Education Reform and Development, 2026, Volume 8, Issue 10
https://bbwpublisher.com/journal/erd
Online ISSN: 2652-5372
Print ISSN: 2652-5364

Predictive Power of Vocational-College Academic Performance on Job-Major Match: Dual-Method SHAP Evidence Based on Logistic Regression and XGBoost

Hongcheng Wu, Bo Li, Tianliang Jiang

School of Mechanical and Electrical Engineering, Xiangxi Vocational and Technical College for Nationalities, Jishou 416000, Hunan, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: The job-major match rate serves as a core indicator for measuring the school-running quality of higher-vocational colleges, and its implicit premise is that academic performance during schooling can predict job-major match outcomes. Taking 1,418 manufacturing-track graduates from a higher-vocational college as a sample, this study constructs 36-dimensional academic features based on Becker's human-capital theory. It adopts logistic regression and XGBoost paired with LinearExplainer and TreeExplainer, respectively, for dual-method SHAP analysis, so as to test the predictive power of academic performance for job-major match. The results show that the test AUC of logistic regression reaches 0.852 while its PR-AUC is only 0.396, and the AUC of XGBoost is 0.817. Only programming-course enrollment (yes/no) ranks first in SHAP importance with a negative direction under both methods. Among the four specific human-capital features, three show significant negative correlations between SHAP values and feature values ($\rho = -0.77$ to -0.36). Academic performance hardly predicts individual differences in job-major match, and the key factors determining job-major match lie outside classroom teaching.

Keywords: Job-major match; Academic performance; SHAP; Explainable machine learning; Higher-vocational college graduates

Online publication: November 25, 2026

1. Problem statement

The job-major match rate is a core indicator for measuring the quality of higher-vocational college education. According to the *Chinese Higher-Vocational Graduate Employment Report* published by MyCOS Research Institute, the job-major relevance of higher-vocational graduates has long fluctuated within the range of 65 %-75 % [1]. Taking the match rate as a basis for judging school-running effectiveness implies a premise: academic performance at school, especially scores of specialized courses, is able to predict job-major match outcomes. If this premise does not hold, such policy logic needs to be re-examined.

Existing studies show that professional identity and family expectations significantly influence willingness for job-major match [2], and professional-interest groups act as the strongest predictor [3-4]. Based on gradient-boosting-tree analysis of 13 828 manufacturing-track graduates, Liu Yunbo et al. [5] found that internship experience (contribution: 43.2 %) ranked first, while academic factors (23.0 %) and institutional factors (13.8 %) had limited effects. Conclusions of these studies converge: key factors determining job-major match lie outside classroom teaching. Nevertheless, those studies rely on self-reported grades, which suffer from bias and information loss. This study adopts full administrative data for verification.

In accordance with Becker's human-capital theory [6], academic performance is operationalized into general human capital (transferable across industries) and specific human capital (valuable only in given industries). Two competing hypotheses are put forward: H1: general human capital of transferable skills is negatively correlated with job-major match; H2 (lock-in hypothesis): specific human capital is positively correlated with job-major match. If H2 is supported, the policy logic of "academic performance → job-major match" is valid.

This study introduces the SHAP framework [7-8] to overcome limitations of traditional regression in multi-dimensional attribution, and treats it as a predictive-attribution tool rather than a causal-inference instrument.

2. Research design

Data are extracted from the information-management system of a higher-vocational college in an ethnic-minority area, covering academic and employment information of 1 418 manufacturing-track engineering graduates of the classes of 2024-2025. The dataset includes 47 classes, 444 courses and 134 182 grade records, involving six highly homogeneous engineering majors. Academic grades are assigned to six skill-oriented directions through keyword matching, standardized by z-score before entering models. Missing records are assigned a value of −2.0 with binary markers to distinguish missing values. After removing low-variance markers, 8-dimensional general-human-capital features, 4-dimensional specific-human-capital features and 24-dimensional control variables are retained, totaling 36 dimensions.

Specific-human-capital features only keep grades of four core courses: Electrical and Electronic Technology, Mechanical Design, Robotics, and Comprehensive Practical Training. The dependent variable, job-major-match degree, is calculated via keyword matching between major names and job-post texts. 231 samples (16.3 %) scoring ≥ 4 are grouped into the high-match group. Alternative thresholds of ≥ 3 and ≥ 5 are adopted to test robustness.

Table 1. Dependent-Variable Measurement Rules and Sample Distribution

Score	Matching Rule	Sample Proportion (%)	Group Assignment
5	≥ 2 matching keywords	9.5	High match (positive class)
4	1 matching keyword	6.8	High match (positive class)
3	Technology/engineering / design-related posts	20.3	Low match (negative class)
2	Other posts	63.4	Low match (negative class)

Note: 231 samples fall into the high job-major-match group (score ≥ 4), accounting for 16.3 %.

Table 2. Variable Definition and Descriptive Statistics (n = 1418)

Category	Variable	Description	Mean	Standard Deviation
General HC	Skill breadth	Average pass rate across three directions ÷ 3	0.477	0.379
	Number of covered skills	Count of studied skill directions	1.431	1.226
	Coefficient of grade variation	std / avg	0.204	0.108
	Programming course (yes/no)	1 = taken	0.491	0.500
	Electrical-electronic course (yes/no)	1 = taken	0.613	0.487
	Robotics course (yes/no)	1 = taken	0.328	0.470
Specific HC	Programming-course average grade	Raw average score	75.19	14.50
	Basic-discipline average grade	Raw average score	82.14	7.73
	Electrical-electronic course grade	Raw average score	74.09	12.56
	Mechanical-design course grade	Raw average score	74.51	12.87
	Robotics-course grade	Raw average score	81.07	10.50
	Comprehensive practical-training grade	Raw average score	76.24	10.79
Control Variables	Total courses / valid courses	—	71.35 / 71.17	13.20 / 12.88
	Average grade/grade standard deviation	—	78.03 / 14.97	9.34 / 5.72
	Pass rate/gender (male = 1)	—	0.876 / 0.913	0.163 / 0.282
	Ethnicity / major/political status/family economic status	One-hot encoding	—	—

Note: 8-dimensional general HC + 4-dimensional specific HC + 24-dimensional control variables = 36 dimensions.

Logistic regression, random forest, XGBoost, LightGBM and gradient-boosting classifiers are adopted. Weight adjustment instead of oversampling is used to address the 16.3 % positive-class imbalance so as not to interfere with SHAP attribution. AUC serves as the primary metric, and 5-fold stratified cross-validation is used to evaluate generalization performance. Dual SHAP implementations are applied: LinearExplainer for logistic regression and TreeExplainer for XGBoost for a robustness check. 200-time bootstrap resampling and 50-time feature perturbation are conducted for robustness testing.

3. Empirical results and analysis

3.1. Limited model predictive performance

The test AUC of logistic regression is 0.852 with PR-AUC of merely 0.396. XGBoost achieves AUC = 0.817 and PR-AUC = 0.380. All models obtain cross-validation AUC above 0.81 (**Table 3**). The gap between AUC and PR-AUC stems from class imbalance: logistic regression reaches 0.99 recall at the cost of 0.31 precision, and roughly two-thirds of positive-class predictions are false positives. Limited differences between linear and ensemble models indicate that grade-related features approximately follow a linear-additive structure.

Table 3. Model-Performance Comparison

Model	Family	Test AUC	PR-AUC	F1-Score	Cross-Validation AUC
Logistic Regression (C = 0.1)	Linear	0.852	0.396	0.471	0.818
Logistic Regression (C = 1.0)	Linear	0.849	0.388	0.491	0.820
Random Forest (max_depth = 10)	Bagging	0.846	0.409	0.523	0.824
XGBoost (max_depth = 6)	Boosting	0.817	0.380	0.400	0.816

LightGBM (max_depth = 6)	Boosting	0.805	0.357	0.408	0.824
Gradient Boosting (max_depth = 3)	Boosting	0.805	0.339	0.270	0.816

Note: n = 1418; positive-class proportion = 16.3 %; CV denotes mean AUC of 5-fold cross-validation.

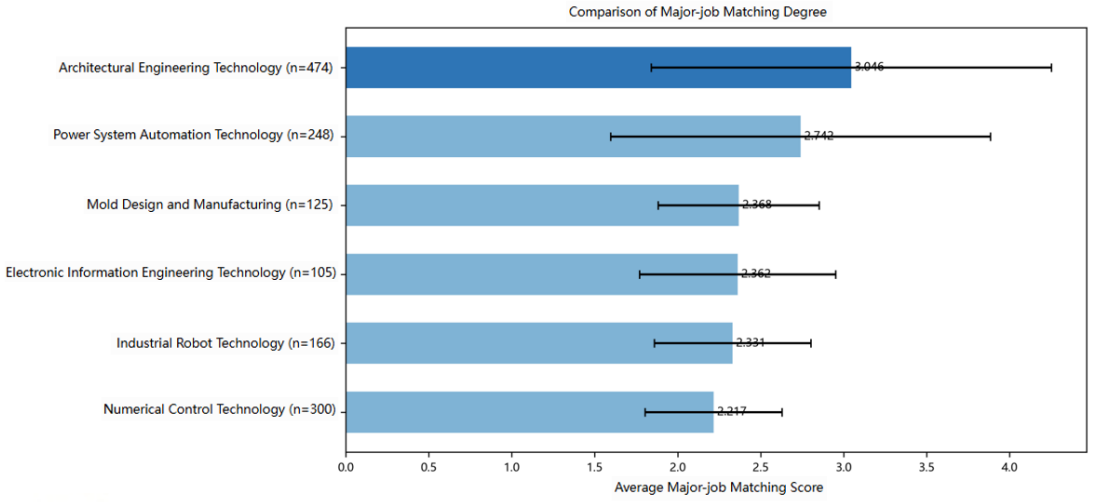


Figure 1. Job-Major-Match-Degree Comparison Across Majors.

SHAP Feature Importance: Academic grades show almost no independent discriminative power. Both explainers consistently rank programming-course enrollment (yes/no) first (LinearExplainer mean(|SHAP|) = 0.465; TreeExplainer mean(|SHAP|) = 2.231, **Figure 2**, **Figure 3**) with a negative direction — students who have taken programming courses are more likely to take jobs outside their major (**Figure 4**). Skill breadth and number of covered skills rank 5th and 6th under LinearExplainer, yet drop to 17th and 23rd under TreeExplainer. Such "attribution collapse" originates from their high multicollinearity (r = 0.78), demonstrating that they have no cross-model-robust independent discriminative power. Among specific-human-capital features, only Fundamentals of Mechanical Design (mean(|SHAP|) = 0.683) enters the top-five features under TreeExplainer. Overall, the most discriminative signal turns out to be a binary variable pointing toward cross-major employment.

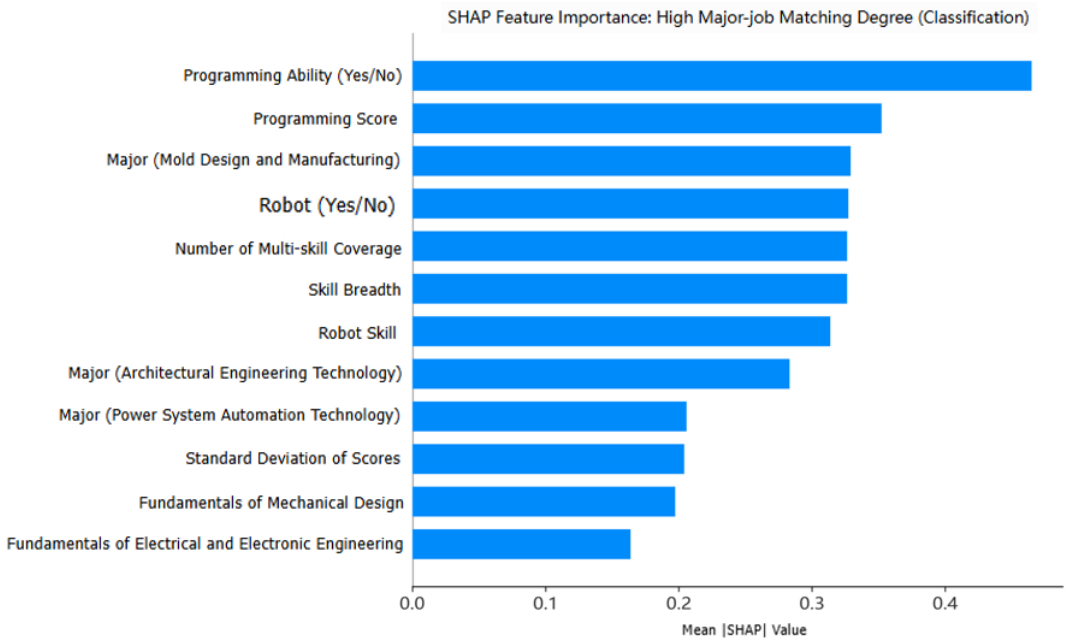


Figure 2. SHAP Feature Importance (LinearExplainer, Top-12 out of all 36 features).

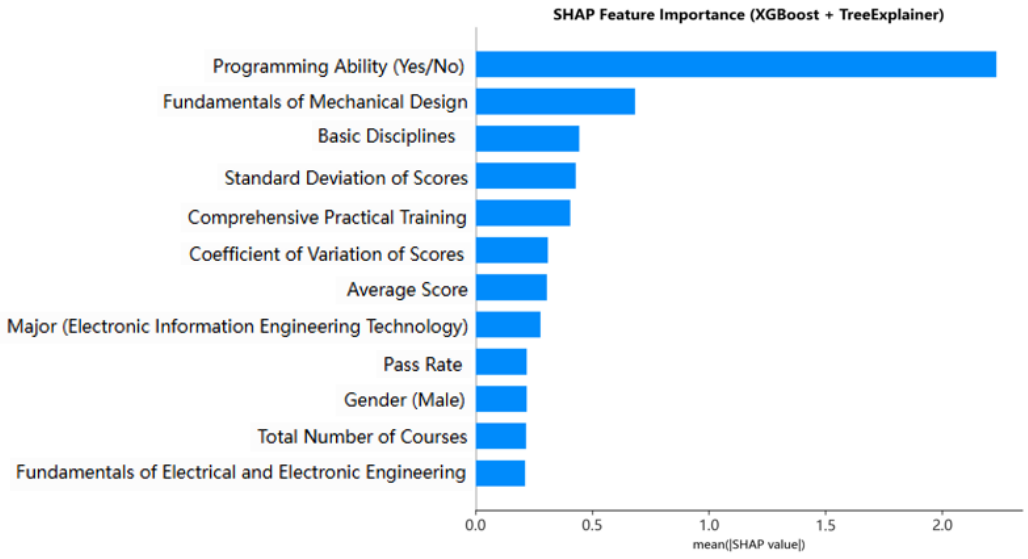


Figure 3. SHAP Feature Importance (XGBoost + TreeExplainer).

3.2 Hypothesis Testing: No Lock-In Effect for Specific Human Capital

Taking high job-major match as the positive class, $\rho < 0$ means higher feature values correspond to a higher tendency for cross-major employment. The Spearman's ρ correlation coefficients between SHAP values and feature values for four specific-human-capital features are: Robotics-course grade −0.77, Fundamentals of Mechanical Design −0.62, Electrical-Electronic Fundamentals −0.36, Comprehensive Practical Training +0.20 (all $p < 0.001$, **Table 4**), three negative and one weakly positive. Programming-course grade yields $\rho = -0.70$ (consistent with H1), and basic-discipline grade $\rho = +0.26$ (reverse effect for general-knowledge courses). Both fall under general human capital. Therefore, H2 is not supported. Marginally, none of the 465

students who took robotics courses fall into the high-match group ($\chi^2 = 134.65$, $p < 0.001$). After conditioning, however, its effect size becomes small and remains negative, consistent with observations made by Wang Xiaogang [9].

Table 4. Spearman Rank Correlation Between Feature Values and Corresponding SHAP Values

Feature	Category	Spearman's ρ	*p*-value
Robotics-course grade	Specific HC	−0.77	< 0.001
Programming-course grade	General HC (transferable)	−0.70	< 0.001
Fundamentals of Mechanical Design	Specific HC	−0.62	< 0.001
Electrical-Electronic Fundamentals	Specific HC	−0.36	< 0.001
Basic-discipline grade	General HC (general knowledge)	+0.26	< 0.001
Comprehensive Practical Training	Specific HC	+0.20	< 0.001

Note: ρ is calculated from SHAP values generated by TreeExplainer; n = 426 (test set).

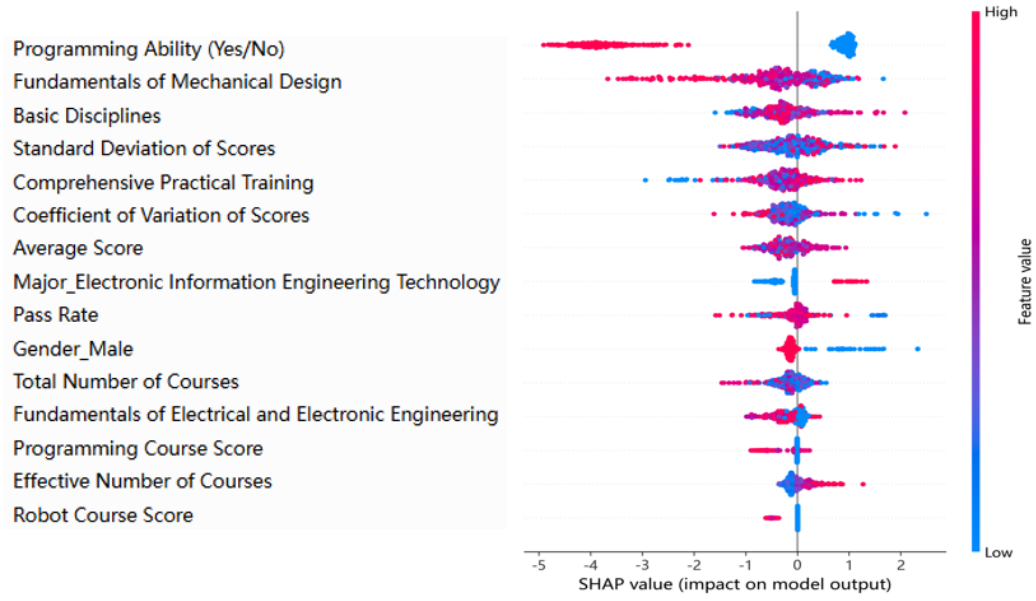


Figure 4. SHAP Beeswarm Plot (TreeExplainer).

3.3. Robustness test

Programming-course enrollment consistently ranks first across 50-time feature-perturbation tests, and its SHAP value maintains a 100-percent-negative direction across 200-time bootstrap resampling. Its rank and sign do not change after switching explainers, verifying cross-method reliability for this top predictive signal (this does not represent real causal effects).

4. Discussion

This study draws a negative-finding conclusion: academic performance hardly predicts individual differences in job-major match. This conclusion is not isolated. Despite large differences in samples, variables, algorithms and dependent-variable measurements, studies by Zhou Xue, Wang Shanqin, Liu Yunbo and other scholars reach consistent conclusions: key factors determining job-major match lie outside classroom teaching. Sun

Fangfang and Li Long [10] interpret this phenomenon from the perspective of structural tensions between skill-formation systems and labor-market institutions. Applying dual-method SHAP for multicollinearity diagnosis also constitutes a methodological contribution of this paper.

Limitations of this study: samples come from a single higher-vocational college; job-major-match measurement based on text keyword matching may underestimate implicit match; cross-sectional data cannot eliminate omitted-variable bias; imputation with −2.0 may produce artificial accumulation in tree-based models, yet the core finding relies on the binary variable "programming-course enrollment" which is immune to such interference; samples do not cover non-engineering majors. The above-mentioned limitations do not reverse the direction of core findings. This study focuses only on job-major matching, and does not rule out the possibility that grades can predict on-the-job performance.

5. Conclusion

Taking Becker's human-capital theory as the analytical framework, this paper conducts analysis on full administrative academic records of 1 418 manufacturing-track higher-vocational graduates using dual-method SHAP, and tests the implicit premise that "academic performance can predict job-major match". The results show that academic performance has extremely limited predictive power for job-major match. Only programming-course enrollment (yes/no) robustly ranks first with a negative effect. Three out of four specific-human-capital features show significant negative effects without any positive lock-in effect. Although model AUC reaches 0.85, PR-AUC is merely 0.40. This finding converges with existing studies across different methods and samples. Therefore, quality evaluation systems shall incorporate extra-class factors such as internship quality, depth of school-enterprise cooperation and professional-identity cultivation into core evaluation indicators.

Funding

Scientific Research Project of Hunan Provincial Department of Education, "Research on Graduate Employment Prediction and Early-Warning Based on Artificial Intelligence Technology" (Project No.: 24C1239)

Disclosure statement

The authors declare no conflict of interest.

References

[1] MyCOS Research Institute, 2024, 2024 Report on Chinese Higher-Vocational Graduates' Employment. Social Sciences Academic Press, Beijing.

[2] Zhou X, 2023, Research on Job-Major-Match Intention and Influencing Factors of Manufacturing-Major Students in Vocational Colleges. Journal of Wuhan Polytechnic, 22(1): 34–42.

[3] Wang S, Wu C, Chen Y, 2019, Application of C4.5 Algorithm in Analysis of Factors Affecting Vocational-College Graduates' Job-Major Match. Journal of Dongguan University of Technology, 26(3): 39–45.

[4] Feng Q, Yue C, 2025, Can Major-Interest Matching of Bachelor-and-Junior-College Graduates Promote Employment Matching? — Empirical Analysis Based on Three-Dimensional Employment Matching of Major, Educational Background and Competence. Journal of East China Normal University (Educational Sciences), 2025(5): 96–114.

[5] Liu Y, Yu Z, Yang P, 2026, Prediction Research on Job-Major Match of Higher-Vocational Manufacturing-Major Graduates — Findings Based on Machine Learning. Tsinghua Journal of Education, 2026(3): 158–166.

[6] Becker GS, 1964, Human Capital: A Theoretical and Empirical Analysis, With Special Reference to Education. University of Chicago Press, Chicago.

[7] Lundberg SM, Lee SI, 2017, A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems 30. Long Beach: NeurIPS, 4765–4774.

[8] Lundberg SM, Erion G, Chen H, et al., 2020, From Local Explanations to Global Understanding With Explainable AI for Trees. Nature Machine Intelligence, 2(1): 56–67.

[9] Wang X, 2018, Analysis on Dilemmas of Job-Major Match for Higher-Vocational Majors of Mechanical Design and Manufacturing. Science-Technology Innovation and Productivity, 2018(8): 111–113.

[10] Sun F, Li L, 2026, Skill Mismatch and Institutional Tension: Generation and Adjustment of Employment Dilemmas for Higher-Vocational Graduates. Vocational and Technical Education, 2026(10): 42–49.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.