Information System Construction and Prediction Model of Leaving Using Big Data Mining and Random Forest

Qiaoyue Zhao*
Sichuan Vocational and Technical College, Suining, Sichuan Province, China

*Corresponding author: qiaoyuezhao@scvtc.edu.cn

Abstract. University talents play an important role in building the core competitiveness of universities and play a decisive role in improving the quality of university education. With the opening up and progress of society, the flow of talents in colleges and universities has become more and more frequent. Therefore, more and more attention has been paid to the analysis of factors affecting the turnover of outstanding college teachers. This paper establishes a random forest model to find out the key factors affecting the turnover of outstanding teachers in colleges and universities, and integrates these variables into an index to help colleges and universities understand which outstanding teachers need to be focused and predict the outstanding teachers on the job, and judge the probability of their departure., Which can formulate effective talent retention measures for the university managers. The experimental results verify that the work itself, work pressure and welfare remuneration are the most important factors affecting the turnover tendency of college teachers, followed by leadership management and development prospects. Relatively speaking, the impact of interpersonal relationships is the least. The evaluation model of university teachers' turnover intention based on the random forest model proposed in this paper can be further applied to the study of teacher turnover intention.

Keywords: Data mining, university teacher turnover, random forest, decision tree, factor analysis.

1. Introduction
With the continuous development of the economy, the construction of universities is also proceeding rapidly. How to effectively ensure the stability of the team of university teachers has become an important issue currently facing. These risks will not only cause the loss of property, but even endanger the university survive. Therefore, the research on the issue of teacher turnover has very important theoretical value and practical significance. In recent years, domestic and foreign scholars have carried out a lot of research work from different angles, and with the gradual deepening of the research on turnover intention, the research objects have been continuously expanded and the research content has also been continuously expanded. However, the discussion on the turnover tendency of college teachers has just begun, and theoretical research is not deep enough, and most of them are concentrated on qualitative analysis [1]. The actual questionnaire survey method is adopted for college teachers, and the research on the quantitative evaluation of teacher turnover intention based on the
statistics of objective data is still in its infancy. Because there are many factors that affect teacher turnover intention, the relationship between each influencing factor and teacher turnover intention is highly non-linear. It is quite difficult to model with traditional methods, which severely restricts the quantitative evaluation of teacher turnover intention. Therefore, it is necessary to use nonlinear scientific theories (such as random forest theory) to construct a relationship model suitable for quantitative evaluation of teacher turnover intention.

This article attempts to use random forest machine learning methods to conduct an empirical analysis on the teacher turnover factors of a university, find out the most critical factors affecting their turnover, remove redundant and useless features, reduce dimensional calculations, avoid over-fitting, and use selection at the same time [2]. The key variables come out to form a new index to help colleges and universities improve their management model, predict the outstanding teachers on the job, and judge the probability of their departure so that college managers can take corresponding measures to retain talents. Finally, put forward effective and feasible suggestions.

2. Algorithm introduction

Random forest is a combined classifier algorithm proposed by Leo Breiman in 2001. It is composed of many single decision trees, and finally the results of all decision trees are combined by the majority voting rule and finally the classification result is given. Random forest is a classifier composed of multiple decision trees \( \{h(x, \Theta_k), k = 1, \ldots, n\} \), where \( \Theta_k \) is a random vector that is independent and identically distributed. Finally, the output result is determined by the comprehensive voting of all decision trees [3]. Figure 1 shows the algorithm flow chart.

![Figure 1. Random forest algorithm flow](image)

Given k classifier sets \( \{h_1(x), h_2(x), \ldots, h_k(x)\} \), input vector \( x \) and output vector \( y \), the interval function is defined as follows:

\[
mg(x, y) = a_{\text{kl}} I(h_k(x) = y) - \max_{j \neq y} avkI(h_k(x) = j)
\]  (1)

In the formula, \( I(\cdot) \) is the indicator function, and \( avk(\cdot) \) is the average value. The interval \( mg(x, y) \) measures the minimum difference between the average number of votes that the set of classifiers classify a particular sample \( x \) against and the average number of votes that misclassify it into other classes; the larger the interval, the better the performance of the classifier. The generalization error of the classifier

\[
PE^* = P_{x,y}(mg(x, y) < 0)
\]  (2)
As the number of trees increases [4], for all random vectors \( \theta, \cdots, P \) \( * \) tends to

\[
P_{x,y}(p_{\theta}(h(x,\theta) = y) - \max_{j \neq y} p_{\theta}(h(x,\theta) = j) < 0) \quad (3)
\]

Marginal function of random forest

\[
m_{r}(x,y) = p_{\theta}(h(x,\theta) = y) - \max_{j \neq Y} p_{\theta}(h(x,\theta) = j) \quad (4)
\]

Strength of s classifier \( \{h(X,\theta)\} \)

\[
s = E_{x,y}m_{r}(x,y) \quad (5)
\]

Assuming \( s \geq 0 \), according to Chebyshev's inequality, from equations (4) and (6), we can get

\[
P \varepsilon \leq \text{var}(m_{r})/s^2 \quad (6)
\]

3. Experimental Design

3.1. Data collection

This article collects teacher information in a human resources department database of a university, and desensitizes sensitive personal information. The explained variable is whether the teacher resigns within six months ("1" represents incumbent; "0" represents resigned within six months), The main explanatory variables are shown in Table 1 and Figure 2.

| Table 1. Comprehensive evaluation index system of university teachers' turnover intention |
|---------------------------------|-----------------|---------------------------------|
| Influencing factors            | Serial number   | Include content                  |
| Leadership management          | 1               | University's values              |
|                                 | 2               | Social image of colleges and universities |
|                                 | 3               | Leadership quality and ability of colleges and universities |
|                                 | 4               | The degree of teacher participation in democratic management |
| Interpersonal relationship      | 5               | Relationship between teachers    |
|                                 | 6               | The relationship between teachers and leaders |
|                                 | 7               | The relationship between teachers and teachers |
| Work itself                     | 8               | The challenge of work            |
|                                 | 9               | Opportunities for success at work |
|                                 | 10              | Learning and further study opportunities |
|                                 | 11              | Teaching arrangement and environmental quality |
|                                 | 12              | teaching technology level        |
| Work pressure                   | 13              | Work intensity of teachers       |
|                                 | 14              | Difficulty of teachers           |
|                                 | 15              | teacher's interest in work       |
|                                 | 16              | Freedom of working hours         |
| Welfare remuneration            | 17              | Fairness of salary distribution  |
|                                 | 18              | Housing, medical and insurance systems for teachers |
|                                 | 19              | Retirement system                |
| Promising                       | 20              | Academic topics                  |
|                                 | 21              | professor                        |
|                                 | 22              | Salary increase opportunity      |
3.2. Predictive analysis

We use the collected university teacher data to establish a random forest prediction model. The comparison between the prediction results and the actual situation is shown in Table 1. From Table 2, it can be seen that the effect of the random forest prediction model is very good, and the prediction accuracy for leaving teachers is 100%, the accuracy of the forecast for current teachers is 99.7%.

| Table 2. Analysis of forecast results |
|---------------------------------------|
| prediction | On-the-job | Resign |
| On-the-job  | 324        | 0      |
| Resign      | 1          | 6      |

3.3. Analysis of the importance of factors affecting the turnover of outstanding teachers

The paper draws out the analysis of the importance of the factors affecting the turnover of outstanding teachers as shown in Figure 3. Since other documents mostly use the Mean Decrease Accuracy indicator, this article also focuses on its elaboration [5]. Figure 1 is the Mean Decrease Accuracy index, which refers to the decrease in average accuracy. If the variable is important, the error of the prediction will increase, that is, the increase of the error is equivalent to the decrease of accuracy. The smaller the accuracy, the more important the variable is. The order of importance of these variables can be seen from the figure. Among them, the more important variables are with the university, working years, performance evaluation, and average monthly working hours.

| Figure 3. The ranking chart of the importance of each factor's influence on the turnover of outstanding teachers |
This article uses weights to explain the impact of input (i.e., six values of leadership management, interpersonal relationships, work itself, work pressure, benefits, and development prospects) on output (i.e., the value of turnover intention), using Xu Hongzhong, etc. The method proposed by scholars calculates the certainty of the six input quantities to the output quantity (see Table 3).

| Leadership management | Interpersonal relationship | Work itself | Work pressure | Welfare remuneration | Promising |
|----------------------|----------------------------|-------------|---------------|----------------------|-----------|
| 0.1642               | 0.1411                     | 0.1797      | 0.1793        | 0.1771               | 0.1586    |

It can be seen from Table 3 that the work itself and work pressure have the greatest impact on the turnover intention of college teachers. Welfare compensation is also a major factor affecting the turnover intention of college teachers, followed by leadership management and development prospects, while the influence of interpersonal relationships is relatively the smallest. This is consistent with the author’s interview results. College teachers generally report that interpersonal relationships are relatively good, benefits are low, and work pressure is increasing, and the work itself is not satisfactory. Therefore, for university administrators, the objective environment for teachers’ work should be improved in terms of hardware, such as good office conditions, advanced teaching configuration, etc., in order to increase the attractiveness and satisfaction of the work itself. From the software, it is necessary to attach importance to emotional investment, implement people-oriented humanized management, and create a harmonious working atmosphere for teachers [6]. At the same time, a sound performance appraisal system and salary incentive system should be established to improve the scientific and effectiveness of salary distribution, and truly enhance its incentive effect. In addition, the self-development needs of teachers should be met, and the career planning of teachers should be organically combined with the development prospects of universities, so as to encourage teachers to contribute to the development and construction of universities.

University managers can formulate relevant policies based on these important influencing factors: (1) Pay close attention to the work demands of teachers with a working experience of 0 to 3 years, and stabilize the workforce; (2) Formulate a reasonable salary system to increase the corresponding income and benefits of teachers, to stimulate the enthusiasm of teachers to ensure that they contribute to the development of universities; (3) While paying attention to the development of teaching quality, there should be enough energy to focus on the needs of teachers, and to improve the work-life balance and job satisfaction of teachers; (4) Improve various systems related to turnover intentions, adopt a series of intervention strategies to reduce the turnover intention of higher-level management teachers, and try to avoid the loss of colleges and universities caused by the turnover of these teachers.

4. Conclusion

With the continuous deepening of research, the in-depth analysis of the reasons for the resignation of university teachers from multiple angles can allow university administrators to formulate corresponding measures more clearly to stabilize the stability of the team of university teachers, and effectively improve it from many aspects and various angles. After investigating the expectations of teachers, they can make reasonable analysis and decision-making, so as to effectively enhance the work confidence of college teachers, better devote themselves to the cause of teaching and let everyone unite and work together to build the university better.
References

[1] Setianto, S. K., & Jatikusumo, D. teacher Turnover Analysis Using Comparison of Decision Tree and Naive Bayes Prediction Algorithms on K-Means Clustering Algorithms at PT. AT. Jurnal Mantik, 4(3) (2020) 1573-1581.

[2] Westland, J. C. Predicting credit card fraud with Sarbanes-Oxley assessments and Fama-French risk factors. Intelligent Systems in Accounting, Finance and Management, 27(2) (2020) 95-107.

[3] Bhuva, K., & Srivastava, K. Comparative Study of the Machine Learning Techniques for Predicting the teacher Attrition. IJRAR-International Journal of Research and Analytical Reviews (IJRAR), 5(3) (2018) 568-577.

[4] Khan, E. A., & Hayat Khan, S. M. Factors Affecting teacher Attrition and Predictive Modelling Using IBM HR Data. Journal of Computational and Theoretical Nanoscience, 16(8) (2019) 3379-3383.

[5] Fan, H., Ma, Z., Li, H., Wang, D., & Liu, J. Enhanced answer selection in CQA using multidimensional features combination. Tsinghua Science and Technology, 24(3) (2019) 346-359.

[6] Nepal, S., Mirjafari, S., Martinez, G. J., Audia, P., Striegel, A., & Campbell, A. T. Detecting Job professor in Information Workers Using Mobile Sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(3) (2020) 1-28.