Midterm Project Final Report

Yifeng He

December 6, 2021

Abstract

This report aims to examine the relationship between employee attrition and five predictors using logit multilevel model. The model indicates that age, monthly income, and relationship satisfaction have negative effects on probability of attrition, while distance from home and years since last promotion have positive effects on attrition. Reasons for the effects are mixed, and more considerations need to be included when applied in real world situations.

Introduction

Employee attrition means the reduction of workforce because of retirement, death, sickness, and relocation, etc. Attrition is a natural process in companies that decreases the work force without much management efforts. Sometimes, however, unpredicted attrition may result in extra cost of continuing the project and training for new workers. To minimize the cost and the possible lost, companies are constantly trying to reduce the attrition. Some companies invest the process of hiring to find the right people, some provide comfortable working environments to employees, and some simply increase employees' salaries to increase their willingness to stay in the company. All those methods will improve attractiveness for employees, but attrition still happens.

This report will use multilevel model to investigate what factors affect attrition and how they influence employees' decision. Based on the analysis, this report will also propose some ways to address the attrition problem.

Method

Data Processing

The dataset used in this report is from Kaggle: IBM HR Analytics Employee Attrition & Performance. The dataset includes responses of attrition from 1,470 subjects and their work-related conditions like total working years, salaries, and satisfaction, etc. After preliminary analysis of the data, I found that the dataset is already well written and cleaned, that it contains no useless information. Then I pick five variables that I am most interested in and listed below:

Variables	Explanation
Age	Age of the employee
Distance From Home	Distance from company to home
Environment Satisfaction	Satisfaction of working environment with 4 levels
Monthly Income	Monthly income of employee
Total Working Years	Working years since the first career
Years Since Last Promotion	Working years since last promotion

Next, I transferred the dataset into the long format that is suitable for exploratory data analysis (EDA).

Exploratory Data Analysis

The radar plot below, figure 1, shows the mean value of the employees who want to leave (green) and to stay (pink) in five selected factors. The outer circle is the largest value of each factor from the dataset. From the plot, employees who choose not to leave have higher average monthly income than those who decide to leave. The same pattern shows in age, as older people have lower average attrition. Mean values for other factors have much less difference.

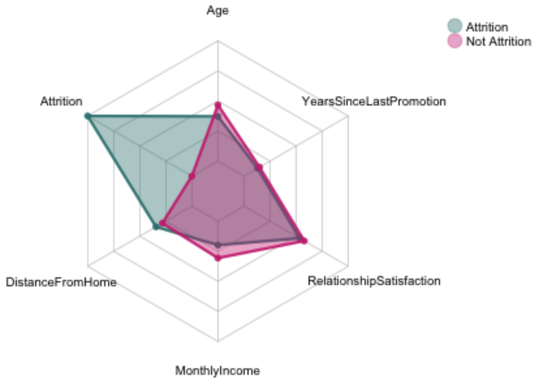


Figure 1: average values for five factors

To further see the difference between attrition and not attrition group, a comparative boxplot shows more information will help. Figure 2 shows the range, first and third quartiles, and median values for all five factors. Log(value) is used to improve the scale for the values. From figure 2, it is clear that Monthly Income, Distance From Home, and Age have the most distinguishable difference; Relationship Satisfaction and Year Since Last Promotion seems to be the same.

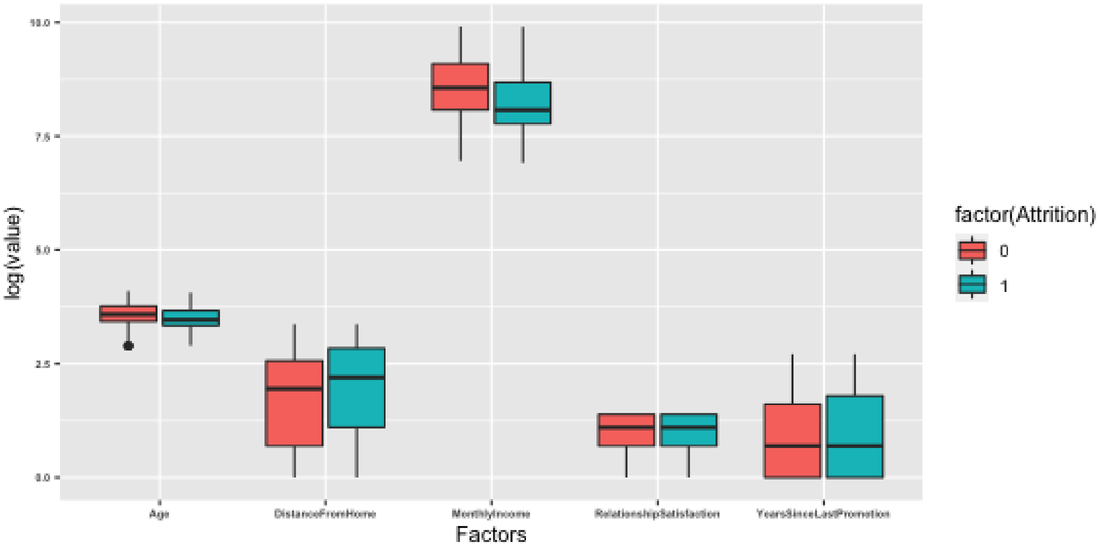


Figure 2: boxplot

Figure 3 shows the relationship between attrition or not and monthly income for all the job roles. It is noteworthy that different jobs react nearly opposite to increased amount of monthly income. Sales executive, healthcare representative, and research director show increasing trend of attrition when monthly income increases, though the slope is small. Other jobs show clear decrease of willingness for attrition when monthly income increases.

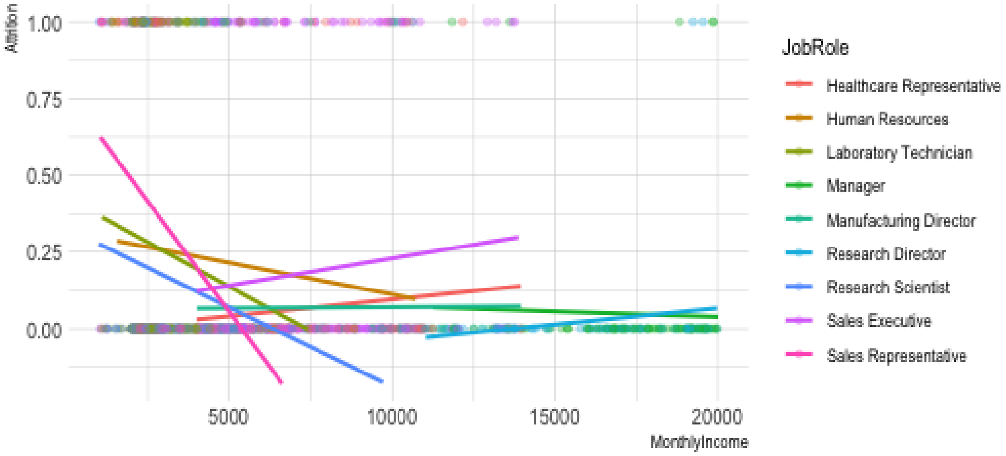


Figure 3: monthly income vs. attrition

Figure 4a shows the relationship between attrition and the distance from work to home. Sales representative shows decreasing trend of attrition when distance from home increases; other jobs show increasing attrition trend when distance from home increases. Figure 4b reveals the connection between the time since last promotion and employee attrition. Sales representative, human resources, and lab technician have reduction in attrition when last promotion time increase; other jobs show the opposite effect that the longer the time from last promotion, the higher the attrition, but the overall slopes are low.

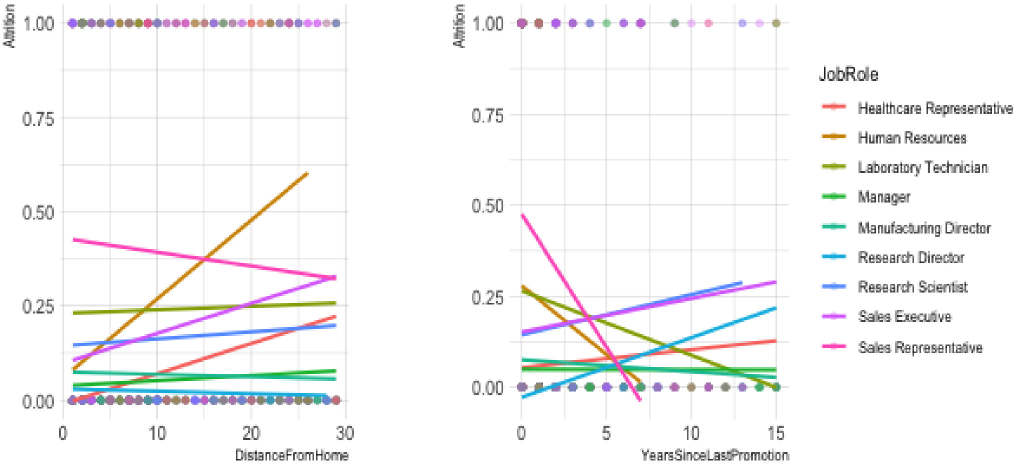


Figure 4: distance from home (a) and years from promotion (b) vs. attrition

Model fitting

With different job roles as categories, the best model to fit the data is multilevel model. I choose 5 predictor variables, they are all continuous variables, and attrition, a binary variable, as the outcome. Since the outcome variable is binary, I will use logit multilevel model. Below is the function:

$$\begin{aligned} fit1 < -glmer(Attrition\ Age + DistanceFromHome + MonthlyIncome + RelationshipSatisfaction + \\ YearsSinceLastPromotion + (1|JobRole), data = attrition, family = binomial(link = "logit")) \end{aligned}$$

The fixed effects of the model are shown in the table below:

	x
(Intercept)	-0.35718
Age	-0.03024
DistanceFromHome	0.02690
MonthlyIncome	-0.00007
RelationshipSatisfaction	-0.10584
YearsSinceLastPromotion	0.03535

Result

Based on the logit multilevel model, the formula can be concluded as below:

$$\begin{aligned} logit(attrition) = -0.36 - 0.03 * Age + 0.03 * Distance - 0.00007 * MonthlyIncome \\ -0.11 * RelationshipSatissfaction + 0.04 * YearSinceLastPromotion \end{aligned}$$

For the above model, -0.36 means the probability of attrition for an employee with average age, average income, and all the rest factors as the average value. And for every one unit increase in age, the log odds of attrition with decrease 0.03 on average. To make it easier to understand, I will take the exp of the coefficients and transfer the model as below:

$$\begin{aligned} Attrition =& exp(-0.36 - 0.03 * Age + 0.03 * Distance - 0.00007 * MonthlyIncome \\ & -0.11 * RelationshipSatisfaction + 0.04 * YearSinceLastPromotion) \\ =& 0.70 * 0.97^{Age} * 1.03^{DistanceFromHomwe} * 1.00^{MonthlyIncome} \\ & * 0.90^{RelationshipSatisfaction} * 1.04^{YearsSinceLastPromotion} \end{aligned}$$

The transferred model indicates that, with average age, distance from home, monthly income, etc., the odds of attrition is 0.7. For one unit increase in age, the odds of attrition will experience a multiplicative effect of 0.97, which means that the probability of attrition will decrease 3.0%, when other predictors take the average values. Increase in distance from home will raise the probability of attrition by 2.7%; monthly income will reduce the probability 0.007%; relationship satisfaction will decrease attrition 10%; and year since last promotion will increase the probability of attrition 3.6%; all the changes of predictor is one unit with other factors as average values.

From the model, it is clear that age, monthly income, and relationship satisfaction have negative effects for the probability of attrition, while distance from home and years since last promotion have positive effect on attrition.

Model Validation

The binned residual plot, in appendix, shows that 95% of the points are within the boundaries, so the model fits good. Since I use logit multilevel model, no other residual plots are needed.

Discussion

The transformed formula of the logit multilevel model shows that increase in age, monthly income, and relationship satisfaction will lead to low probability of attrition; on the opposite, distance from home and years since last promotion will increase the likelihood of attrition.

Though all the factors influence the probability of attrition, the effects are very different. Take monthly income as an example, one unit of income increase only reduce the probability of attrition by 0.007%. However, the increase of monthly income is usually by thousands, so when the income level increases, the reduction in attrition should be relatively large. As shown in figure 1, the difference between monthly income for attrition and not attrition groups are the most distinguishable factor. The large gap between the medians in figure 2 also indicates that monthly income plays a significant role in affecting employment attrition.

Other results are also easy to understand. Longer commute distance means the commute time is longer, the cost of going to work is higher, thus reduce the likelihood to stay in the company; longer time from promotion demonstrate that there is no room for improvement and employees may leave for other opportunities. Older people tend to stay in the company for they are less likely to move; employees with good intimate relationships usually satisfy with their environments. However, there is no casual relationships between these predictors and the attrition results. Unforeseen factors like family emergency or accident may also affect the attrition. Any single predictor also can not predict the results of attrition cause there is always complicated interactive factors that lead to the result.

Future studies can include more factors into the model, and also examine the interactive effects between these variables. However, this correlated model never implicate true causational relationship between attrition and any of the factors. The real-world situation is much more complex than the model prediction, and time and efforts must be spend to reduce employee attrition.

Citation

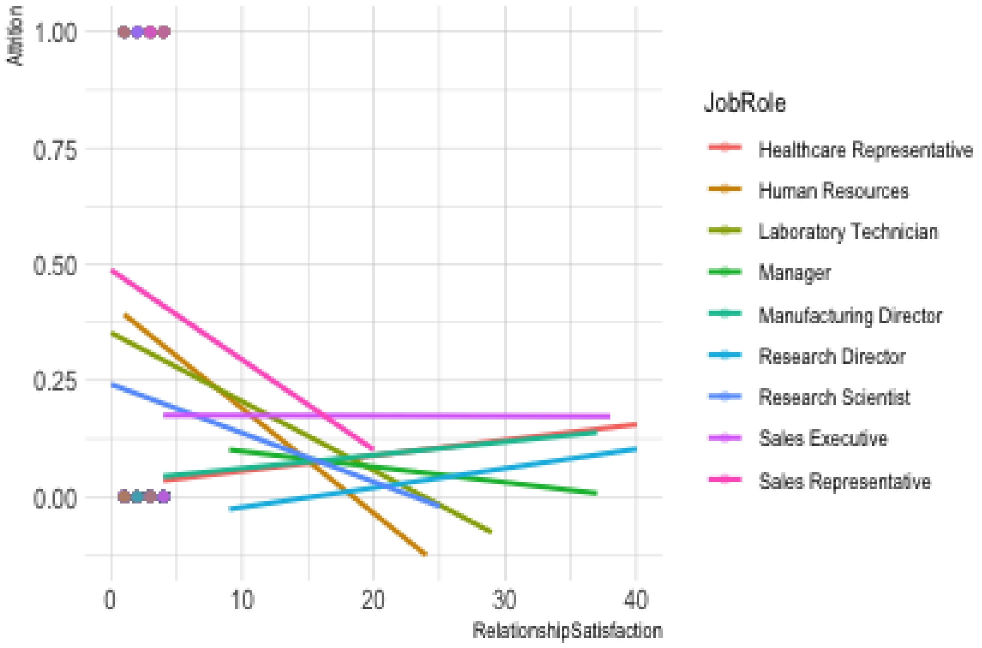
The citation is as below:

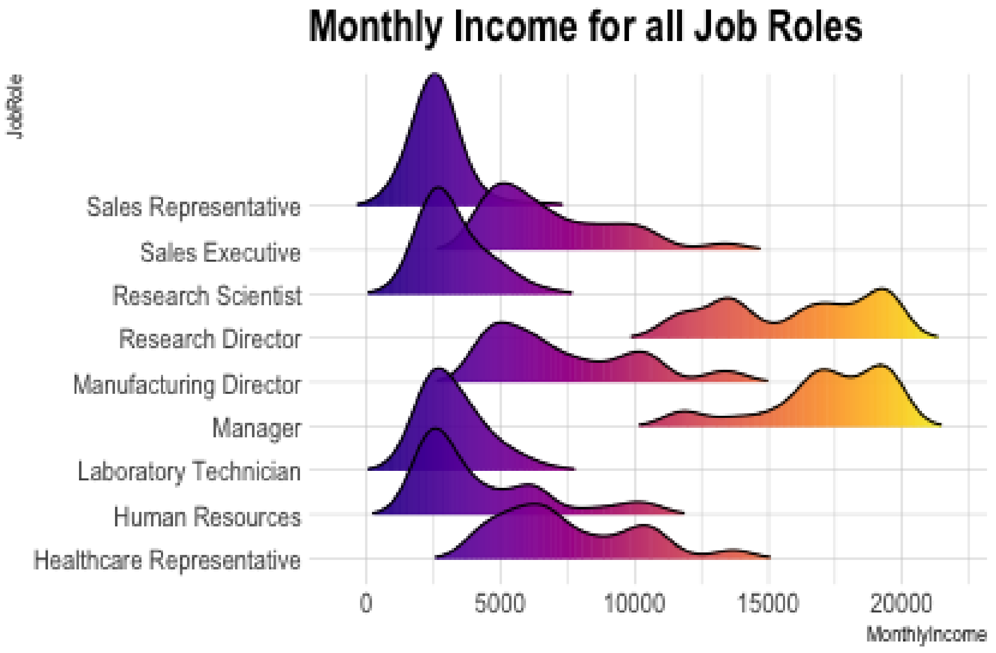
Alao, D.O., & Adeyemo, A.B. (2013). ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS.

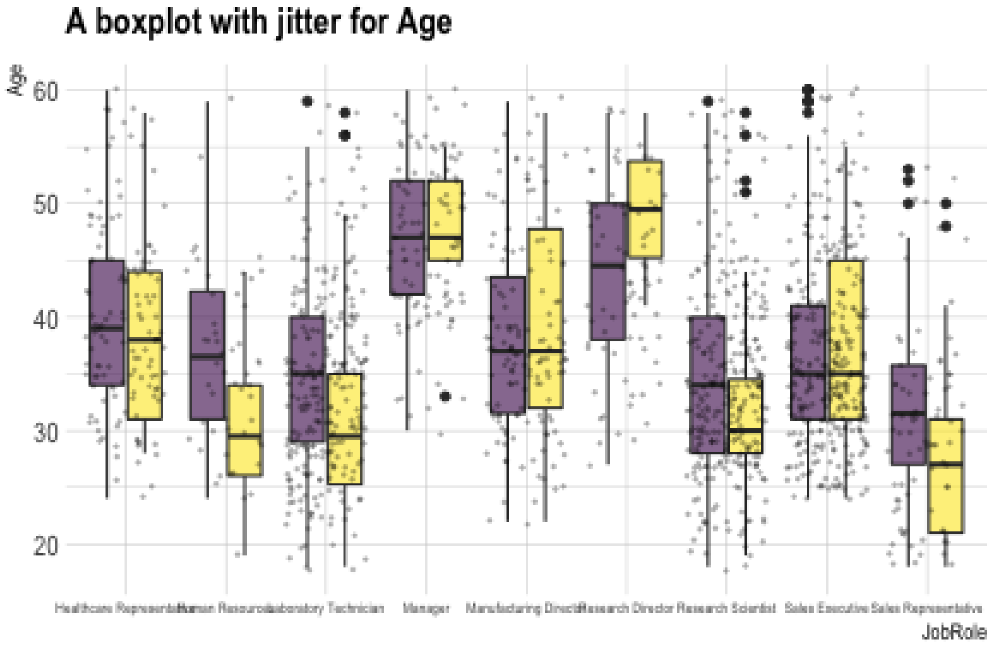
Khan, Afaq & Khan, Sumaira. (2019). Factors Affecting Employee Attrition and Predictive Modelling Using IBM HR Data. Journal of Computational and Theoretical Nanoscience. 16. 3379-3383. 10.1166/jctn.2019.8296.

Appendix

EDAs







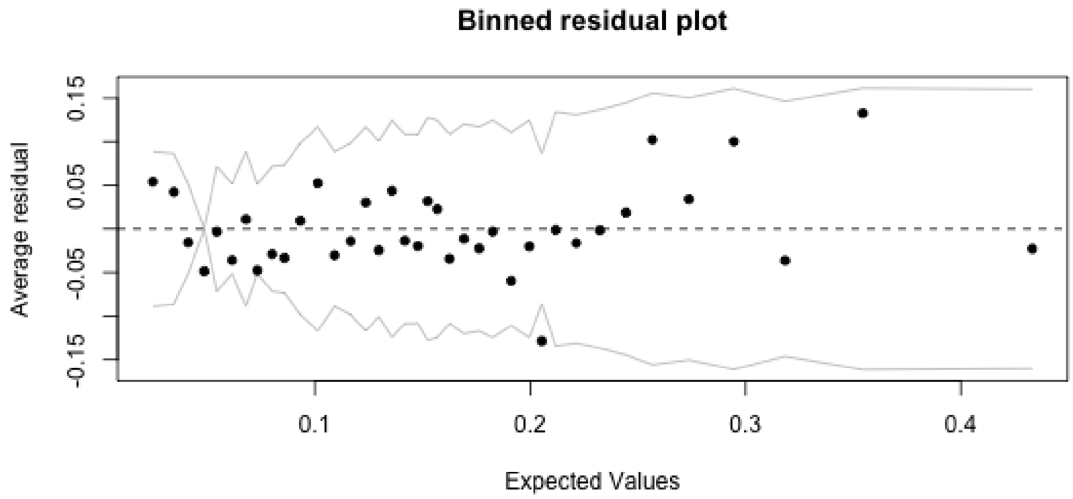


Figure 5: binned residual plot

Effects of the model

```
## $JobRole
##                           (Intercept)
## Healthcare Representative -0.50934250
## Human Resources            0.35917172
## Laboratory Technician      0.36487056
## Manager                   -0.12672190
## Manufacturing Director    -0.52355774
## Research Director         -0.44323561
## Research Scientist        -0.08915748
## Sales Executive            0.24789375
## Sales Representative       0.86089095
##
## with conditional variances for "JobRole"
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.357	0.496	-0.721	0.471
Age	-0.030	0.010	-3.082	0.002
DistanceFromHome	0.027	0.009	3.095	0.002
MonthlyIncome	0.000	0.000	-1.396	0.163
RelationshipSatisfaction	-0.106	0.067	-1.576	0.115
YearsSinceLastPromotion	0.035	0.027	1.320	0.187