EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

A credible predictive model for employment of college graduates based on LightGBM

Yangzi He[1], Jiawen Zhu[2], Weina Fu[1,3,*]

[1]College of Information Science and Engineering, Hunan Normal University, Changsha, 410081, Hunan Province, China
[2]School of Education and Science, Hunan Normal University, Changsha, 410081, Hunan Province, China
[3]Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha, 410081, Hunan Province, China

Abstract

INTRODUCTION: "Improving the employment rate of college students" directly affects the stability of the country and society and the healthy development of the industry market. The traditional graduate employment rate model only predicts the future employment rate based on changes in historical employment data in previous years.
OBJECTIVES: Quantify the employment factors and solve the employment problems in colleges and universities in a targeted manner.
METHODS: We construct a credible employment prediction model for college graduates based on LightGBM.
RESULTS: We use the model to predict the employment status of students and obtain the special importance which is important to employment of college students.
CONCLUSION: The final result shows that our Model performs well in the two indicators of accuracy and model quality.

Keywords: employment rate of college students, predict model classification, characteristics prediction Accuracy.

Received on 31 December 2021, accepted on 15 February 2022, published on 17 February 2022

doi: 10.4108/eai.17-2-2022.173456

* Corresponding author. Email: fuwn@hunnu.edu.cn

1. Introduction

In recent years, with the expansion of college enrolment, higher education has gradually transformed into popular education. On the one hand, more people have the opportunity to receive higher education, which improves the overall quality of the people; on the other hand, it also creates greater employment pressure for college students. As the national production and living environment has been affected by the new crown epidemic in recent years, many small and medium-sized enterprises are unable to maintain it. Under this situation: How to predict the relationship between the employment of college students and the trend of national development strategies, so that colleges and universities can adjust their employment strategies and guide college students in a timely manner correct employment values and guide college students on the right path to employment. Successful employment is not only a focus of attention of all sectors of society, but also a livelihood issue related to millions of individuals and families.

Predicting the employment rate of college students is an important research aspect of graduate employment quality evaluation. With the continuous development of deep learning, the research on employment rate prediction has become diverse: for example, Xi et al. have used neural networks to dynamically adjust parameters and successfully predicted the employment direction of parts

in Germany[1]. Wang et al. used a basic decision tree algorithm to construct an employment prediction model based on the behavioral data of college students[2]. Qi's research on the employment rate prediction of college students based on the gray system uses the gray system to fit past data to obtain the prediction results of the employment rate of college students[3].

In practical applications, because of the existence of a variety of important employment factors, how can they be better applied to the problem of college students' employment rate prediction. At present, the data of previous years are mainly used for fitting. Although this prediction method is relatively simple and direct, and the amount of calculation is small, there are also problems such as small amount of data and over-fitting. Information such as activities and innovative practices are excavated and analysed to predict the future employment destination of students, and then conduct employment guidance and decision-making. Most of these methods ignore the increasing speed of student data samples and the diversification of employment information, so it is difficult to handle large-scale complex data. Secondly, in the past, data mining methods often ignored important factors affecting employment, and only focused on the overall accuracy of prediction. Although these unemployed data samples accounted for a small proportion, it was the university employment department that made decisions that helped increase the employment rate. Focus. At the same time, the forecast results are inaccurate in the overall social employment environment.

Therefore, this article introduces a variety of important factors into the employment rate prediction model, taking into account the complementary importance of the impact factors, through the important feature information in the important factors, and at the same time through the mutually exclusive feature bundling through LightGBM to predict the employment rate. The model constructed by the method can better reflect the employment situation. The experimental verification shows that the model proposed in this paper effectively overcomes the problem of over-fitting the prediction results in traditional methods.

The main contributions of this paper are as follows.

(i) This paper proposes a novel method for predicting the employment rate of graduates. By analysing multiple classification feature information in students' personal information, binding mutually exclusive features to enable modal interaction information, thus constructing employment based on LightGBM Forecasting system, in which the distributed gradient boosting framework based on decision tree algorithm can effectively coordinate the bundling of certain features to reduce the dimensionality of features, reduce the consumption of finding the best segmentation point, and improve the prediction accuracy and efficiency of the prediction model.

(ii) Based on the specificity of students' employment situation, this paper proposes a personal employment prediction model based on multi-feature fusion. Taking into account the strong relevance of the characteristics related to the employment of students, this paper is based on the multi-feature fusion model to capture the features and the feature classification that has a strong correlation with the employment label. Obtaining these more significant influencing factors can give related suggestions increasing the employment rate.

(iii) The LightGBM graduate employment prediction credible system proposed in this paper is verified in the employment dataset. The results show that this method can effectively improve the accuracy of graduate employment rate prediction. At the same time, the system will promote the further expansion of employment rate prediction in the field of smart education management, help college employment departments understand the existing employment advantages of students, screen out students with difficulties in finding employment, timely adjust employment training strategies, promote graduates' motivation for employment, and improve the overall employment rate of graduates in colleges and universities.

2. Related work

In this section, we mainly review the related work of employment prediction and category feature theory related to this article.

2.1. Application of predictive model in the field of intelligent education management

In order to deal with college students and their complex and severe forms of employment, follow the national implementation of the "Internet +" action plan and the national big data strategy, improve college employment management, guidance and service levels, and gradually develop personalized and customized employment services and the "smart employment" of precise employment management and guidance is the only way to use modern information technology to lead the employment of college students to a new level. Under the current employment environment, pre-judgment of employment situations has an important impact on colleges' adjustment of employment training strategies and the promotion of college's employment rate.

On the one hand, since the reform and opening up, colleges and universities have expanded their enrollment year by year, and the contradiction between college students' employment and social needs has increased day by day. In addition, the number of graduates has repeatedly set new highs in the past few years, and the

imbalance between supply and demand in the human resource market has directly caused the current severe employment forms. On the other hand, the contradiction between supply and demand in the job market is not only reflected in the contradiction between the number of people and the number of occupations, but also refers to a structural disorder. For example, graduates cannot find ideal jobs, while some work units cannot recruit suitable personnel. Therefore, how to predict the student's employment probability based on the student's personal information plays a very important role in job selection and employment in a changing employment environment. How to accurately obtain the student's employment probability has become the essential strategy for the employment management department of colleges and universities to formulate strategies and implement the employment training process.

For example, Hou et al. proposed a predictive model of college student performance based on educational data, using students' basic attributes and prerequisite course performance as features to predict students' academic performance[4]. Shi et al. proposed a study on the impact of student behavior on academic performance based on data mining, mining the relationship between campus behavior habits and academic performance, using the daily behavior data of students on campus to model, and analyzing the classification results from the overall and performance levels of students' behavioral characteristics predict students' academic performance[5]. Xia et al. proposed a college student employment prediction model based on campus big data, using campus big data such as one card to construct a prediction model to classify student employment status[6].

2.2. Categorical characteristics

In regression, classification, clustering and other machine learning algorithms, the calculation of distance between features or the calculation of similarity is very important. When extracting text features, using one-hot encoding will extend the value of discrete features to Euclidean space. Among them, a certain value of the discrete feature corresponds to a certain point in Euclidean space[7]. Using one-hot encoding for discrete features will make the distance calculation between features more reasonable. When classifying category features, one-hot code is often used as a bag-of-words model, which does not consider the order between words[8]; secondly, it assumes that words are independent of each other, and often words are independent of each other. Influencing each other from time to time. Therefore, the final features obtained by using one-hot encoding are discrete and sparse. For example, a list of features [0,1,0,1][0,1,0,1], which means that it has two values 0 or 1, then one-hot will use two bits to represent this feature, [1,0][1,0] means 0, [0,1][0,1 means 1, the first two [1,0...][1,0...] in the output of the above example are also It means that the feature is 0.

Gradient-Boosting-Decision-Tree (GBDT) is a popular machine learning algorithm with many effective implementations, such as XGBoost and PGBRT. But when the feature dimension is high and the amount of data is large, these two algorithms are still unsatisfactory in terms of efficiency and scalability. On the one hand, for each feature, they need to scan all data instances to estimate the information gain of all possible segmentation points, which is very time-consuming. G et al. proposed an efficient gradient boosting decision tree algorithm. Experiments on multiple public datasets show that under the condition of achieving almost the same accuracy, the training process using LightGBM is more than 20 times faster than the traditional GBDT[9].

3. Proposed theory

3.1. Insufficient utilization of dataset information

3.1.1. Insufficient utilization of dataset information

Affected by the epidemic, the employment environment at home and abroad has been impacted both large and small. There are many uncertain factors in the employment rate, and the development law of the employment rate does not necessarily follow the historical trajectory. Jobs in certain industries may be affected by national policies, and the demand may rise or fall significantly. However, existing models often predict the employment rate in the next few years by fitting based on past employment history data, or they do not take into account important employment factors such as changes in employment background and the employment advantages of college students. None of these shortcomings can well reflect the actual employment situation of college graduates, which makes the decision of the government and colleges and universities to improve the employment of fresh graduates to cause a certain interference risk.

3.1.2. Poor learning effect of category features

There are a large amount of text data in the dataset, and for the predictive model, the first thing to do is to convert it to multi-dimensional 0/1 features through one-hot encoding, which reduces the efficiency of space and time[10]. But for decision trees, one-hot coding is used, especially when there are a large number of categories in the category features, there will be an imbalance in sample segmentation, which affects the learning of the decision tree.

(i) Sample segmentation imbalance problem: Using one-hot encoding means that only one vs rest segmentation method can be used on each decision node, resulting in very small segmentation gain. In this series of features, only a small number of samples are 1 and a large number of samples are 0.

At this time, the segmentation of the samples will produce imbalance, which means that the segmentation gain will also be small. Because the proportion of the smaller segmented sample set to the total sample is too small, no matter how large the gain is, it can be almost ignored after multiplying by the ratio; the larger split sample set is almost the original sample set, and the gain is almost zero

(ii) Influencing the learning of decision trees: When the one-hot encoding divides the category features, it will divide the data into many scattered small spaces. However, statistical information is used when learning decision trees. In these small data spaces, statistical information is inaccurate and the learning effect will deteriorate.

3.2. Improved classification features and employment rate prediction model features and properties

Decision tree iterative training can not only get the optimal model, but also the model has the advantages of good training effect and not easy to overfit. In each iteration of GBDT, it needs to traverse the entire training data multiple times. If the entire training data is loaded into the memory, the size of the training data will be limited; if it is not loaded into the memory, it will consume a lot of time to read and write the training data repeatedly. In order to solve the problems encountered by GBDT in large amounts of data, GBDT can be used in predictive model applications better and faster. Therefore, LightGBM is used to improve the prediction model.

3.2.1. Gradient-based One-Side Sampling

In order to speed up the training speed of the gradient boosting decision tree (GBDT) model without compromising accuracy, LightGBM optimizes the decision tree algorithm on the traditional GBDT algorithm. The classic GBDT generally only uses the first-order negative gradient of the loss function, but it also uses the first and second-order negative gradient of the loss function to calculate the residual of the current tree, and use the result to fit the next round of the new tree[10]. One-sided gradient sampling (GOSS for short) is a decision tree algorithm based on Histogram.

Using GOSS can reduce a large number of data instances with only small gradients, so that when calculating information gain, only the remaining data with high gradients can be used. Compared with XGBoost (decision tree algorithm of pre-sorting method), it traverses all feature values. This algorithm saves general execution time, and it adopts the leaf growth strategy of maximum depth limit to make the overall execution of the algorithm more efficient. Algorithm 1 is as follows.

Algorithm 1. Gradient-based One-Side Sampling

Input: I: training data, d : iterations
Input: a : sampling ratio of large gradient data
Input: b : sampling ratio of small gradient data
Input: *loss*: loss function, L: weak learner
models $\leftarrow \{\}$, fact $\leftarrow \frac{1-a}{b}$
topN $\leftarrow$ a $\times$ len (I), randN $\leftarrow$ b $\times$ len (I)
for $i = 1$ **to** d **do**
  preds $\leftarrow$ models.predict (I)
  g $\leftarrow loss(I, \text{preds}), \text{w} \leftarrow \{1,1,\dots\}$
  sorted $\leftarrow$ GetSortedIndices(abs(g))
  topSet $\leftarrow$ sorted[1:topN]
  randSet $\leftarrow$ RandomPick(sorted[topN:len(I)], randN)
  usedSet $\leftarrow$ topSet + randSet
  w[randSet] $\times=$ fact ▷ Assign weight $fact$ to the small gradient data.
  newModel $\leftarrow$ L(I[usedSet], $-$g[usedSet], w[usedSet])
  models.append(newModel)

3.2.2. Classification feature segmentation

When segmenting categorical features, One-hot Code will segment the data into many scattered small spaces. However, statistical information is used when learning decision trees. In these small data spaces, statistical information is inaccurate and the learning effect will deteriorate. Level-wise decision tree growth strategy believes that each node of each layer must be divided. Leaf-wise is a more efficient strategy, each time from all the current leaves, find the leaf with the largest split gain, and then split, and so on[11].

Therefore, compared with Level-wise, Leaf-wise can reduce more errors and get better accuracy when the number of splits is the same. The disadvantage of Leaf-wise is that it may grow a deeper decision tree, resulting in overfitting. Therefore, LightGBM adds a maximum depth limit on top of Leaf-wise, the data will be divided into two larger spaces, and further learning will be better. Prevent over-fitting while ensuring high efficiency.

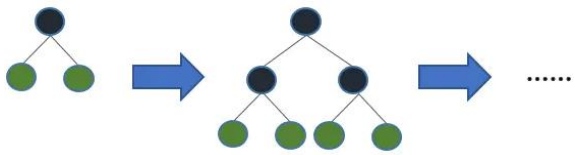

Figure 1. Level-wise tree growth

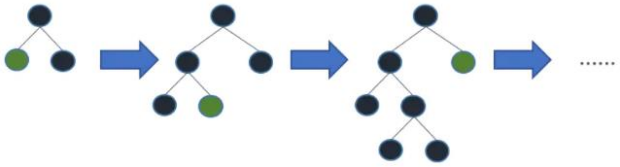

Figure 2. Leaf-wise tree growth

3.2.3. Classification feature segmentation

The aforementioned GOSS can accelerate model training by reducing the number of samples, while EFB can further reduce the data size by reducing the number of features. For example, there may be many features in a recommendation system, reaching tens of thousands of dimensions or even more, but many of these features are mutually exclusive. Mutually exclusive means that some features rarely have non-zero values at the same time, similar to one-hot features. LightGBM bundles these features together to form a new feature to reduce the number of features and improve training speed. The Algorithm 2 is as follows.

Algorithm 2. Merge Exclusive Features

```
Input: numData: number of data
Input: F: One bundle of exclusive features
binRanges ← {0}, totalBin ← 0
for f in F do
   totalBin += f.numBin
   binRanges.append(totalBin)
newBin ← new Bin(numData)
for i = 1 to numData do
   newBin[i] ← 0
   for j = 1 to len(F) do
      if F[j].bin[i]≠ 0 then
         newBin[i] ← F[j]-bin[i] + binRanges[j]
Output: newBin, binRanges
```

3.2.4. Direct support for category features

Compared with traditional machine learning, LightGBM can support direct input of category features, without pre-conversion into multi-dimensional 0-1 features, thus improving time efficiency and space efficiency, and reducing the conversion by changing the decision rules of the decision tree algorithm. -1 feature link, which has increased the speed by nearly 8 times[12].

3.3. Feature Selection

Because the dataset itself is high-dimensional data with classification features as the main content, in order to avoid the dimension disaster of high-dimensional data and reduce the running time of the model, we need to evaluate the application of features to determine whether each feature should enter the model.

(i) Near zero variance variables need to be eliminated directly: near zero variance variables can be regarded as constant items. For example, the value of a feature in all samples is a constant (such as 0 or missing). Such features obviously can not provide information and can be deleted directly.

(ii) Processing of classification variables: the reality is full of classification variables. The usual method is to encode the classification variables alone. However, when there are too many categories in a classification, using one hot will aggravate the dimension disaster. The ultimate goal of features is to distinguish different samples. More features will be complex. In this way, there will be many categories, and the samples of a few categories will be insufficient. We identify whether there is overlap between categories on the dataset and split the categories.

(iii) Treatment of collinearity problem: the direct correlation of variables may lead to data snooping, inaccurate accuracy of the model and the illusion of too high prediction results.

(iv) Treatment of missing values: because there are too few samples containing missing values, if the missing values are regarded as a category alone, additional noise will be introduced and the quality of the model will be affected. Therefore, we eliminate the missing values.

The evaluation of feature importance with random forest is mainly based on the influence factors of each feature on each tree in the random forest, then take the average value, and finally compare the contribution of different features. As shown in Table 1.

Table 1. Data feature importance

Feature name	Importance
Sex	222
Major	119
Academy	45
Political outlook	28
Education	27
Type of Employment difficulties	23
Urban and rural students	18
Nation	2

Finally, we select the above 8 influential features from the 92 features in the data source as the main features of the data set. The data after feature selection has strong applicability to the existing model or traditional non classified feature model, which can greatly reduce the data dimension, reduce the running time of the model, maintain the data balance and improve the quality of the model.

3.4. Employment rate prediction model based on LightGBM

First, perform data cleaning on the employment history data of postgraduate graduates after desensitization in colleges and universities, and then obtain the student data to be processed, and mark the samples with employment label as employment as positive samples, and label the samples with employment label as waiting for employment as negative sample. Among them, the label value of employment is 1, and the label value of waiting to be employed is 0.

(i) Data acquisition module: obtaining the student data to be processed and mark the positive sample according to the employment label in the student data.
(ii) Feature box module: the employment of graduates is related to a variety of factors, such as the major studied, the source of urban and rural students, the source of previous students and other information. Therefore, this module is used to box the features according to the characteristics of the student data.
(iii) Proportion calculation module: used to calculate the proportion of the number of positive samples in each sub box to the total number of samples.
(iv) Similarity calculation module: which is used to establish a trend change relationship according to the proportion of boxes and positive samples, analyse the trend change, and screen the features whose similarity is greater than the first threshold.
(v) Prediction module: after training the model, use the model for prediction.

3.5. Dataset and its evaluation criteria

In order to verify the accuracy of the newly constructed model in this paper, the employment rate of college graduates is verified by the graduate employment dataset of the Graduate School of Hunan Normal University. In order to further verify the effectiveness of this model in predicting the employment rate of the current year, the employment dataset is still used for verification. The table 2 is as follows.

Table 2. Data feature description

Feature name	Data type
Employment	Category
Education	Category
Sex	Category
Major	Category
Academy	Category
Types of employment difficulties	Category
Political outlook	Category
Nation	Category
Urban and rural students	Category

Class1 is employment, marked as 1, class2 is unemployed, marked as 0. This paper takes the employment information sample of fresh graduates in a year as the test set and the rest as the training set.

Confusion matrix, also known as error matrix, measures the classification accuracy of a classifier, with n rows and N columns (n represents the number of categories). Each column of the confusion matrix represents the prediction category, and the total number of each column represents the number of data predicted as the category; each row represents the real belonging category of the data, the total number of data in each row represents the number of data instances of the category, and the value in each column represents the number of real data predicted as the category TP indicates the number of samples with correct classification, and FP indicates the number of samples with wrong classification. Take the second classification as an example:

Table 3. Classification result confusion matrix

Real	Prediction	
	Positive example	Counter example
Positive example	TP	FN
Counter example	FP	TN

Accuracy, Recall and F1 score are used as the evaluation indexes of the model. Calculate the precision and recall under each category through the statistical value of the confusion matrix, where precision is the proportion of the positive sample in the positive example determined by the classifier; Recall is the proportion of predicted positive cases in the total positive cases, and accuracy represents the proportion that the classifier judges correctly for the whole sample.

$$\text{precision}_k = \frac{TP}{TP+FP} \quad (1)$$

$$\text{recall}_k = \frac{TP}{TP+FN} \quad (2)$$

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

According to the above calculation results, the F1 score under each category is counted, and the calculation formula is as follows.

$$f1_k = \frac{2 \cdot \text{precision}_k \cdot \text{recall}_k}{\text{precision}_k + \text{recall}_k} \quad (4)$$

Calculate the mean value through F1 score under each category solved above, and finally get the final evaluation result. The calculation formula is as follows.

$$\text{score} = \left(\frac{1}{n}\sum f1_k\right)^2 \quad (5)$$

4. Results and analysis

In order to verify the performance of the model proposed in this paper, it is evaluated from the following two aspects: 1 Cross validation using historical data, 2 Model quality assessment.

It can be seen from the experimental results in Table 3 that due to the multidimensional and complexity of the data, this method can not only effectively identify and predict the employment situation of graduates, but also whether the employment situation can be significantly improved according to the massive data with high-dimensional characteristics.

Since this paper uses the two feature importance indicators in LightGBM as the measurement basis between data features, the two measurement indicators of dimension reduction rate and CPU time have been significantly improved, as shown in Figure 3.

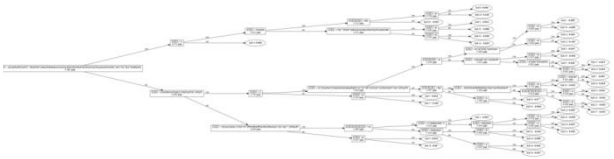

Figure 3. The 3rd decision tree

We divide 20% of the data sets into test sets and use the trained prediction model for prediction. The accuracy of employment forecast is as high as 84%. Compared with the traditional prediction model method, the experiment shows that in the data set with classification features as the main content, the model in this paper is superior to the traditional prediction model in terms of prediction accuracy, model quality and running time, as shown in Table 4.

Table 4. Model prediction index

Model	Model accuracy score	Precision	Recall	F1 Score	Time
LightGBM	0.83	0.85	0.95	0.90	10.4s
Radom Forest	0.60	0.99	0.86	0.92	17.53s
SVM	0.62	0.80	0.62	0.68	587.02 s

In order to further verify the robustness of the algorithm, the loss function curve of the model is drawn, as shown in Figure 12. In each index, the AUC curve of this algorithm is about 20 rounds of iteration, and the prediction accuracy is the highest. Therefore, it can be seen that this model has a certain anti fitting effect. It is proved that this model can effectively deal with unbalanced data classification.

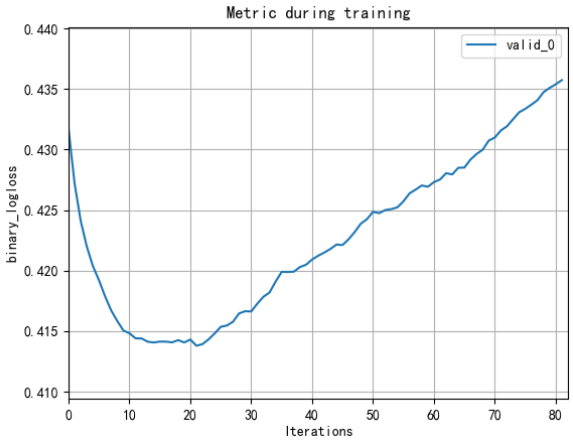

Figure 4. Loss function

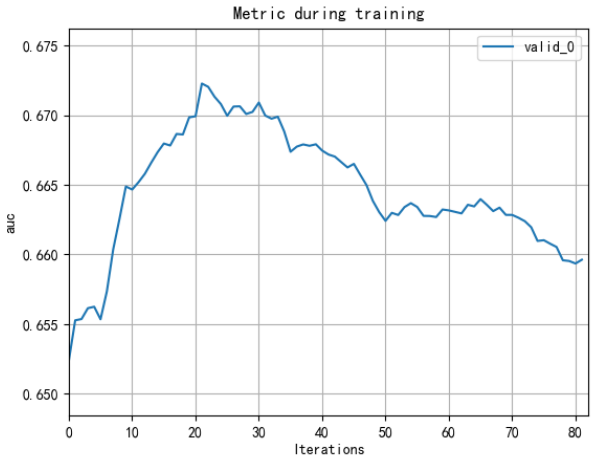

Figure 5. Accuracy during training process

In addition, according to the ranking of characteristic weights in Figure 6, in addition to gender characteristics, specialty and college characteristics have a great impact on employment, mainly because some trump majors of the University have a great impact on the surrounding areas, there are many special recruitment and natural employment impact factors. The master's degree or doctoral degree is not a particularly important factor in the model. Therefore, the master's degree or above does not play an absolute factor in the first employment. Among them, the proportion of ethnic characteristics is very small, which has little impact on employment. Due to the rapid promotion of China's integrated urbanization policy, we know in Figure 6 that the category of employment difficulties and the source of urban and rural students also have little impact on employment. Take the characteristics with high weight as the employment factor, so as to increase the relevant guidance and help to the employed students, focus on the improvement of employment

factors, and further improve the overall employment satisfaction.

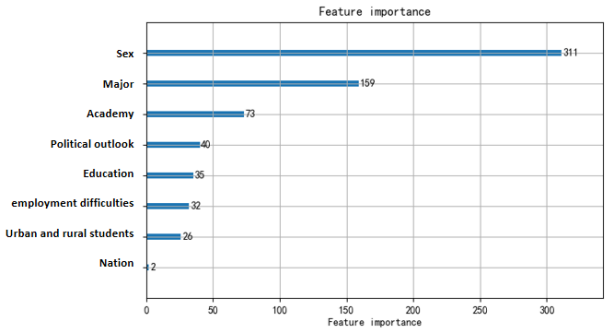

Figure 6. Characteristic influence factor

Finally, we used the employment data of other teacher training colleges to verify the prediction quality of the model again, and the dataset collected the employment information of graduates of a university in the past 4 years, a total of 16173 data volumes, and selected the important employment factors obtained from the above experiments for model quality verification. Based on the data in Table 5, we can see that the model can be effectively applied to solve the problem of employment forecasting in other universities.

Table 5. Model prediction index

Model accuracy score	Precision	Recall	F1 Score
0.85	0.90	0.93	0.91

Again, we rank important factors, with the highest ranking being the gender factor, followed by majors and colleges. Considering that both datasets use employment information from teacher training colleges and universities, educational institutions tend to prioritize female job seekers when recruiting for teacher positions. Therefore, gender factors are an important factor affecting the employment situation of teacher training colleges and universities. In the existing teacher job hunting, women are more favored by recruiters than men.

5. Summary

At present, the research of artificial intelligence is applied to various fields, and the prediction of employment rate in the field of intelligent education management is the focus of the research. The existing prediction model of graduate employment rate ignores the internal relationship between a variety of discontinuous values[13], so how to deal with the integration of classification features is the key to employment rate prediction.

In this paper, we introduce LightGBM to fuse a variety of classification features, because in this employment information database, classification text features account for the main part of dataset features. In the past, research on employment rate often focused on general employment data, ignoring the employment situation of students themselves and did not take into account the change of external employment Landscape[14], Students' own employment advantage is the key to realize employment. Therefore, this paper applies the classification feature mechanism in LightGBM to the employment classification feature fusion, so as to better obtain the relationship between students' personal characteristics and employment, and effectively solve the problem of ignoring the relationship between dataset classification features in the traditional employment prediction process.

Experiments show that the proposed model has achieved better recognition performance in the employment dataset of the University. It reduces the employment prediction error of college graduates, the employment prediction time period of college graduates, and obtains the ideal employment prediction results of college graduates.

References

[1] Patuelli R, Reggiani A, Nijkamp P, et al. Neural networks for regional employment forecasts: are the parameters relevant [J]. Journal of Geographical Systems, 2011, 13(1): 67-85.

[2] Wang Yaru. Research on the employment prediction model and application of college students based on decision tree algorithm [D]. Wuhan: Central China Normal University, 2018.

[3] Qi Hongqiang, Zhang Fukun, Gao Dakun, Wang Huiqiang. The employment rate prediction of college students based on the gray system[J]. Modern Electronic Technology, 2019, 42(11): 174-177.

[4] Hou Jie. University student performance prediction based on education data [D]. Dalian University of Technology, 2020.

[5] Shi Jing. Research on the influence of student behavior on academic performance based on data mining [D]. Central China Normal University, 2017.

[6] Xia Pengbin. Employment prediction of college students based on campus big data [D]. Central China Normal University, 2020.

[7] Zhu Wenqi. Research on the calculation method of user similarity in recommendation system[D]. Chongqing: Chongqing University, 2014.

[8] Ge J, Qiu Y. Concept similarity matching based on semantic distance[C]//2008 fourth international conference on semantics, knowledge and grid. IEEE, 2008: 380-383.

[9] Lu Tongshuang, Wang Hongguo, Liu Yinggang, et al. A method for predicting employment destinations of college students based on stereo data[J]. Computer Integrated Manufacturing System, 2019, 25(No 4): 1032-1036.

[10] Huang Z, Liu G. Prediction model of college students entrepreneurship ability based on artificial intelligence and fuzzy logic model[J]. Journal of Intelligent & Fuzzy Systems, 2021 (Preprint): 1-12.

[11] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30: 3146-3154.
[12] Li Zhanshan, Yao Xin, Liu Zhaogeng, et al. Feature selection algorithm based on LightGBM [J]. Journal of Northeastern University (Natural Science Edition), 42(12): 1688.
[13] Fan Yayun, Feng Jingjing, Zhang Xiaowei. The application of Markov model in predicting the employment quality of college graduates[J]. Enterprise Technology and Development, 2018 (2): 230-231.
[14] Wu Gongcai, Zheng Hemin. Research on the Application of Data Mining in the Precision Employment of College Graduates[J]. Electronics World, 2018 (9): 84-84.