Project of Employee Attrition

Zonghao Li

5/12/2021

1. Abstract

Employee attrition is one of the key factors that plague companies and in this project we will analyze the following. Firstly visualize and explore the analyses of some important variables, such as univariate analysis in terms of monthly income, distance from home, and whether or not they work overtime. Then we analyze the factors of employee attrition, explore the degree of influence of each variable by constructing an effective model, and make predictions of whether employees will leave or not. Finally, we conclude and provide appropriate recommendations for employers.

When constructing different models (logit, decision tree, random forest), there are two dimensions to measure the accuracy of models and then select the best model to make predictions next. Then we can achieve the following main conclusions: (1) The main reasons for employees to quit jobs are their monthly income, age and overtime. (2) Certain aspects of life, such as the distance from home to workplace, are also a relatively important reason for employees to leave jobs. (3) The attrition rate of employees with high job involvement but low monthly income is surprisingly almost the same as the attrition rate of employees with high monthly salaries in the same category. (4) Among different job roles, sales representative owns the highest attrition rate.

2. Introduction

As we all know, the development of the enterprise cannot be separated from the employees' contribution to the enterprise, and only when the employees carry out normal work, the enterprise can run smoothly. Employees, as the human resource reserve of the enterprise, are the important support for the effective implementation of the enterprise development strategy. Nevertheless, employee turnover (or "employee churn") is a common phenomenon in enterprises at present, and it is getting worse with the development of society, which is a tough problem for business managers. Admittedly, it is the employee's right to resign, but this can be an expensive issue for companies. The true cost of replacing an employee is often considerable. A study by the Center for American

Progress found that companies typically pay about one-fifth of an employee's salary to replace that employee, and the cost increases significantly when it comes to replacing executives or the highest paid employees.In other words, the cost of replacing an employee remains high for most employers. This is due to the time spent interviewing and finding a replacement, the signing bonus, and the lost productivity during the months it takes for the new employee to settle into the new role.

On the one hand, the mobility of employees has the rationality and necessity. For a company, it can optimize the structure of personnel within the company, and for the society, it can realize the rational allocation and full utilization of human resources. However, there is a common phenomenon of high staff turnover in a plenty of enterprises today for various reasons, which seriously restricts the development of enterprises, which are diverse such as low employee satisfaction, employee distrust of the company, high work pressure, weak corporate cohesion, and other contrasting work environments, which can be summarized as employee personal factors, corporate factors and market environment factors, etc.

3. Methods

3.1 Description of data set

In this case study, a HR dataset was sourced from IBM HR Analytics Employee Attrition & Performance which contains employee data for 1,470 employees with various information about the employees. We will use this dataset to predict when employees are going to quit by understanding the main drivers of employee churn. This is a fictional data set created by IBM data scientists". Its main purpose can be to demonstrate the IBM Watson Analytics tool for employee attrition. IBM (International Business Machines), is the world's largest company dealing with information technology and business solutions. IBM's various information systems have become the most reliable means of information technology in many important business areas such as finance, transportation, commodity distribution, government and education.

In this section, we will provide data visualizations that summarizes or extracts relevant characteristics of features in our dataset. Let's look at each column in detail, get a better understanding of the dataset, and group them together when appropriate.

3.1.1 Simple summary

Table1. Summary of variables (Age & Monthly income)

```
##       Age        MonthlyIncome
##  Min.   :18.00   Min.   : 1009
##  1st Qu.:30.00   1st Qu.: 2911
##  Median :36.00   Median : 4919
##  Mean   :36.92   Mean   : 6503
##  3rd Qu.:43.00   3rd Qu.: 8379
##  Max.   :60.00   Max.   :19999
```

We can get: the average age of the employees of the company is around 36 or 37 years old. The monthly income is about $4900, the median is used here, the average may cause bias.

3.1.2 Distribution of attrition among age, number of companies worked, years at company and monthly income

Figure1. Distributions of attrition(1)

From the four specific tables in Figure 1, we can get:

(1) Employees who are younger have a higher attrition rate, mainly in their 30s and before 40 years old.

(2) Employees who have worked for a greater number of companies are more likely to leave jobs.

(3) The longer employees have worked for a firm, the less likely they are to leave.

(4) Employees with low monthly income are more likely to quit jobs.

3.1.3 The Distribution of gender, job level, education and department

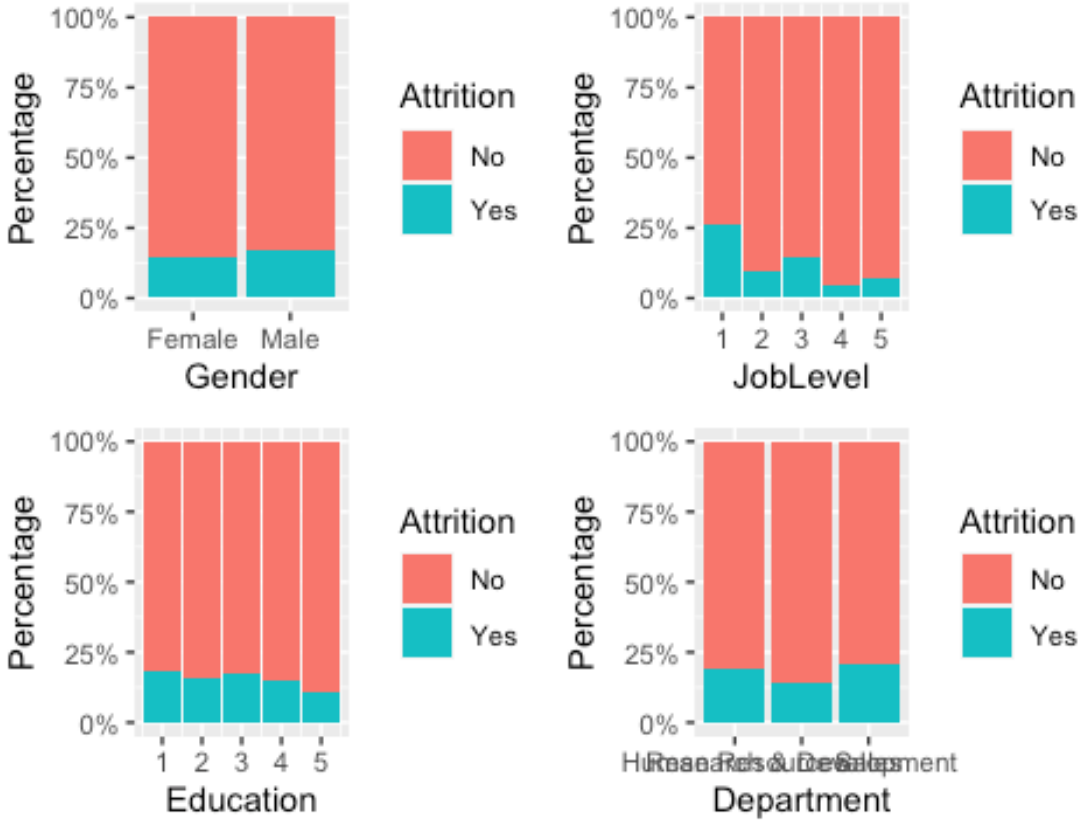


Figure2. Distributions of attrition(2)

From the four specific tables in Figure 2, we can get:

(1) The attrition rate of men is slightly higher than that of women.

(2) The higher the job level, the less likely for employees to quit jobs, and the high attrition is mainly concentrated in the newcomers of job level 1.

(3) There is not much correlation between education and attrition rate, but the attrition rate of employees with particularly high education is relatively low.

(4) Sales department has a higher attrition rate compared to the other two departments.

3.1.4 The Distribution of overtime, work-life balance, business travel, distance from home to workplace

Figure3. Distributions of attrition(3)

From the four specific tables in Figure 3, we can get:

(1) The more overtime worked, the higher the attrition rate.

(2) Employees who perceive a work-life balance of 1 have a higher attrition rate.

(3) Employees who travel a lot have a higher attrition rate.

(4) The farther away from home to workplace, the higher the attrition rate.

3.2 Methods

We can see that the attrition variable that would be predicted takes only two values: Yes (attrition) or No (no attrition). This prevents us from utilizing a simple linear regression, instead we can use logistic regression as one of our choices. Another two methods will be decision tree and random forest.

Before building models, we should firstly recheck the data set and remove several unnecessary factors, such as employee count, employee number, standard hours. After that divide this data set into train set and test set, and then we can build these three models and make some predictions.

To evaluate the accuracy of the model, we will judge it in two dimensions. The first dimension is the rooted mean squared error (RMSE). The RMSE is used to measure the deviation of the observed value from the true value and it is a very common measure of the accuracy of the constructed model. The second dimension is the ROC curve, which is a curve plotted with TPR (True positive rate, i.e., specificity) as the y-axis and FPR (False positive rate, i.e., sensitivity) as the x-axis, and is mainly used to evaluate the merit of a binary classifier. The performance performance of this classifier is represented by the area under the ROC curve (AUC). An ideal ROC curve will fit tightly in the upper left corner, so the larger the AUC, the better the classifier and the more effective the model will be.

4. Results

Following the methods described in 3.2, we can obtain the results below.

4.1 Three models

Model1: Logit model

```
## Area under the curve: 0.8491

## [1] 3.871633
```

From the results, in the aspect of AUC, the high value of AUC affirms the predictive power of the model.

Model2: Decision tree

```
## Area under the curve: 0.6837

## [1] 0.3519016
```

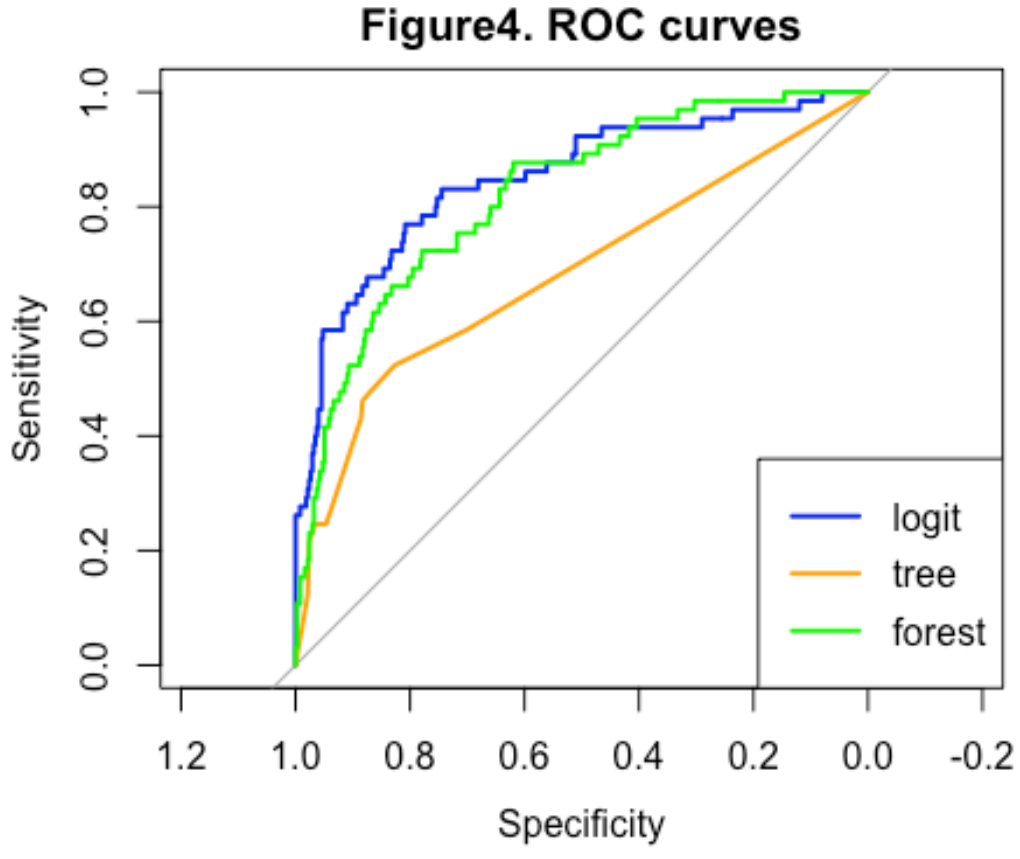
Although the value of RMSE is much lower than logit model, the values of AUC is relatively low, so this method is not considered to be a very good model for predicting attrition.

Model3: Random forest

```
## Area under the curve: 0.8214

## [1] 0.3156632
```

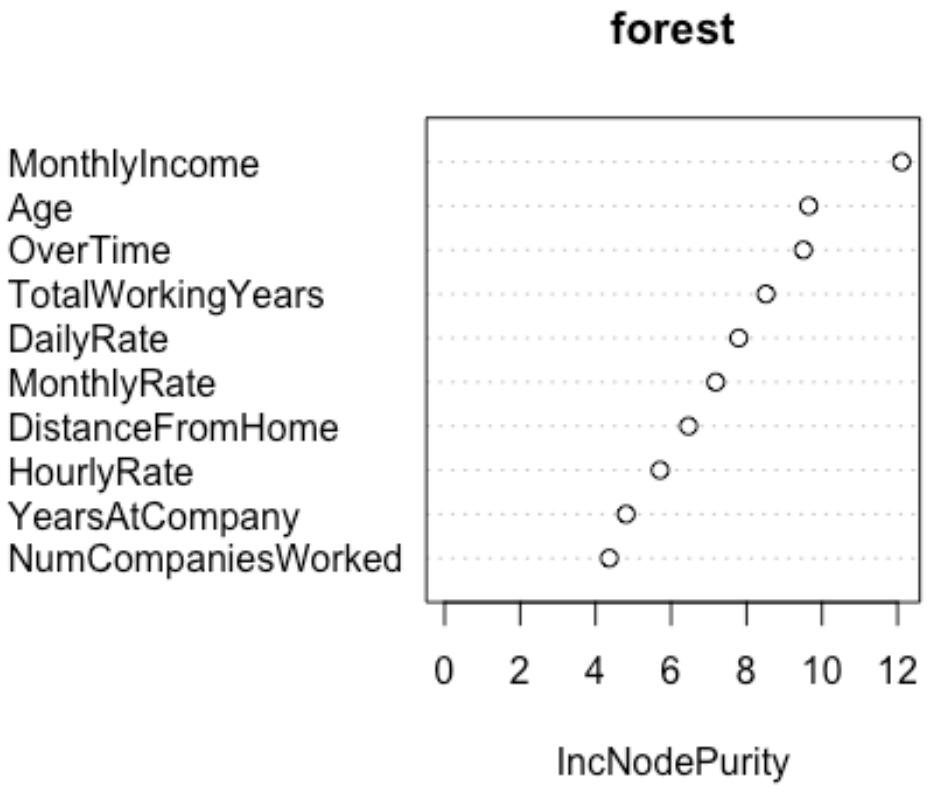
Also, we can plot the above three ROC curves in the same figure, as shown in Figure 3-1 below.



It is obvious that the ROC curves of logit model and random forest are similar to each other, which can also be verified by their close AUC numerical results above. Combined with these three RMSE values, random forest model is the relatively convincing model to be chosen to predict the attrition of employees.

4.2 The importance of variables

Under this model, we can sort all variables by importance as shown below (select top 10 here).

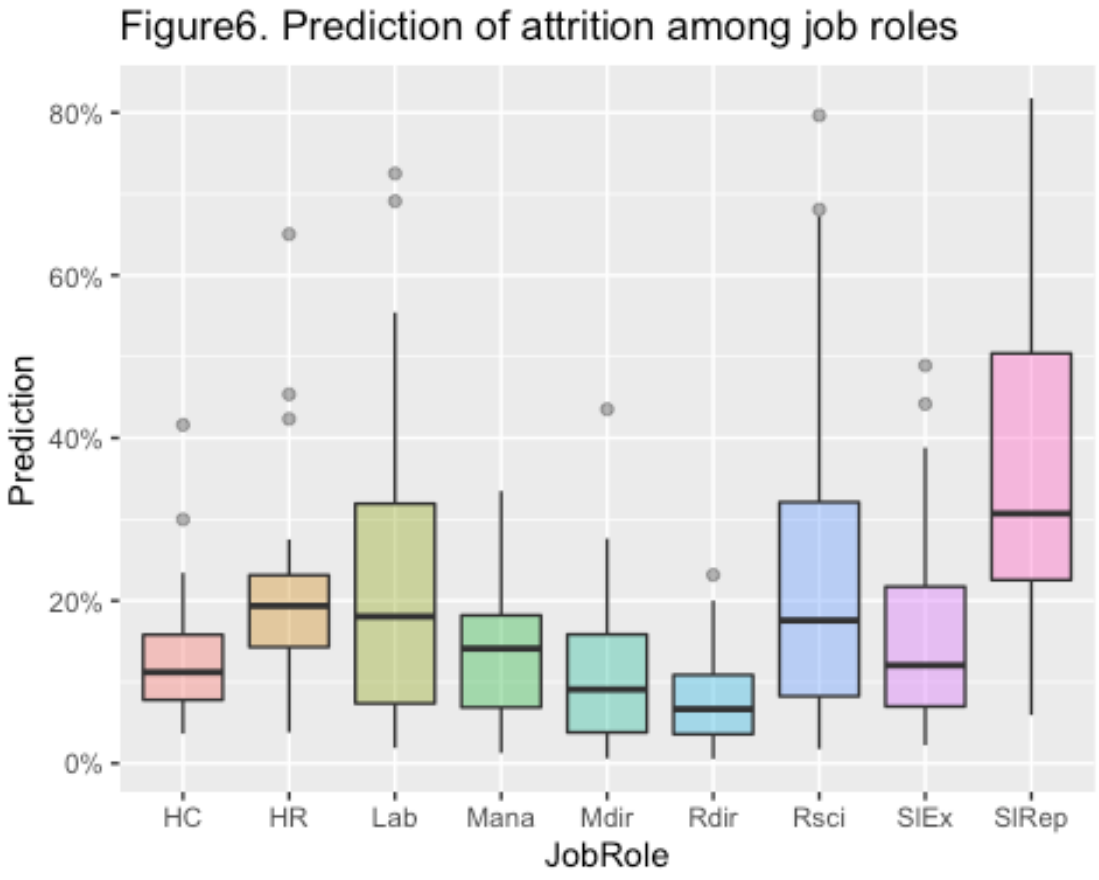


The variables at the top of the list are: monthly income, age, overtime and total working years.

4.3 Make a prediction–the attrition rate of employees with high job involvement and low monthly income

The first three graphs in Figure 5 shows that the relationship between attrition rate and monthly income is negative, while the last graph turn this relationship to smooth or a little positive when job involvement is very high.

4.4 Predict which positions will have the highest attrition rate



The top few positions with high attrition rates are: Sales representative, research scientist, human resources and laboratory technician.

5. Conclusions

5.1 Interpretations of main results

(1) The main reasons for employees to quit jobs are their monthly income, age and overtime. This makes sense since more income and less overtime appeals to more employees, and when people get older they will less likely to change their job.

(2) Certain aspects of life, such as the distance from home to workplace, are also a relatively important reason for employees to leave jobs. It is obvious that the longer distances incur both time costs and money costs (transportation). People may find

very tired after working all day, plus taking much time on the way, so they prefer a job with shorter distance from home to workplace.

(3) Graph4 in Figure4 indicates that the attrition rate of employees with high job involvement but low monthly income is surprisingly almost the same as the attrition rate of employees with high monthly salaries in the same category, and it is even a little less likely for them to leave jobs. This may be because such employees have a sense of belonging to the company or because the company's other benefits are nice, making them willing to stay with the company even if their income is low.

(4) Among different job roles, sales representative owns the highest attrition rate. On the one hand, the pressure of this job is high and the elimination rate is high. On the other hand, some of the employees with strong ability have accumulated a lot of contacts in their work, so they would choose to go out to work alone. So, it is not suitable to be a long-term job.

5.2 Main lessons

In summary, this project will be able to generate the following recommendations for employers.

(1) In order to control the employee attrition rate, employers can focus on how much monthly income and overtime system. The number of overtime can be reduced appropriately, or different management programs can be implemented according to different types of overtime situations, instead of a one-size-fits-all approach.

(2) In many cases the pursuit of higher income is still a direct reason for jumping ship. Therefore, employers should create an equal opportunity and fair competition for employees, so as to improve the sense of fairness in income.

(3) Among sales department, employers should strengthen the training of sales managers 'leadership. They can also classify sales staff into different levels, such as sales representative, senior sales representative, sales specialist, etc., so that different categories of employees can be recognized and promoted in their respective fields, thus enhancing their income.

(4) Firms should shape an excellent own corporate culture and create a harmonious working atmosphere, which is always meaningful.

Appendix

In section 4.2, the importance of all of the variables are listed below:

```
##                             Overall
## Age                       9.6478068
## BusinessTravel            2.2299315
## DailyRate                 7.7937739
## Department                1.2818195
## DistanceFromHome          6.4626911
## Education                 1.9097923
## EducationField            2.7307460
## EnvironmentSatisfaction   3.2308153
## Gender                    0.8276560
## HourlyRate                5.7095065
## JobInvolvement            3.3466143
## JobLevel                  2.2321711
## JobRole                   3.8252457
## JobSatisfaction           2.4605963
## MaritalStatus             2.6885297
## MonthlyIncome            12.1094207
## MonthlyRate               7.1863148
## NumCompaniesWorked        4.3638968
## OverTime                  9.5059693
## PercentSalaryHike         4.2485581
## PerformanceRating         0.4331882
## RelationshipSatisfaction  2.9127274
## StockOptionLevel          3.4863176
## TotalWorkingYears         8.5149539
## TrainingTimesLastYear     3.3462566
## WorkLifeBalance           3.8523794
## YearsAtCompany            4.8104530
## YearsInCurrentRole        2.9069108
## YearsSinceLastPromotion   3.1557509
## YearsWithCurrManager      4.1177518
```