Research on Student Academic Performance Prediction Methods

Chenghao Yan

Beijing No.80 High School, Beijing, China

Chenghao2022@163.com

Abstract. Student academic performance prediction can not only detect students' academic problems in advance, but also optimize teaching methods and provide students with personalized teaching methods, considering the complex relationship between academic performance and other factors, this paper uses linear regression and random forest to predict student academic performance.

Keywords: Academic Performance Prediction; Linear Regression; Random Forest.

1. Introduction

With the development of big data, how to find potential information and value from a large amount of data has become the focus of research. Big data analytics is an effective way for people to extract deep information. In the context of education big data, student academic performance prediction is the key content of education data analysis. Using different variables such as students' historical performance, and student behavior to study the relationship between variables and grades, form a prediction model, evaluate students' learning effects, and then infer students' future learning performance, to provide an important basis for academic warning, adjustment of teaching strategies, optimization of educational resource allocation, and personalized customization of students' learning plans. Student academic performance prediction can also enable teachers to take personalized teaching guidance in advance according to the expected results, intervene in students' learning, and help students improve the learning process. There are three research methods to achieve performance prediction. The first is performance prediction based on traditional machine learning, such as the use of decision trees, random forests, artificial neural networks, and Bayesian learning for learning performance prediction, Wang Yaru uses the Pearson product-moment coefficient to study the correlation between students' behavior data and student grades in the online learning platform, uses a decision tree to predict student achievements, and obtains the correlation between students' online learning characteristics in decision trees. The predicted performance was analyzed and verified [1]. Chang uses the Dynamic Bayesian Network to represent uncertainty in time series data and proposes a modeling method based on the state of students' acceptance of knowledge, which can effectively explain the relationship between students' degree of acquisition of knowledge and dynamic behavior [2]. Tang Jie et al. clustered students based on learning behavior and used convolutional neural networks to fuse students' learning behavior and course information to predict whether students could complete the course [3]. Based on machine learning methods (logistic regression, support vector machine, random forest, Naive Bayes, XGBoost), Holst builds predictive models based on students' performance and assesses whether students are at risk of dropping out [4]. Wang Fengqin modeled based on MOOC online data and classroom learning performance data, constructed a performance prediction model through K-means algorithm, proposed data preprocessing methods in order to eliminate the differences between multiple data, and finally gave the principle of K-means optimization algorithm based on genetic algorithm. Abdelhafez Ahmed Hoda utilizes support vector machines, multilayer perceptrons, the Naive Bayes model, and decision trees to predict how students will perform in a particular course. The second is an academic performance prediction study based on recommendation methods, and the third is an academic performance prediction research based on deep learning, Zhang Qizeng et al. based on student behavior data and historical performance data, using oversampling technology with neural network model to predict learning performance, but due

to the black-box effect of the deep learning model, the weight of the influencing factors cannot be identified [5]. Cao Hongjiang proposed a learning performance prediction model based on Long Short-Term Memory (LSTM) to model the state of students' knowledge structure according to the temporal sequence of students' historical achievements and the forgetting characteristics of the learning process, and integrate emotional characteristics and behavioral characteristics to measure learning performance through fully connected neural networks [6]. Aiming at the problems of lagging prediction, sparse data and single characteristics in traditional classroom performance prediction, Guo Zongxin proposed a traditional classroom performance prediction model that integrates self-attention mechanism and deep matrix decomposition, so that the model can focus more on useful information and improve the predictability of the model. First, this paper introduces linear regression, random forest, support vector machine, and decision tree method. Secondly, uses linear regression and random forest methods to study the relationship between different variables and achievements, establishes an academic performance prediction model, and compares and analyzes the accuracy of the academic performance prediction model constructed by linear regression and random forest.

2. Introduction to Traditional Machine Learning Performance Prediction

Machine learning includes linear regression, decision trees, random forests, artificial neural networks, Bayesian learning, and more. Linear regression is a statistical analysis method that determines the quantitative relationship between two or more variables that are interdependent. Linear regression is suitable for data sets with small data volume and simple variable relationships, and its results have good interpretability, which is conducive to decision analysis [7]. The model of linear regression is $Y(x)=w1x1+w2x2+w3x3+wnxn+b$.The following conditions need to be met to use linear regression. (1)The independent variable has a linear relationship with the dependent variable, and the relationship is straight. (2) There is no multicollinearity problem between independent variables. (3) Whether the dependent variable conforms to the normal distribution, and the residual e obeys the normal distribution N (0, $\sigma 2$). (4) Whether the dependent variable values are independent of each other. (5) Whether the variance is homogeneous. Therefore, the disadvantages of linear regression are as follows. (1) Nonlinear data cannot be modeled. (2) It is greatly affected by the extreme values or outliers in the data. If some outliers are particularly large or small, the fitted straight line will be biased towards the outliers, resulting in a decrease in the accuracy of the prediction. (3) Not suitable for highly complex data.

Random forest is one of the commonly used supervised learning algorithms that can solve both regression and classification problems. Even if some data is missing, random forests can maintain high classification accuracy. Random forest is a classifier that uses multiple trees to train and predict samples. The base classifier of a random forest is a decision tree. The decision tree is to classify sample data by arranging samples from the root node to a leaf node, including feature selection, decision tree generation, and decision tree pruning. When random forests are applied to classification or regression problems, it is a classifier integrated by multiple classification trees or regression trees, using different features of multiple subsamples to form multiple decision trees, and then predicting the final class label for each sample [8]. When there are too many noises or features in the data, the decision tree is prone to overfitting the training set. The construction process of the random forest is shown in Figure 1, with N representing the number of samples and M representing the number of features:

(1)There is a total of N samples in T, and there are N randomly selected samples. Train a decision tree with selected N samples as samples at the root node of the decision tree, and make predictions with undrawn use samples to evaluate their errors.

(2)When each sample has M attributes when each node of the decision tree needs to be split, m attributes are randomly selected from these M attributes and $m<M$. m is used to determine the decision result of a node on the decision tree. For each node, m attributes are randomly selected to train a new training set, and M sub-models are trained.

(3)In the process of the decision tree, each node should be split according to step 2 until it can no longer be split.

(4) According to steps 1~3 to establish many decision trees, many decision trees constitute a random forest, each decision tree will have a voting result, and the final voting result of most categories, is the final model prediction results.

(5) For classification problems, the voting method is adopted, and the classification category of the sub-model with the most votes is the final prediction category; For regression problems, take the average of these decision trees to get the final predicted value.

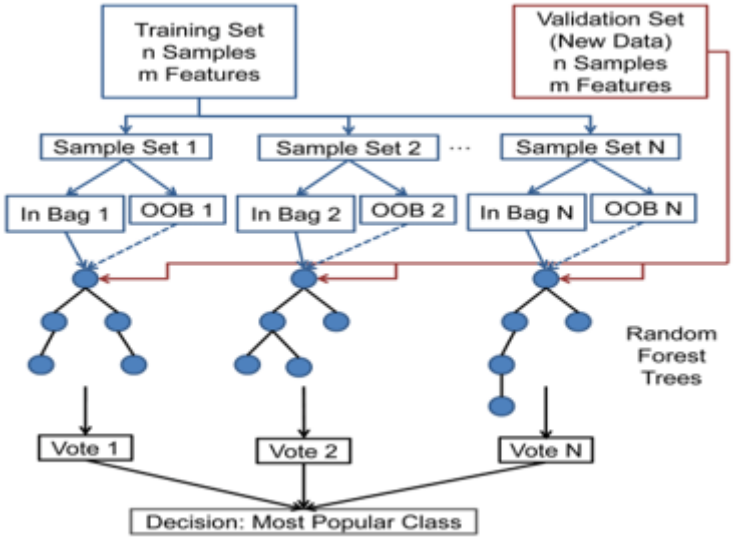

Figure 1. The process of constructing a random forest

For random forest data with inconsistent sample sizes across categories, the results of the information gain are biased towards those features with more numerical values. Many decision trees in a random forest slow down the algorithm and are ineffective for real-time predictions. In addition, for small data or low-dimensional data (data with fewer features), it is not possible to produce a better classification.

Support vector machines represent instances as points in space, separating data points with a straight line to "optimally" distinguish between the two types of points (as shown in Figure 2). Support vector machines are essentially nonlinear methods. In small sample sizes, support vector machines can grasp the nonlinear relationship between the data and features, so the nonlinear problem can be solved. In addition, support vector machines can solve the classification problem of high dimensions. However, support vector machines are sensitive to missing data, parameter tuning, and the choice of kernel functions. Support vector machines is no universal solution to nonlinear problems, so kernel functions must be carefully chosen. The classical support vector machine algorithm only gives the algorithm of two-class classification, and in the practical application of data mining, it is generally necessary to solve the classification problem of multiple classes.

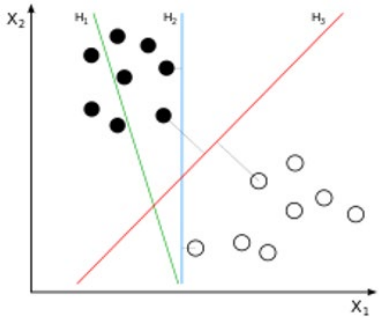

Figure 2. Support vector machine SVM

3. Data Analysis

3.1 Data Sources

In order to effectively ensure the accuracy of the prediction results, the data in this paper are derived from a school data platform and combined with field research and literature statistics. This paper includes 395 sets of data on 32 factors, including gender, age, the school's address, number of family sizes, parents' academic qualifications, length of study, number of failures, whether students received additional education, paid courses, extracurricular activities, and health status.

3.2 Model Selection and Result Analysis

In this paper, 75% of the data is selected as the training set and 25% of the data is used as the test set. Typical factors were selected and the relationship between age, the school's address, parents' academic qualifications, number of failures, and grades of students was analyzed by linear regression and random forest.

(1) The relationship between age and grades.As shown in Figure 3, the scores are mainly concentrated between 15-19 years old, which is in line with the age group of students. For each age, grades meet a normal distribution, so age is more correlated with grades.

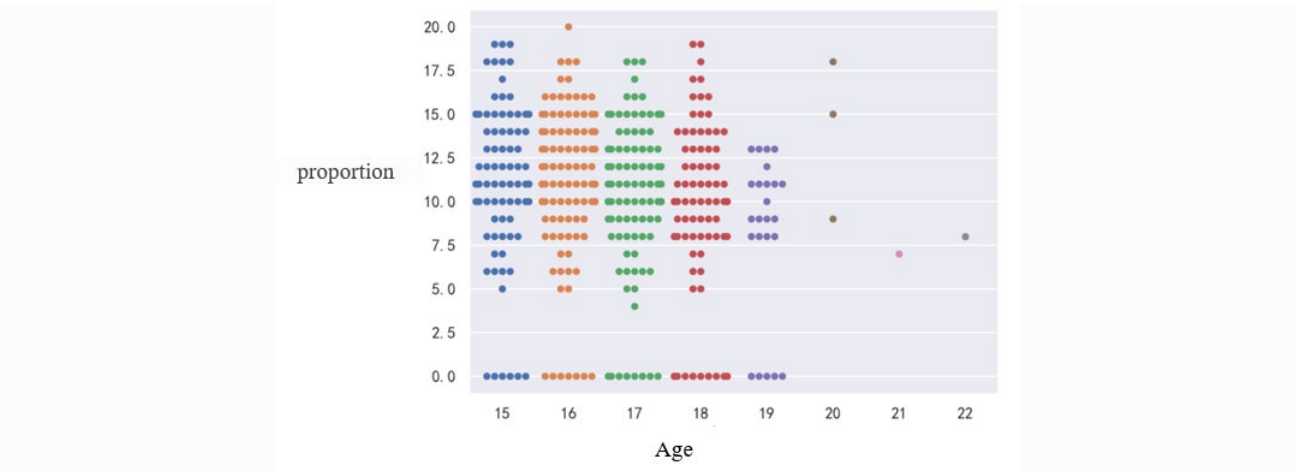

Figure 3. Graph of age vs grades

(2) The relationship between the school's address and grades. Divide the school's address into urban and rural areas, and U represents the city.It can be seen that the score of the students enrolled in the urban school is high, therefore, the urban and rural areas have a strong correlation with the grade, and the performance of the students enrolled in the urban school is higher than that of the students enrolled in the rural area (as shown in Figure 4)

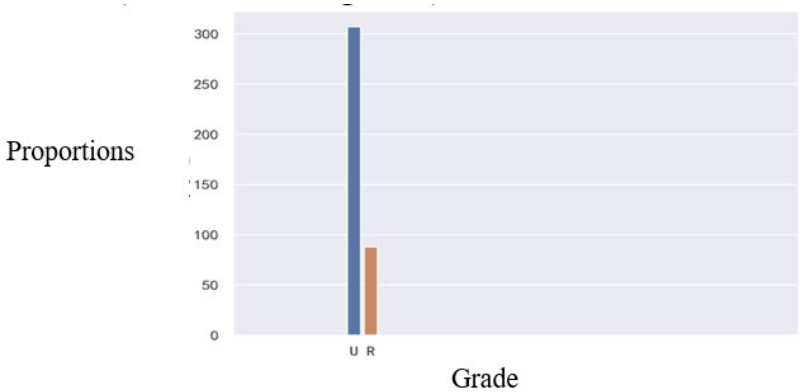

Figure 4. The relationship between t school's address and the grades

(3) The relationship between the number of failures and grades. The fewer the number of failures, the higher the score. The linear law between the number of failures and the score is satisfied. (Shown in Figure 5).

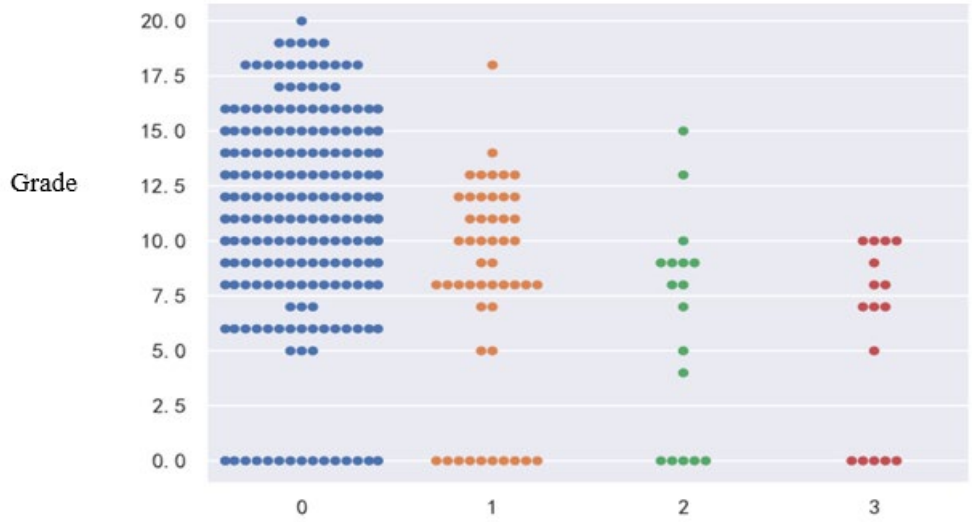

Figure 5. The relationship between the number of failures and grades

(4) The relationship between parents' academic qualifications and grades: 0 indicates that parents have no experience in education, 1 indicates that parents have primary education in the fourth grade, 2 indicates grades 5 to 9, 3 indicates secondary education, and 4 indicates higher education. As shown in Figure 6, as the education level of parents increases, student achievement also improves, so there is a clear correlation between parents' academic qualifications and grades.

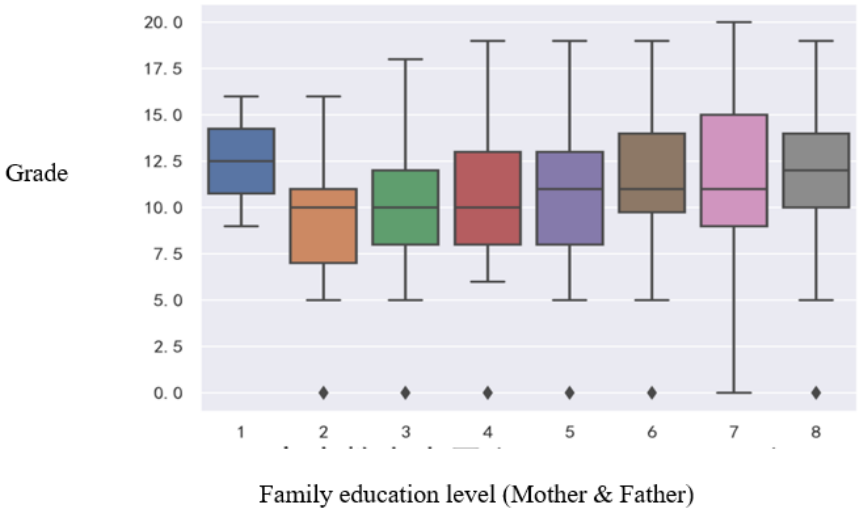

Figure 6. Diagram of the relationship between parents' academic qualifications and children's grades

The independent variables of this article are the factors that affect grades, and the dependent variable grades. By analyzing part of the data, it can be obtained that there is a linear relationship between part of the independent variable and the dependent variable, part of the dependent variable conforms to the normal distribution, and the values of the dependent variable are independent of each other. Eligible for using linear regression. At the same time, the amount of data in this study is large,

and in order to process multidimensional data, reduce the amount of calculation of selected features, and improve the prediction efficiency, random forests are also used for prediction.

4. Model Evaluation

This paper uses python coding. For random forests, the number of decision tree models is set to 100, that is, the effects of 100 decision trees are integrated, and the results of different decision trees are integrated to obtain the results of the final random forest. The key code for training and prediction of the model is as follows:

```
model.fit(X_train, y_train)
        predictions = model.predict(X_test)
        print(predictions)
        # Error criteria
        mae = np.mean(abs(predictions - y_test))
        rmse = np.sqrt(np.mean((predictions - y_test) ** 2))
```

The accuracy of the model prediction is judged by calculating the root mean square error. The accuracy of the budgets of the two models is different (as shown in Figure 7). The prediction using random forest fits optimally, and the root mean square error is minimal and less than 2, so it can be shown that the model has large applicability to the prediction of academic performance.

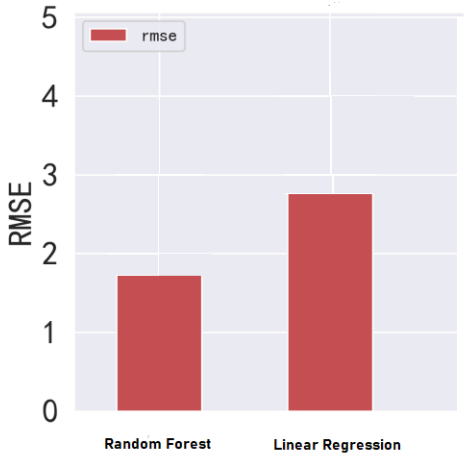

Figure 7. Root mean square error (RMSE) for random forests and linear regression

5. Summary

Due to the complex relationship between academic performance and other factors, this paper adopts random forest prediction model and linear regression to predict academic performance, and this paper believes that academic performance is also affected by the actual environment and subjective factors. Therefore, in the future research process, the accuracy of the data and the integration of processing methods should be improved, and environmental and subjective factors should be added to further improve the accuracy of predictions.

References

[1] Wang Yaru, Research on Online Learning Behavior Analysis and Performance Prediction Based on Decision Tree, Henan Normal University, 2020.

[2] Chang K, Beck J, Mostow J, et al. A Bayes net toolkit for student modeling in intelligent tutoring Systems, Proceedings of the International Conference on Intelligent Tutoring Systems. Berlin, Heidelberg, 2006: 104-113.

[3] Wenzheng Feng, JieTang. Understanding dropouts in moocs. In Proceedins of the 33rd AAAI conference on Artificial Intelligence.2019.517-524.

[4] Hlosta M, Zdrahal Z, Zendulka J. Ouroboros: early identification of at-risk students without models based on legacy data, Proceedings of the Seventh International Learning Analytics &Knowledge Conference. Vancouver, BC, ACM, 2017: 6-15.

[5] Zhang Qizheng, Dai Hanbo. Research on Student achievement prediction model based on data preprocessing technology. Journal of Hubei University (Natural Science Edition), 41 (2019): 101-108.

[6] Cao Hongjiang, Xie Jin, Study on academic performance prediction and its Influencing factors based on LSTM. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 22(2020).

[7] Sun Rongheng. Applied Mathematical Statistics (3rd Ed.). Beijing: Science Press, 2014, pp.204-206.

[8] Liu Yong, Xing Yanyun. Research and Application of text classification based on improved Random Forest algorithm. Applications of Computer Systems, 28(2019), pp.220-225.