Analyzing Rural Online Education Talent Development with Data and Machine Learning

Te-Hsin Hsieh*, Xueye Lai

School of International Business, Xiamen University Tan Kah Kee College, Xiamen, China

*195197110@qq.com, 2413947747@qq.com

Abstract. In the rapidly evolving landscape of information technology, online education has emerged as a vital tool, particularly in rural areas. The "14th Five-Year Plan for Educational Informatization and Cybersecurity Development" projects substantial growth in China's online education sector by 2025. This study investigates the effectiveness of online talent training in rural regions, utilizing data analysis and machine learning techniques to explore its impact on adaptability. Thirteen features including gender, age, education level, and network type were analyzed. Results show online training positively influences adaptability in rural students, with variations in feature impacts. Models like Random Forest and XGBoost excel in predicting adaptability levels. These findings provide valuable insights for talent development in rural areas, supporting local economic growth.

Keywords: Online education, Rural areas, Talent training, Data analysis, Machine learning.

1 Introduction

In the ever-evolving landscape of information technology, online education emerges as an innovative educational paradigm undergoing rapid global expansion, particularly showcasing significant potential and pivotal roles in rural areas. According to the "14th Five-Year Plan for Educational Informatization and Cybersecurity Development" [1], by 2025, the user base of online education in China is projected to reach 300 million, with the market size expected to surpass 800 billion yuan. Against this backdrop, talent cultivation through online education in rural areas has become a crucial component driving the implementation of strategies for educational equity and rural revitalization. Fueled by technological advancements and national initiatives, the online education model in rural areas exhibits unprecedented vigor and innovative potential.

From a global perspective, the flourishing growth of the online education market is particularly pronounced in rural areas. Data indicates a rapid increase in online education participation rates in rural China from 2017 to 2020, with rural students' online learning time surging by over 50% [2]. This trend not only underscores the strong demand for educational resources in rural areas but also unveils the immense potential

Y. Feng et al. (eds.), *Proceedings of the 4th International Conference on Internet, Education and Information Technology (IEIT 2024)*, Atlantis Highlights in Social Sciences, Education and Humanities 26,
https://doi.org/10.2991/978-94-6463-574-4_47

for leveraging modern information technology to optimize resource allocation and innovate teaching methodologies. Nationally, rural online education is experiencing explosive growth, offering more personalized and efficient learning pathways and educational resources to rural communities.

However, amidst these developmental opportunities, talent cultivation in rural online education faces formidable challenges. According to the "2020 Report on the Current Status and Development of Rural Online Education in China" [3], issues such as inefficiency persist within the rural online education sector. Behind this phenomenon lie both historical legacy issues of uneven distribution of educational resources and current challenges stemming from uneven development in information technology[4]. Therefore, exploring the effectiveness of online talent training in rural areas is not only an effective approach to addressing educational inequality but also a key driver for advancing educational modernization and rural revitalization.

In order to enhance the quality of online education in rural areas, this study aims to investigate the effectiveness of online training. Employing methods of data analysis and machine learning prediction, we will delve into the impacts of various training characteristics on students' adaptability. Through the collection and analysis of data encompassing 13 features including gender, age, educational level, network type, and adaptability levels, we will establish predictive models to assess the influence of different features on the adaptability of rural online education students. This endeavor will provide scientific grounds for enhancing the quality of online education in rural areas and fostering the sustainable development of rural talent.

2 Data Collection and Feature Analysis

2.1 Description of Data Features

Features are attributes or characteristics that describe data samples, providing input information for models to make predictions or classifications. In machine learning, features are typically the independent variables of the model used to capture differences and patterns among samples. The selection and extraction of features are crucial for the performance of the model; good features can enhance the accuracy and generalization ability of the model. In this study, features such as gender, age, educational level, network type, among others, are used to describe students' personal backgrounds, educational environments, and online learning situations, aiming to predict their adaptability levels.

In this study, adaptability level is designated as the target column, serving as an indicator of students' adaptability in the online education environment. The task of the model is to predict or classify students' adaptability levels based on the given features, thereby providing guidance and recommendations for educators and decision-makers.

In this study, we utilized a dataset comprising 13 features and 1 target column. These features cover students' personal background information, educational environments, and online learning situations, including gender, age, educational level, institution type, whether studying IT-related courses, location, power outage level, financial condition, network type, connection type, class duration, whether the institu-

tion has its own learning management system, and the primary devices used in the classroom. The target column indicates students' adaptability levels in the online education environment.

2.2 Univariate Analysis Results

In univariate analysis, the gender feature shows that males and females account for 55% and 45% of the total population, respectively. Regarding institution type, non-governmental institutions comprise 68.3% of the total, indicating a significant proportion. In terms of age, respondents' ages are mainly concentrated between 11 and 25 years old, reflecting a typical distribution of respondent ages. Class duration is primarily distributed between 1 and 3 hours, indicating respondents' preferences and acceptance of course durations. Most respondents attend schools and universities, indicating higher educational levels. Concerning financial condition, most respondents are in moderate financial condition, which may influence their degree and mode of participation in online education. The usage rates of mobile data and mobile phones are high, at 57.7% and 84.1%, respectively, indicating widespread use of mobile networks and mobile devices in online education. Adaptability levels show that approximately 51.9% of respondents indicate medium adaptability, 39.8% indicate low adaptability, and the rest indicate high adaptability, providing important insights into respondents' adaptability. Overall, these univariate analysis results provide important clues and insights into understanding respondent characteristics and behaviors.

2.3 Multivariate Analysis Results

In multivariate analysis, a myriad of complex relationships among different features in the dataset are revealed. Firstly, by examining the correlation between gender and age, it is evident that females' ages are predominantly distributed between 11 and 20 years old, while males are mainly concentrated between 11 and 15 years old and 21 and 25 years old. Further exploration of the relationship between gender and being an IT student demonstrates that males are more inclined to participate in the survey as IT students compared to females. Additionally, investigating the association between IT students and educational level reveals that universities have the highest number of IT students, with 30 and 27 IT students in universities and schools, respectively. Analyzing the correlation between gender and adaptability level indicates that among the surveyed individuals, 71 male respondents exhibit a high level of adaptability, whereas only 29 female respondents exhibit a similar level. Furthermore, both males and females overall demonstrate a medium level of adaptability. Finally, a study on the relationship between age and adaptability level reveals that respondents aged between 11 and 15, 21 and 25, and 26 and 30 tend to have relatively higher levels of adaptability, while no respondents aged between 1 and 5 exhibit a high level of adaptability. These multivariate analysis results offer a comprehensive understanding of the interrelationships among features in the dataset, aiding in uncovering the underlying meanings and patterns behind the data.

3 Model Establishment and Evaluation

3.1 Data Preprocessing

In the data preprocessing stage, features were encoded using label encoding to convert categorical data into numerical form. The dataset was then split into training and testing sets, with 70% for training (843 samples) and 30% for testing (362 samples). This ensures the model has enough data for both learning and evaluation.

3.2 Model Selection

The study compared and evaluated five distinct machine learning algorithms: logistic regression, K-nearest neighbors (KNN), random forest, XGBoost, and CatBoost. Logistic regression, a traditional method, operates effectively for binary classification tasks, employing a logistic function to transform predictor variables into probabilities. KNN, an instance-based technique, determines classification by referencing the labels of nearby data points in the feature space, offering flexibility without the need for pre-training. Random forest, an ensemble learning algorithm, aggregates predictions from multiple decision trees to enhance accuracy and robustness. XGBoost, a gradient boosting tree algorithm, iteratively trains decision tree models, refining predictions by adjusting to residuals from preceding models[5]. Lastly, CatBoost, designed explicitly for classification, proficiently manages categorical features and missing data, making it suitable for diverse datasets encountered in talent training models.

3.3 Cross-validation Results and Model Comparison

In this study, five selected machine learning models underwent 5-fold cross-validation[6], and their average accuracy scores on the validation set were calculated. The result shows that the random forest model performed the best in cross-validation, achieving 0.896, followed by the XGBoost and CatBoost models with scores of 0.891 and 0.886, respectively. In contrast, the logistic regression model exhibited relatively poor performance, with a score of only 0.688. This indicates that the random forest, XGBoost, and CatBoost models may be more suitable for constructing talent training models. The random forest model demonstrates good performance in handling high-dimensional features and large-scale data, while the XGBoost and CatBoost models excel in addressing classification problems, effectively capturing the complexity of the data relationships. Therefore, in selecting the final model, we will focus on the performance and applicability of these three models and further optimize and tune the models to ensure their robustness and generalization capability.

4 Results Discussion and Analysis

4.1 Analysis of Feature Influence on Adaptability

Figure. 1 illustrates the ranking of feature importance, which is based on the feature importance ranking results of the random forest model, enabling analysis of the impact of each feature on adaptability. From the results, it is evident that the features "Class Duration," "Financial Condition," and "Age" have the most significant influence on adaptability. This indicates that class duration, financial condition, and age may be crucial factors affecting students' adaptability in online education. Additionally, features such as gender, school type, and network type also have some impact on adaptability, albeit relatively minor.

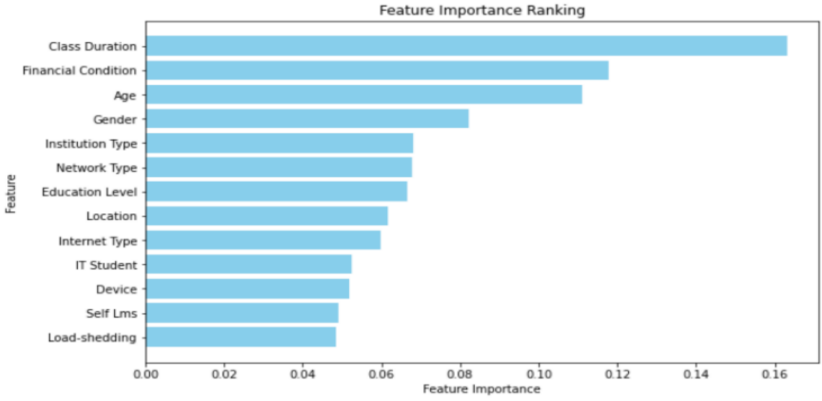

Fig. 1. Feature Importance Ranking

4.2 Analysis of Important Features

Class duration, financial condition, and age are among the key factors influencing students' adaptability in online education. Here is an in-depth analysis of the specific relationships between these three features and adaptability:

Class duration in online education is a significant indicator of students' daily study time. Longer class durations may imply that students require more extended study time and effort to complete course content, thus placing higher demands on their adaptability[7]. Research findings indicate that class duration significantly influences students' adaptability, with students who have longer class durations often demonstrating higher adaptability. This could be attributed to the fact that prolonged learning helps students better adapt to the online learning environment, enhancing learning efficiency and self-discipline.

Financial condition reflects the economic situation and resource support level of students' families, which has a significant impact on students' learning conditions and environments[8]. Students with better financial conditions may have better learning devices, more stable network connections, and richer learning resources, thereby improving their learning performance and adaptability in online education. Conversely,

students with poorer financial conditions may face challenges such as inadequate learning resources and unstable networks, which could affect their learning outcomes and adaptability.

Students' age is also one of the crucial factors influencing their adaptability in online education. Through the above research, it can be observed that students aged 11 to 15, 21 to 25, and 26 to 30 exhibit relatively higher levels of adaptability. This may be because students in these age groups have richer educational and life experiences, enabling them to better adapt to the online learning environment. In contrast, younger or older students may lack sufficient learning experience or adaptability, resulting in lower levels of adaptability.

5 Conclusion

This study delves into students' adaptability in online education in rural areas, identifying class duration, financial status, and age as crucial factors. Longer classes demand higher self-management skills, while better finances and older age correlate with enhanced adaptability. Random forest, XGBoost, and CatBoost models excel in predicting adaptability, with random forest leading in accuracy. Recommendations include optimizing class schedules, providing financial aid, tailoring courses by age, and leveraging machine learning for personalized support.

Reference

1. Ministry of Commerce of the People's Republic of China. (2022). "14th Five-Year Plan for the Development of E-commerce."
2. General Administration of Customs of the People's Republic of China. (2023). "China's Cross-border E-commerce Import and Export Situation in January 2023."
3. iResearch Consulting. (2023). "Analysis Report on the Development Trends of China's Cross-border E-commerce Market in 2023."
4. Research Team on the Development of Rural E-commerce Talent in China. (2020). "Report on the Current Situation and Development of Rural E-commerce Talent in China in 2020." Beijing: China Agriculture Press.
5. Sagi, O. and Rokach, L., 2021. "Approximating XGBoost with an interpretable decision tree." Information Sciences, 572, pp.522-542.
6. Malakouti, S.M., 2023. "Improving the prediction of wind speed and power production of SCADA system with ensemble method and 10-fold cross-validation." Case Studies in Chemical and Environmental Engineering, 8, p.100351.
7. International Institute of Trade and Economic Cooperation. (2019). "China Cross-border E-commerce Development Report (2019)."
8. Ren, Y. and Wei, R., 2023. "Research on the Development Path of Rural E-commerce under the Background of Rural Revitalization." Agricultural Outlook (1673-3908), 19(12).