The effect and contribution of e-book logs to model creation for predicting students' academic performance

Zhao, Fuzheng

Kumamoto, Etsuko

Yin, Chengjiu

The effect and contribution of e-book logs to model creation for predicting students' academic performance

Fuzheng Zhao
Graduate School of System Informatics
Kobe University
Kobe, Japan
zhaofz635@gmail.com

Etsuko Kumamoto
Information Science and Technology Center
Kobe University
Kobe, Japan
kumamoto@kobe-u.ac.jp

Chengjiu Yin
Information Science and Technology Center
Kobe University
Kobe, Japan
yin@lion.kobe-u.ac.jp

***Abstract*—As a kind of data that can reflect learning status, e-book logs have been widely used in learning analytics, especially for the prediction of academic performance. However, the best prediction model cannot be found without determining the contribution of e-book logs to the prediction performance of the model and its creation process. To this end, this study used the scikit-learn, a free software machine learning library, to analyze learning performance of 234 participants by learning behavior logs, which were collected by an e-book system. Finally, six prediction models containing Decision Tree, Random Forests, XGBoost, Logistic Regression, Support Vector Machines, and K-nearest Neighbors were created. Also, the contribution of e-book logs on the establishment of different prediction models was obtained by three feature importance calculation methods, i.e., the impurity-based feature importance, coefficients feature importance, and permutation feature importance. Based on statistical results, it was concluded that the Decision Tree and Random Forests had the best prediction performance, which was compared to the other four models, with prediction performance scores ranging from 0.7 to 0.8. Besides, the four data features of Prev, Highlight, Maker, and Next were found to have the greatest impact on model prediction creation.**

Keywords—e-book logs; learning analytics; academic performance; prediction

I. INTRODUCTION

Predicting student academic performance provides access to building positive feedback of their learning status for instructors [1], which allows allocating instruction precisely based on each student's performance [2]. With the increasing of numerous studies, some data and its features have been used to make predictions [3].

Parallel to this, some learning analytics techniques were taken to the predicting practices [4], which has provided an opportunity to explore the relationships between learning behavior and learning performance [5]. Notably, with the increasing abundance of e-publication, an amount of reading behavior data has created novel evidence for predicting learning performance [6].

However, there has been limited to the effect and contribution of learning log data to model creation for predicting students' academic performance [7]. As a result, roles and strategies of log data to predict academic performance are peculiarly prone to be confused. Consequently, the effectiveness and accuracy of predicting models are not adequately supportive for educational practices [8]. In this study, the effect and contribution of e-book logs were examined and analyzed in the procession of various model creation for predicting student academic performance.

II. METHODOLOGY

A. Data Collection

To test the impact and contribution of e-book logs on different model creation. 234 participants were recruited to participate in an experiment. In this experiment, participants were asked to read a learning material for a specified period, followed by a post-test to make a judgment about the learning achievement. The study consisted of data collection, analysis, predictive model creation, evaluation, and feature importance of data. The data collection was conducted through the e-book system, and the learning data were stored in log form, as shown in Table I. Finally, 11 features were composed of Prev, Next, Marker, Highlight, Bookmark, Backtrackrate, Readtime, Readpages, Mobile, Tablet, and PC.

First, the missing values are processed. Through data analysis, we found that there were a small number of missing values in the data. To maximize the validity of the data, we used the operation of removing missing values and finally obtained valid data for 229 participants. Also, the study performed encoding categorical features for gender, and represented men and women as 1 and 0, respectively. Also, standardization was employed to unify the values of each feature of the data, compressing the values from 0 to 1. To avoid a class imbalance problem, the passing score is set to 70 points.

TABLE I. AN EXAMPLE OF E-BOOK LOGS

User	Action Name	Learning Material	Page Number	Action Time	Device
Student 1	Next	Education technology	15	2019/7/5 8:40	PC
Student 1	Prev	Education technology	15	2019/7/5 8:42	Mobile

B. Models Creation

The Decision Tree, Random Forests, XGBoost, Logistic Regression, Support Vector Machines, and K-nearest Neighbors are selected to construct prediction models for at-risk student identification and the preparatory functions in scikit-learn were used to complete the prediction model creation. The basic process to create models was successively consisted of train and test set splitting, fitting models, tuning parameter, measuring models, and computing feature importance. The first step is known to divide the data into test dataset and training dataset by *train_test_split* function, and the specific ratio of splitting is 3:7, where test dataset accounted for 30%. The second step is to fit the model with the preparatory functions provided by scikit-learn, such as *DecisionTree_Classifier* is a class capable of performing classification. The successive step is to tune parameters suitable with models. Notably, the learning curve and grid search were the way to search for optimal parameters. After being tuned, the model can be measured by five metrics: Accuracy, F-score, Recall, Precision, and AUC. The final step came to the computation of feature importance, including the impurity-based feature importance, coefficients feature importance, and permutation feature importance.

C. Models Evaluation

This study follows the principle of confusion matrix to evaluate the performance of the prediction model from Accuracy, F-score, Recall, Precision, and AUC. The confusion matrix table consists of the prediction dimension and the actual dimension, and is divided into four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP means that the prediction is positive and the actual is positive, FP means that the prediction is positive and the actual is negative, FN means that the prediction is negative and the actual is positive, TN means that the prediction is negative and the actual is negative. Based on the confusion matrix table, and the performance metrics formulas are shown in (1), (2), (3), and (4).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{P+N}} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{F1 score} = 2\times \frac{\text{Precison}\times\text{Recall}}{\text{Precison +Recall}} \quad (4)$$

Accuracy is defined as the number of judgments that are correct out of all judgments, i.e., those that are positive and those that are negative. The recall is defined as the percentage of samples that are predicted to be positive out of all positive samples. The ROC (Receiver Operating Characteristic) curve is a curve with FP and TP as the axes, and the area under the ROC curve is called the AUC, which is a numerical value to evaluate the classifier visually, the larger the value the better. F1-score is a weighted summation average of Precision and Recall.

D. Feature Importance Calculation

The Scikit-learn provides three mainstream methods for calculating feature importance: the impurity-based feature importance, coefficients feature importance, and permutation feature importance. For the Decision Tree, Random Forests, and XGBoost, the impurity-based feature importance is more suitable. In terms of Logistic Regression and SVM, where weight-based coefficients are a way to identify which is the best feature to contribute to the prediction model. When calculating the feature importance of KNN models, the pre-determined feature importance method provides an evaluation insight through R2 scores.

III. Result

A. Evaluation of Predictive Performance in Models

In Figure 2, two distinct clustering groups emerged in the prediction performance of the six models, one containing Decision Tree and Random Forest and the other containing XGBoost, SVM, Logistic Regression, and KNN. The average score of the former group was about 0.1 higher than that of the latter group. According to the score range of 0.7 to 0.8, the predictive ability of the former group is moderate and acceptable.

However, for the latter group, since the metric scores of these models ranged from 0.58 to 0.69 and did not exceed the acceptable predictive power value of 0.7, the latter group was considered to have low predictive power. The results showed that the models that present the best predictive performance are Decision Tree and Random Forest. It is also obvious that the optimal predictive models supported by the e-book system can provide a classification to students' academic performance, however, the predictive effectiveness does not reach above 90, which is usually viewed as the optimal performance for a near-perfect model.

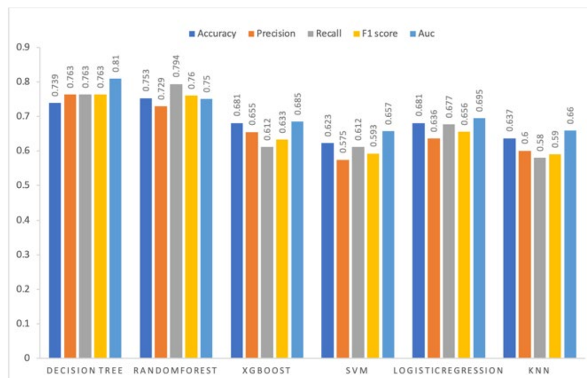

Fig. 1. The prediction performance of models

The best metric for the model predicting evaluation. Usually, the bigger the AUC, the greater the distinction between true positives and true negatives, because it compares the true positive rate against the false positive rate. The score of 0.5 is an indicator. In the evaluation of the decision tree model, the AUC reached the optimal value of 0.81, indicating that the model has a good classifying ability. By contrast, the random forest model has a good performance in the recall, with a maximum score of 0.794. The recall gives information about how the model's performance concerning false negatives is. In other words, the random forest model did not miss many true positives, and most false negatives were classified.

B. Contribution of Data Features to Models

According to the statistical results and score distributions in Figure 3, we found that the contribution of each feature varies according to the different models. For example, the data feature with the greatest feature importance in the Decision Tree is Prev, reaching the score of 0.116. Second, the same phenomenon occurs in tree models with closely similar algorithmic principles, where the feature contribution does not have coherence. For example, in the Decision Tree, Random Forest, and XGBoost, the features that contribute most to them are Prev (0.116), Prev

(0.212), and Highlight (0.19). Third, for the overall picture of the data features, the Prev, Highlight, Maker, and Next were found to have the greatest impact on model prediction creation.

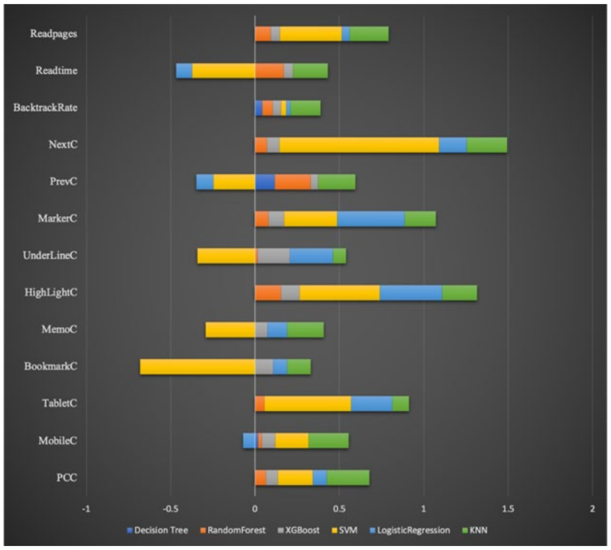

Fig. 2. The statistical distribution of feature importance to models

In the decision tree, Prev and BacktracRate ranked the top two with 0.116 and 0.043, respectively. That indicates that whether to repeatedly read and backtrack the learning material can be a key feature in predicting the learning outcome. It should be noted that the features that affect Random Forest best are similar to decision trees, with the top three scoring features being Prev (0.212), Read time (0.17), and Highlight (0.153).

For XGBoost, the lean material is emphasized by Highlight (0.19), Underline (0.11), and Bookmaker (0.106) as the features that have a differentiating effect on the prediction results. These three features are also in the top three in Logistic Regression, while the ranking is slightly different, with the order of Maker (0.401), Highlight (0.371), and Underline (0.256).

In SVM, there is a significant difference between the first feature and the second feature, which has the greatest impact on prediction performance. next ranks first with a score of 0.948, whose value is 0.268 larger than the second feature. it is worth noting that this large difference does not exist in the other models, highlighting the leading position of Next in terms of predictive power. The KNN is the most specific one with near-average feature importance, which means that each feature plays an almost equal role in the prediction results.

IV. CONCLUSION

For the prediction performance of the five metrics, although Decision Tree and Random Forest are two separate models, whose mathematics behind them are almost similar. So, it was concluded that the tree models created by e-book logs have good predictive performance. Also, the Decision Tree and Random Forest are different because of the differences in specific evaluation metrics, and they differed in the best evaluation metrics. In contrast, the Random Forest performs well on recall, indicating that the model has a good predictive function for false negative. Specifically, it has good performance in predicting the performance of truly failing students.

Considering the contribution of each feature in e-book logs to model creation, the variation which has the most significant effect on all models did not appear. In terms of the relationship between each model and all features, Prev, Highlight, Maker, and Next ranked first in most models with the highest feature importance scores respectively. Note that a feature with a large contribution to model creation like Prev, which has the largest contribution in one model, while does not show the same advantage in other models. This may be determined with the characteristics of the model feature selection.

For those students who read whether to mark or not and the length of reading time have obvious differences in grades, which also reflects a certain correlation between the degree of reading conscientiousness and study habits. There have been still some issues left in this study. For example, the relationship between model and data category needs to be further investigated. In the future, this issue will be focused on. Also, the data features that contribute the most to the model remain to be studied.

ACKNOWLEDGMENT

Part of this research work was supported by the Grants-in-Aid for Scientific Research Nos. 21H00905 and 18H04125 from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan. We would like to thank all the people who prepared and revised previous versions of this document.

REFERENCES

[1] Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., & Liao, S. N. (2018). Predicting academic performance: a systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 175-199).

[2] Mouri, K., Okubo, F., Shimada, A., & Ogata, H. (2016). Bayesian network for predicting students' final grade using e-book logs in university education. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)* (pp. 85-89). IEEE.

[3] Nakajima, T., Shinohara, S., & Tamura, Y. (2013). Typical functions of e-Textbook, implementation, and compatibility verification with use of ePub3 materials. *Procedia Computer Science*, *22*, 1344-1353.

[4] Yin, C., Yamada, M., Oi, M., Shimada, A., Okubo, F., Kojima, K., & Ogata, H. (2019). Exploring the relationships between reading behavior patterns and learning outcomes based on log data from e-books: A human Factor Approach. *International Journal of Human–Computer Interaction*, *35*(4-5), 313-322.

[5] Yin, C., & Hwang, G. J. (2018). Roles and strategies of learning analytics in the e-publication era. *Knowledge Management & E-Learning: An International Journal*, *10*(4), 455-468.

[6] Huang, A. Y., Lu, O. H., Huang, J. C., Yin, C. J., & Yang, S. J. (2020). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interactive Learning Environments*, *28*(2), 206-230.

[7] Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., ... & Hirokawa, S. (2017). Learning analytics for e-book-based educational big data in higher education. In Smart sensors at the IoT frontier (pp. 327-350). Springer, Cham.

[8] Zhao F, Hwang GJ, & Yin C. (2021). A Result Confirmation-based Learning Behavior Analysis Framework for Exploring the Hidden Reasons behind Patterns and Strategies. *Educational Technology & Society*, 24 (1), 138–15.