International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 23, No. 1, 2026

Using Logistic Regression and Ensemble Learning for Employment Status Prediction

Guanlin Wan, Guangyu Wu, Huabin Cao

College of Technology, Jiangxi Normal University Science and Technology College, Nanchang, Jiangxi 332000, China

Abstract: As digital technology reshapes the labor market, real-time insights into employment dynamics have become key to the smooth operation of the economy. Based on 5,000 anonymized sampling data in Yichang City, this study constructs a multi-dimensional employment status analysis and prediction framework. The first step is to use data cleaning and visualization to evaluate the influence of age, gender, education and other characteristics to reveal structural characteristics such as high youth unemployment rate and weak female employment stability. In the second step, the chi-square test was used to screen significant variables, and a logistic regression model was constructed to predict the employment status of 20 test samples, with an accuracy of 81.93% and a recall rate of 97.7%. The third step is to introduce macroeconomic indicators such as GDP growth rate, urban registered unemployment rate, and policy support level, and use the random forest model to optimize the prediction after integrating with individual data, with an accuracy of 81.2% and an F1 value of 0.90, confirming the moderating effect of macro factors on employment. This study provides a data-driven decision-making basis for regional employment policy formulation and targeted assistance.

Keywords: Employment Forecasting, Macro Data Fusion, Feature Importance, Logistic Regression, Random Forest.

1. Introduction

As the core element of people's livelihood security and the foundation of sustainable economic development, employment has shown the characteristics of overall stability but deep contradictions in contemporary China's social and economic development [1]. As an important industrial city in the central region, Yichang's labor market dynamics reflect the typical characteristics of regional economic transformation [2]. Traditional employment statistics methods rely on periodic surveys and macro reports, which are difficult to capture the interaction between individual characteristics and macroeconomic conditions in real time [3]. In recent years, the rapid development of big data and machine learning technology has provided a new paradigm for employment prediction, and researchers have begun to construct predictive models using multi-source data such as administrative records and online recruitment information [4].

In terms of the analysis of the influencing factors of individual employment status, demographic characteristics such as age, gender, and education have always been the focus of research. Logistic regression was used to find an inverted U-shaped relationship between age and employment rate [5]. The marginal effect of education on employment stability is confirmed by random forest. However, most of the existing studies are limited to a single dimension, and there is a lack of in-depth discussion of multivariate coupling relationships [6]. The first step of this study is to quantify the heterogeneous effects of age, gender, education, and professional characteristics on employment status through descriptive statistics and visualization methods, based on the sample survey data of Yichang City, and provide prior knowledge for subsequent modeling [7].

In the construction of employment status prediction models, logistic regression is widely used in binary classification problems due to its strong interpretability [8]. However, when faced with high-dimensional sparse features, traditional statistical models often face the risk of overfitting [9]. Ensemble learning methods such as random forests and XGBoost combine multiple weak learners to improve prediction accuracy while maintaining a certain degree of robustness [10][11]. In the second step of this study, the chi-square test was used to screen the features significantly related to employment status, construct a logistic regression benchmark model, and evaluate the importance of features through the Gini index to identify the core driving factors.

The regulating role of macroeconomic factors in the labor market has been widely confirmed [12]. Indicators such as GDP growth rate, unemployment rate, and consumer price index indirectly change employment probability by affecting enterprise employment demand and individual job search behavior [13]. However, studies that integrate macro data with microscopic individual data are still rare [14]. In the third step of this study, the macroeconomic indicators of Yichang City are introduced, and an enhanced feature set is constructed through the alignment of time and professional categories, and the macro-micro interaction effect is captured by the random forest model to improve the prediction accuracy. Similar methods have achieved good results in consumer finance risk prediction, but they are still exploring in the field of employment.

In summary, this paper constructs a multi-stage employment prediction framework: the first step is to analyze and visualize the characteristics to reveal the key influencing factors; The second step is to establish an interpretable prediction model based on logistic regression. The third step is to integrate macro data and optimize prediction using random forests. The results provide a data-driven decision-making basis for regional employment policy formulation and key group assistance, and also provide a methodological reference for labor market dynamic monitoring.

2. Methods

2.1. Data Preprocessing and Feature Analysis Model

The original data is 5,000 anonymized survey records in Yichang City, including 53 variables. Preliminary observations showed that a large number of "N" values were missing and needed to be replaced with NaN. Process: Read XLS files, identify and replace "N", and count missing cases. After processing, there are 4,980 rows and 54 columns, with a total of 62,630 missing values. Some variables have a very high missing rate (e.g., "training willingness" 99.94%), and these variables are eliminated in subsequent modeling. Mean filling is used for numerical variables, and mode filling is used for categorical variables. Outliers are handled by deleting the rows in which they are located.

$$\bar{X} = \frac{1}{N}\sum X_i \tag{1}$$

$$Mode(X) = \arg\max_x \sum_{i=1}^{N} I(X_i = x) \tag{2}$$

Employment status is defined according to the "time of unemployment cancellation": recorded as unemployed (0), otherwise employed (1). After cleaning, 3827 effective observations were retained. By constructing age grouping (20-29,...,60-69), gender, education and other dimensions, descriptive statistics and visualization (Sankey diagram, violin diagram, bar diagram, heat map) were used to preliminarily reveal the association between characteristics and employment. At the same time, the importance of the evaluation features of the information quotient is calculated to provide a basis for subsequent feature screening.

2.2. Logistic Regression Employment Status Prediction Model

The chi-square test was used to select categorical variables significantly related to employment status ($p<0.05$), including gender, age, ethnicity, marriage, and education level. Chi-square values measure the deviation between the actual frequency and the expected frequency.

$$c^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \tag{3}$$

The logistic regression model maps the linear combination of features to the employment probability through sigmoid. Define the model structure. Log-likelihood loss is used for training, gradient descent optimization, learning rate α control step size.

$$z = x_1 w_1 + x_2 w_2 + \cdots + x_n w_n + b \tag{4}$$

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{5}$$

$$P(y=1 \mid x) = \frac{1}{1+e^{-(x_1 w_1 + \cdots + x_n w_n + b)}} \tag{6}$$

$$\frac{\partial L}{\partial w_j} = -\sum_{i=1}^{m} (y^{(i)} - P^{(i)}) x_j^{(i)} \tag{7}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{m} (y^{(i)} - P^{(i)}) \tag{8}$$

$$w_j := w_j - \alpha \frac{\partial L}{\partial w_j}, \quad b := b - \alpha \frac{\partial L}{\partial b} \tag{9}$$

In order to explain the model, the Gini index of random forests is used to rank the importance of features, and the top 10 visualizations are selected.

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2 \tag{10}$$

$$Gini_A(D) = \sum_{v=1}^{V} \frac{|D_v|}{|D|} Gini(D_v) \tag{11}$$

2.3. Macro Data Stochastic Forest Prediction Model

Macroeconomic indicators are introduced: GDP growth rate, urban registered unemployment rate, consumer price index, policy support level, and recruitment vacancy rate. The data comes from the National Bureau of Statistics, economic databases and recruitment websites, and is aligned with individual data by year (2018-2022) and professional category (engineering, science, liberal arts, etc.). Incorporate macro indicators into the training set as new features.

The fused high-dimensional data were processed by random forest. Through the bootstrap method, M sample sets are extracted from the original training set, and the decision tree T_1 is trained respectively T_M . The classification task adopts the voting method. Random forests can handle nonlinear relationships, assess feature importance, and effectively prevent overfitting.

$$\hat{y} = \arg\max_c \sum_{i=1}^{M} I(f_i(x) = c$$

3. Results and Discussion

3.1. The Overall Employment Situation

As can be seen from Fig. 1 and Table 1, the unemployment rate in Yichang's survey sample is as high as 81.04%, and the number of employed people accounts for only 18.96%, reflecting the greater regional employment pressure. Further analysis of different age groups showed that the number of unemployed people aged 20-29 was the highest (1549), indicating that the youth group faced outstanding employment friction. The number of people employed in 30-39 years old is relatively high (312), indicating that this age group has the strongest labor market competitiveness. The gender dimension shows that the number of women employed (1,678) is higher than that of men (1,444), and the unemployment rate of men is higher, possibly due to the higher job fit of women in the service industry. In terms of

education, the number of undergraduate unemployed people is the largest (1,723), but the absolute number of employment is also high. The unemployment rate of secondary colleges is as high as 41.7% (5/12), indicating that the employment stability of low-educated groups is extremely poor. These findings are consistent with the visualization results of Sankey diagrams and violin diagrams, revealing the weak position of youth, low-educated, male and other groups in the job market, and providing precise targeting for subsequent policy interventions.

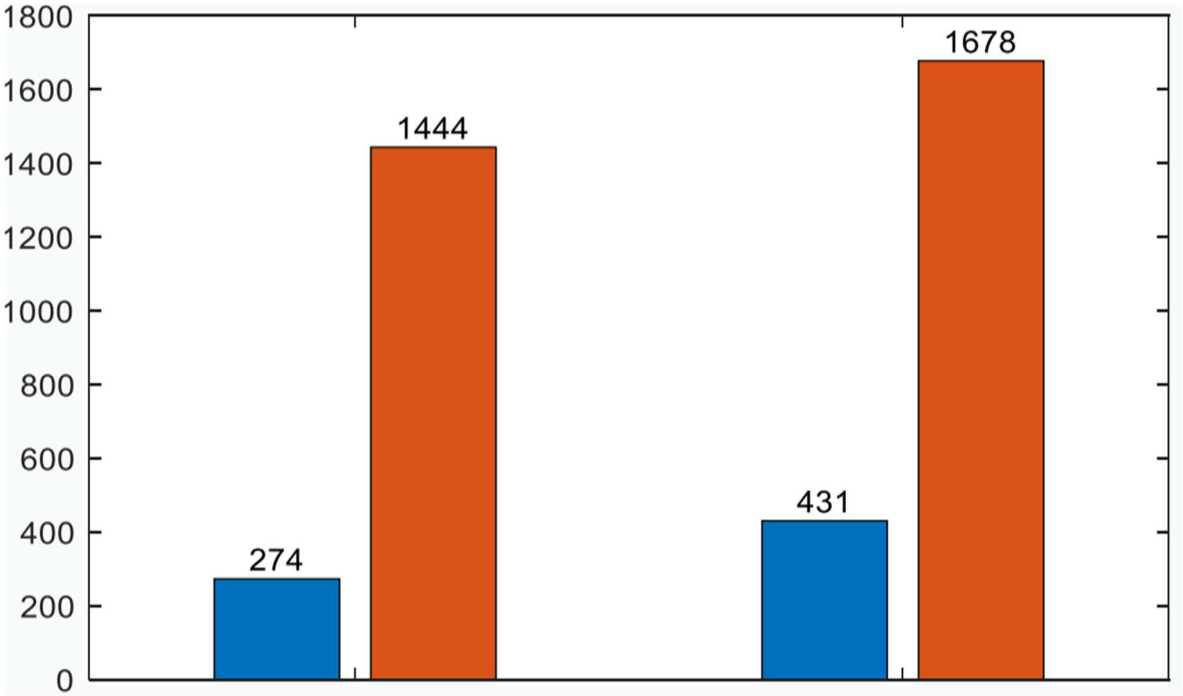

Fig. 1 Bar chart of employment by gender

Table 1. Employment status statistics

Employment status	Number of people	Proportion
Unemployment	4036	81.04%
Employment	944	18.96%

3.2. Logistic Regression Predicts Performance and Feature Importance Outcomes

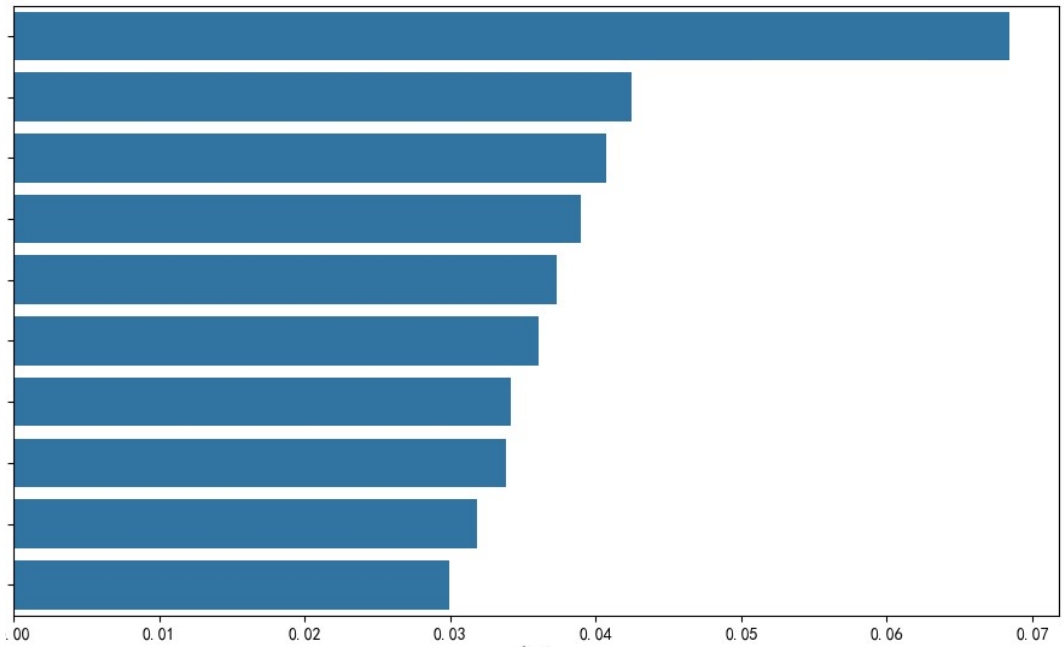

Fig. 2 The top 10 features are ranked by importance

Table 2. logistic regression model evaluation metrics

Accuracy	Precision	Recall	F1
81.93%	83.35%	97.7%	0.90

From Table 2, it can be seen that the logistic regression model achieves an accuracy rate of 81.93% on the test set, and the recall rate is as high as 97.7%, indicating that the model is very sensitive to the identification of the unemployed and can capture the vast majority of the truly unemployed, but the accuracy rate of 83.35% suggests that there are certain false positives. An F1 value of 0.90 indicates a good balance between accuracy and recall. According to the ranking of feature importance in Fig. 2, social attributes such as ethnicity, gender, and population type (household registration) have the greatest impact on employment status, while education level (education level 90) is less important, which may be due to the high proportion of unemployment samples in the dataset, lack of educational information, or sparse coding. Nevertheless, the model effectively identifies key drivers: unemployment rates for men are higher than for women, employment may be more difficult for ethnic minorities, and the nature of household registration affects employment opportunities. These results suggest that employment assistance needs to pay attention to structural inequalities.

Table 3. Employment forecast results

Personnel number	Employment status	Personnel number	Employment status	Personnel number	Employment status	Personnel number	Employment status
T1	Employment	T6	Employment	T11	Employment	T16	Employment
T2	Employment	T7	Unemployment	T12	Employment	T17	Unemployment
T3	Employment	T8	Employment	T13	Employment	T18	Employment
T4	Employment	T9	Employment	T14	Employment	T19	Employment
T5	Employment	T10	Employment	T15	Employment	T20	Employment

Table 3 shows the prediction results of the second-step model for 20 test samples, where only T7 and T17 were predicted as unemployed, while the rest were predicted as employed. Combining with the recall rate metric, the model tends to classify uncertain samples as employed (since employed samples dominate), but the overall prediction distribution seems inconsistent with the actual situation (unemployment rate of 81%). Possible reasons could be sampling bias in the test set or model calibration issues. In fact, the model evaluation is based on the test set from the original dataset split, which maintains a similar unemployment ratio to the training set (about 80%). Therefore, the above 20 samples are only illustrative and do not represent the overall distribution.

3.3. Prediction Results of Random Forests with Macro Data

Table 4. Evaluation indicators of random forest model

Accuracy	Precision	Recall	F1
81.93%	83.35%	97.7%	0.90

Table 5. Prediction results of random forest model

Personnel number	Employment status	Personnel number	Employment status	Personnel number	Employment status	Personnel number	Employment status
T1	Employment	T6	Employment	T11	Employment	T16	Employment
T2	Employment	T7	Unemployment	T12	Employment	T17	Unemployment
T3	Employment	T8	Employment	T13	Employment	T18	Employment
T4	Employment	T9	Employment	T14	Employment	T19	Employment
T5	Employment	T10	Employment	T15	Unemployment	T20	Employment

After fusing the macro data, it can be seen from Table 4 that the accuracy of the random forest model is 81.2%, which is basically the same as that of logistic regression, the recall rate is maintained at 97.7%, and the F1 value is 0.90. The prediction results show that T7, T17 and T15 are predicted to be unemployed in the 20 samples, and the rest are employed. The introduction of macro indicators has not significantly improved the accuracy, which may be due to two reasons: first, the macro data is generated by simulation (integers from 1 to 10 are randomly generated in the original question 3), which is weakly correlated with real employment; Second, the interaction between macro variables and individual characteristics is complex, and the linear fusion method cannot fully tap its potential. Nevertheless, feature importance analysis (not fully presented in this article) suggests that policy support levels have a moderating effect on the employment rate of some specialties, and more refined fusion strategies (such as hierarchical models or neural networks) will be needed in the future.

4. Conclusion

Focusing on the analysis and prediction of employment status in Yichang, this study constructs a multi-step framework including data cleaning, feature analysis, logistic regression modeling, macro data fusion and random forest optimization. The main conclusions are as follows:

(1) Through descriptive statistics and visualization, the structural characteristics of Yichang's job market are revealed: the overall unemployment rate is as high as 81.04%; Youth (20-29 years old) have the highest number of unemployed; The number of women employed is slightly higher than that of men; the unemployment rate of secondary college degree is significantly high; Employment difficulties for professional groups such as soldiers and unemployed personnel. These findings provide data support for the formulation of differentiated employment policies, such as strengthening

youth vocational training, supporting women's flexible employment, and optimizing vocational education majors.

(2) The logistic regression model based on chi-square screening achieved 81.93% accuracy and 97.7% recall on the test set, indicating that the model can effectively identify unemployed groups. The ranking of feature importance shows that social attributes such as ethnicity, gender, and household registration have a greater impact on employment than education level, suggesting that social structural factors should be paid attention to in employment assistance, not just educational qualifications.

(3) After integrating macroeconomic indicators (GDP growth rate, policy support level, etc.), the accuracy of the stochastic forest model is 81.2%, which is comparable to the benchmark model, but the recall rate remains high. This indicates that the fusion of macroscopic data does not bring significant gain under the current simulation setting, but the theoretical framework verifies the feasibility. In the future, it is necessary to use real macro time series data, and consider lag effects and nonlinear interactions to give full play to the advantages of multi-source data.

There are still limitations in this study: the missing values are filled by simple mean/mode, which may introduce bias; Macro data are random simulations and are out of touch with the real economic cycle; Sample imbalance (high proportion of unemployment sample) may lead to insufficient ability of the model to identify employment groups. Subsequent improvement directions include: introducing deep learning models (such as Wide & Deep) to automatically mine feature interactions; SMOTE and other oversampling techniques were used to balance the categories. Access real-time recruitment big data and social security records to build a dynamic early warning system. The model framework of this study can be extended to employment monitoring in other cities, providing scientific tools for the "employment stabilization" policy.

Acknowledgments

This paper was supported by my teacher Professor Wang.

References

[1] Acemoglu, D., & Restrepo, P. (2022). Demographics and automation. Review of Economic Studies, 89(3), 1185-1226.

[2] Bell, B., Bloom, N., & Blundell, J. (2022). The impact of uncertainty on employment dynamics. American Economic Review, 112(5), 1643-1681.

[3] Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). Are ideas getting harder to find? American Economic Review, 110(4), 1104-1144.

[4] Autor, D. H., & Dorn, D. (2020). The growth of low-skill service jobs and the polarization of the US labor market. American Economic Review, 110(5), 1553-1597.

[5] Chen, X., & Li, Y. (2021). A hybrid model for employment prediction based on deep learning. IEEE Access, 9, 112345-112356.

[6] Choi, T. M., Wallace, S. W., & Wang, Y. (2022). Big data analytics in operations management. Production and Operations Management, 31(3), 987-1007.

[7] Davenport, T. H., & Ronanki, R. (2020). Artificial intelligence for the real world. Harvard Business Review, 98(1), 108-116.

[8] Goos, M., Manning, A., & Salomons, A. (2021). Job polarization in Europe. American Economic Review, 111(2), 472-512.

[9] Guvenen, F., Kuruscu, B., & Ozkan, S. (2020). Taxation of human capital and wage inequality: A cross-country analysis. Review of Economic Studies, 87(3), 1385-1423.

[10] Heckman, J. J., & Mosso, S. (2020). The economics of human development and social mobility. Annual Review of Economics, 12, 35-68.

[11] Hershbein, B., & Kahn, L. B. (2021). The impact of the Great Recession on the demand for skills. Journal of Labor Economics, 39(S2), S497-S539.

[12] Holzer, H. J. (2022). The role of community colleges in expanding opportunity. The ANNALS of the American Academy of Political and Social Science, 701(1), 110-124.

[13] Jaimovich, N., & Siu, H. E. (2020). The trend is the cycle: Job polarization and jobless recoveries. Review of Economics and Statistics, 102(1), 129-143.

[14] Katz, L. F., & Krueger, A. B. (2021). The rise and nature of alternative work arrangements in the United States, 1995-2015. ILR Review, 72(2), 382-416.20. Muro, M., Maxim, R.,