Why best employees leave?

Human Resources Analytics by Exploration Data Analysis and Modeling

Jiamin Zhong, Ruonan Zhang, Jiayi Geng

December 21, 2017

Abstract

Employees leave for lots of reasons, but some can be controlled. Does salary really matter the most when reducing turnover rate? Our research reveals the important influences of turnover based on data analysis and modeling are usually satisfaction level, workload and performance, and provide suggestions to manager who wants to keep good employees from leaving.

Introduction

For business managers, few things are as costly and disruptive as good people walking out the door. According to existing research, employee turnover already costs US companies $160 billion a year. One Society for Human Resource Management publication predicted that direct employee replacement costs can reach as high as 50 percent to 60 percent of an employee's annual salary.

Employees quit their job for many reasons. They change careers, find upwardly promotions, or go back to school. But most of the reasons why employees quit their job are lays inside their office. To reduce the turnover rate, we cannot simply attribute these to unsatisfactory wages. whole bunch of influences should be considered including workloads, promotions, teamwork or such.

Our project aims at discovering the importance of possible leaving reasons for employees in a simulated situation concerns a big company, by conducting data analysis and modeling based on HR data resources. The goal is to bring up reproducible research method that could be applied in practical situation.

Materials and Methods

Data description

Our research based on a data set named HR_comma_sep, which is a simulated data set uploaded by user from Kaggle and has been widely used for modeling. The Human Resources Analytics is a dataset providing informations on the situation and work aspect of 14999 employees. There are 10 variables in this dataset, 8 of which are numeric and 2 of which are

Table 1: Numerical Variables Summary.

satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left
Min. :0.0900	Min. :0.3600	Min. :2.000	Min. : 96.0	Min. : 2.000	Min. :0.0000	Min. :0.0000
1st Qu.:0.4400	1st Qu.:0.5600	1st Qu.:3.000	1st Qu.:156.0	1st Qu.: 3.000	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.6400	Median :0.7200	Median :4.000	Median :200.0	Median : 3.000	Median :0.0000	Median :0.0000
Mean :0.6128	Mean :0.7161	Mean :3.803	Mean :201.1	Mean : 3.498	Mean :0.1446	Mean :0.2381
3rd Qu.:0.8200	3rd Qu.:0.8700	3rd Qu.:5.000	3rd Qu.:245.0	3rd Qu.: 4.000	3rd Qu.:0.0000	3rd Qu.:0.0000
Max. :1.0000	Max. :1.0000	Max. :7.000	Max. :310.0	Max. :10.000	Max. :1.0000	Max. :1.0000

categorical. Among the categorical variables, the `Sales` variable is nominal and the `Salary` variable is ordinal. The detailed information for these 10 variables are as follows:

Response:

left : It is a variable that shows whether the employee terminated a year later. It has a binary class of 0 and 1.

Numerical Predictors:

satisfaction_level: This attribute ranges from 0 to 1, denoting the overall satisfaction of the employee. It could represent how an employee is satisfied with this company. We assume that with a higher satisfaction level, the employee will not tend to leave.

Last_evaluation:This attribute ranges from 0 to 1, denoting the latest yearly evaluation. It shows how the employee is evaluated by his company. We assume that with a higher evaluation score, the employee will not tend to leave.

number_project: This attribute is the number of projects conducted by the employee. In this dataset, it is from 2 to 7. The more projects the employee conduct, the more valuable he may be in the company. But with too many projects, one would have a heavy workload.

average_montly_hours: This attribute is the average monthly hours the employee spend in the company. It ranges from 96 to 310 hours with a mean of 201.1 hours.

time_spend_company: This attribute represents the tenure in years of the employee. In this dataset, it is from 2 to 10 years.

Work_accident : This attribute is the work accidents happened within the past 2 years. 0 represents No and 1 represents Yes. Among all the employees, 2169 of them encountered a work accident. promotion_last_5years: This attribute denotes the promotions of the employee has within the past 5 years. Among all the employees, 319 of them got a promotion.

The table 1 is a summary of the numerical variables in this dataset.

Categorical Predictors:

sales: This attribute is the department that the employee belongs to. It has 11 categories.

salary: This attribute is the level of salary the employ received. It contains 3 level.

From the result above, we can see that normally in this company, the overall attrition rate is about 24%, the satisfaction level is around 62% and the performance average is around 71%.

We can also see that on average people work on 3.8 projects a year and about 200 hours per months.

Data Preprocessing

Before we proceed to further analysis and modeling, we should check the for missing values first.

In this case, there is no missing value in this data set.

Next, we modify the response variable `left` from its original use as a numerical variable, to a categorical variable with 1 for leave and 0 for not leave.

```
hr_data$left = as.factor(hr_data$left)
```

In addition, we also set the 2 categorical variables, `sales` and `salary`, as factor in the same way.

Our Method

The first thing to do is to train-test split the data.

```
hr_idx = createDataPartition(hr_data$left, p = 0.50, list = FALSE)
hr_trn = hr_data[hr_idx, ]
hr_tst = hr_data[-hr_idx, ]
```

We leveraged logistic classification, tree and random forest for modeling. The reason we choose those three models is our concentration of variable importance, which would be more interpretable by using those methods mentioned above. But still we would take predicting accuracy into account, since we believe higher accuracy stands for a better present of information. We would value the variable importance of the model with best performance.

The final models for each method we used are as follows:

Logistic classifier

```
set.seed(1)
HR_glm = glm(left ~ .,
             data = hr_trn,
             family = "binomial")
```

Decision Tree

For the decision tree model, we firstly train an unpruned classification tree using all of the predictors. We set a relatively small CP value in order to select the best size of tree in the next step.

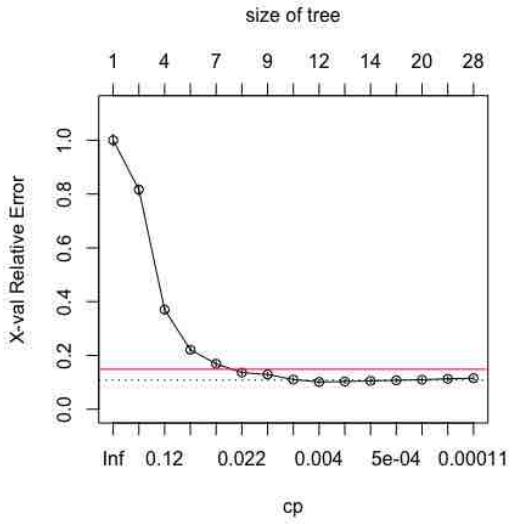


Figure 1: CP for decision tree

```
tree_mod = rpart(left~.,
                 data = hr_trn,
                 method = "class",
                 cp = 0.0001)
```

Instead of using the original model, we tried to figure out the optimal choice of CP. The cptable could give us all candidate tree sizes; for each tree size, we compute the corresponding cross-validation error, and then pick the subtree with the CV errors. There are usually two approaches to determining the optimal tree size. We chose to pick the optimal tree size which has the smallest CV error.

The figure 1 is the plot of relative error and cp values. The relative error would decrease dramatically when cp value reaches 0.022.

In this case,We use prune.rpart() to obtain this final tree model from our original tree.

```
tree.min = prune.rpart(tree_mod,
                       CP.min,
                       nodes = 4)
```

Random Forest

We further use random forest method. Random forest will grow multiple trees which are then combined to yield a single consensus prediction.Combining a large number of trees can often result in dramatic improvements in prediction accuracy, at the expense of some loss interpretation. To build this model, we first use the `train` function in caret package and find the best mtry. Then we set several groups of "ntree" and fit the model.

```
set.seed(430)
mtry = seq(1,7)
```

Table 2: Logistic Classification Variable Importance

	term	estimate	std.error	statistic	p.value
2	satisfaction_level	-4.0759465	0.1374890	-29.645616	0
7	Work_accident	-1.4445667	0.1250425	-11.552602	0
18	salarylow	2.0688336	0.1855559	11.149383	0
6	time_spend_company	0.2410186	0.0217020	11.105844	0
4	number_project	-0.3035947	0.0296620	-10.235149	0
19	salarymedium	1.5539103	0.1865303	8.330605	0

```
rf = train(
  left~.,
  data = hr_tst,
  method = 'rf',
  metric = "Accuracy",
  tuneGrid = expand.grid(.mtry = mtry),
  trControl = trainControl(method = 'cv',number = 5)
)
```

Below is the final model we choose, which gives us the highest accuracy in the combinations we tried.

```
rf_mod = randomForest(left~.,
                      data = hr_trn,
                      mtry = 7,
                      ntree = 500)
```

Results

Logistic Classifier

First let's look at the predicting accuracy of Logistic classifier. [code of predicting accuracy]

```
## [1] 0.7950393
```

There are no explicit quantitate expressions of the variable importance for glm classifier, but we can make use of the glm regression result. The critical value shows the significance of each predictors, thus revealing the importance of them while fitting the model. Below are the six predictors with the highest absolute critical value. [summary of glm model, sorted and subset]

We found out that varImp would also return the z-value directly.

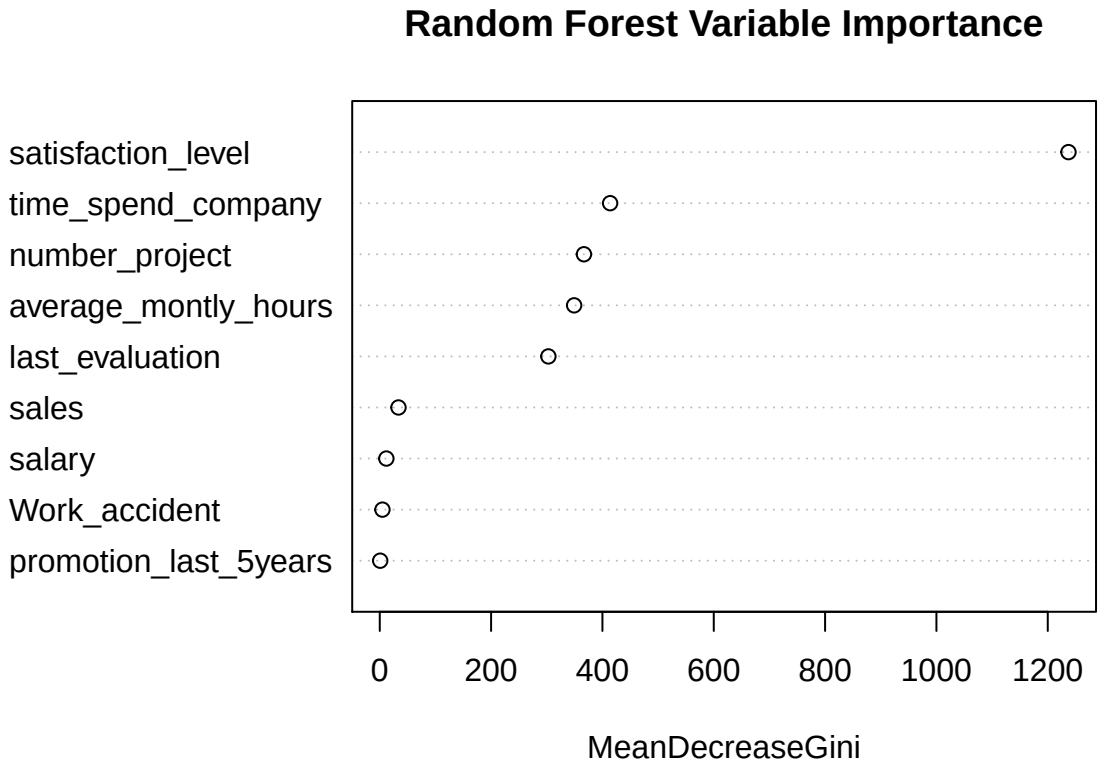
Tree

By using the decision tree model, we can easily produce a variable importance plot. It indicates that the level of satisfaction of the employee with the company mostly influences whether the employee will leave the company.

	Overall
average_montly_hours	1105.409939
last_evaluation	861.424293
number_project	1244.558658
salary	44.668469
sales	1.325118
satisfaction_level	1872.993683
time_spend_company	1113.874477
Work_accident	6.934380
promotion_last_5years	0.000000

Random Forest

Lastly, we build random forest model and generate a variable importance plot accordingly. We find that the satisfaction level has the greatest influence on wether employees retain or not. This result meets with the outcomes of the previous two models.



Discussion

Let's take a closer look of the result of our models.

Logistic Classification

For logistic regression, by the significance shown above, the satisfaction level has the largest z value, which means it has strong performance when predicting the possibility of turnover of an employee. Besides satisfaction, the predictors of work accidents, time spent in company, low salary and number of project also shows large significant. By this specific model, we concluded that those factors are the main drive for turnover in this company.

More specifically, we are interested to see that the working age variable has a positive coefficient, which means turnover rates are higher among experienced employees. Although it's not good news to managers, this result fits our expectation, as employees with more working experience have more choices of transfer.

Now let's see how it goes in other methods.

Decision Tree

Decision tree has a better performance compared to logistic classifier, it has higher accuracy. We also get the variables of importance, the top five of them are satisfaction level, average monthly working hours, number of projects in past two years, latest evaluation and working years in company. Part of the results match the conclusion from logistic classifier, but this time we did not see the importance of work accidents and salary level.

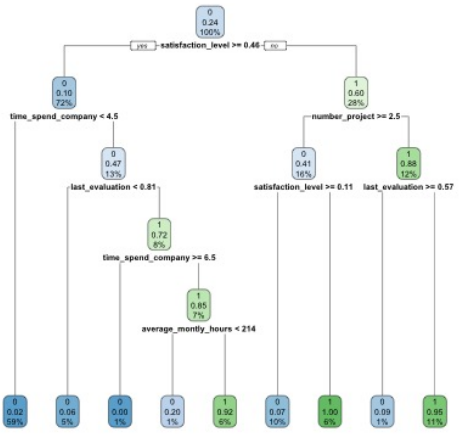


Figure 2: Decision Tree

The figure /@ref(fig:space-advert) shows the plot of our final tree model. As we can see from the figure, in the first case, people who are not an experienced staff, have a relatively high satisfaction level (higher than 0.46) and high evaluation score by the company and spend much time every month in his or her work would tend to leave the company. In the second case, with a low satisfaction level, even if the employee have some projects to work with, he or she may leave to obtain a higher satisfaction. In the third case, with a low satisfaction level and evaluation by the employer and having little things to deal with, this group of employee

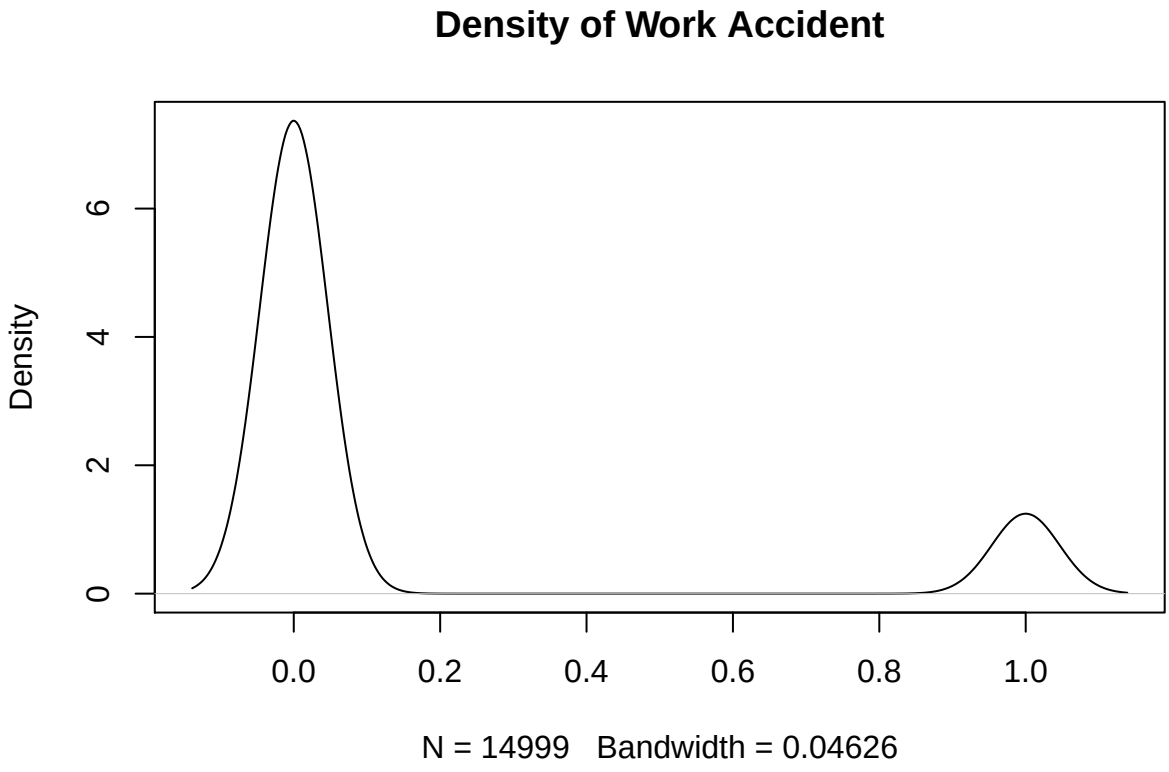
also leave the company. Therefore, the groups in first case are valuable employee. As far as we concerned, this group of people have strong enthusiasm and ability work, which will be valued highly in this company as well as other companies which are hiring. As a result, he or she might be reached by other better opportunities and thus leave this company.

Random Forest

A well-tuned random forest model shows the highest accuracy. The most importance variables are satisfaction level, latest evaluation, number of project, average monthly working hours, and working years in company.

Now it's clear that satisfaction level plays an important role in all three models, we can conclude that it is highly relative to the turnover of the company. From decision tree and random forest model, the number of project and average monthly working hours both rank high, these two variables stand for the workload, it is reasonable to see that work stress drives employees away.

The 'Work_accident' have contradictory performance in random forest model and glm model. It has remarkable significance for its coefficient in glm, but did not rank high in the importance plot of random forest. With a closer look to the data, we found out that this variable is largely skewed.



For glm model it could be treated as outliers, which has great influence on the original model. This influence would be defused by ensemble method such as random forest. From this aspect we consider the result of random forest to be better, since work_accident contains less information due to unbalance, we cannot assert that it has great influence on employees' leaving.

For this dataset, we observed that salary is not as important as workloads and satisfaction while an employee decides his/her leaving, which is different from our guessing (but we do know that employees with lower salary are more likely to leave).

Conclusion

Based on our research of this human resource data, we found out that it is the high satisfaction level that keeps good employees from walking away. What also matters are the reasonable workloads and employees' performance. Keeping a good employee takes more than providing higher payment. For this big company, managers should pay attention to the workload of their valuable employee as well as their satisfaction. The methodology we provided in this report could be a reference for the managers when they are wondering if their valuable employees would leave for the next year.In addition, managers may consider proper promotions as well as a better career path for those valuable employees who have a higher probability to leave.