Information Systems and Economics (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2026.070112
ISSN 2523-6407 Vol. 7 Num. 1

An Interpretable Machine-Learning Framework for Employment Status Prediction and Job Matching

Yuxin Zhu[1, a, *]

[1] *School of Foreign Studies, China University of Political Science and Law, Beijing, China*
[a] *shuangxinxin2006@gmail.com*
**Corresponding author*

Keywords: Employment status prediction; binary logistic regression; support vector machine; random forest; imbalanced classification; job matching

Abstract: Accurate employment-status prediction is essential for targeted labor-market services, especially when individual socio-demographic attributes, household conditions, and industrial restructuring jointly shape employment opportunities. Based on an anonymized survey from Yichang, China, this study develops a data-driven framework for employment diagnosis, prediction, and person-job matching. After data screening, 4,980 valid observations were retained. A binary logistic regression model was first used to identify interpretable determinants of employment status, including gender, age, disability status, education, household registration, residence pattern, family size, and industrial-structure indicators. Machine-learning classifiers were then compared under an imbalanced-label setting, including logistic regression, decision tree, random forest, Naive Bayes, and support vector machine. The results show that education and age are strong positive predictors, while larger household size and some structural-transition factors are associated with lower employment probability. After resampling and parameter tuning, SVM and random forest achieved the best predictive performance, with accuracy of 0.93, recall of 0.97, and F1-score of 0.95. Finally, a resume-job matching module is proposed by combining resume parsing, skill ontology, and weighted similarity ranking. The study contributes an integrated and interpretable workflow for regional employment monitoring, risk identification, and personalized job recommendation.

1. Introduction

Employment is a central outcome of regional economic development and social governance. In local labor markets, employment status is not determined by a single factor; it is shaped by individual human capital, household constraints, mobility conditions, industrial demand, and the availability of suitable vacancies. The spread of digital labor-market services also creates new opportunities to transform survey records and administrative information into decision-support tools. However, the value of such tools depends on whether they can combine predictive accuracy with interpretable evidence for policy and service design. This requirement is particularly important for

regional employment governance, where decision makers need to identify vulnerable groups, explain why risks emerge, and connect people with suitable job opportunities.

The literature provides two relevant streams for this problem. The first stream focuses on employment, employability, and labor-market transformation. Studies of occupational change and employability emphasize that education, skills, mobility, and institutional conditions jointly affect employment outcomes [1]. Applied econometric and data-science research further shows that large-scale individual records can reveal heterogeneous labor-market patterns that are difficult to capture through descriptive statistics alone [2, 3]. The second stream concerns prediction methods. Logistic regression remains valuable because its coefficients can be interpreted through odds ratios and therefore translated into policy-relevant explanations [4]. Machine-learning models, including random forest and support vector machine, can capture nonlinear boundaries and feature interactions in complex tabular data [5, 6].

A further methodological issue is class imbalance. In employment-status prediction, the group requiring intervention may be smaller than the employed majority. Under such conditions, accuracy alone can overstate model quality because a model may perform well on the majority class while failing to identify vulnerable cases. Resampling methods and precision-recall-oriented evaluation are therefore needed to provide a more balanced assessment [7-9]. Beyond prediction, employment services also require matching technologies. Information retrieval, entity matching, semantic resources, natural-language processing, and recommender-system methods provide useful foundations for comparing resumes with job postings and ranking suitable vacancies [10-15].

Although these studies provide important foundations, three gaps remain in the context of the original Yichang employment dataset. First, many prediction studies emphasize classification performance but provide limited interpretation of socio-demographic mechanisms. Second, explanatory statistical models often identify influential factors but stop short of operational prediction and matching. Third, job matching is frequently treated as a separate recommendation problem, while regional employment analysis needs a continuous workflow from diagnosis to prediction and service delivery. These gaps motivate a framework that combines interpretable regression, machine-learning prediction, multisource optimization, and resume-job similarity ranking.

This study addresses four related tasks: descriptive employment profiling, interpretable binary logistic regression, machine-learning-based employment prediction, and resume-job matching. These tasks are integrated into a unified research framework to answer the following question: how can survey-based employment data be transformed into an interpretable and operational framework for employment-status prediction and personalized job recommendation?

The contributions are fourfold. First, the study summarizes a cleaned sample of 4,980 respondents and identifies key demographic, household, and industrial factors associated with employment status. Second, it uses binary logistic regression to provide interpretable evidence on how gender, age, education, household registration, residence pattern, family size, and industrial-structure variables are associated with employment odds. Third, it compares five classification models under class imbalance and shows that SVM and random forest provide the most stable predictive results after tuning. Fourth, it extends the prediction task to a practical matching framework in which resume attributes, job-posting attributes, and skill ontology are combined through a weighted similarity measure.

2. Data and Research Design

This study uses anonymized survey data collected in the Yichang region. The original dataset contained approximately 5,000 respondents. After removing abnormal records and observations

with unusable values, 4,980 valid samples were retained. The dependent variable is employment status, coded as 1 for employed respondents and 0 for unemployed or non-employed respondents. Explanatory variables include gender, age, disability status, educational attainment, cultural-education score, household-registration type, residence pattern, family size, population type, and two industrial-structure indicators.

Because the paper combines social-science interpretation with predictive modeling, the research design follows a two-stage logic. The first stage uses binary logistic regression to estimate the direction and relative magnitude of employment determinants. The second stage evaluates supervised classifiers for employment-status prediction. Class imbalance is treated through resampling, and model quality is evaluated by accuracy, precision, recall, and F1-score rather than by accuracy alone.

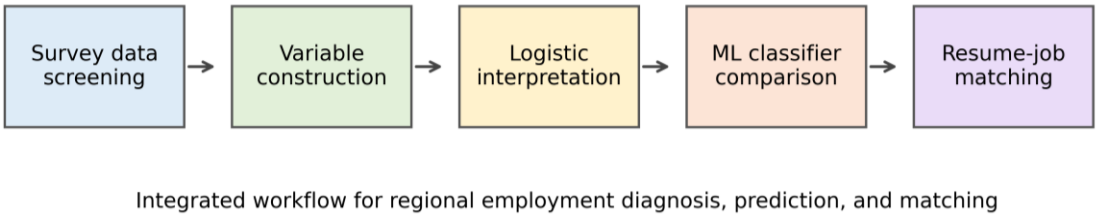

Figure 1: Research Workflow for Employment Prediction and Job Matching

The complete analytical route is shown in Figure 1. The workflow begins with survey-data screening and variable construction, proceeds to interpretable regression and machine-learning comparison, and ends with a matching-oriented application module. To provide a clearer description of the cleaned dataset used in this workflow, Table 1 presents the profile of the valid employment survey sample, including demographic characteristics, educational background, age distribution, ethnicity, military status, and residence status.

Table 1: Profile of the Valid Employment Survey Sample

Dimension	Observed distribution	Analytical implication
Valid sample	4,980 respondents	Cleaned from about 5,000 original records
Gender	Male 44.70%; female 55.30%	Female respondents slightly outnumbered male respondents
Marital status	Unmarried 48.84%; married 50.00%; widowed 0.08%; divorced 1.08%	The sample is concentrated in unmarried and married groups
Education	Graduate 2.31%; bachelor 55.82%; junior college 41.87%	Higher-education groups dominate the sample
Age	26-30: 51.30%; 31-35: 22.46%; 36-40: 11.10%	Young and early-career adults are the main respondents
Ethnicity	Han 90.02%; other ethnic groups 9.98%	Han respondents form the majority
Military status	No military service 96.34%; retired soldiers 1.00%; active service 0.04%; other 2.22%	Most respondents have no military-service record
Residence status	Permanent resident 91.56%; out-migrant 8.44%	Most respondents were permanent residents, indicating that the sample mainly reflects the employment conditions of locally resident individuals.

Figure 2 visualizes the most important sample-distribution characteristics. The sample contains slightly more female respondents than male respondents, is dominated by bachelor and junior-college groups, and is concentrated in the 26-30 and 31-35 age ranges.

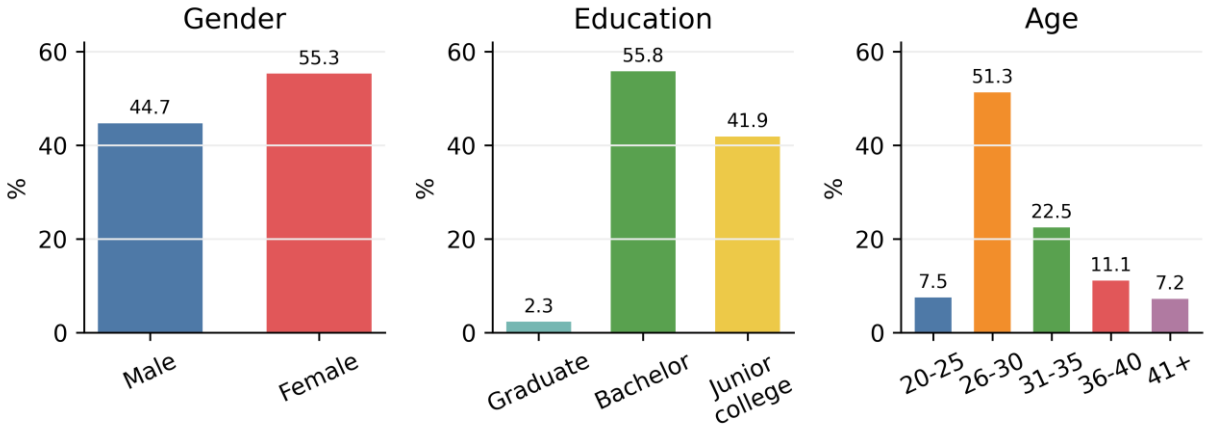

Figure 2: Demographic and Education Profile of the Valid Sample

3. Methodology

3.1. Interpretable Logistic Regression

Binary logistic regression estimates the probability that respondent i is employed. Let y_i be the employment label, and X_i be the feature vector. The model is specified as follows:

$$\Pr(y_i = 1 \mid X_i) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}\right)\right]} \tag{1}$$

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} \tag{2}$$

The odds ratio exp(beta) is used to interpret the relative change in employment odds associated with a one-unit change or category shift in a predictor. The logistic loss minimized during training is:

$$L(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i)\right] \tag{3}$$

The regression stage is included not only as a baseline classifier but also as an explanatory model. It helps identify which socio-demographic and household variables should be treated as priority factors in employment support.

3.2. Machine-Learning Classifiers

Five classifiers are evaluated: logistic regression, decision tree, random forest, Naive Bayes, and support vector machine. Random forest aggregates multiple bootstrapped decision trees and predicts the class by majority voting. SVM searches for a separating hyperplane that maximizes the margin between classes, which is useful when employment status is influenced by multiple interacting attributes. Decision tree and Naive Bayes are retained as transparent and computationally efficient baselines.

$$\hat{y}_{RF}(x) = \text{majority\_vote}\left\{T_1(x), T_2(x), \ldots, T_B(x)\right\} \tag{4}$$

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{5}$$

The initial target distribution is imbalanced: the non-employed class is much smaller than the employed class. Therefore, resampling is used before model comparison. This treatment reduces majority-class bias and makes recall and F1-score more informative.

The initial target distribution is imbalanced: the non-employed/other class contains fewer records than the employed class. Therefore, resampling is used before model comparison to obtain a more balanced training distribution. As shown in Figure 3, the class distribution changes from 60 non-employed/other records and 125 employed records before resampling to 100 records in each class after resampling. This treatment reduces majority-class bias and makes recall and F1-score more informative.

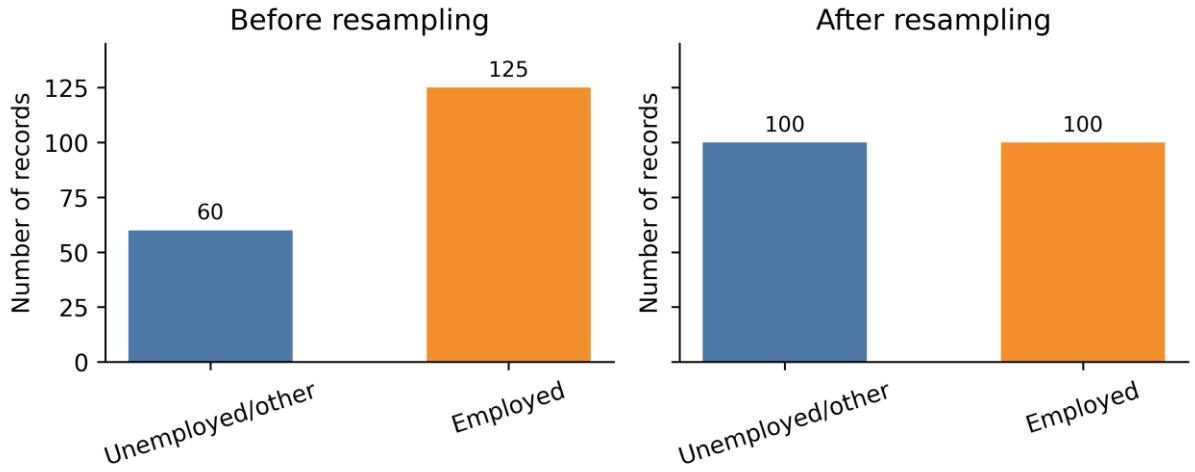

Figure 3: Target-Class Distribution Before and After Resampling

3.3. Multisource Optimization and Matching Framework

In addition to individual-level survey variables, macroeconomic and labor-market information, including industrial rationalization and industrial upgrading, is incorporated to improve regional interpretation. The rationalization index reflects the allocation consistency between output and employment across industries, while the upgrading index captures the shift toward higher-value and innovation-oriented sectors.

$$DAI = w_1 RRI' + w_2 HUI' + w_3 ESI \tag{6}$$

where DAI denotes the dynamic adaptation index, *RRI'* and *HUI'* are normalized industrial-structure indicators, *ESI* is an eco-industrial synergy indicator, and w_1, w_2, and w_3 are weights. These contextual variables support regional interpretation rather than replacing individual-level predictors.

For practical job placement, the study proposes a resume-job matching module. Candidate resumes and job postings are parsed into structured fields, including job title, major, degree, professional experience, and technical skills. A domain skill ontology is then used to expand exact matches into semantic matches. The final similarity score is a weighted sum of field-level similarities:

$$\mathrm{sim}(x_j, x_r) = \sum_{i=1}^{m} \beta_i \alpha_i (x_j, x_r) \tag{7}$$

where x_j is a job object, x_r is a resume object, α_i is the similarity on feature *i*, and β is the feature weight. Exact major or skill matches receive the highest score; semantically related hypernym or hyponym matches receive partial credit; unrelated terms receive zero.

Figure 4 shows the proposed resume-job matching framework, which integrates resume parsing, job-posting features, skill ontology, and weighted similarity ranking.

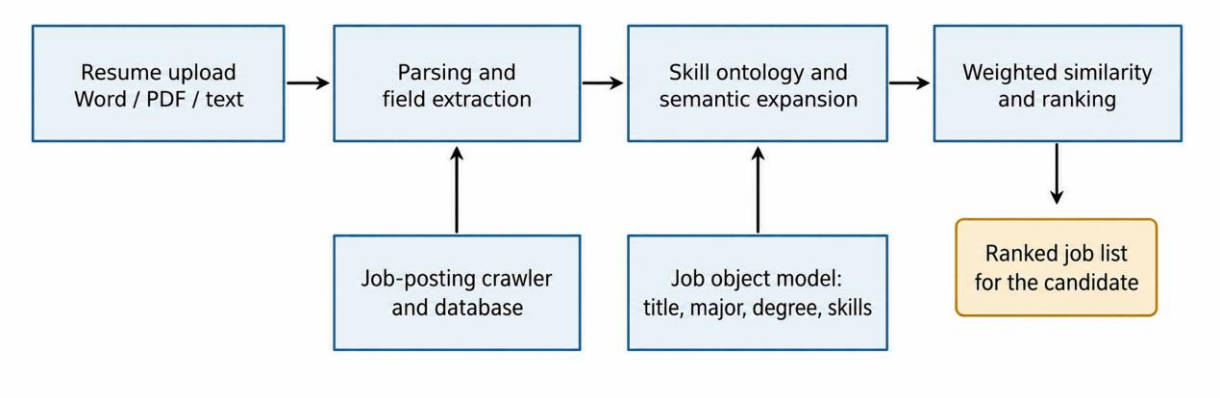

Figure 4: Resume-Job Matching Framework Based on Parsing, Ontology, and Similarity Ranking

4. Empirical Results

4.1. Logistic Regression Findings

Table 2 summarizes the core odds-ratio findings retained from the original regression output. The results indicate that gender, age, disability status, education, hukou type, family size, and industrial structure are strongly associated with employment status. Female respondents show lower employment odds than male respondents. Age groups between 21 and 35 have substantially higher odds than the 16-20 reference group, suggesting that early labor-market experience improves employability. Education is one of the strongest predictors: bachelor and graduate groups have much higher employment odds than the lower-education reference group.

Table 2: Core Logistic Regression Results for Employment Status

Predictor	Odds-ratio range	Direction	Interpretation
Female	0.349-0.361	Negative	Employment odds are lower than for male respondents
Age 21-25	2.091-2.294	Positive	Higher odds than the 16-20 reference group
Age 26-30	2.283-2.650	Positive	Early-career respondents show stronger employment probability
Age 31-35	3.010-3.568	Positive	The odds increase further with labor-market experience
Disabled respondent	2.172-2.380	Positive	May reflect targeted support policies and sample structure
Cultural-education score	1.973-2.012	Positive	Higher human capital improves employment odds
Junior college	3.182-3.845	Positive	Higher than the lower-education reference group
Bachelor degree	4.755-6.687	Positive	Strong education premium
Graduate education	5.354-7.400	Positive	Highest observed education-related odds
Non-agricultural hukou	1.223-2.067	Positive	Urban registration improves access to jobs
Three-person household	0.343-0.421	Negative	Larger household burden may restrict job mobility
Industrial rationalization	0.199-0.257	Negative	Structural adjustment may reduce low-skill positions
Industrial upgrading	1.035-1.051	Positive	Upgrading creates demand for higher-skilled labor

Figure 5 further displays selected odds ratios from the logistic model. The horizontal reference line at 1 separates positive and negative associations. Education-related variables and prime working-age groups show the strongest positive effects, whereas female gender, larger household size, and industrial rationalization show lower employment odds.

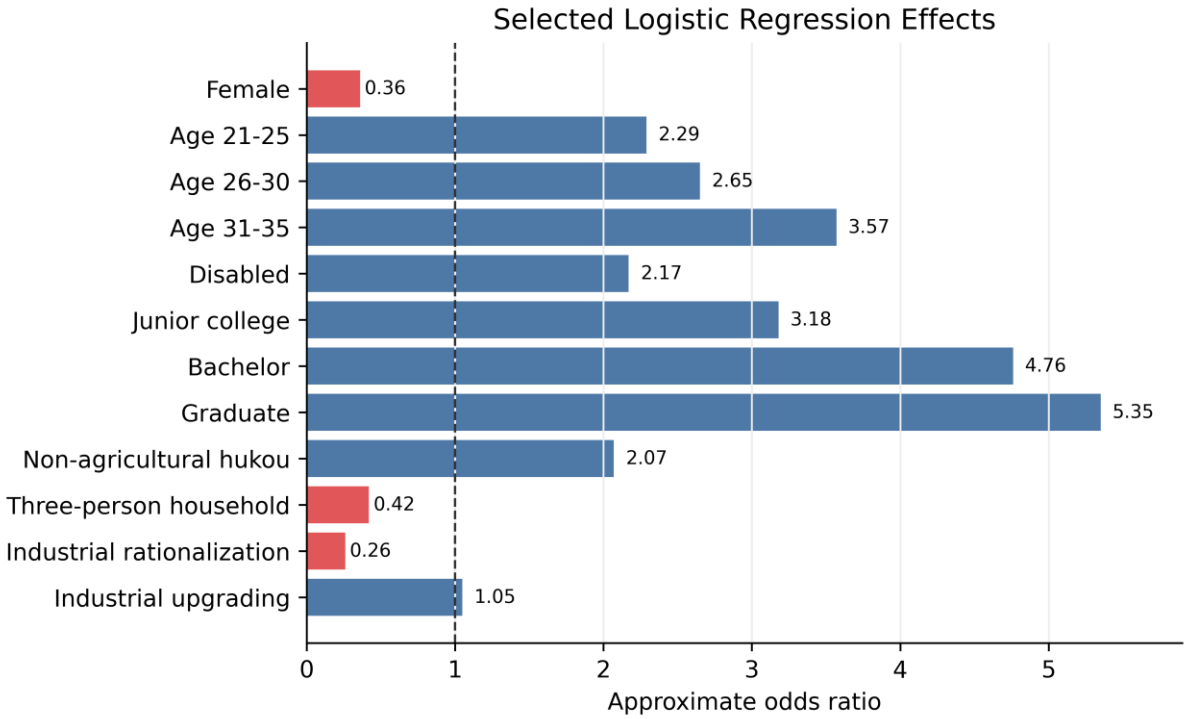

Figure 5: Selected Logistic Regression Odds Ratios

The negative coefficient for industrial rationalization requires careful interpretation. A more rational industrial structure may remove redundant or low-productivity jobs in the short run, thereby lowering employment odds for some groups. By contrast, industrial upgrading is positively associated with employment, which is consistent with the creation of higher-quality positions in technology-intensive and service-oriented sectors.

4.2. Classifier Performance

Table 3 reports the tuned performance of the five classifiers. SVM and random forest perform best, both reaching 0.93 accuracy, 0.97 recall, and 0.95 F1-score. Logistic regression and Naive Bayes also remain competitive, but their recall and F1-score are lower than those of SVM and random forest. The decision tree is the weakest model after tuning, which may reflect instability caused by high-dimensional categorical variables and class imbalance.

Table 3: Tuned Classification Performance

Classifier	Accuracy	Precision	Recall	F1-score
Logistic regression	0.91	0.93	0.93	0.93
Decision tree	0.84	0.87	0.90	0.88
Random forest	0.93	0.93	0.97	0.95
Naive Bayes	0.91	0.93	0.93	0.93
Support vector machine	0.93	0.93	0.97	0.95

Figure 6 visually compares the tuned performance of the five classifiers in terms of accuracy, precision, recall, and F1-score.

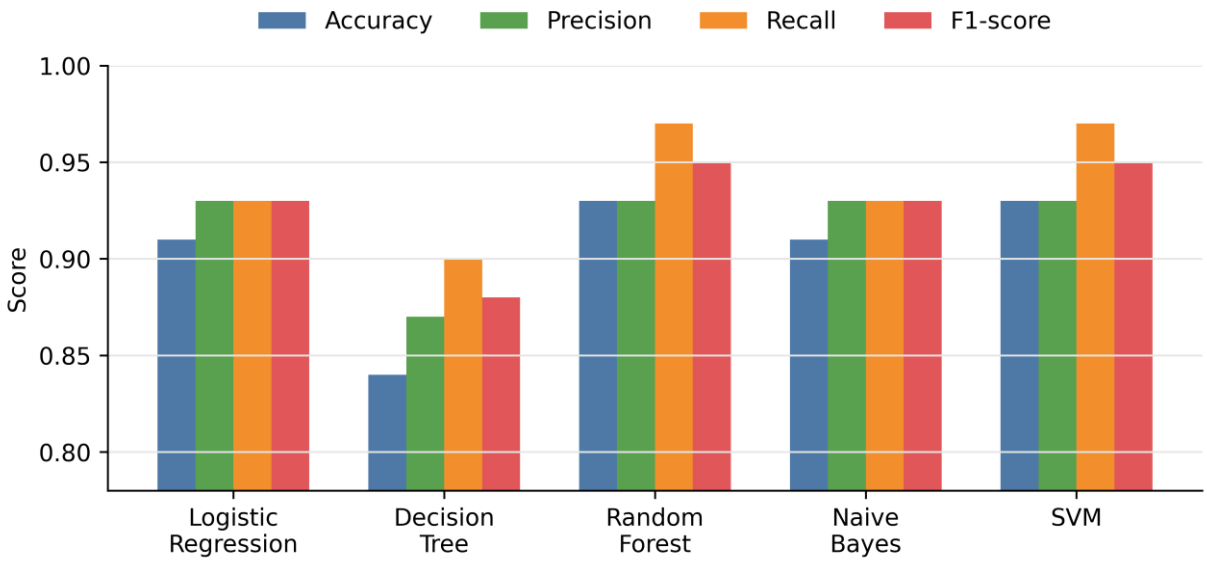

Figure 6: Comparison of Tuned Classifier Performance

The high recall of SVM and random forest is particularly important for employment-risk monitoring. In this context, failing to identify non-employed or vulnerable individuals may lead to missed service opportunities. A model with high recall and F1-score is therefore more useful than a model that only maximizes overall accuracy.

4.3. Multisource Consistency Results

A multisource consistency experiment was further conducted to compare logistic regression, decision tree, and random forest when survey labels were linked with administrative records. The confusion-matrix proportions show that tree-based methods improve class recognition after balancing. Random forest produced the strongest class-level recognition, with approximately 0.977 correct recognition for class 0 and 0.946 for class 1 in one reported configuration, while decision tree performed between logistic regression and random forest. Table 4 presents the multisource consistency experiment results extracted from the original model comparison.

Table 4: Multisource Consistency Experiment Extracted from the Original Results

Model	Correct class-0 proportion	Correct class-1 proportion	Implication
Logistic regression	0.925	0.868	Transparent baseline but weaker nonlinear adaptation
Decision tree	0.958	0.920	More flexible and interpretable than linear regression
Random forest	0.977	0.946	Best reported consistency under the multisource setting

5. Discussion

The empirical results support three substantive interpretations. First, human-capital variables dominate the explanation of employment status. Education and age capture accumulated knowledge, labor-market readiness, and job-search adaptability. Second, household and registration variables reveal structural constraints that are not purely individual. Hukou type and family size influence employment access through mobility, social-network resources, and care obligations. Third,

industrial transformation has asymmetric effects: upgrading increases opportunities in emerging sectors, whereas rationalization may temporarily compress employment in lower-productivity sectors.

From a modeling perspective, the comparison demonstrates the value of using interpretable and predictive models together. Logistic regression is suitable for explaining factor effects, but SVM and random forest provide stronger predictive performance when nonlinear boundaries or feature interactions exist. This combined design prevents the paper from becoming either a purely explanatory regression study or a purely black-box prediction exercise.

The proposed matching framework translates prediction into an operational service scenario. Instead of only estimating whether a person is likely to be employed, the system can recommend suitable jobs by comparing structured resume attributes with job-posting attributes. Ontology-based skill expansion is useful because labor-market terms are rarely identical across resumes and vacancies; for example, one job advertisement may use a broad term while a candidate lists a narrower technical skill. Weighted semantic similarity can reduce this vocabulary mismatch.

6. Conclusion

This study developed an integrated and interpretable framework for employment-status prediction and job matching based on regional survey data from Yichang, China. After data screening, 4,980 valid observations were used to examine the relationship between employment status and individual characteristics, household conditions, residence-related factors, and industrial-structure indicators. Binary logistic regression was first applied to identify key determinants of employment status, while five machine-learning classifiers were further compared to improve predictive performance under an imbalanced-label setting.

The empirical results show that education, age, household-registration type, household size, and industrial-structure indicators are closely associated with employment outcomes. Higher education levels and prime working-age groups are linked to higher employment odds, suggesting the important role of human capital and labor-market readiness. In contrast, larger household size and some structural-adjustment factors may reduce employment probability, indicating that employment risk is shaped not only by individual ability but also by household constraints and regional industrial transformation. Among the compared classifiers, support vector machine and random forest achieved the strongest overall performance after resampling and parameter tuning, both reaching an accuracy of 0.93 and an F1-score of 0.95.

Beyond employment-status prediction, this study also proposed a resume-job matching module that combines resume parsing, job-posting features, skill ontology, and weighted similarity ranking. This design extends the analytical framework from employment-risk identification to practical job recommendation, providing a potential decision-support tool for regional employment services. The proposed framework can help local agencies identify vulnerable groups, evaluate the employment effects of industrial restructuring, and improve the precision of job-placement services.

Several limitations should be acknowledged. First, the dataset is limited to one regional labor market, so the findings may not be directly generalized to other cities or provinces. Second, the cross-sectional nature of the data limits causal interpretation. Third, some important variables, such as detailed work experience, occupational category, contract type, salary expectation, and real job-search behavior, were not fully available. Future research should incorporate longitudinal data, broader regional samples, richer administrative records, and real recruitment-platform data to further validate and improve the proposed employment prediction and job-matching framework.

References

[1] Frey C B, Osborne M A. The future of employment: How susceptible are jobs to computerisation[J]. Technological Forecasting and Social Change, 2017, 114: 254-280.
[2] Mullainathan S, Spiess J. Machine learning: An applied econometric approach[J]. Journal of Economic Perspectives, 2017, 31(2): 87-106.
[3] Varian H R. Big data: New tricks for econometrics[J]. Journal of Economic Perspectives, 2014, 28(2): 3-28.
[4] Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression[J]. Critical Care, 2005, 9(1): 112-118.
[5] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
[6] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.
[7] Chawla N V, Bowyer K W, Hall L O, Kegelmeyer W P. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.
[8] Fawcett T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861-874.
[9] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets[J]. PLoS ONE, 2015, 10(3): e0118432.
[10] Elmagarmid A K, Ipeirotis P G, Verykios V S. Duplicate record detection: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(1): 1-16.
[11] Singhal A. Modern information retrieval: A brief overview[J]. IEEE Data Engineering Bulletin, 2001, 24(4): 35-43.
[12] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia: A crystallization point for the Web of Data[J]. Journal of Web Semantics, 2009, 7(3): 154-165.
[13] Hirschberg J, Manning C D. Advances in natural language processing[J]. Science, 2015, 349(6245): 261-266.
[14] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734-749.
[15] McQuaid R W, Lindsay C. The concept of employability[J]. Urban Studies, 2005, 42(2): 197-219.