Research on Human Resource Effect Based on Data Mining

Abstract

Nowadays, most of the human resource evaluation services are within the enterprise, and college students, as the fresh blood of the society, do not have a human resource evaluation model suitable for themselves. Therefore, it is of great significance to study a human resource model suitable for college students' self-adjustment. The data used in this article is human resource performance data from the Kaggle platform. The evaluation score, number of projects, working hours and work errors are used as independent variables. Three correlation models are established through correlation algorithms, and the importance of independent variables is analyzed. Variables have different degrees of impact on retention, promotion and employee level.

Keywords: human resources; correlation algorithm; importance analysis

Problem description

In the practice of human resource management, performance evaluation is a process of measuring and evaluating the performance level of an organization or employee within a predetermined time frame. Performance evaluation has the basic value of human resource management. In the attraction and allocation of employees, the influence and flow of employees, the salary system and the human resources work system, the correlation between the scores of performance evaluation and the results of performance evaluation becomes other human resources The starting point of practice, and determines the use effect of these practical activities.

Through the scientific performance evaluation system, it is accurately analyzed that the performance gap of employees is caused by insufficient knowledge and skills, which can help organizations and individuals to conduct targeted training, development and motivation on employees' lack of knowledge and skills from their respective perspectives. At the same time, performance evaluation also has a "beacon"-style benchmark value, which can help employees understand the organization's goals and codes of conduct, and help employees formulate their own career development plans.

It is precisely based on the potential strategic value of the performance evaluation chain

that various types of enterprises have adopted various performance evaluation systems or methods to enhance their competitiveness. However, existing human resource assessments are mostly for the healthy competition of employees within the enterprise or the structural adjustment of the enterprise itself, and are not suitable for students who have not entered the workplace. Therefore, this topic mainly studies the evaluation models related to college students such as retention and promotion. I hope to give some suggestions to students who are about to enter the workplace.

Data description

The data comes from the human resources analysis section of the Kaggle platform https://www.kaggle.com/ludobenistant/hr-analytics. Kaggle is a data science competition platform that mainly provides developers and data scientists with machine learning competitions, hosting databases, writing and sharing codes.

The data set includes 14,999 pieces of data about human resource performance information, each of which includes 10 attributes such as employee level, working hours, number of projects involved, department, salary, and whether there is any mistake. These 10 attributes and their specific meanings are shown in Table 1.

Table 1 Human Resources Data

Attribute name	Property content
satisfaction_level	from 0 to 1
last_evaluation	from 0 to 1
number_project	from 2 to 7
average_monthly_hours	from 96 to 310
time_spend_company	from 2 to 10
work_accident	0-no 1-yes
left	0-no 1-yes
promotion_last_5years	0-no 1-yes

sales	accounting, hr, IT, management, marketing, product_mng, RandD, sales, support, technical
salary	high, medium, low

Data preprocessing

Among the 10 attributes in the original data set, working hours and monthly working hours are similar and basically in a linear relationship. Therefore, only the attribute of monthly working hours is selected in the analysis. In addition, the two attributes of department and salary are not applicable to this analysis and are also excluded, so this article will analyze the remaining 7 attributes and perform preprocessing.

Model selection

After selecting the data to be studied, draw a correlation coefficient diagram for these 7 attributes, as shown in Figure 1.

As can be seen from Figure 1, each data is related to multiple data, so this article chooses to use correlation analysis to analyze and process the data.

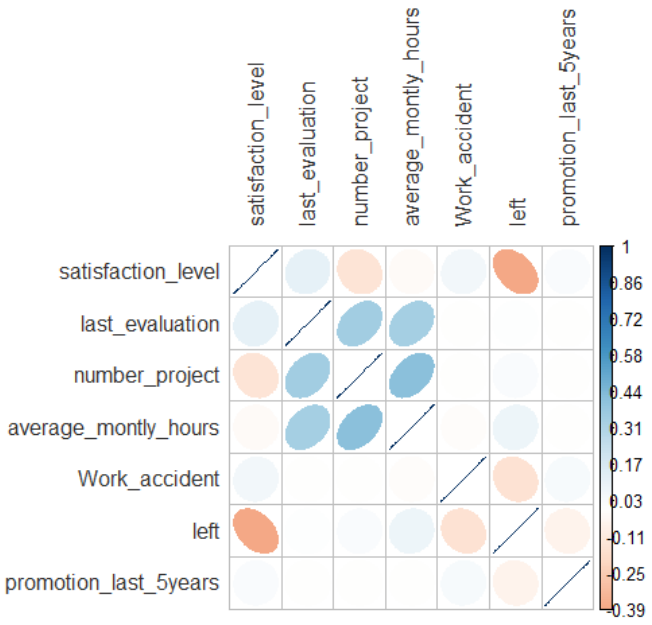

Fig.1 Graph of Correlation coefficient

Data binning

A total of 4 data were binned in the preprocessing. Divide employee levels into 3 levels: 1st, 2nd, 3rd; divide performance evaluation into 4 levels: excellent, good, moderate,

and poor; divide the number of participating projects into 3 levels: less, normal, more; The monthly working hours are divided into 3 levels: short, normal, and long. The data after binning is shown in Figure 2.

	satisfaction_level	last_evaluation	number_project	average_montly_hours	Work_accident	left	promotion_last_5years
1	3rd	Medium	less	short	0	1	0
2	2nd	Excellent	normal	long	0	1	0
3	3rd	Excellent	more	long	0	1	0
4	2nd	Excellent	normal	normal	0	1	0
5	3rd	Medium	less	short	0	1	0
6	3rd	Poor	less	short	0	1	0
7	3rd	Good	more	normal	0	1	0
8	1st	Good	normal	long	0	1	0
9	1st	Excellent	normal	normal	0	1	0
10	3rd	Medium	less	short	0	1	0

Fig.2 Data after Cut

Algorithm principle

Principle of Association Algorithm

Association rules are used to mine the correlation between valuable data items from a large amount of data, which is an important topic in the field of data mining. Like the implication formula of X⇒Y, association rules are the discovery of frequent patterns in transactions, itemsets and objects in relational databases, along with process dependencies, correlations and even possible causal structures.

In the field of data mining, Apriori algorithm is a classic algorithm for mining association rules. The Apriori algorithm uses a bottom-up method, starting from 1-frequent sets, and gradually finding high-order frequent sets.Known theorem: If the item set X is a frequent set, then its non-empty subsets are all frequent sets.

It can be obtained from the theorem that, given a k-frequent set item set X, all k-1 subsets of X are frequent sets, that is to say, two k-1 frequent itemsets can be found, they Only one item is different, and it is equal to X when connected. This proves that the k-candidate set generated by concatenating the k-1 frequent set covers the k-frequent set. At the same time, if the item set Y in the k-candidate set contains a certain k-1 subset that does not belong to the k-1 frequent set, then Y cannot be a frequent set and should be cut from the candidate set. This is one of the core ideas of the Apriori

algorithm.

Most of the association rules are not causal, it shows the correlation between the two. Therefore, in this article, the process of studying the impact of the four performance evaluation independent variables on the three dependent variables uses correlation analysis. The forerunner of its association rules are performance evaluation indicators, which specifically include evaluation scores, number of projects, working hours, and work errors, and the subsequent ones are whether to stay, whether to be promoted, and employee level.

Basic flow of association algorithm

When the transaction database D is scanned for the first time, a 1-frequent set is generated. On this basis, 2-frequent sets are generated through connection and pruning. And so on, until no more frequent sets can be generated. In the kth cycle, that is, when k-frequent sets are generated, k-candidate sets are first generated. Each item set in the k-candidate set is for two items that are different from each other and belong to k-1 frequent sets. The k-candidate set is screened to produce k-frequent sets.

Analysis of the importance of random forest algorithm

The random forest function package can be used to analyze the importance of variables. There are 4 methods for measuring the importance of variables: importance score, importance() function, average precision reduction and Gini index. In the function package randomForest in the R language, there is a parameter that evaluates the importance of variables, importance, which can analyze the importance of the data variables used in the modeling process.

In random forest, the main principles of analyzing the importance of variables are:

(1) For each random tree in the random forest model, OOB data is used when calculating its data error outside the bag, which is recorded as Oe1;

(2) Randomly add noise perturbation to the feature P of all samples in the out-of-bag data, so that the value of the sample at P can be changed randomly, and then calculate its out-of-bag data error and record it as Oe2;

(3) Suppose there are M trees in the random forest, then the importance of the feature

$P = \Sigma(Oe2-Oe1)/M.$

Algorithm implementation

Association algorithm implementation

The algorithm of the association rule analysis part is divided into three parts in total, which are the performance evaluation score, the number of projects involved, the average monthly working hours, and whether there is work negligence. These four factors (hereinafter referred to as the 4 factors) affect the level of employees, whether to stay, Whether the impact of promotion.

Analysis of the influence of 4 factors on whether employees will stay in office:

Performance is the preprocessed data set. The output on the right side of the rule is whether to stay left=0 and left=1. The left side includes last_evaluation, number_project, average_montly_hours, work_accident four attributes, set the maximum support to 0.005, confidence For 0.8, sort the rules and eliminate redundant rules.

The analysis of the influence of 4 factors on employee rank and whether it is appreciated is similar to the above-mentioned code flow, using the same association rules and eliminating redundant rules to analyze it.

Implementation of Random Forest Importance Analysis Algorithm

Random forest can give the effect of each independent variable on the dependent variable, and rank the importance of the independent variable. Run a random forest, 4 factors as independent variables, make the importance of independent variables to whether to stay.

Similarly, the analysis code for the importance of the 4 factors to the level of employees and whether they are appreciated is the same as the above, so I will not repeat them here.

Calculation results

Association analysis results

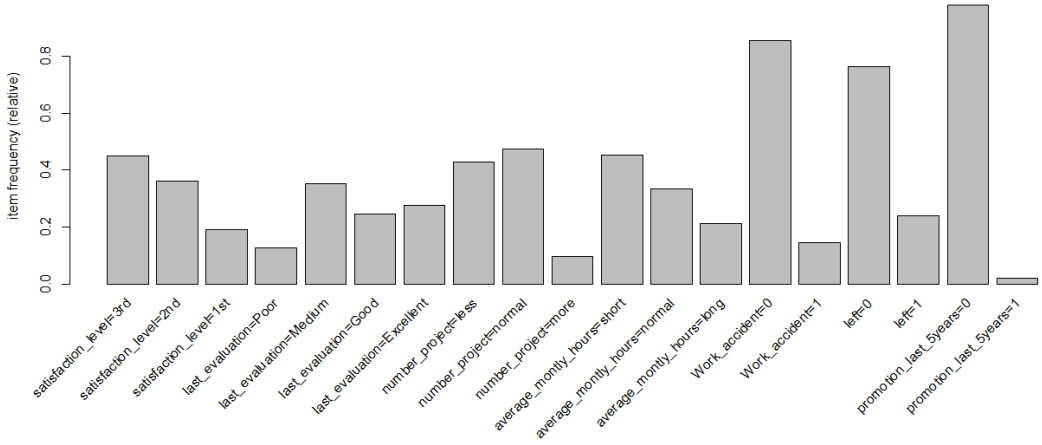

Fig.3 Frequent itemsets

The influence of independent variables on whether to retain

The association rules are shown in Figure 4.

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{last_evaluation=Excellent, number_project=more, average_montly_hours=long, Work_accident=0}	=> {left=1}	0.027868525	0.9609195	0.029001933	4.036077	418
[2]	{last_evaluation=Excellent, number_project=more, average_montly_hours=long}	=> {left=1}	0.029335289	0.9544469	0.030735382	4.008891	440
[3]	{last_evaluation=Good, number_project=more, average_montly_hours=long}	=> {left=1}	0.020601373	0.9142012	0.022534836	3.839850	309
[4]	{number_project=more, average_montly_hours=long, Work_accident=0}	=> {left=1}	0.048403227	0.8974042	0.053936929	3.769299	726
[5]	{number_project=more, average_montly_hours=long}	=> {left=1}	0.050736716	0.8890187	0.057070471	3.734078	761
[6]	{last_evaluation=Excellent, number_project=more, Work_accident=0}	=> {left=1}	0.031802120	0.8383128	0.037935862	3.521102	477

Fig.4 Association rules of left

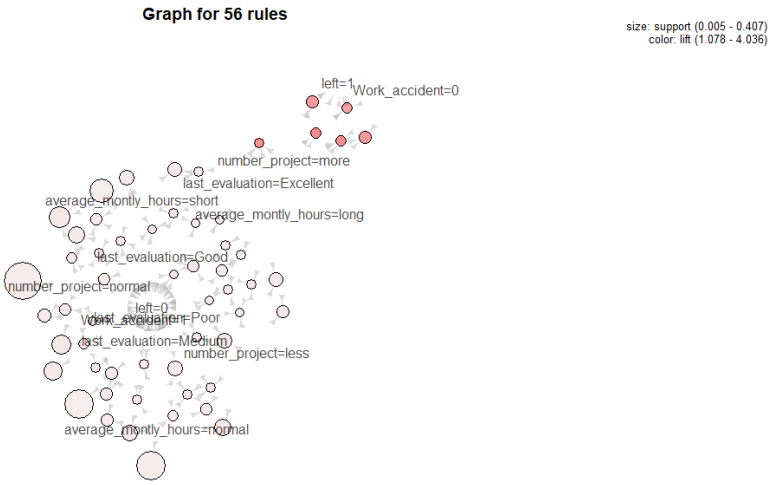

Fig.5 The association rules graph of left

It can be seen from Figure 5 of the relationship rule that people who have good performance evaluation, participated in many projects, long average working hours and no mistakes in work are more likely to be retained by the company; while those who have average performance in working hours and participating in projects are more likely to be retained by the company. Most will not be retained. And whether there will be mistakes at work, it has little effect on whether or not to stay. This shows that at work, the main factor for the boss to decide whether to keep an employee lies in the employee's effort and enthusiasm, and the higher tolerance for making mistakes.

The influence of independent variables on promotion

The association rules are shown in Figure 6.

```
     lhs                                rhs                          support confidence    coverage
     lift count
[1]  {number_project=more}           => {satisfaction_level=3rd} 0.082938863  0.8699301 0.095339689
 1.939082  1244
[2]  {number_project=more,

      average_montly_hours=long}     => {satisfaction_level=3rd} 0.054870325  0.9614486 0.057070471
 2.143077   823
[3]  {last_evaluation=Good,

      number_project=more}           => {satisfaction_level=3rd} 0.030202013  0.9151515 0.033002200
 2.039881   453
[4]  {last_evaluation=Excellent,

      number_project=more}           => {satisfaction_level=3rd} 0.038535902  0.9160063 0.042069471
 2.041786   578
[5]  {number_project=more,

      average_montly_hours=normal}   => {satisfaction_level=3rd} 0.017334489  0.8024691 0.021601440
```

Fig.6 Association rules of promotion

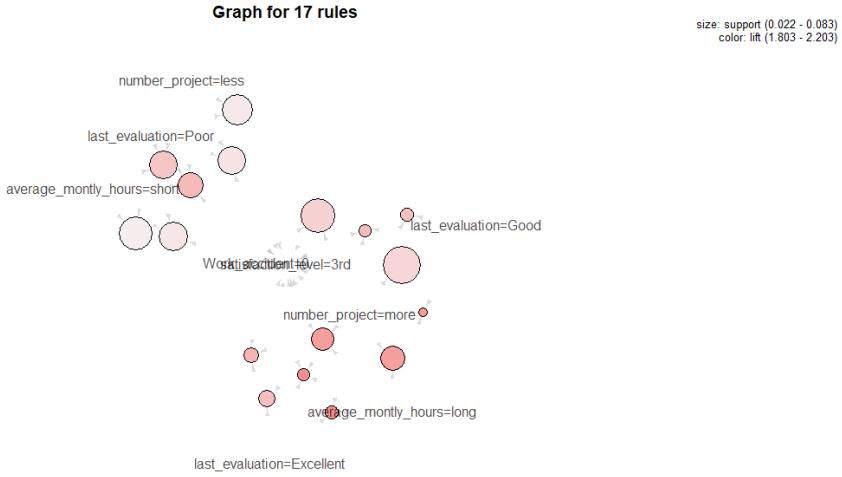

Fig.7 The association rules graph of promotion

It can be seen from Figure 7 of the relationship rule that people with average performance evaluation scores, few projects involved, short average working hours, and mistakes in work are less likely to be promoted; and they are more prominent in one aspect, but perform in other aspects. Ordinary people have little chance of being promoted. This shows that at work, the conditions for being promoted are much stricter than the conditions for staying. The decisive factor for promotion lies in every aspect, and the chance of promotion is only when all aspects are better.

The influence of independent variables on employee rank

The association rules are shown in Figure 8.

```
    lhs                             rhs                          support confidence  coverage
 lift count
[1] {}                           => {promotion_last_5years=0} 0.9787319  0.9787319 1.0000000 1.00
00000 14680
[2] {Work_accident=0}            => {promotion_last_5years=0} 0.8391893  0.9810600 0.8553904 1.00
23787 12587
[3] {number_project=normal}      => {promotion_last_5years=0} 0.4639643  0.9765647 0.4750983 0.99
77857  6959
[4] {average_montly_hours=short} => {promotion_last_5years=0} 0.4420961  0.9793236 0.4514301 1.00
06045  6631
[5] {number_project=less}        => {promotion_last_5years=0} 0.4207614  0.9795126 0.4295620 1.00
07977  6311
[6] {number_project=normal,

     Work_accident=0}            => {promotion_last_5years=0} 0.3928929  0.9777667 0.4018268 0.99
90138  5893
```

Fig.8 Association rules of level

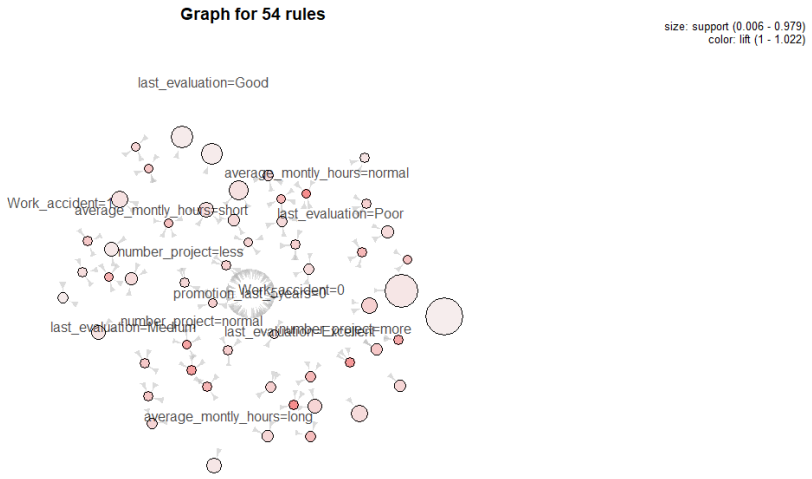

Fig.9 The association rules graph of level

It can be seen from Figure 9 of the relationship rule that people with average performance evaluation scores, few participating projects, and short average working hours have a lower rank in the company; especially the number of participating projects and average working hours have a more obvious impact. The higher the number of projects and the longer the working hours, the higher the position. This shows that in work, ability is an important factor in determining the level of a position, and the ability of an employee, the effort and time spent directly determine the level of an employee. Paying attention to improving one's work ability is the key to determining the level of the position.

Results of analysis of importance of independent variables

MeanDecreaseAccuracy is a measure of the degree of reduction in the accuracy of random forest prediction when the value of a variable is changed to a random number. The larger the value, the greater the importance of the variable. MeanDecreaseGini calculates the influence of each variable on the heterogeneity of the observed value at each node of the classification tree through the Gini index, so as to compare the importance of the variable. The larger the value, the greater the importance of the variable.

From Figure 10 to Figure 12, we can see that the importance of variables measured by MeanDecreaseAccuracy and MeanDecreaseGini will be slightly different, but the difference will not be very large.

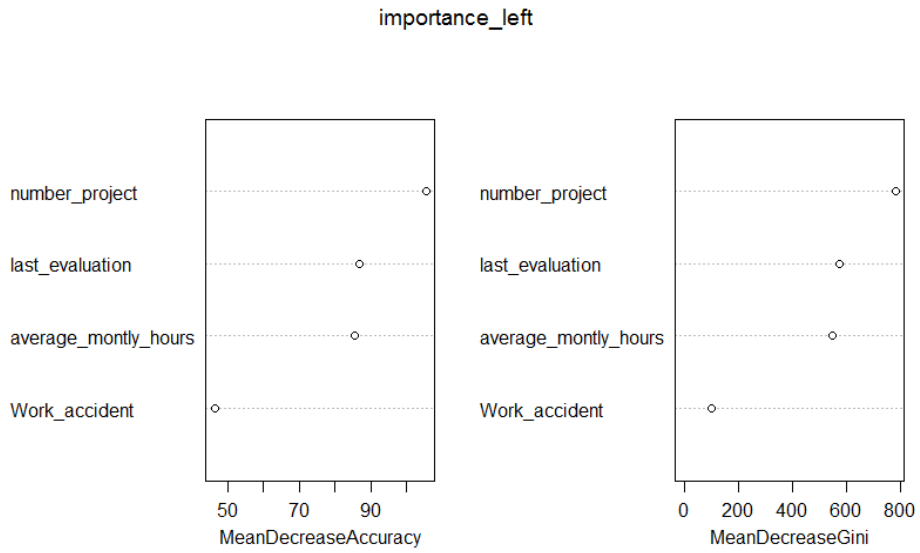

Fig.10 Analysis of the Importance of Independent Variables of left

It can be seen from Figure 10 that the main factors for the boss to decide whether to retain an employee at work are the number of projects the employee has participated in, the job evaluation score, and the average working time. In other words, retention is the most important factor in the ability to work, followed by high quality of work , Work actively and hard, and have a high tolerance for mistakes. Mistakes at work are not enough to be an important factor not to be retained.

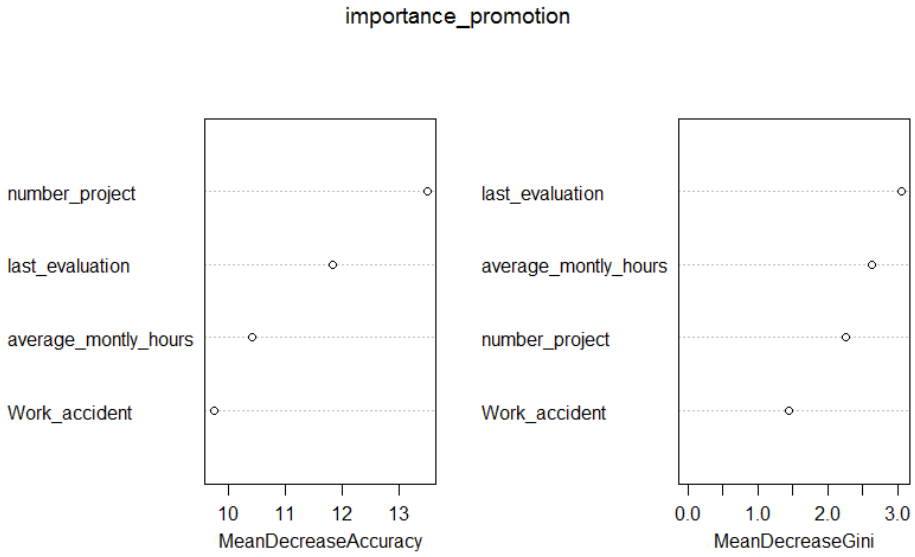

Fig.11 Analysis of the Importance of Independent Variables of promotion

From Figure 11, it can be seen that whether you can be promoted at work, performance evaluation scores, number of projects involved, and average working hours are all very

important. Even the number of mistakes that hardly play a role in deciding whether to stay is also whether you can be promoted. This occupies a proportion. This shows that the conditions and requirements for promotion are very strict, and it is no longer necessary to stay in office as long as they pass the pass. The most important factor in deciding a promotion lies in the ability to work and the quality of the work, while at the same time the work is also active. The mistakes in the work have little influence, but they may become an influential factor at the critical moment of promotion.

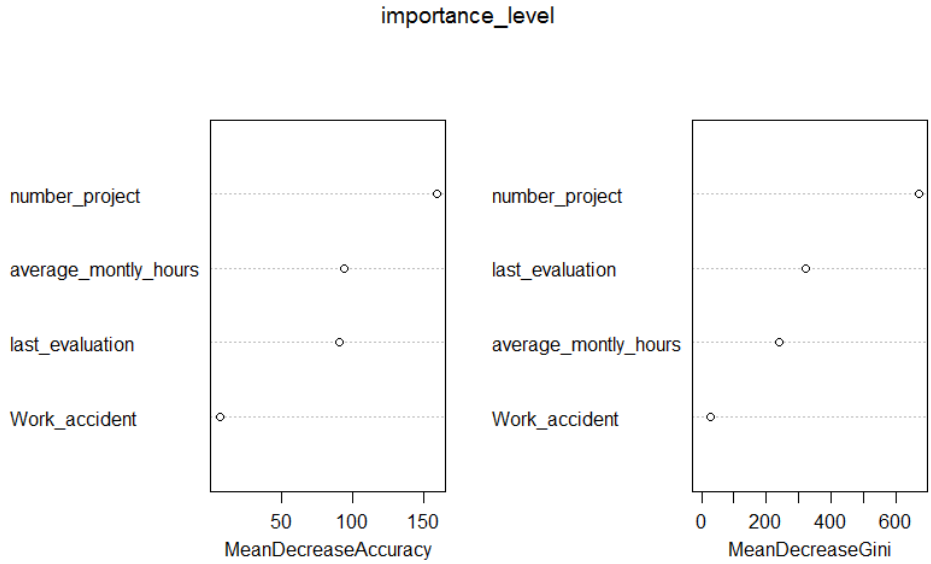

Fig.12 Analysis of the Importance of Independent Variables of level

It can be seen from Figure 12 that the most important factor in determining the job level is the number of projects involved, and this factor far exceeds the second and third job evaluation scores and average working hours. Participating in many projects means having diversified or highly professional working ability. Therefore, work ability is an important factor in determining the level of a position. In terms of raising the job level, the effect of effort is relatively inconspicuous. Employees with strong work ability are more likely to take a new level of job level.

Main conclusions

Factors affecting whether to stay in office

(1) The main factors for deciding whether to stay in office are: number of projects involved, job evaluation score and average working hours

At work, the boss decides whether to retain employees, mainly based on work ability, work quality and work attitude. For college students who are still in the internship period, they should do more, think more, and think more. In deciding to stay, the company has a high tolerance for mistakes, so employees who have just stepped into the work environment should not be afraid of making mistakes. Even if there is a mistake at work, as long as the attitude is correct and the work is serious, the chance of being retained is high.

(2) The conditions for deciding whether to stay in office are relatively loose

For the company, the retention of new employees is mostly not a highly competitive selection, so the conditions for retention are relatively loose. In terms of work ability, work attitude, and work quality, it is not required to be top-notch in every aspect. As long as they are in the middle and upper class, or are particularly prominent in one aspect, they will be very retained.

Factors influencing promotion

(1) The important factors for deciding whether to get promoted are: number of projects involved, job evaluation score and average working hours

Whether or not he can be promoted at work, performance evaluation scores, number of projects involved, and average working hours are all very important. This means that the employee's work ability, work attitude, and work quality are the criteria for the boss to judge whether he can be qualified for a higher position .

(2) The conditions for deciding whether to be promoted are relatively strict

Different from retention, promotion is a highly competitive selection activity, and the attitude of passing by will be quickly eliminated in this activity. The competitive form of promotion is destined to be very strict. Even if there is a level of competence among competitors, work errors that usually have little impact will play a key role at an important juncture.

Factors Affecting Staff Level

The important factor in deciding whether to get promoted is the number of projects involved.

The most important factor in determining the job level at work is the number of projects involved, and it is far more important than the job evaluation score and the average working time. Diversified or highly professional working ability determines the possibility of the employee to undertake multiple projects. Therefore, working ability is the most important factor in raising the level of the position.

Reference

[1] Li Chun. Universal human resource performance evaluation system [J]. Sichuan: Southwestern University of Finance and Economics, 2012.

[2]Han Botang. Comparison and selection of human resource performance evaluation methods[J]. Beijing: Beijing Institute of Technology, 2002.

[3] Shi Fuquan. An Empirical Study on the Performance Evaluation of Enterprise Human Resource Management Information System [D]. Nanjing: Nanjing University, 2016.

[4] Paul Teetor. Classical Examples of R Language [M]. Machinery Industry Press. 2013

[5]Gareth James. Introduction to Statistical Learning [M]. Machinery Industry Press. 2015