[colnames(Attrition)=="TotalWorkingYears"] = "TWorkYea"

colnames(Attrition)[colnames(Attrition)=="TrainingTimesLastYear"] = "TrainLast"

colnames(Attrition)[colnames(Attrition)=="WorkLifeBalance"] = "W\_L\_Ban"

colnames(Attrition)[colnames(Attrition)=="YearsAtCompany"] = "Y\_company"

colnames(Attrition)[colnames(Attrition)=="YearsInCurrentRole"] = "Y\_CurRole"

colnames(Attrition)[colnames(Attrition)=="YearsSinceLastPromotion"] = "Y\_LastProm"

colnames(Attrition)[colnames(Attrition)=="YearsWithCurrManager"] = "Y\_CurManager"

Attrition.ori = Attrition

# divide raw data based on data type

temp <- unlist(lapply(Attrition, is.numeric))

Attrition.numeric = Attrition[, temp]

Attrition.factor = Attrition[, !temp]

library(tidyr)

library(ggplot2)

#histogram of numerical variables

ggplot(gather(Attrition.numeric), aes(value)) +

geom\_histogram(bins = 10) +

facet\_wrap(~key, scales = 'free')

#bar plot of categorical variables

library(scales)

temp <- gather(Attrition.factor, key = type\_col, value = categories)

ggplot(temp, aes(x = categories)) +

geom\_bar(aes(y = (..count..)/tapply(..count..,..PANEL..,sum)[..PANEL..])) +

ylab("Percentage") +

facet\_wrap(~ type\_col, scales = "free") +

scale\_y\_continuous(labels = percent)

# note that "StdHours", "EmpCount", "EmpNum" are not useful.

# Only "MonthIncome" will be used. All other three Rates will be ignored.

Attrition$StdHours = NULL

Attrition$EmpCount = NULL

Attrition$EmpNum = NULL

Attrition$DailyRate = NULL

Attrition$HourRate = NULL

Attrition$MonthRate = NULL

# note that "Over18" is not useful.

Attrition$Over18 = NULL

Attrition.used = Attrition

# This section has been repeated six times to find out the most significant variables

# based on the logistic regression results

train.size = floor(0.5\*nrow(Attrition.used))

#train.size = floor(0.6\*nrow(Attrition.used))

#train.size = floor(0.7\*nrow(Attrition.used))

train.index = sample(nrow(Attrition.used), train.size)

train.df = Attrition.used[train.index, ]

temp = names(Filter(is.factor, train.df))

train.df[temp][is.na(train.df[temp])] = 'U'

temp = names(Filter(is.numeric, train.df))

for (i in temp) {

train.df[is.na(train.df[,i]), i] = mean(train.df[,i], na.rm = TRUE)

}

# Logistic regression, to find out which parameter dominates the results

logit.reg <- glm(Attrition ~., data = train.df, family = "binomial")

summary(logit.reg)

# from these six results in the Excel file, we got significant variables as below:

# Travel, DistHome, EnvSatis, JobInvolve, JobLevel, JobSatisfaction, ComWorked

# OverTime, RelSatis, StkLevel, W\_L\_Ban, Y\_LastProm

# Gender(this one is not that important as above ones, but still very important and interesting)

#

# for Y\_CurRole, physically, it should be equal to Y\_LastProm. But the records does not show it in this way. So

# this variable is removed.

# Retrieve data from Attrtion.ori into Attrition.used in following sections.

temp = c("Attrition", "Travel", "DistHome", "EnvSatis", "JobInvolve", "JobLevel", "JobSatisfaction",

"ComWorked", "OverTime", "RelSatis", "StkLevel", "W\_L\_Ban", "Y\_LastProm", "Y\_company",

"Y\_CurManager", "Gender")

Attrition.dct = Attrition.ori[, (colnames(Attrition.ori) %in% temp)]

# ----------------------------- 2 ---------------------------

# resample when repeat is required. change ratio when required.

# Decision tree

train.size = floor(0.5\*nrow(Attrition.dct))

train.index = sample(nrow(Attrition.dct), train.size)

train.df = Attrition.dct[train.index, ]

valid.df = Attrition.dct[-train.index, ]

temp = names(Filter(is.factor, train.df))

train.df[temp][is.na(train.df[temp])] = 'U'

temp = names(Filter(is.factor, valid.df))

valid.df[temp][is.na(valid.df[temp])] = 'U'

temp = names(Filter(is.numeric, train.df))

for (i in temp) {

train.df[is.na(train.df[,i]), i] = mean(train.df[,i], na.rm = TRUE)

}

temp = names(Filter(is.numeric, valid.df))

for (i in temp) {

valid.df[is.na(valid.df[,i]), i] = mean(valid.df[,i], na.rm = TRUE)

}

library(rpart)

library(rpart.plot)

train.cf <- rpart(Attrition~., data=train.df, method="class")

prp(train.cf)

library(lattice)

library(caret)

train.cf.pred.train<- predict(train.cf, train.df, type='class')

confusionMatrix(train.cf.pred.train, train.df$Attrition)

train.cf.pred.valid<- predict(train.cf, valid.df, type="class")

confusionMatrix(train.cf.pred.valid, valid.df$Attrition)

library(pROC)

train.cf.pred2.valid = predict(train.cf,valid.df,type="prob")

train.cf.pred2.valid.roc <- roc(valid.df$Attrition, train.cf.pred2.valid[,1])

plot.roc(train.cf.pred2.valid.roc)

auc(train.cf.pred2.valid.roc)

# Logistic regression

Attrition.lr = Attrition.dct

temp = c("EnvSatis", "JobInvolve", "JobLevel", "JobSatisfaction",

"RelSatis", "StkLevel", "W\_L\_Ban")

Attrition.lr[temp] = lapply(Attrition.lr[temp], factor)

train.df = Attrition.lr[train.index, ]

valid.df = Attrition.lr[-train.index, ]

logit.reg <- glm(Attrition ~., data = train.df, family = "binomial")

summary(logit.reg)

logit.reg.pred.train <- predict(logit.reg, train.df)

confusionMatrix(as.factor(ifelse(logit.reg.pred.train>0.5,"Yes","No")), train.df$Attrition)

logit.reg.pred.valid <- predict(logit.reg, valid.df)

confusionMatrix(as.factor(ifelse(logit.reg.pred.valid>0.5,"Yes","No")), valid.df$Attrition)

logit.reg.pred.valid.roc = roc(valid.df$Attrition, logit.reg.pred.valid)

plot.roc(logit.reg.pred.valid.roc)

auc(logit.reg.pred.valid.roc)