Research on the Construction and Prediction of a Model for Evaluating the Employment Ability of College Students Based on Big Data Analysis

Shang Sun[1], Di Yang[2], Juan Hu[3]

[1]School of Economics and Management, Anhui University of Science and Technology, Huainan 232001, Anhui, China
[2]School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan 232001, Anhui, China
[3]Huainan Vocational and Technical College, Huainan 232001, Anhui, China

Corresponding author: Shang Sun (e-mail: sunshang2024@126.com)

ABSTRACT To address the limitations of current employability assessments, including static isolation and detachment from market demand, this paper constructs a big data-driven dynamic assessment and prediction model that aligns student competencies with the evolving technical requirements of intelligent manufacturing, electromagnetic engineering, antenna-related systems, and high-end machinery control. Specifically, on-campus academic behavior data and external recruitment market data are integrated to construct a multidimensional indicator system covering professional competence, practical ability, career potential, and personal traits. Multi-source heterogeneous data are quantified through feature engineering and natural language processing, and prediction models are constructed using ensemble learning algorithms, including Random Forest and XGBoost. The SHAP framework is applied to improve interpretability and identify key contributors to employability prediction. Experimental results show that the model achieves 91.2% accuracy in assessing employment competitiveness levels and reduces the RMSE of starting salary prediction to 2,800 yuan, corresponding to a mean absolute percentage error of 10.5%, outperforming traditional benchmark methods. The study verifies the effectiveness of multi-source data fusion and machine learning in employment assessment, and it provides decision support for aligning talent profiles with smart manufacturing and electromagnetic-technology-related industrial demands.

INDEX TERMS employability assessment, machine learning, multi-source data fusion, electromagnetic engineering, smart manufacturing

I. INTRODUCTION

With the acceleration of higher education and the profound transformation of the labor market—driven by the shift toward high-tech manufacturing and intelligent weaving systems—the employment of college students has become a critical issue for social stability and the optimal allocation of human resources within these modernized industrial sectors [1-2]. However, the current employment ability assessment methods commonly used by universities mainly rely on academic transcripts, resumes, and limited subjective interviews, which have prominent problems such as one-sided assessment dimensions, static and isolated data sources, and serious disconnect between evaluation results and rapidly evolving market demands. This model is difficult to comprehensively, dynamically, and forward-lookingly reflect students' true employment competitiveness, resulting in a lack of accurate data support for employment guidance and a delayed feedback loop in talent cultivation [3-4].

Regarding the assessment of college students' employment ability, scholars at home and abroad have conducted extensive research from multiple disciplines such as education, psychology, and management. Parrella et al. used a cross-sectional research design and distributed questionnaires to all agricultural students through customized research tools to assess the importance, development status, and related influencing factors and experiences of their employment skills [5]. Carter et al., based on the employment ability self-assessment matrix from the UK Biosciences Centre, studied the degree of embedding of employment ability skills in higher education courses in the UAE through self-administered questionnaires filled out by course managers [6]. Zhang et al. used quantitative methods to study the relationship between basic skills and employability of Pakistani students and examined the moderating effect of "university reputation" on the two [7]. Li explored the role of mental health education in colleges and universities in improving students' employability, analyzed its importance in the educational environment and how it promotes students' employment skills and overall career preparation [8]. However, most existing studies are still limited to single data sources or cross-sectional static analysis, lacking systematic integration of multi-source heterogeneous big data, and failing to effectively construct dynamic prediction models from historical data to future employment outcomes, resulting in insufficient comprehensiveness and predictability of the assessment.

To overcome the above limitations, data science and machine learning methods have provided new technical paths for employability assessment research in recent years. With the continuous growth of the number of graduates and the intensification of competition for high-quality positions, employability has become a key indicator for measuring the performance of colleges and universities. To this end, Muhammad et al. used machine learning models to predict students' employability before graduation [9]. In order to improve the employment rate of colleges and universities, Li et al. used machine learning algorithms to construct an employment early warning model based on the academic and non-academic factors of undergraduate physics majors in N colleges and universities in a certain region, which effectively realized the prediction and guidance of graduate employment [10]. However, these methods often fall short in terms of feature engineering depth, model complexity, and interpretability, failing to fully explore the deep connections between academic behavior, practical records, and market dynamics. Therefore, this paper aims to introduce a more systematic technical solution: by integrating multi-source big data and comprehensively utilizing advanced machine learning techniques such as natural language processing, feature engineering, and ensemble learning, a model can be constructed that integrates multi-dimensional information and achieves dynamic and accurate predictions, thereby overcoming the shortcomings of traditional methods in terms of data breadth, analytical depth, and timeliness.

The main purpose of this study is to build a dynamic assessment and prediction model of the employability of college students based on big data analysis. Method: Firstly, off-campus recruitment market data and on-campus academic behavior data are collected and integrated; secondly, a quantitative assessment index system is constructed based on the feature engineering and text mining method; then classification and regression prediction model is constructed based on ensemble learning algorithm, random forest and XGBoost; finally, through model fusion and interpretability analysis, an assessment system with high accuracy and operability is formed. Result: Experimental results show that the model has obvious advantages in assessment of employment competitiveness and prediction of key employment destination. The model provides effective empirical evidence and decision-making tools for precise employment guidance, enabling universities to refine talent training models that meet the high-precision technical standards and intelligent manufacturing requirements of the modern industry.

II. MULTI-SOURCE HETEROGENEOUS DATA ACQUISITION AND FUSION

For numerical structured data, we calculated each student's core course average GPA, grade percentile rank, and interdisciplinary course breadth index based on course selection. These cumulative indicators were computed using only courses completed prior to the prediction time point (e.g., for senior-year prediction, only first-three-year grades were used). For internship records, we designed weighted scoring rules based on institution size, industry relevance, and duration, generating an "Internship Experience Quality Score". Campus behavior data were processed through time-window analysis to extract features such as "percentage of regular study time", "frequency of borrowing non-professional books", "repeated viewing rate of course videos", and "on-time assignment submission rate". These derived features serve as quantitative proxies for learning habits, self-regulation, and intellectual engagement.

For unstructured texts (internship reports, project descriptions, job postings), we first constructed an industry-specific skill lexicon based on China's National Vocational Qualification Standards, supplemented by high-frequency terms extracted from over 450,000 job descriptions from Zhaopin.com and 51job.com using TF-IDF. The final lexicon was manually curated to cover 12 industry sectors with 3,200 skill keywords. Key technical terms were weighted using TF-IDF, and semantic similarity was quantified using a Chinese pre-trained word2vec model (trained on Chinese Wikipedia and general domain text; 300-dimension) to map skill terms into vectors, with cosine similarity yielding the "market skills matching score". All text processing used Python's jieba and gensim.

All features were normalized prior to model training: min-max scaling to [0,1] was applied to bounded or outlier-prone features (e.g., internship duration, competition awards, library visit frequency), while Z-score standardization was used for approximately normally distributed features (e.g., core course GPA percentile, interdisciplinary breadth index). These features were then organized into a three-level evaluation index system—professional competence, practical ability, career potential, and personal traits as first-level dimensions—which serves an organizational rather than prescriptive weighting purpose. The actual contribution of each feature, whether from structured GPA data or unstructured text, is automatically learned by the subsequent ensemble models (Random Forest, XGBoost), eliminating any manual weighting logic and ensuring transparent, data-driven fusion across heterogeneous data sources.

III. TRAINING OF ENSEMBLE LEARNING PREDICTIVE MODELS

According to different prediction targets, we built models for classification prediction and regression prediction respectively.

Classification Model:

Predicting students employment competitiveness level (e.g., split students into 3 levels of "high", "medium", "low"). We selected random forest as the core algorithm for the following reasons. Random forest constructs an ensemble of decision trees and aggregates their predictions via majority voting. It naturally handles high-dimensional features, provides built-in feature importance measures, and is robust to overfitting, making it well-suited for stable classification tasks.

Regression Model:

Predicting students expected starting salary (a continuous numerical value). We selected extreme gradient boosting trees as the core algorithm for regression problem. Compared with other gradient boosting trees, XGBoost is based on a gradient boosting framework and can minimize the loss function by iteratively adding new decided trees. It is suitable for regression problem with advantages of high accuracy, good computational efficiency and ability to capture effective interactions between features.

Model training is to learn the mapping relationship between features and target variables from the data. We first split the pre-processing complete data into training set, validation set and test set with ratio of 7:2:1. The training set will be used to fit the model initially. The validation set will be used to monitor the training of the model and do the hyper-parameter tuning during the training process.

The tuning part will use a strategy of combining grid search and random search to tune the key parameters of different models. For Random Forest, the main parameter tuning part would be the number of trees, maximum depth and minimum number of samples per leaf node. While for XGBoost, the main parameter tuning part would be learning rate, maximum tree depth and the subsampling ratio of the training set per iteration. We will use cross validation to evaluate the performance of different parameter combination on validation set and finally select the model parameter combination with best validation set performance.

In order to get higher prediction performance and more stable results, we tried a model fusion strategy based on the optimization of base model training. For classification problem, we used voting ensemble method and combined the prediction result of optimized Random Forest model and similar high performance Gradient Boosting Decision Tree classifier through hard vote. It leveraged the difference of data pattern learned from different models to further improve the reliability of final classification decision. All model training and evaluation were conducted under the strict training-validation-test partitioning to ensure that the reported test performance accurately reflects the model's generalization ability on unseen data.

IV. MODEL INTERPRETABILITY APPLICATION ANALYSIS

This study uses the SHAP framework as the core interpretability analysis tool. Table 1 lists the five global core features influencing employability predictions calculated based on SHAP values and their contribution examples:

TABLE 1. Global Feature Importance Ranking (Top 5)

Rank	Key Feature	Mean	Impact Direction
1	Market Demand Skill Match Score	0.141	Positive
2	High-Quality Internship Experience Score	0.125	Positive
3	Core Course GPA Ranking Percentile	0.102	Positive
4	Weighted National Competition Awards	0.088	Positive
5	Interdisciplinary Course Breadth Index	0.075	Positive

SHAP values are based on cooperative game theory, which can consistently and fairly allocate the contribution of each feature to the model's individual prediction results.

The analysis unfolds on two levels:

Global Interpretation:

By calculating and analyzing the average absolute value of the SHAP values for all sample features, we obtained the global ranking of feature importance upon which the model relies. For example, the analysis might reveal that "market demand skill matching" and "high-quality internship experience" are the two most important global factors influencing the prediction results, validating the rationality of this evaluation index system from a data-driven perspective.

Local Interpretation:

For the prediction results of individual students (e.g., why they are assessed as highly competitive or have a low predicted starting salary), SHAP can generate detailed contribution waterfall plots or force graphs. These graphs clearly show the specific values by which each of the student's specific features (e.g., "GPA 3.6," "lack of project experience," "Python skills") pushes the final predicted value up or down from the baseline (the average predicted value for all students), thus achieving "white-box" interpretation of the prediction results.

Systematically integrating the above interpretability analysis results is a key step in this research moving from "prediction" to "decision support."

Specific applications include:

Generating Personalized Diagnostic Reports:

Automatically generating an employability assessment report for each student. The report not only includes a competitiveness level and projected starting salary, but also, using SHAP (Shape, Advantage, Benefit) as its core explanation, clearly identifies core strengths and key weaknesses. For example, the report might state: "Your academic performance (Feature A) contributed +15 points, but insufficient project experience (Feature B) resulted in a major negative contribution of -20 points."

Providing precise improvement suggestions:

Based on weakness analysis, the model can push personalized skill enhancement suggestions. For example, for students who lost points due to "low target industry skill matching," the system can connect to a market knowledge base and recommend specific online courses or relevant certifications; for students with "low participation in campus activities," it may suggest joining relevant clubs or participating in practical activities.

Optimizing talent development programs:

Through cluster analysis of the weakness characteristics of all students, university administrators can identify potentially common weaknesses in the curriculum (such as a general lack of certain practical skills among students in a particular major), thus providing quantitative data for curriculum reform and investment in practical teaching resources.

Through interpretability analysis, the model constructed in this study can not only "predict the future" but also "explain the present" and "guide action," ultimately forming a complete closed loop from data perception to intelligent prediction and then to precise intervention, which significantly enhances the practical application value of the research results in college employment guidance.

V. MODEL PREDICTIVE PERFORMANCE VALIDATION

A. EXPERIMENTAL DESIGN AND ANALYSIS

To rigorously evaluate model performance, we adopted a time-aware data split: students were sorted by graduation year and divided chronologically into training (70%, earliest cohorts), validation (20%), and test (10%, most recent cohorts) sets. All features were constructed using only information available up to the prediction time point (Section 2). Hyperparameter tuning was performed using time-series forward chaining on the training and validation sets—training on earlier windows and validating on subsequent non-overlapping windows—rather than random shuffling, eliminating data leakage. A composite evaluation system (classification, regression, ranking) was defined, with logistic regression as baseline. Ablation experiments quantified the contribution of academic, behavioral, and market data to prediction accuracy, generalization, and model superiority.

B. HORIZONTAL COMPARISON OF MULTIPLE ALGORITHMS AND BENCHMARK MODEL VALIDATION

To verify the true effectiveness of the system, after one year of experimental implementation, we compared the main output indicators of the experimental class (using comprehensive evaluation) and the control class (using traditional sports evaluation). The comparison focused on the main outputs that should be achieved in "ideological and political goal guidance" and "integration of vocational training, courses, competitions, and certification."

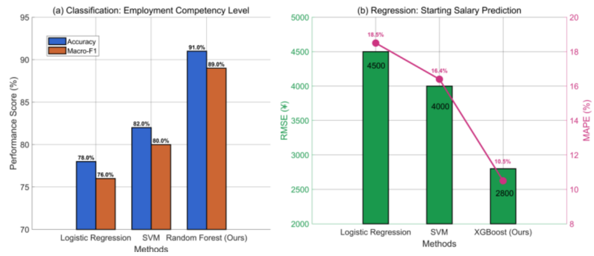

FIGURE 2. Horizontal Comparison of Multiple Algorithms and Evaluation of Benchmark Models

Figures 1 (a-b) respectively demonstrate the significant performance advantages of the ensemble learning model constructed in this paper compared to traditional benchmark models in classification and regression prediction tasks. In employment competitiveness level prediction, the accuracy (91%) and macro F1 score (0.89) of the random forest model are improved by 13 percentage points and 0.13 points respectively compared to the logistic regression benchmark. In starting salary prediction, the RMSE (2800 yuan) and MAPE (10.5%) of the XGBoost model are reduced by 30% and 36% respectively compared to the SVM model, fully demonstrating the superior performance of ensemble learning algorithms in handling multi-source, nonlinear employment data, with a 30% reduction in RMSE and a 36% reduction in MAPE compared to the SVM baseline.

C. MULTI-SOURCE DATA CONTRIBUTION ANALYSIS AND MODEL ROBUSTNESS TEST

To explore the specific contribution of multi-source heterogeneous data to the model's prediction performance, this experiment uses ablation analysis to construct four comparative feature sets by gradually introducing different data sources. Under the premise of keeping other conditions consistent, the same XGBoost model was trained using data containing only academic data, data combining academic and campus behavior data, data combining academic and market demand data, and fully integrated data.

Performance changes were compared based on independent test sets to quantitatively evaluate the value of each data dimension and the robustness of the model.

TABLE 2. Multi-Source Data Contribution Analysis and Model Robustness Assessment

Feature Set	Description	Accuracy (%)	Macro-F1	RMSE (CNY)	MAPE (%)
A	Academic Data Only (GPA, Core Course Scores)	69.5	0.665	3950	16.2
B	A + Campus Behavior Data	79.8	0.785	3450	14.1
C	A + Market Demand Data	85.2	0.832	3150	11.8
D (Ours)	A + B + C (Full Fusion Data)	91.2	0.890	2800	10.5

In Table 2, multi-source data fusion plays a crucial role in improving model performance. Compared to the baseline (A) which only uses academic data, the introduction of campus behavior data (B) and market demand data (C) improved prediction accuracy by 10.3 and 15.7 percentage points, respectively; while the fusion of all data (D) achieved the best performance, with an accuracy of 91.2%. SHAP analysis further showed that market demand matching and project practice experience were the two features that contributed the most. These results demonstrate that the model framework integrating academic, behavioral, and market data significantly outperforms a single data source, exhibiting stronger representational capabilities and robustness.

VI. CONCLUSION

This paper builds and verifies a machine learning evaluation model that leverages multi-source big data to enable the multi-dimensional quantification and dynamic prediction of student employability within the high-precision technical landscape of the modern industry. By aligning academic profiles with the evolving mechanical and digital requirements of intelligent weaving systems, the model facilitates a more accurate assessment of talent readiness for advanced industrial roles. The experimental results demonstrate that the model achieves effective prediction performance in employment competitiveness grading and key outcome forecasting, with a 91.2% classification accuracy and a MAPE of 10.5% for starting salary—substantially outperforming conventional baseline models.

Of course, there are still some deficiencies in this research. On one hand, it is necessary to further enhance the breadth and timeliness of data collection. On the other hand, it is also necessary to further verify the universality of the model in the different types of universities.

Future research will introduce real-time data from online learning trajectories and direct feedback from the intelligent manufacturing sector, utilizing deep learning architectures to enhance the timeliness of predictions and achieve a more personalized, data-driven evolution of employment guidance.

By integrating the specific mechanical and digital competency requirements of modern weaving systems into these complex models, we can ensure that guidance services remain synchronized with the high-precision demands of the industry.

AUTHOR CONTRIBUTIONS

Conceptualization –Shang Sun, Di Yang and Juan Hu; methodology – Shang Sun, Di Yang and Juan Hu; investigation – Shang Sun, Di Yang and Juan Hu; writing-original draft preparation – Shang Sun, Di Yang and Juan Hu. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This work was supported by the research topic of Anhui provincial entrepreneurship research institute in 2025 (grant number cyyj2025-06), the General Project of Humanities and Social Sciences Research of the Ministry of Education in 2023 [grant number 23JDSZ3081], the key research project of Anhui University of Science and Technology's school-level teaching reform in 2023 [grant number JG202344], as well as the key research project of Anhui University of Science and Technology's school-level teaching reform in 2023 [grant number JG202343].

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] M. Yuan, H. Mao, J. Zhang, et al., "The influence of social support on college students' employment anxiety: the mediating role of self-concept and self-criticism," Psychological Monthly, vol. 19, no. 6, pp. 47–49, 2024, doi: 10.19738/j.cnki.psy.2024.06.015.

[2] Y. Li and Zhu, "On the impact of artificial intelligence news writing on the employment of media major college students," Journalism and Communications, vol. 12, no. 06, pp. 1475–1480, 2024, doi: 10.12677/jc.2024.126225.

[3] W. Wang and J. Wang, "A study on the employment status of foreign language major college students in the post-epidemic era—taking Dalian University of Foreign Languages as an example," Modern Education Forum, vol. 4, no. 10, pp. 65–66, 2021, doi: 10.12238/mef.v4i10.4115.

[4] J. Ding, "Research on the Practice Path of Ideological and Political Education in College Students' Employment and Entrepreneurship Series Courses under the Background of "Great Ideological and Political Courses"," Journal of Jiangxi Electric Power Vocational and Technical College, vol. 38, no. 4, pp. 35–39, 2025, doi: 10.3969/j.issn.1673-0097.2025.04.010.

[5] J. A. Parrella, C. Esquivel, H. R. Leggette, et al., "Preparing agricultural leaders: An assessment of agricultural students' perceived importance and development of employability skills," The Journal of Agricultural Education and Extension, vol. 30, no. 2, pp. 233–255, 2024, doi: 10.1080/1389224X.2023.2179086.

[6] D. Carter, W. Minhas, and G. Al Hafidh, "Exploring employability skills in UAE higher education programs using a self-assessment matrix," Journal of Applied Research in Higher Education, vol. 17, no. 1, pp. 495–509, 2025, doi: 10.1108/JARHE-05-2024-0220.

[7] H. Zhang, A. Khaskheli, S. A. Raza, et al., "Linkage between students' skills and employability: Moderating influence of university reputation," Corporate Reputation Review, vol. 27, no. 4, pp. 229–248, 2024, doi: 10.1057/s41299-023-00169-9.

[8] L. Li, "Analysis of the relationship between college students' mental health education and employability improvement," Frontiers in Educational Research, vol. 7, no. 3, pp. 34–39, 2024, doi: 10.25236/FER.2024.070306.

[9] H. B. Muhammad, A. M. Muhammad, and S. D. Abdullahi, "Machine learning for predicting students' employability," UMYU Scientifica, vol. 2, no. 1, pp. 1–9, 2023, doi: 10.56919/usci.2123_001.

[10] Y. Li , H. Chen, G. Zhu, et al., "Research and application of early warning model for employment of physics major college graduates based on machine learning algorithm," Physics and Engineering, vol. 35, no. 3, pp. 142–152, 2025, doi: 10.3969/j.issn.1009-7104.2025.03.025.