International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1821-1831

https://doi.org/10.70517/ijhsa464158

Use big data analysis methods to conduct in-depth research on future job market skills requirements

Wenting Yuan[1] and Lei Bao[1,*]
[1] Yunnan Vocational College of Mechanical and Electrical Technology, Kunming, Yunnan, 654100, China
Corresponding authors: (e-mail: lwzcsq@126.com).

Abstract With the development of the times, the competition in the future job market will become more and more intense. Therefore, how to use big data analysis methods to accurately predict the demand for occupational skills is of great significance. In this paper, Baidu search index from May 5, 2020 to December 31, 2024 is selected as the research data, and six influencing factors that are more important to the future demand for vocational skills in the job market are screened out as the explanatory variables of the empirical model. The random forest model was used to fit and predict the demand for vocational skills in the future job market, and finally, the PDP and SHAP value interpretable machine learning techniques were used to analyze the interpretable factors influencing the demand for vocational skills in the future job market. The results of the study show that the differences in core technical demand skills, communication skills, and the ability to apply interdisciplinary skills have the most significant impact on the model, with an average SHAP value of 0.102269, 0.032896, and 0.015822, respectively, which shows that the core professional technical demand has an important competitiveness in the future job market.

Index Terms random forest, SHAP interpretability, occupational skill needs, big data analysis

I. Introduction

According to the latest labor statistics, the overall employment rate in the current job market continues to grow. Current employment segments in several countries have some variations between regions and industries, but overall, unemployment rates remain relatively low [1]-[3]. In an ever-changing society, understanding and adapting to the needs and opportunities of the job market is key to successful career development. At present, especially after the outbreak of the epidemic, the growth in the number of fresh graduates, people from all walks of life will face greater employment pressure, but there is no growth in the supply of social jobs, resulting in the supply of talent in the job market is greater than the demand, and the social contradiction has become more prominent, and the job market requires that job-seekers master a wide range of vocational skills [4]-[7]. In addition, driven by globalization and technological innovation, many industries have shown a trend of rapid growth and great potential. Some of these industries include multiple sectors such as technology, healthcare, and green energy, which are affected by the growing global economy, aging population, technological advances, and environmental awareness, the future job market will face unprecedented changes and challenges [8]-[10]. The trend of skills demand in the future job market is constantly changing, and understanding and acquiring relevant vocational skills in advance can better adapt to future career development, and by analyzing the skills demand in the future job market, we can make an important contribution to the development of educational institutions and vocational training, as well as enterprises [11]-[13].

Nowadays, due to the in-depth development of online job search services, the data of job market has increased dramatically, and the literature [14] points out that the use of “big data” in online job search platforms can realize the monitoring and analysis of the job market, assessment of the skill demand in the job market, matching of job-seeking skills, and prediction of the skill demand, and so on. Based on big data, literature [15] applies mining tools to explore the intrinsic relationship between occupational skills from member profiles, and there are a large number of correlation rules with high confidence, enhancement value and support. Therefore, the correlation data can be analyzed to explore the future demand for occupational skills in the job market.

In today's information age, big data has become one of the core resources for all industries. However, simply having large-scale data is not enough to bring business value; rather, it needs to be transformed into useful insights and strategies through big data analytics. Literature [16] uses big data methodology combined with artificial intelligence to quantitatively analyze job information and resume information collected from online job search websites, which can accurately identify labor demand and supply. Literature [17] integrated big data analytics, natural language processing, and autoregressive integrated moving average model to analyze and predict job

market demand work skills using data from job boards as inputs, noting that vocational skills such as business development skills, data visualization, verbal communication, Microsoft Excel, and stress resilience are most in demand in Bangladesh, which promotes graduate employment. Literature [18] proposed a big data analytics-supported job demand analysis for college students, which was analyzed by keyword capturing the job text messages through word frequency-inverse document frequency algorithm, and analyzing them by splitting the text words under Word2Vec tool, and introducing XGBoost model for predicting the demand. Literature [19] uses text mining methods to extract data from job postings in order to characterize job skills and needs, which is important for the training of early stage researchers and future R&D personnel.

This paper firstly introduces the random forest model in detail, and also introduces the SHAP interpretation framework to improve the interpretability of the model. Then Baidu search index from May 5, 2020 to December 31, 2024 was selected as the search data for empirical analysis. Then a series of data preprocessing work was carried out, including the treatment of missing values, scientific coding of categorical variables, and data normalization and other key steps, aiming to transform the data into a more suitable form for model construction, and also the treatment of unbalanced data and feature selection to improve the accuracy of subsequent modeling. Finally, based on the machine learning prediction results, the PDP and SHAP values are used to provide an interpretable analysis of the factors influencing the demand for occupational skills in the future job market.

II. Big data analysis methodology

II. A. Random Forest Model

Random Forest is an integrated learning algorithm based on decision trees. Named after its dendritic form, the decision tree classifier is a very simple and straightforward way of categorizing or labeling things: just ask a set of questions and then answer them, and the model gets better as those questions are answered. For categorization problems, using information entropy can quantitatively describe the amount of information in a sample set, or the uncertainty that occurs in a particular class of samples [20]. After dividing the samples using a certain attribute, the sample set becomes ordered, the information entropy of the divided samples decreases, and based on the different degrees of the decrease, differences in the importance of the attribute can be determined.

(1) Information entropy

Let S be a collection of N samples. Set S can be divided into M classes $C=\{C_1,C_2,...,C_M\}$, if n_m is the sample size in class $C_m(i=1,2,...,M)$, then the information entropy of sample S is defined by Eq:

$$I(S)=-\sum_{m=1}^{M}P_m\log_2 P_m \tag{1}$$

where $P_m=n_m/N$ is an estimate of the probability that the sample belongs to C_m.

(2) Information gain

Assuming that any sample of the sample set can be represented by a K-dimensional attribute $A=\{A_1,A_2,...,A_K\}$, attribute $A_i(i=1,2,...,K)$ according to the different values of the sample will be divided into V_i classes of sample S, that is, V_i subsets $\{S_{i1},S_{i2},...,S_{ivi}\}$. where subset $S_{iv}(i=1,2,...,V_i)$ has $n_{iv}(v=1,2,...,V_i)$ in the number of samples belonging to the class $C_m(i=1,2,...,M)$ of n_{ivm}, then the information entropy of the samples after the use of A_i on S for a division of the sample for a time is Eq:

$$I(S,A_i)=-\sum_{v=1}^{v_i}\frac{n_{iv}}{N}I(S_v)=-\sum_{v=1}^{v_i}\frac{n_{iv}}{N}\sum_{m=1}^{M}\left(\frac{n_{ivm}}{n_{iv}}\log_2\frac{n_{ivm}}{n_{iv}}\right) \tag{2}$$

After dividing the sample S using attribute A_i, the sample becomes ordered, i.e., the information entropy decreases. The amount by which the information entropy decreases compared to the pre-division is called the information gain, which is defined by the equation:

$$G(S,A_i)=I(S)-I(S,A_I) \tag{3}$$

(3) Information gain rate

In order to change the unfavorable effect of the information gain criterion preferring multi-attribute data, the information gain rate has been proposed, which is defined as Eq:

$$G_{Ratio(S,A_i)}=\frac{G(S,A_i)}{I(A_i)} \tag{4}$$

where $G(S,A_i)$ is the information gain after dividing the sample using attribute A_i, which is calculated as Eq:

$$I(A_i)=-\sum_{v=1}^{V_i}\frac{n_{iv}}{N}\log_2\frac{n_{iv}}{N} \tag{5}$$

The optimal splitting effect is mainly achieved in decision trees by the following centralized algorithm:

One fatal disadvantage of decision trees is that the model it trains is very easy to overfit, which will lead to, when you adopt your model to predict a new point, the model is more likely to be affected by the noise in the data rather than the data itself. The solution to the overfitting problem is to use an integration method, and Random Forests are a good integration method. Random Forest (RF), as the name suggests, is a relatively new machine learning model that builds a forest in a randomized way, with many decision trees inside the forest, and each decision tree is independent. Compared to other existing classification algorithms, Random Forest has good regression and classification performance. From a computational point of view, random forests are attractive because they naturally handle regression and (multi-class) classification, are relatively fast to train and predict, and rely on only one or two tuning parameters. There is a built-in generalized error estimate that can be used directly for high-dimensional problems and can be easily executed in parallel. Statistically, random forests are attractive because they have additional features to offer, such as importance-variable metrics, differential classification weighting, missing value imputation, visualization, outlier detection, and unsupervised learning.

The specific implementation process is as follows: k samples are extracted from the initial training sample set N using bootstrap method $(k < N)$. Secondly, the corresponding decision tree models are built for the k extracted samples to obtain k classification results. Finally, the k sample results are voted on, and the minority is obeyed by the majority to get the final classification results. The classification decisions are:

$$H(x) = arg\max_y \sum_{i=1}^{k} I(h_i(x) = Y) \tag{6}$$

where, $H(x)$ is the combination classification model. h_i is the decision classification model. Y is the output variable (target variable). $I(h_i(x) = Y)$ is the schematic function.

II. B.SHAP Interpretability

SHAP is a method to address the interpretability of a model.The SHAP model is based on the Shapley value, a “game” where there are multiple individuals, each trying to maximize their own outcome. The method determines the importance of an individual by calculating the contribution of that individual to the cooperation, and the Shapley value is the average contribution of the eigenvalues to the prediction over all possible combinations.

$$\phi_i(f, x') = \sum_{z' \subseteq \{x'_1, \ldots, x'_n\} \{x'_i\}} \frac{|z'|!(M - |z'| - 1)!}{M!} * [f(z' \cup x'_i) - f(z')] \tag{7}$$

z' denotes the feature set, x' denotes the feature of instance x, and M denotes the number of features. $f(z')$ denotes the predicted value, which is modeled by masking the i th feature and applying a random value of that feature in the calculation of that value.

The SHAP model represents the Shapley value as a class of additive feature imputation and represents the prediction as a linear function of the binary variable with the formula shown in (8):

$$g(z') = \Phi_0 + \sum_{i=1}^{M} \Phi_i x'_i \tag{8}$$

where M in $z' \in \{0,1\}^M$ is the number of features in the simplified input, $\Phi_i \in R$, Eq. x_i represents the i th sample, x_{ij} is the j th column of features in sample i, y_i is the prediction result of sample i, and y_{base} is the prediction average of all samples, then SHAP value obeys Eq. (9):

$$y_i = y_{base} + f(x_i, 1) + f(x_i, 2) + \cdots + f(x_i, k) \tag{9}$$

where, $f(x_i, j)$ is x_{ij} the SHAP value calculated through the SHAP model. That is if j is 1, then the value of $f(x_i, 1)$ represents the SHAP value of the first feature of the sample i which represents the contribution to the final prediction result y_i [21]. The SHAP value of each feature of the sample then represents how the feature affects the model's prediction result when it takes that value. That is, when the value of $f(x_i, 1)$ is greater than 0, it indicates that the feature has a strengthening effect on the predicted value of the model, and vice versa, it indicates that the feature has a weakening effect on the predicted value of the model.

The schematic diagram of SHAP model interpretation is shown in Figure 1. Machine learning models are generally a black box, such as a certain model to carry out some prediction tasks, first of all, the model input some known conditions (Age=65, Sex=F, BP=180, BM=40), and then the model according to the input to train, and ultimately, the trained model can get the condition output prediction results (Output=0.4), however, so the model can only get the However, this model can only get the final result, as for how the model is calculated internally, how the known conditions of the input affect the prediction result, we have no way to understand. The SHAP framework model helps us to solve these problems by understanding how the known conditions affect the final prediction and the

direction of the known conditions on the prediction.SHAP can be used for debugging models, assisting in feature engineering, guiding the direction of data collection, guiding people to make decisions, and building trust between models and people.SHAP can solve the problem of multicollinearity, which is the problem of taking into account the effect of not only a single variable but also the effect of the variables on each other. variables, but also the synergies between variables.

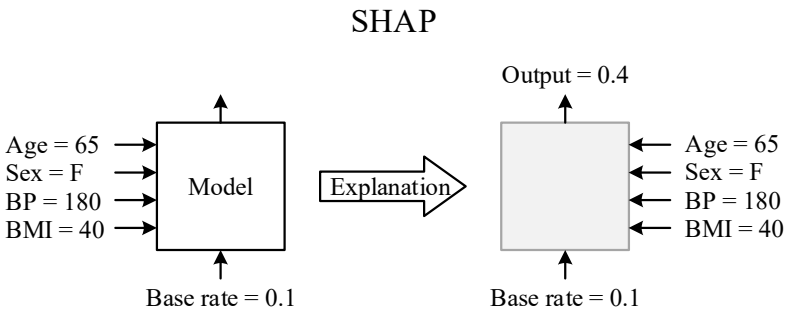


Figure 1: Schematic diagram of SHAP model explanation

III. Analysis of vocational skills needs for the future job market

III. A. Data sources

The official employment demand index used in this study comes from the CHER index released by the China Employment Research Institute in conjunction with the Zhilian Recruitment website, and the sample time period is from the second quarter of 2023 to the fourth quarter of 2024.The CIER index is calculated as follows: job supply index/employment demand index. From the formula of the index, it can be learned that the CIER index is negatively correlated with the employment demand index. The network search index can reflect the attention of the employed to the industry or the job, the higher the attention, to a certain extent, can indicate that the job supply of talent is large. In order to facilitate observation and comparison, the inverse time series of CIER index DS_CIER is taken as the variable to be predicted. In addition, according to the Prospect Industry Research Institute data show that at the end of 2024, Baidu search engine market share compared to 2023, although there is a decline, but still occupies the leading position in China's search engine market, the market share of 71%, so Baidu search index has a very significant representative. 2024 China search engine market share is shown in Figure 2.

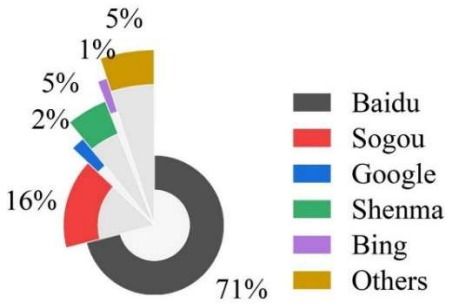


Figure 2: 2024 search engine market share

Therefore, the keyword web search data related to the demand for vocational skills in the future job market involved in this study are based on the Baidu Search Index as the data source. The Baidu index algorithm is specifically expressed as follows: the data is based on the search volume of netizens obtaining information in Baidu by typing in keywords, using keywords for statistical purposes, and using scientific methods to calculate the weighting of keywords' search frequency in Baidu web searches. Baidu search index can reflect the degree of Internet users' attention to keywords, and its changing trend. According to the difference of data source, search index exists pc search index and mobile search index, this study will be the sum of the two types of index as the search index data. In this paper, the Baidu search index of each keyword at different times is selected as the search data, and the python program is used to crawl the daily data of 45 keywords' comprehensive search volume on PC and mobile, the time span is from May 5, 2020 to December 31, 2024, with a total of 83,655 data points.

III. B. Data pre-processing

III. B. 1) Missing value processing

When dealing with missing values, there are two methods, the deletion method and the filling method. The deletion method, due to its simplicity and ease of implementation, performs better in cases where the percentage of missing values is relatively low. The fill method, on the other hand, fills in the missing values by some rule or method, such as mean, plurality, or great likelihood estimation, aiming to maintain the integrity and accuracy of the data. The missing values were viewed using python and the data set was missing as shown in Table 1.

Table 1: The situation of missing datasets

Variable	The number of missing values	Proportion of missing values
Education_Level	1522	0.1485
Marital_Status	753	0.1079
Income_Category	1106	0.7399

The feature variable Income_Category reflects the annual income status of users. For the treatment of its missing values, the plurality is selected as a substitute value, so as to keep the overall trend of the dataset stable. In addition, the characteristic variable Education_Level represents the educational level of the customer, which is divided into six different categories. A more refined approach was used to fill in the missing values of this category variable. Specifically, the proportions of each education level category in the total sample were first counted, and then missing values were assigned according to these proportions. This approach helps to preserve the original characteristics of the education level distribution. Finally, for the characteristic variable Marital_Status, a similar approach to Education_Level was used, whereby missing values were filled in based on the percentage of people in each marital status, thus ensuring the completeness and accuracy of the data. Next, the focus is on the data de-weighting process. Given that the column CLIENTNUM would not be applied in subsequent analyses, the choice was made to simply remove it for reasons of simplicity and redundancy reduction. This series of steps aims to ensure that the data is of a high quality level in order to guarantee the reliability of the results obtained in the construction of the predictive model.

III. B. 2) Imbalance data processing

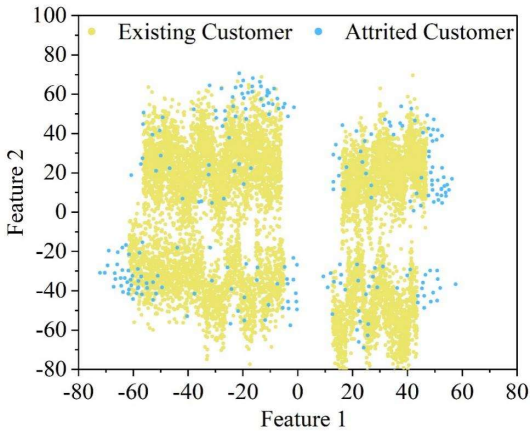
In performing categorical modeling, it is critical to ensure that the samples are balanced. This focuses on whether the number of samples from different categories is relatively equal. If the samples are unevenly distributed, it may trigger the problem of inaccurate model training or unreliable prediction results, while balancing the samples helps to prevent overfitting. In addition, checking the balance of the samples is also important for improving the robustness of the model. When the sample distribution is imbalanced, the model may have difficulty adapting to the new data environment. Therefore, by adjusting the number of samples to make them balanced, the robustness of the model can be effectively enhanced, which in turn improves the model's ability to adapt to new data. Problems such as customer churn and fraud prediction often involve the occurrence of a small number of events, in which churned customers usually occupy a minority. Due to this unbalanced distribution, the related datasets usually show an unbalanced state as well. In the case of the dataset in this paper, the churn-to-non-churn ratio is approximately 1:4, and the data is highly unbalanced and needs to be processed. Data imbalance treatment methods usually fall into two categories: undersampling and oversampling. The ultimate goal of both methods is to adjust the dataset so that the ratio of samples of different categories in it reaches 1:1. In this paper, we choose to use the SMOTE algorithm to deal with the imbalance of the dataset. After the SMOTE algorithm, the number of samples in the training set before and after the oversampling treatment is shown in Table 2. It can be seen that the ratio of the number of churned job-seeking users to the number of non-churned job-seeking users has successfully reached a balance of 1:1. This step plays a crucial role in data processing, effectively improving the imbalance of data distribution, which in turn improves the training effect and prediction accuracy of the model.

Table 2: The number of training set samples before and after oversampling processing

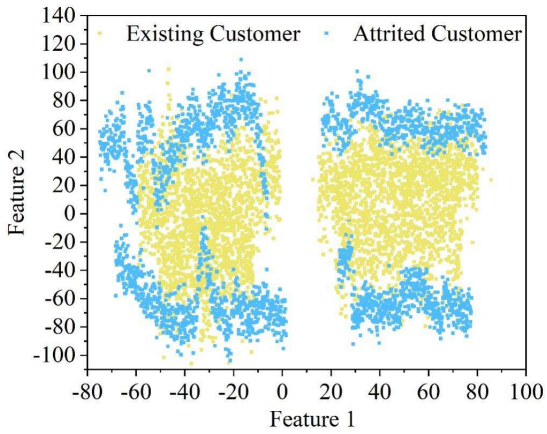
	No loss of job users	The number of loss of job-seeking users
Before the balancing process.	5936	1033
After balancing treatment	5836	5836

In order to more intuitively show the difference in the distribution of the samples of churned job-seeking users and non-churned job-seeking users before and after the oversampling process, the high-dimensional data is downsized

to two-dimensional space with the help of the t-SNE model, and then a scatter plot is drawn for the visual analysis, and the changes in the distribution of the samples before and after the oversampling process are shown in Fig. 3 (Fig. a shows the distribution of the samples before the oversampling process, and Fig. b shows the distribution of the samples after the oversampling process). From the figure, it can be clearly observed that after the SMOTE oversampling process, the sample distribution of churned job-seeking users and non-churned job-seeking users is more balanced. This change reflects the effectiveness of the oversampling method in balancing the distribution of categories in the dataset.



(a) The sample distribution before oversampling



(b) The distribution of samples after oversampling

Figure 3: The changes in sample distribution before and after oversampling

III. C. Data Processing and Keyword Index Synthesis

In this study, the inter-correlation analysis method is used to determine whether the keywords lead, lag or synchronize with the benchmark index data. Since it is necessary to forecast the employment demand index data, and the 45 keywords do not have the same lead-lag relationship with the future job market occupational skills demand index data, the leading keywords that have a predictive effect on the future job market occupational skills demand index data are needed to be extracted in this link. Combined with the interrelationship analysis method mentioned above, the interrelationship analysis results, i.e., the number of interrelationships between each word in the keyword word list and the DS_CIER index, are calculated by Eviews software. It should be noted that since the number of searches on the internet search engine reflects the attention of the job seeker to the position, the job seeker will search for it only if he/she wants to know about it. And searching for the position can indicate to a certain extent that the job seeker is interested in the position or wants to engage in the position, so the number of network searches should be positively correlated with the employment demand index, so the keywords that are negatively correlated with the employment demand index are excluded here. Secondly, among the correlation coefficients of the calculated results, the maximum value of the absolute value of the correlation coefficients between the keyword data and DS_CIER and the corresponding leading order are found. The leading keywords and their correlation coefficients are shown in Table 3. The results of the inter-correlation analysis of vocational skills demand keywords on DS_CIER index are shown in Table 4. The correlation analysis is shown in Figure 4. It can be seen that the absolute maximum value of the correlation coefficient between the keyword search data and the future vocational skills demand index data is 0.7265, which is located in the first three periods, so “core technology demand skills” is classified as the first prediction index.

Table 3: Leading keywords and their correlation coefficients

Key words	Leading order	Correlation coefficient
Core technical requirements skills	Lead by 2 ranks	0.7265
Communication skills	Lead by 8 ranks	0.5493
Complex problem solving skills	Lead by 8 ranks	0.7261
Emotional management skills	Lead by 9 ranks	0.5563
New tools application skills	Lead by 7 ranks	0.6931
Trade crossing skills	Lead by 8 ranks	0.5264

Table 4: The requirement keywords for professional skills are analyzed by the DS_CIER

Number of leading (lagging) periods	The correlation coefficient between the lag value and the DS_CIER index	The correlation coefficient between the leading value and the DS CIER index
1	0.6165	0.6508
2	0.593	0.7005
3	0.6301	0.7116
4	0.5488	0.7164
5	0.4869	0.7155
6	0.446	0.6845
7	0.3474	0.6862
8	0.3272	0.6549
9	0.2315	0.68391
10	0.2022	0.6629
11	0.1842	0.5873
12	0.1129	0.5065
13	0.094	0.4903
14	-0.016	0.4239
15	-0.0325	0.3397

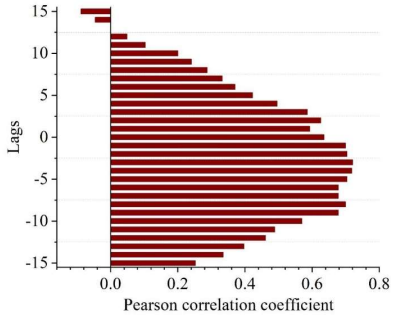


Figure 4: Cross-correlation analysis

Normally, a correlation coefficient above 0.5 can be regarded as weak correlation, and above 0.6 can be regarded as strong correlation. In order to ensure the relevance of the data and enhance the accuracy of the web search index in predicting the demand index for vocational skills in the future job market, this paper only selects keywords with correlation coefficients greater than 0.6, and finally obtains three leading keywords: core technology demand skills, cross-cutting skills in the industry, and communication skills. After determining the leading keywords, the index synthesis of keyword search data is carried out according to the misplaced weighted synthetic index method described in the previous section. Synthesizing the web search synthetic index first requires the keyword data to be misaligned by the corresponding number of periods, and the number of misalignments is the leading order of the keyword. The effect of fitting the web search synthetic index to the DS_CIER index trend is shown in Figure 5. Comparing the screened web search synthetic index with the actual DS_CIER index trend, it can be found that although the web search synthetic index can not be consistent with the DS_CIER index in every time period, the overall view of the two sequences have roughly the same trend.

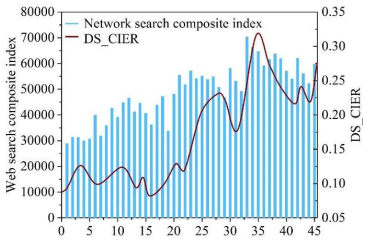


Figure 5: The trend fitting effect

III. D. Random Forest Predictive Modeling

III. D. 1) Eigenvalue Tuning

The construction of a random forest model for predicting the future market demand index of occupational skills is also implemented using R. The optimization of the model's parameters is implemented in two steps. Firstly, the optimization of the number of features of the decision tree model is carried out. In the regression model, we determine two variables, one is the index of employment demand index of the same period last year, and the second variable is the synthetic index of web search keywords. So the prediction model is constructed when the number of features is 1 and 2 respectively, and the mean square error of the model prediction results is calculated in both cases, and the model fitting when the number of variables is different is shown in Table 5. It can be found that when the number of variables is 2, the mean square error of the model prediction results reaches lower.

Table 5: The model of the variable number is consistent

The number of variables	1	2
Mean square error (RMSE)	0.036952	0.035122

III. D. 2) Quantitative Tuning of Decision Trees

The number of feature numbers of the random forest can be determined as 2. Next the number of decision trees in the random forest is optimized to determine the size of the random forest. A plot of the number of decision trees versus model error is shown in Figure 6. It can be seen that as the number of decision trees increases, it decreases rapidly and the model error stabilizes when the number of decision trees is greater than 500. Therefore the number of decision trees is set to 500.

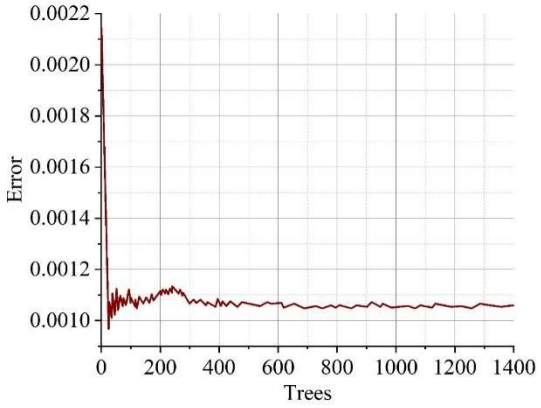


Figure 6: The relationship between the number of decision trees and model errors

III. D. 3) Optimal Random Forest Regression Prediction Models

After tuning the values of the eigenvalues and the number of decision trees of the random forest prediction model, the relevant information of the optimal random forest prediction model is shown in Table 6. The prediction object of this study is the value of employment demand index, so the category of the random forest model is regression, and the number of decision trees and the number of eigenvalues are selected after tuning to be 500 and 2, respectively.In addition, the residual squared mean value of the model is 0.001, which is small, which indicates that the model is fitted well. The goodness of fit of the explanatory rate of the variables is 79.23%, which is a good fit.

Table 6: Information related to the optimal random forest prediction model

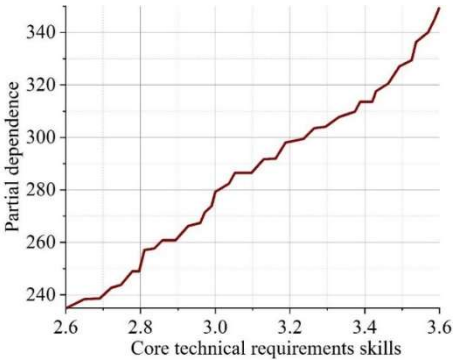
Random forest model type	Return	The number of decision trees	500
Mean squared residuals	0.001	The explanatory rate of the variable	79.23%

IV. Feature interpretability analysis

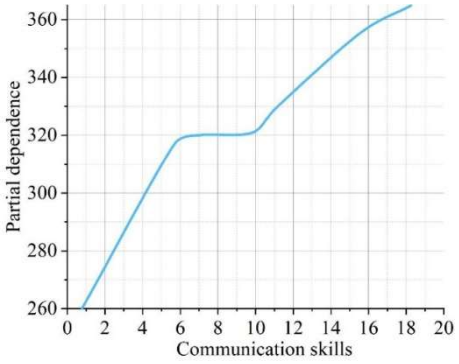
Based on the results of machine learning prediction results and feature importance analysis in Section 3, this chapter uses a total of 2 types of interpretable machine learning techniques, Partial Dependency Plots (PDPs) and SHAP, respectively, to conduct interpretability analysis of the factors influencing the demand for occupational skills in the future job market.

IV. A. PDP results

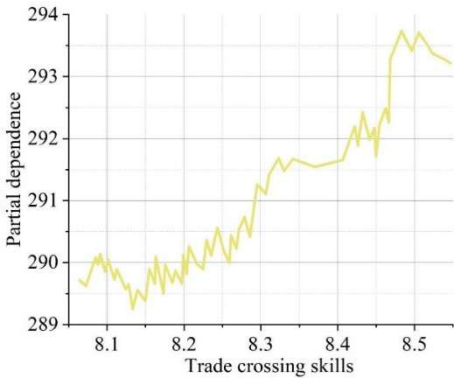
This paper employs a partial dependency plot (PDP) based on a random forest model to conduct an interpretability analysis of six feature variables influencing future occupational skill demand in the job market. To conduct an explainability analysis of a specific feature variable using PDP, it is necessary to ensure that the correlation between that feature and the remaining features is relatively low. First, PDP is used to conduct an explainability analysis of three important feature variables influencing future labor market skill demands. The PDP diagrams of the random forest model regarding important feature variables for future labor market skill demands are shown in Figure 7 (Figures a–c represent core technical skill demands, communication skills, and industry-crossing skills, respectively). Figure (a) shows that within the sample interval, career development increases as core technical skill requirements increase.



(a) Core technical requirements skills



(b) Communication skills



(c) Trade crossing skills

Figure 7: The PDP of the important eigenvariables of the random forest model

PDP's determination of the future job market demand for occupational skills by each machine learning model on the role of the characteristic variables is shown in Table 7 (+ means positive, - means negative). Because the machine learning model of the selected sample will be random sampling and lead to some bias, so its results have some differences. The results of the machine learning models using PDP to determine all the feature variables are: core technical skills, communication skills, complex problem solving skills, emotion management skills, new tool application skills, and interdisciplinary cross-skills, which are positively correlated with the demand for occupational skills in the future job market.

Table 7: The PDP determines the direction of the machine learning model

Variable	Random forest model	XGBoost model	LightGBM model	Comprehensive judgment
Core technical requirements skills	+	+	+	+
Communication skills	-	+	+	+

Complex problem solving skills	+	+	+	+
Emotional management skills	+	+	+	+
New tools application skills	+	+	-	+
Trade crossing skills	+	+	+	+

IV. B. SHAP results

The Random Forest model on the SHAP values of occupational skills for the future job market is shown in Figure 8. The coordinates of the horizontal axis correspond to the size of the SHAP values of individual features under all sample data points, and the coordinates of the vertical axis indicate the ordering of the SHAP values of all features, with the SHAP values decreasing from the top to the bottom. Secondly, the vertical axis serves as a positive and negative dividing line to indicate that there are positive and negative SHAP values for individual features under different sample points, with positive values representing the positive impact of individual features on corporate bond credit spreads under the sample points, and negative values representing the negative impact of individual features on the demand for occupational skills in the future job market under the sample points.

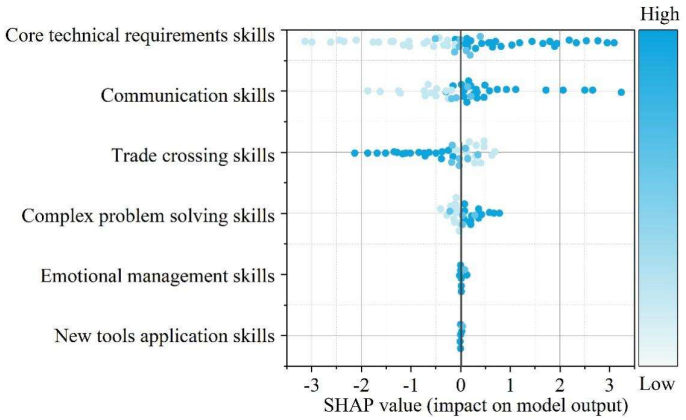


Figure 8: Random forest model is about the future job market career skill SHAP value

The ordering of the average SHAP values of all the features affecting the output of the Random Forest model is shown in Table 8, where the values are averaged over all the sample points for each feature, and it can be seen that the average SHAP value of the core technology needs is the largest, and that the average SHAP value of the skills of applying new tools is the smallest.

Table 8: The ranking of the average SHAP values of all features that affect the output

Variable	Average SHAP value	Importance ranking
Core technical requirements skills	0.102269	1
Communication skills	0.032896	2
Trade crossing skills	0.015822	3
Complex problem solving skills	0.005278	4
Emotional management skills	0.004312	5
New tools application skills	0.004536	6

Combined with the graph and table, it can be seen that the differences in core technology demand skills, communication skills and interdisciplinary skills application abilities have the most significant impact on the model, with average SHAP values of 0.102269, 0.032896 and 0.015822, respectively.In addition, the graph results also show that core technology demand skills, communication skills and interdisciplinary skills application abilities have a significant positive impact on the future job market occupational skill demand, and their SHAP values increase as the eigenvalues increase.

V. Summary

In the context of the big data era, this article big data analysis method for the future job market occupational skills demand prediction, the results of the study for the current talent training has a positive reference significance. The conclusions of the article are as follows:

The relevant information can be found through the optimal random forest prediction model. The residual squared mean of the model is 0.001, which is a small value. The goodness of fit of the explanatory rate of the variables is 79.23%, which indicates that the model has a good fit.

By sorting the average SHAP values of all the features affecting the output of the random forest model, it can be concluded that the average SHAP value of core technology demand can be seen to be the largest at 0.102269, thus indicating that the mastery of professional core technology is an important factor in improving the competitiveness of job seekers in the future job market.

Funding

This research was supported by the Chief Technician Training Program of the Xingdian Talent Program.

References

[1] Choi, M., Cho, K. J., & Kim, M. S. (2018). Analysis on Regional and Industrial Disparity of Employment in Korea. Journal of Korean Society of Industrial and Systems Engineering, 41(4), 34-41.

[2] Kumar, S., & Pattanaik, F. (2020). Regional disparities in employment intensity of Indian industries: A state-level analysis. Emerging Economy Studies, 6(1), 23-38.

[3] Marukawa, T. (2017). Regional unemployment disparities in China. Economic Systems, 41(2), 203-214.

[4] Yang, S., Yang, J., Yue, L., Xu, J., Liu, X., Li, W., ... & He, G. (2022). Impact of perception reduction of employment opportunities on employment pressure of college students under COVID-19 epidemic–joint moderating effects of employment policy support and job-searching self-efficacy. Frontiers in Psychology, 13, 986070.

[5] Figueiredo, H., Biscaia, R., Rocha, V., & Teixeira, P. (2017). Should we start worrying? Mass higher education, skill demand and the increasingly complex landscape of young graduates' employment. Studies in Higher Education, 42(8), 1401-1420.

[6] Cao, Y., Cheng, S., Tucker, J. W., & Wan, C. (2023). Technological peer pressure and skill specificity of job postings. Contemporary Accounting Research, 40(3), 2106-2139.

[7] Lee, S. H. (2020). Skills development driven by labor market demand. Anticipating and Preparing for Emerging Skills and Jobs: Key Issues, Concerns, and Prospects, 271-277.

[8] CHUMWATANA, T., & HPONE, A. K. K. (2025). BRIDGING THE IT SKILL GAP WITH INDUSTRY DEMANDS: AN AI-DRIVEN TEXT MINING APPROACH TO JOB MARKET TRENDS USING LARGE LANGUAGE MODEL. Journal of Theoretical and Applied Information Technology, 103(6).

[9] Alibasic, A., Simsekler, M. C. E., Kurfess, T., Woon, W. L., & Omar, M. A. (2020). Utilizing data science techniques to analyze skill and demand changes in healthcare occupations: case study on USA and UAE healthcare sector. Soft Computing, 24, 4959-4976.

[10] Pavlova, M. (2019). Emerging environmental industries: impact on required skills and TVET systems. International Journal of Training Research, 17(sup1), 144-158.

[11] Przytuła, S. (2018). Global labor market trends and their significance for the future employees' competences. Journal of Intercultural Management, 10(4), 5-38.

[12] Yahya, A. E., Yafooz, W. M., & Gharbi, A. (2024). Mapping Graduate Skills to Market Demands: A Holistic Examination of Curriculum Development and Employment Trends. Engineering, Technology & Applied Science Research, 14(4), 14793-14800.

[13] Akyazi, T., del Val, P., Goti, A., & Oyarbide, A. (2022). Identifying future skill requirements of the job profiles for a sustainable European manufacturing industry 4.0. Recycling, 7(3), 32.

[14] Nomura, S., Imaizumi, S., Areias, A. C., & Yamauchi, F. (2017). Toward labor market policy 2.0: the potential for using online job-portal big data to inform labor market policies in India. World Bank Policy Research Working Paper, (7966).

[15] Wan, J., Chen, B., & Si, H. (2017, August). Mining and measurement of vocational skills and their association rules based on big data. In Proceedings of the 1st International Conference on Digital Technology in Education (pp. 59-63).

[16] Vankevich, A., & Kalinouskaya, I. (2021). Better understanding of the labour market using Big Data. Ekonomia i prawo. Economics and law, 20(3), 677-692.

[17] Hoque, M. R., & Islam, M. A. (2021). Predicting Job Skills in Demand: A Big Data Approach. Dhaka University Journal of Business Studies, 221-239.

[18] Wei, Y., Zheng, Y., & Li, N. (2023). Big Data Analysis and Forecast of Employment Position Requirements for College Students. International Journal of Emerging Technologies in Learning, 18(4).

[19] Maer-Matei, M. M., Mocanu, C., Zamfir, A. M., & Georgescu, T. M. (2019). Skill needs for early career researchers—a text mining approach. Sustainability, 11(10), 2789.

[20] Credit Kevin. (2021). Spatial Models or Random Forest? Evaluating the Use of Spatially Explicit Machine Learning Methods to Predict Employment Density around New Transit Stations in Los Angeles. Geographical Analysis,54(1),58-83.

[21] Zhizheng Wu,Shengzheng Wang,Leyao Li & Yongfeng Suo. (2025). An interpretable ship risk model based on machine learning and SHAP interpretation technique. Ocean Engineering,335,121686-121686.