Early Warning Modelling of Employee Turnover Based on Psychological Analysis and Xgboost Algorithm

Zhang Biwei[1,2], Chen Yiwen[1,2]

1.Cas Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, 100101, China

2.Department of Psychology, University of Chinese Academy of Sciences, Beijing, 100049, China

Keywords: Psychological analysis, Employee turnover, Early warning model, Xgboost

Abstract: employee's turnover behavior contrary to the enterprise's will often means great losses to the enterprise. It is of great strategic significance for enterprises to reduce employee turnover as much as possible at the least cost. This paper proposes a new employee turnover early warning model and application process. Firstly, the factors related to employee turnover are preliminarily selected based on psychological analysis. Secondly, the attributes related to employee resignation are further selected by data preprocessing and Chi-square test. Then, the Xgboost algorithm is used to build an employee turnover early warning model. Finally, the early warning model is tested in the actual scenario of an Internet enterprise. The test results show that the effectiveness of the early warning model is 92.5%, which is recommended for large-scale application in the actual landing work.

1. Introduction

From the perspective of psychology, employee turnover can be divided into two categories. First, the enterprise does not want the employee to leave, but the employee insists on leaving. Second, the enterprise wants the employee to leave. The discussion scope of employee turnover behavior in this paper is limited to the first kind, which often means great losses to enterprises. First of all, it takes a certain amount of human and financial resources for the company to train new employees to become skilled employees in business lines. Secondly, it takes some time for new employees to grow into skilled employees, and the company will pay the waiting cost. Third, after employees grow into backbone talents, they usually master some business secrets of the company. After backbone employees leave, the company may face the risk of disclosure. For this reason, the company will need to pay a certain price, such as paying compensation for the non-competition agreement to the resigned employee, or adjusting the relevant business logic of the company, so that the trade secrets mastered by the resigned employee no longer have great value. Reducing the turnover rate of key employees has important economic value and practical significance for modern enterprises [1-2].

There are many reasons for employee turnover. For example, the salary is lower than demand, the promotion is difficult, the work is too tired, the work content is inconsistent with their expectations, the work responsibilities are frequently adjusted, dissatisfied with the work style of the leaders, the company favors one over the other in the implementation of rewards and punishments, the relationship between colleagues is not harmonious, the working atmosphere is poor, the internal consumption in the work is serious, headhunters recommend better positions, and etc.

The diversity of employee turnover brings great difficulties to establish an employee turnover early warning model. Many literatures have studied employee turnover from the perspective of psychology. It is discussed in [3] about the differences between employees with different personality fields from the perspective of psychology. It is discussed in [4] about the psychology reason of employee turnover from the aspects of family background, educational background, working years and industry background. The reasons for the turnover of employees in the Internet financial industry is discussed in [5]. The reasons for the turnover of employees engaged in sales

  DOI: 10.25236/iccem.2021.046

are discussed in [6].

The above research results do have some value, but there is still room for improvement. Firstly, the above research mainly studies the reasons for employee turnover from the perspective of psychology, and its prediction accuracy still needs to be improved. Secondly, large Internet enterprises have a large number of employees. They rely solely on manpower to analyze whether all employees have resignation risk. Their workload is very huge, and new technologies (such as artificial intelligence, AI) are needed to reduce the manual workload.

This paper will jointly use psychology science and data mining technology to establish an early warning model of employee turnover intention. Firstly, combined with the principles of psychology principle and data mining theory, the prediction of employee turnover intention is modeled as a binary classification problem, and the profit function of the prediction model are given. Secondly, based on the employee turnover data of an Internet company in 2020, the characteristic fields are extracted from the original data by using psychology analysis and Chi-square correlation analysis. Thirdly, the Xgboost model and logistic regression model of employee turnover intention early warning are constructed. Finally, the early warning model is applied to the actual data to test the model performance.

2Problem Statement and scheme

A.Process of employee turnover early warning modeling and application

We propose a data-driven employee turnover early warning modeling process, which is divided into three stages.

1) Modeling stage, make a preliminary analysis of the reasons for employees' resignation according to the principles of psychology and provide suggestions for subsequent data collection - collect the data of on-the-job employees and resigned employees of a large Internet enterprise within a certain period of time - data preprocessing - correlation analysis and eliminate partial fields - establish an employee resignation early warning model by using data mining algorithm.

2) Test stage, the employee resignation early warning model is applied to test the data set - test the effectiveness of the model - repeatedly optimize the model.

3) Application stage, extract the data of all current employees - input the employee resignation early warning model - output the list of employees with high turnover intention in the recent period - carry out personalized psychological intervention and material care for these employees - accurately intercept the resignation demands of some employees and reduce the turnover rate of employees.

B.Measurement of employee's turnover intention

Generally, people have the following consensus on the quantitative description of employee's turnover intention.

1) Let turnover intention be expressed by variable Y, Y should usually be modeled as a continuous variable, and Y belongs to [0, 1) interval. Y=0 means employee resignation is an impossible event, and Y=1 means employee resignation is an inevitable event.

2) There are many factors affecting the value of Y.

3) It is very difficult to find a formula to accurately describe the quantitative relationship between Y and the above factors.

However, starting from the actual demand to reduce employee turnover, it is not necessary to make an accurate measurement of Y. Generally, the problem can be simplified as dividing the employees to two category called as *low turnover risk* (no additional psychological intervention and material care are required) and *high turnover risk* (additional psychological intervention and material care are required to prevent the employee from leaving). Therefore, Y can be modeled as a two-dimensional variable, that is, Y=1 means *low turnover risk* and Y=0 means *high turnover risk.*

C.Classification model for predicting employee turnover intention

Suppose that in 2020, denote n to be the total number of employees in a company, and m to be the number of employees who have resigned. Let the above N employees to form set A. It can be divided into two categories: 1) positive class, including M samples. 2) Negative class, including N-M samples. The purpose of employee turnover intention prediction modeling is to mine the

characteristics of two classes in set A and build a binary classification model F.

Assuming that all the current employees of the company form set B, model F will be used to predict which employees in set B belong to the positive category, that is, *high turnover risk*, so as to provide a reference basis for subsequent employee care work.

Assuming that the binary model F predicts that *K* employees belong to the positive category, the perfect prediction result is that all *K* employees really resigned. However, this is only possible in ideal. The actual situation will have the following four prediction results, as shown in Table 1.

Tab. 1 A confusion matrix for employee turnover prediction

	Predicted as positive class	Predicted as negative class	Sum
Actually belong to positive class	*a*	*b*	*M*
Actually belong to negative class	*c*	*d*	*N-M*
Sum	*K*	*N-K*	

In Table 1, *a* denotes that employees with high turnover risk are correctly predicted as positive (True Positive, TP), *b* denotes that employees with high turnover risk are incorrectly predicted as negative (False negative, FN), *c* denotes that employees with low turnover risk are incorrectly predicted as positive (False Positive, FP), and *d* denotes that employees with low turnover risk are correctly predicted as negative (True Negative, TN).

Model F will be used for personalized employee care and retention. The company will care for and retain (*a*+*c*) employees who are predicted as positive class. Let *q* denotes the cost of retaining an employee, and *p* denotes the profit of retaining an employee. If the model is a perfect prediction, the total profit Q^{opt} can be expressed as

$$Q^{\text{opt}} = N \times (p - q) \tag{1}$$

However, the actual model cannot obtain perfect benefits because of FN and FP.

1) For TP case, the retention cost paid by the company is $a \times q$ and the profit obtained is $a \times p$.

2) For FP case, the *c* employees have low turnover risk and do not need to pay retention costs. The retention cost is $c \times q$ and the profit obtained is zero.

3) For FN case, the *d* employees who fail to pass judgment, their turnover risk is high, but the company does not care for and retain them, and their turnover behavior will bring losses $d \times p$.

The total profit *Q* of prediction model F can be expressed as

$$Q = a \times p - (a + c) \times q \tag{2}$$

The efficiency of the prediction model is defined as the ratio of the actual total profit *Q* to the perfect predicted total profit Q^{opt}.

$$\beta = \frac{Q}{Q^{\text{opt}}} = \frac{a \times p - (a + c)q}{N \times (p - q)} \tag{3}$$

The maximum value of β can be obtained by optimizing the proportion of FN and FP of prediction model F.

3Psychological analysis and data preprocessing

A.Data Collecting under Psychological Analysis

Data source: collected from an Internet fast-moving goods enterprise in China in 2020. There are 10000 employees, of which 2182 have resigned (marked as *positive category*) and 7818 still on duty (marked as *negative category*). 8000 employees were randomly selected as training set A and the remaining 2000 employees as test set B.

According to the results of psychological analysis, the main reasons for employees' turnover are divided into three categories.

1) External factors: social values, economy, law, transportation and market competitors.

2) Internal factors of the organization: poor salary and welfare, dissatisfaction with the leadership style of the boss, lack of promotion and development opportunities, heavy workload, high pressure, lack of attention, unable to give full play to their talents, etc.

3) Personal factors: family factors, personality traits, professional fields and personal achievement motivation factors.

Based on the above considerations, 129 attributes are collected in both set A and B.

B.Data Preprocessing

Many attributes are difficult to directly use for modeling, so data preprocessing is required first.

1) Duplicate data audit: It is to conduct data audit on the overall key indicators, for example: check whether each user has a unique record, check the accuracy of the data, and delete duplicate data.

2) Singular value data: Singular values are reflected in the data in the form of outliers, that is, they deviate greatly from most normal values, and are identified and deleted by the histogram or scatter diagram of the variables.

3) Severe missing features: For missing data, evaluate the difficulty and value of complementing, and identify and delete the severely missing features.

4) Re-encoding of strings: Such as “Graduated school”, “Gender”, “Education level”, etc., re-encode the strings into different variables in advance.

C.Chi-square Test and Feature Elimination

There are a total of 129 data set attributes. If all are used for modeling, the modeling time is very long. On the other hand, some interference attributes will reduce the accuracy of the model. Therefore, it is necessary to perform a correlation analysis on the data attributes first, and eliminate the weaker correlation attributes, and only retain the highly correlated attributes for subsequent modeling.

Table 2 Shows the Strong Correlation Attributes Retained after Chi-Square Test.

Tab. 2 Attributes retained after Chi-square test

Employee number	Working years	Commuting distance	Attendance
Gender	Job type	Last promotion time	Overtime per month
Age	Job rank	Relatives work in the same enterprise	Weekend overtime
Marital status	Education	Assessment result	Honorary title
…	…	…	…

4Algorithm modelling and model evaluation

A. Principle of Xgboost Algorithm

Xgboost algorithm [7] is based on traditional Boosting, using CPU multi-threading, introducing regularization items, adding pruning, and controlling the complexity of the model.

Compared with GBDT, Xgboost has the following improvements:

1) GBDT uses traditional CART as the base classifier, while Xgboost supports linear classifiers, which is equivalent to introducing logistic regression (classification problem) and linear regression (regression problem) with L1 and L2 regularization terms.

2) GBDT only uses the first-order derivative when optimizing. Xgboost does a second-order Tailor expansion on the cost function, and introduces the first-order derivative and the second-order derivative.

3) When the sample has missing values, Xgboost can automatically learn the splitting direction.

4) Xgboost learns from RF and supports column sampling, which not only prevents overfitting, but also reduces calculations.

5) The Xgboost cost function introduces a regularization term to control the complexity of the model. The regularization term includes the number of all leaf nodes, and the sum of squares of the L2 modulus of the Score output by each leaf node. From the perspective of Bayesian variance, the regular term reduces the variance of the model and prevents the model from overfitting.

B. Training model

There are about 80,000 users in training data set A. the variable “*whether turnover*” is taken as the dependent variable. 16 variables such as “*working years, assessment results* and *average monthly overtime*” are taken as independent variables (see Table 2). The proportion of positive and negative samples is about 22%: 78%. Through repeated optimization, the prediction model F of employee turnover intention is finally output.

The model call is shown in Figure 1. The operating environment of the model is shown in Figure 2. The parameter settings of the model are shown in Figure 3.

```
import numpy as np
from sklearn.model_selection import train_test_split
import xgboost as xgb
import pandas as pd
import matplotlib
%matplotlib inline
def GetNewDataByPandas():
    wine = pd.read_csv("/Data/UCI/wine/wine.csv")
    wine['alcohol**2'] = pow(wine["alcohol"], 2)
    wine['volatileAcidity*alcohol'] = wine["alcohol"] * wine['volatile acidity
    y = np.array(wine.quality)
    X = np.array(wine.drop("quality", axis=1))
    # X = np.array(wine)

    columns = np.array(wine.columns)

    return X, y, columns
```

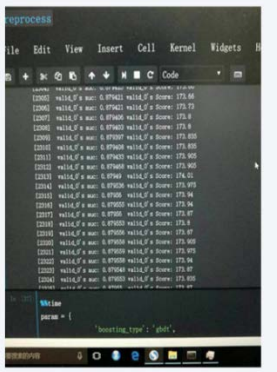


Fig.2 Boost Call

Fig. 2 Xgboost Runtime Environment

```
param = {'max_depth': 7, 'eta': 1, 'silent': 1, 'objective': 'reg:linear'}
param['nthread'] = 4
param['seed'] = 100
param['eval_metric'] = 'auc'
```

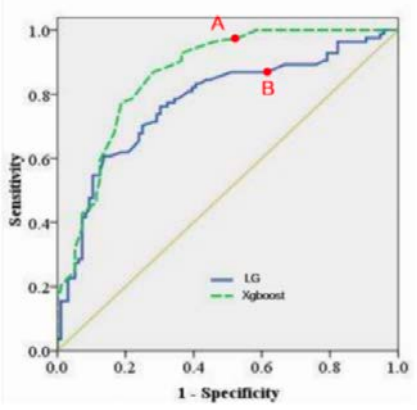
Fig.3 Xgboost Tuning Parameter

C. Performance comparison between Xgboost model and LR model

Figure 4 shows the ROC curve (green) of the prediction model F constructed by the Xgboost algorithm on the test dataset B. In order to make a comparison of model performance, the ROC curve (blue) of the prediction model constructed using the Logistics regression (LR) algorithm is also given. It can be seen that the prediction performance of the Xgboost model is better than that of the LR model.

D. Evaluation of Model Efficiency

The evaluation expression of employee turnover predication model is show in (3). The parameter is chose as p=10,0000 and q=50,000. The maximum value of β is solved in the Xgboost model and the LR model, and the optimal decision threshold is shown in point B and point A in Fig. 7. Point A (93.2% TP rate, 51.1% FP rate), point B (85.3% TP rate, 61.4% FP). The results of β are shown in Figure 5. The prediction performance of the Xgboost model is better than that of the LR mode, which is almost 92.5%. The effect is relatively satisfactory.



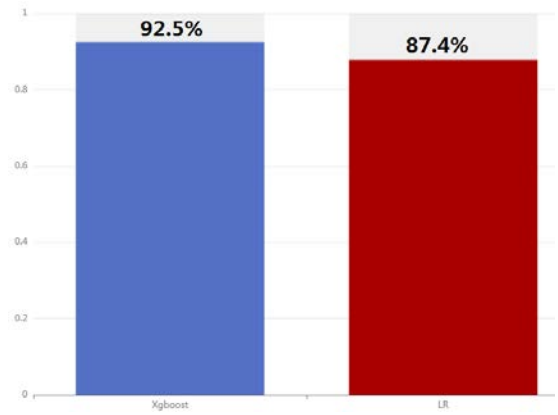


Fig.5 Curves of Xgboost and Lr

Fig. 5 Comparision of Model Efficiency B

2. Conclusion

The test results based on the actual data show that the Xgboost prediction model can effectively predict employees' turnover intention, and the effectiveness rate of the model is 92.5%. It can provide a more reliable basis for employees' accurate care and retention work, and it is recommended to apply it on a large scale in practical work. Xgboost prediction model points out that employee with different backgrounds have different turnover intention. The top three factors leading to employees' high turnover intention are: average monthly overtime, rank promotion time

and distance from home, which are basically consistent with the results given by psychological analysis. It is worth noting that the actual test results show that the income does not rank high among the factors leading to high turnover intention, which potentially shows that the income distribution mechanism of the enterprise is basically reasonable.

References

[1] N Santhanam, Kumar J R , Kumar V , et al. Employee turnover intention in the milieu of human resource management practices: moderating role of work-life balance[J]. International Journal of Business Innovation and Research, 2021, 24.

[2] Mathisen J , Nguyen T L , Jense J H , et al. Reducing employee turnover in hospitals: estimating the effects of hypothetical improvements in the psychosocial work environment[J]. Scandinavian Journal of Work, Environment & Health, 2021.

[3] Tab A , Sttt A , Dtnn A , et al. Psychosocial influences on psychological distress and turnover intentions in the workplace[J]. Safety Science, 137.

[4] Ntow M , Abraham D K , Bonsu N O , et al. Psychosocial Risk and Turnover Intention: The Moderating Effect of Psychological Wellbeing[J]. Advances in Safety Management and Human Performance, 2020.

[5] Suhartatik A , Junaedi C M , Novianti P M . THE IMPACT DISTRIBUTIVE JUSTICE, PROCEDURAL JUSTICE, INTERACTIONAL JUSTICE, EMPLOYEE ENGAGEMENT AND JOB SATISFACTION ON TURNOVER INTENTION. 2020.

[6] Urbanaviciute I , Massoudi K , Toscanelli C , et al. On the Dynamics of the Psychosocial Work Environment and Employee Well-Being: A Latent Transition Approach[J]. International Journal of Environmental Research and Public Health, 2021.

[7] T. Chen, S. Singh, B. Taskar, and C. Guestrin. Efficient second-order gradient boosting for conditional random fields. In Proceeding of 18th Artificial Intelligence and Statistics Conference (AISTATS’15), volume 1, 2015. Shandong normal University. 3, 11-14.