Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2022.030404
ISSN 2790-0983 Vol. 3, Issue 4: 27-35

Predicting the Employment Destinations of University Students Based on Machine Learning Algorithms

Weijun Xie[*]
Qinghai Normal University, Qinghai, China
xieweijun520@126.com
[*]*corresponding author*

Keywords: Machine Learning, University Students, Employment Destinations, Employment Prediction

Abstract: The research on the classification prediction of students' employment destinations in higher education not only opens up new application areas for classification algorithms, but also represents a new attempt to introduce machine learning algorithms into the analysis of employment guidance and the development of teaching systems. The purpose of this paper is to study the employment destination prediction of college students based on machine learning algorithms. The factors influencing the employability of university students are analysed and the XGBoost model in decision trees is explored. A graduate employment prediction algorithm based on HMIGW feature selection and XGBoost algorithm is proposed to predict the employment situation of the class of 2022 and the type of employment, and the experimental results show that the algorithm is able to obtain relatively accurate conclusions on graduate employment prediction.

1. Introduction

Currently, as China's higher education enters a new stage of popularity, students are also facing an unemployment crisis after graduation. The reason is that, besides the influence of the expansion of colleges and universities and changes in China's labour market demand, the obvious lack of employment guidance for vocational college students is also an important factor [1-2]. How to improve the employment rate and employment quality of students is a very critical and urgent task [3]. If we can analyze and mine the data of higher vocational students' cultivation process to get the factors affecting students' employment and the prediction model of employment, the career guidance departments of higher vocational institutions can carry out targeted career guidance work according to the mined model [4-5].

Nooshin Pordelan has incorporated the advanced technology of the "Internet+" era into school-enterprise cooperation in employment, enabling schools and enterprises to forecast talent needs, build professional incubation platforms, publish and track employment trends, and provide employment guidance to students. Jo ão M. Fernandes analysed institutional performance in relation to student satisfaction and their readiness for future employment. The questionnaire was designed for students eligible to graduate and the survey was administered through a student portal on the university's website, to which a total of 750 undergraduate students seeking a degree (the target population) were invited to participate. Findings The descriptive results of this study suggest that while student satisfaction with all academic programmes may be relatively similar, there are differences in perceived career aspirations based on the academic programme chosen. Most notably, the findings also suggest that students' career expectations did not negatively affect their satisfaction with their higher education institution (HEI) [7]. Therefore, it is of relevance to study the prediction of employment destinations of HEI students based on machine learning algorithms.

Based on the employment data of graduates from the Department of Finance of College M in the past four years, this paper collects students' employment data through the academic affairs system to discover the influence of students' place of origin, on employment units and employment areas, etc., analyses the accuracy of the prediction rate of the algorithm in this paper, and explores the information that has guidance value for the employment work in colleges and universities, so as to provide education departments and college employment guidance and other relevant departments and school leaders in carrying out employment guidance work, to The paper also provides valuable information for the design of talent training programmes and employment decisions.

2. Research on the Prediction of Employment Destinations of University Students Based on Machine Learning Algorithms

2.1. Factors Influencing the Employability of College Students

There are many factors that influence the employability of university students, and employability is divided according to the influence of the internal and external environment, one being the internal factors influenced by the graduate's personal and family, and the other being other external factors such as social employment trends and the macroeconomic environment of the school [8]. Of these, the trainee's ability to organise his or her skills and the school work guidance obtained through the competence factor play an irreplaceable role in employment.

2.2. Decision Trees

A decision tree is essentially a generator for category classification. The algorithm for building a decision tree classifier is usually divided into two steps, one is to construct a decision tree and the other is to prune the constructed decision tree [9-10].

At the beginning, all data mining samples are in the root node, and using a recursive approach, the samples are divided by specified attributes, and a corresponding number of branches are developed based on the attribute value, which must be discrete, or if the attribute value is continuous, it must first be discretized by processing this attribute value to make it discrete [11-12]. In the second stage of decision tree pruning, it is to cut out the isolated points and noise generated in the training data, so that the constructed decision tree can be compared one by one with each attribute value of the sample data, so as to achieve the purpose of classifying the unknown sample data and mining it [13].

2.3. XGBoost

XGBoost is a library of model implementation tools based on gradient stepping decision trees. In the XGBoost model, it uses an iterative increasing training method in order to train the decision tree. The core idea is that each time we train a tree it is the tree from the previous training plus the tree from the current training, which can be understood as each training step corrects the results of the previous training step and adds the corrected results as part of this training step [14].

XGBoost itself has a large number of adjustable superparameters, and we cannot perform multiple trials to adjust the values for each parameter, so we only select a few key parameters for tuning, and take their default values for all other parameters. The hyperparameters we selected for tuning are the learning rate, the number of iterations, and the total number of trees. The learning rate is the number of steps along the gradient that are updated each time in the optimisation algorithm. The number of iterations is the total number of iterations required in the training. The total number of trees is the number of subtrees that need to be created in the XGBoost algorithm [15].

2.4. Feature Selection Algorithm

The process of feature selection is to extract all the features from the dataset, with the total number of features denoted as N. From the N features, the most effective set of features for solving the problem is selected according to certain rules and becomes the optimal feature subset, with the size of the optimal feature subset denoted as n ($n<N$).

The process of feature selection is one in which some features are combined into a feature subset from all features of the initial data set, and the selected feature subset will make some performance evaluation value obtain the optimum [16-17]. In the early stages of research on feature selection algorithms from a statistical and data processing perspective, there was relatively more research, and most problems involved a relatively limited number of features in the dataset. With the rapid development of big data on the Internet, the amount of data from all walks of life is getting larger and larger, and the reasonable selection of feature selection algorithms is becoming more and more critical [18].

3. Investigation and Research on the Prediction of Employment Destinations of University Students Based on Machine Learning Algorithms

3.1. Student Data Collection

In this paper, the data collected are the school data of 118 undergraduate graduates from the Finance Department of M College Class of 2022. For these raw behavioural data, they are characterised by scattered sources, unstructured and heterogeneous. We need to pre-process the raw data to ensure data quality, data availability and strengthen the correlation between data, with a view to formulating a specification of student multidimensional behavioural data in the digital campus environment based on the existing data standards and specifications.

The data sources for this study include the Digital Academic System (DAS) and the Web Centre. Among them, the digital academic system mainly holds information such as students' basic information and subject grades. According to the HMIGW algorithm proposed in this paper, the relevance valuation I of each feature and category Y is calculated separately as shown in Table 1.

Table 1. Relevance Estimation I for Each Character and Category Y

Characteristic attribute	I	Characteristic attribute	I
Comprehensive achievements	578	registered residence	201
Professional achievements	376	TOEFL/IELTS/GMAT	254
Number of clubs	499	Gender	16
Professional popularity	482	School category	18
Political outlook	211	Nominal family	4

3.2. Graduate Employment Prediction Algorithm Based on HMIGW Feature Selection and XGBoost

(1) Algorithm flow

The algorithm flow is shown specifically in Figure 1, where data collection is carried out followed by data pre-processing, which involves the following processes:missing values, outliers, etc. The employment prediction model is established by successfully constructing the dataset.

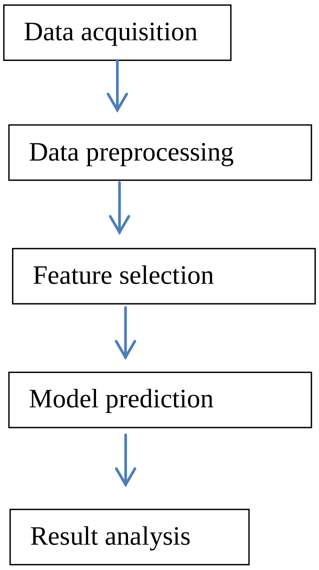

Figure 1. Algorithm flow

(2) Feature selection algorithm HMIGW

The algorithm in this paper includes two stages, filtering and wrapping, respectively as follows:

Screening for redundant and irrelevant features, and calculating an information measure for each feature and evaluating correlation Ix.

For data series, X=(X1, X2,... Xm) entropy calculation formula is as follows:

$$H(X) = -\sum_{x_i \in X} P(x_i) \log(p(x_i)) \tag{1}$$

Where p(xi) represents the probability density of xi in its parent data sequence.

The entropy is positively correlated with the degree of stability between variables, and the greater the mutual information is, the closer the variables are. Mutual information can be expressed in the form of entropy as shown in Equation 2:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2}$$

I remember (X; Y) is the correlation valuation Ix of feature X.

Feature selection strategy.

The action space operations are sorted in descending order based on the Ix evaluation obtained in the filtering, and the feature space is added by a direct policy. Each time a feature is added, a feature and its corresponding X1, X2,... , Xm(m is the operand included in the subset of operations), then the prediction is made using the XGBoost algorithm, and if AI < AI-1, the execution is repeated until the end.

4. Analysis and Research of College Students' Employment Direction Prediction Based on Machine Learning Algorithm

4.1. Algorithm Comparison

Table 2 lists the feature selection results using the hmigw resource selection algorithm in the dataset, and then compares the results after five cross-checks using the random Forest and XGBoost algorithms, respectively.

Table 2. Comparisons between XGBoost and random forest model

Model name	Precision	Recall	F1 value	training time (s)
XGBoost	0.93	0.94	0.91	0.06
Random forest	0.88	0.89	0.88	0.10

Figure 2 shows the accuracy of random Forest algorithm and XGboost algorithm in each of the five times of cross-validation. The abscissa represents the serial number of cross-validation, and the ordinate is the accuracy.

Combined with the ratio of different indexes of the model trained by two different algorithms in Table 2, from the three aspects of prediction accuracy, recall and 1F value, it can be concluded that the practicality of XGBoost model is closer to the real value than the random forest model.

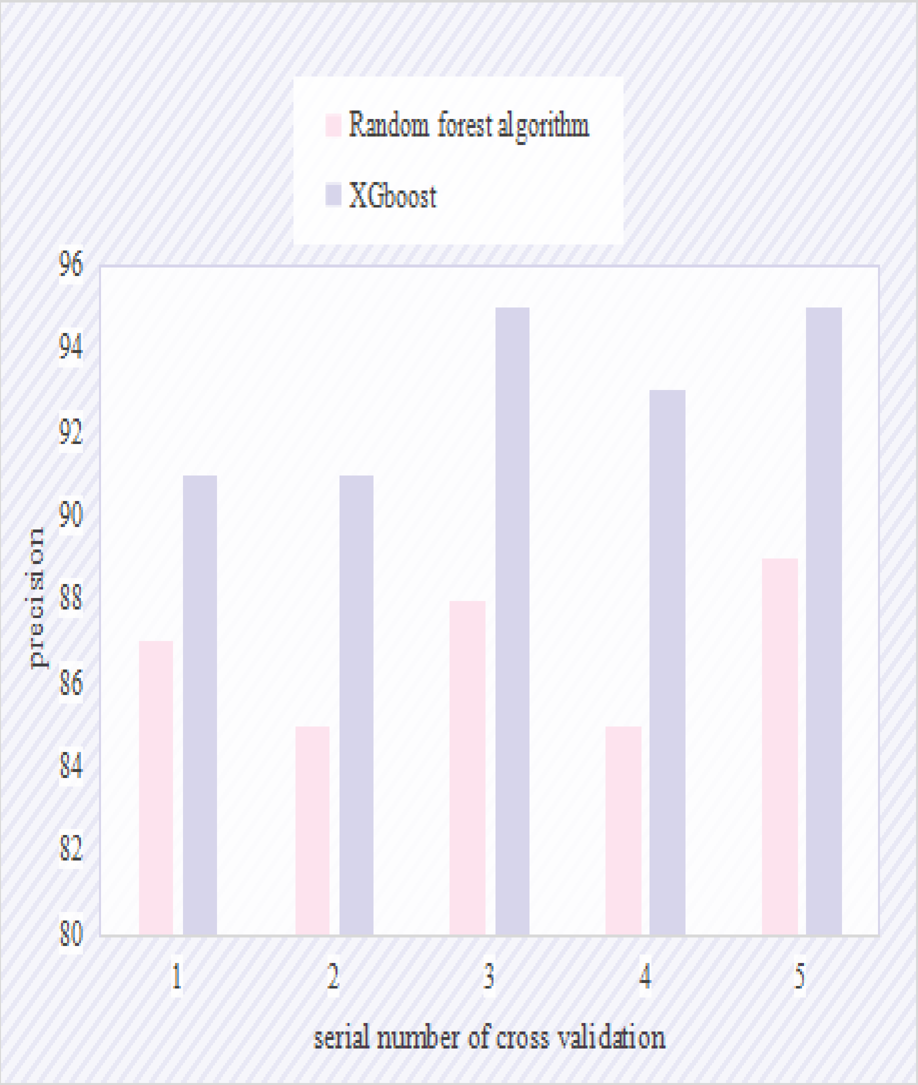

Figure 2. Comparison of XGBoost and random forest performance based on HMIGW

4.2. Application Analysis of Employment Forecast Model

The role of employment prediction model is to predict the employment situation of college students, to understand the employment trend of students in turn, and make corresponding countermeasures. The more time away from students to prepare for employment, the more accurate the prediction results, the more conducive to education personnel career guidance. Therefore, it is of great significance to study the accuracy of employment prediction model in various stages of college students' employment data. Firstly, the data of the students of grade 2022 were divided into 8 data sets according to the time series from the first semester to the eighth semester, and the employment prediction model was used for prediction, and the accuracy of the model was compared with the real results, as shown in Figure 3.

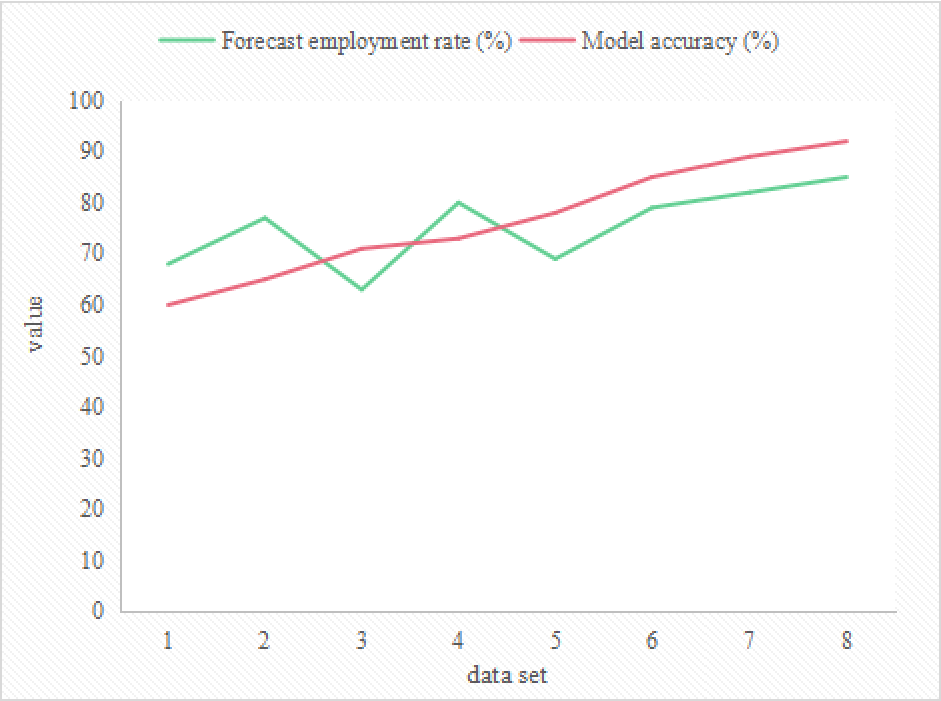

Figure 3. Dynamic Chart of the Predicted Employment Value and Accuracy of 2022 Students

As can be seen from the figure, with the study of each semester, the accuracy of predicting students' employment shows an overall upward trend. Among them, the unstable prediction results in the first four semesters are due to the incomplete data in all aspects of students just entering college life, which tends to be stable after the fifth semester.

5. Conclusion

This article USES the present hot machine learning algorithm to the traditional data management system for data analysis, on the test set can get employment situation forecast accuracy at more than 90% of the test results, it further demonstrates the use of machine learning algorithms for data analysis and forecast the use value and application prospect in the field of employment. Model on the test data set effect is good, but because of inadequate sample size, model no way to further improve, if applied to the reality in the future you can use more training samples to adjust parameters, at the same time can be further subdivided according to the demand parameters, can get a more detailed prediction result makes the system more available.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Ishwank Singh, A. Sai Sabitha, Tanupriya Choudhury, Archit Aggarwal, Bhupesh Kumar Dewangan: Mapping Student Performance with Employment Using Fuzzy C-Means. Int. J. Inf. Syst. Model. Des. 11(4): 36-52 (2020) https://doi.org/10.4018/IJISMD.2020100103

[2] Theodoros Lappas: Mining Career Paths from Large Resume Databases: Evidence from IT Professionals. ACM Trans. Knowl. Discov. Data 14(3): 37:1-37:38 (2020) https://doi.org/10.1145/3379984

[3] Sophia Alim, Ibrahim AlShourbaji: Professional uses of Facebook amongst university students in relation to searching for jobs: an exploration of activities and behaviours. Int. J. Soc. Media Interact. Learn. Environ. 6(3): 200-229 (2020) https://doi.org/10.1504/IJSMILE.2020.109225

[4] Alexander A. McKenzie: David Corydon Walden's Five Careers. IEEE Ann. Hist. Comput. 44(3): 70-79 (2022) https://doi.org/10.1109/MAHC.2022.3193812

[5] Andrea Delgado, Verónica G. Melesse Vergara, Andrea Schneibel: Careers in STEM: A Latina Perspective. Comput. Sci. Eng. 24(3): 81-85 (2022) https://doi.org/10.1109/MCSE.2022.3188195

[6] Nooshin Pordelan, Simin Hosseinian, Hamid Heydari, Sadaf Khalijian, Mohammad Khorrami: Consequences of teleworking using the internet among married working women: Educational careers investigation. Educ. Inf. Technol. 27(3): 4277-4299 (2022) https://doi.org/10.1007/s10639-021-10788-6

[7] João M. Fernandes, António Costa, Paulo Cortez: Author placement in Computer Science: a study based on the careers of ACM Fellows. Scientometrics 127(1): 351-368 (2022) https://doi.org/10.1007/s11192-021-04035-5

[8] Marek Kwiek, Wojciech Roszka: Academic vs. biological age in research on academic careers: a large-scale study with implications for scientifically developing systems. Scientometrics 127(6): 3543-3575 (2022) https://doi.org/10.1007/s11192-022-04363-0

[9] D. Venkata Vara Prasad, Srinivas Gumparthi, Lokeswari Y. Venkataramana, S. Srinethe, R. M. Sruthi Sree, K. Nishanthi: Prediction of Stock Prices Using Statistical and Machine Learning Models: A Comparative Analysis. Comput. J. 65(5): 1338-1351 (2022) https://doi.org/10.1093/comjnl/bxab008

[10] Praveen Lalwani, Manas Kumar Mishra, Jasroop Singh Chadha, Pratyush Sethi: Customer churn prediction system: a machine learning approach. Computing 104(2): 271-294 (2022) https://doi.org/10.1007/s00607-021-00908-y

[11] Amirpasha Mozaffari, Michael Langguth, Bing Gong, Jessica Ahring, Adrian Rojas Campos, Pascal Nieters, Otoniel José Campos Escobar, Martin Wittenbrink, Peter Baumann, Martin G. Schultz: HPC-oriented Canonical Workflows for Machine Learning Applications in Climate and Weather Prediction. Data Intell. 4(2): 271-285 (2022) https://doi.org/10.1162/dint_a_00131

[12] Filippos Giannakas, Christos Troussas, Akrivi Krouska, Cleo Sgouropoulou, Ioannis Voyiatzis: Multi-technique comparative analysis of machine learning algorithms for improving the prediction of teams' performance. Educ. Inf. Technol. 27(6): 8461-8487 (2022) https://doi.org/10.1007/s10639-022-10900-4

[13] Muhammad Shoaib, Nasir Sayed, Nedra Amara, Abdul Latif, Sikandar Azam, Sajjad Muhammad: Prediction of an educational institute learning environment using machine learning

and data mining. Educ. Inf. Technol. 27(7): 9099-9123 (2022) https://doi.org/10.1007/s10639-022-10970-4

[14] Abdolreza Nazemi, Friedrich Baumann, Frank J. Fabozzi: Intertemporal defaulted bond recoveries prediction via machine learning. Eur. J. Oper. Res. 297(3): 1162-1177 (2022) https://doi.org/10.1016/j.ejor.2021.06.047

[15] Marwah Sattar Hanoon, Alharazi Abdulhadi Abdullatif B, Ali Najah Ahmed, Arif Razzaq, Ahmed H. Birima, Ahmed El-Shafie: A comparison of various machine learning approaches performance for prediction suspended sediment load of river systems: a case study in Malaysia. Earth Sci. Informatics 15(1): 91-104 (2022) https://doi.org/10.1007/s12145-021-00689-0

[16] Mohammed Zubair M. Shamim, Sattam Alotaibi, Hany S. Hussein, Mohammed Farrag, Mohammad Shiblee: Diagnostic Accuracy of Smartphone-Connected Electrophysiological Biosensors for Prediction of Blood Glucose Level in a Type-2 Diabetic Patient Using Machine Learning: A Pilot Study. IEEE Embed. Syst. Lett. 14(1): 27-30 (2022) https://doi.org/10.1109/LES.2021.3096717

[17] T. Vaisakh, R. Jayabarathi: Analysis on intelligent machine learning enabled with meta-heuristic algorithms for solar irradiance prediction. Evol. Intell. 15(1): 235-254 (2022) https://doi.org/10.1007/s12065-020-00505-6

[18] Abdalrhman Milad, Sadaam Hadee Hussein, Ahlam R. Khekan, Mohammed Rashid, Haitham Al-Msari, Tan Huy Tran: Development of ensemble machine learning approaches for designing fiber-reinforced polymer composite strain prediction model. Eng. Comput. 38(4): 3625-3637 (2022) https://doi.org/10.1007/s00366-021-01398-4