Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

Feature analysis for predicting students' performance from reading patterns in an e-learning system

Shohei KIKUCHI[a*], Taro TEZUKA[b*]

[a]*Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan*
s.kikuchi1120@gmail.com
[b]*Faculty of Library, Information and Media, University of Tsukuba, Japan*
tezuka@slis.tsukuba.ac.jp

Abstract: Final grade scores of students were predicted based on how they accessed teaching content provided by an e-learning system, BookRoll. In order to train machine learning models, features were designed heuristically, and various machine learning methods were trained and compared. The result showed that random forest and AutoML performs well. Analyzing trained random forest predictors revealed that time-related features contribute significantly to the performance of the regressor.

Keywords: Feature analysis, feature engineering, educational data, e-learning

1. Introduction

Predicting the final grade scores of students from how they interacted with an e-learning system is important for developing and improving systems for online learning. If some functions in the system affect students' final grade scores significantly, it will be worthwhile to put more effort in extending those functions. Also, teachers can guide students to use the e-learning system in ways that would increase their expected final grade scores. In this work, logs recorded on an e-learning system that provides online teaching material, were used to predict the final grade scores of students. The system used for logging is BookRoll (Flanagan et al., 2017; Ogata et al., 2017). The data was provided for the 5th ICCE workshop on Learning Analytics (LA) & Joint Activity on Predicting Student Performance.

2. Methods

Before training machine learning models, we carefully designed features (attributes) that would contribute to making good predictions. In addition to features already provided in the data sets, some extra features were constructed heuristically.

Event time was split into parts representing year, month, day, day of the week, and hour within the day. This is to see time in different time scales. By counting the number of events happened within each time scale, we obtained new attributes representing how often events happened in a certain time frame. For example, a feature named "hour_8" represents how many times the student accessed the e-learning system between 8 am till 9 am of any day. A feature named "day_5" represents how many times the student accessed in day 5 of any month. These features were added based on a hypothesis that times that students access the system correlates with how motivated or involved they are to the course.

Similarly, the number of times the student accessed a certain page was turned into a feature. For example, "pageno_10" represents how many times the student accessed page 10 of the teaching material. This is based on a hypothesis that well-performing students would be checking out more relevant or difficult part of the teaching material, whereas less-performing students might be looking at less relevant or easier part of the material.

The number of times the student conducted a specific process is turned into a feature to. For example, "processcode_3" represents how many times the student performed the process labeled by 3.

After adding these features, we trained machine learning models. We evaluated models with different levels of complexity. Specifically, we compared LASSO, Elastic Net, multilayer perceptron regressor (MLP), kernel ridge regression, random forest regression, AutoSklearn (Efficient and Robust AutoML for Scikit-learn, abbreviated as AutoML), and gradient boosting regression.

3. Results

We conducted 3-fold cross-validation for 10 times, using different partitioning each time. Two datasets were trained and evaluated separately. For each partition, we trained models using the training part (2/3 of the whole data set), then computed RMSE (root mean squared error) for the test data part (1/3 of the whole data set) using the trained regressor.

3.1 Dataset 1

Figures 1 and 2 illustrate how final grade scores are distributed. Most students scored between 75 and 100, indicating much deviation from a fitted Gaussian distribution. This suggests that it might be difficult to make predictions based on an assumption that data follows a Gaussian distribution, such as linear regression.

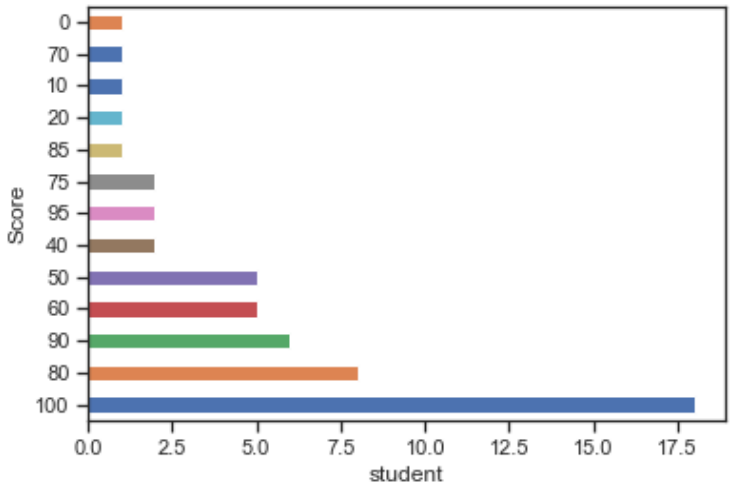


Figure 1. Score distribution by a bar graph. (for dataset 1)

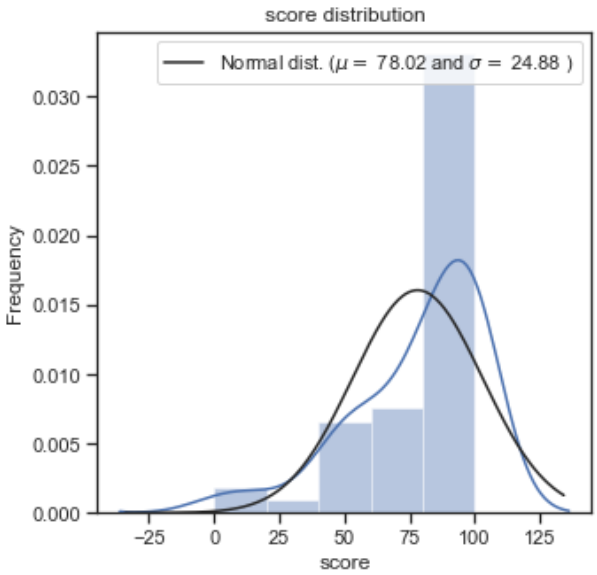
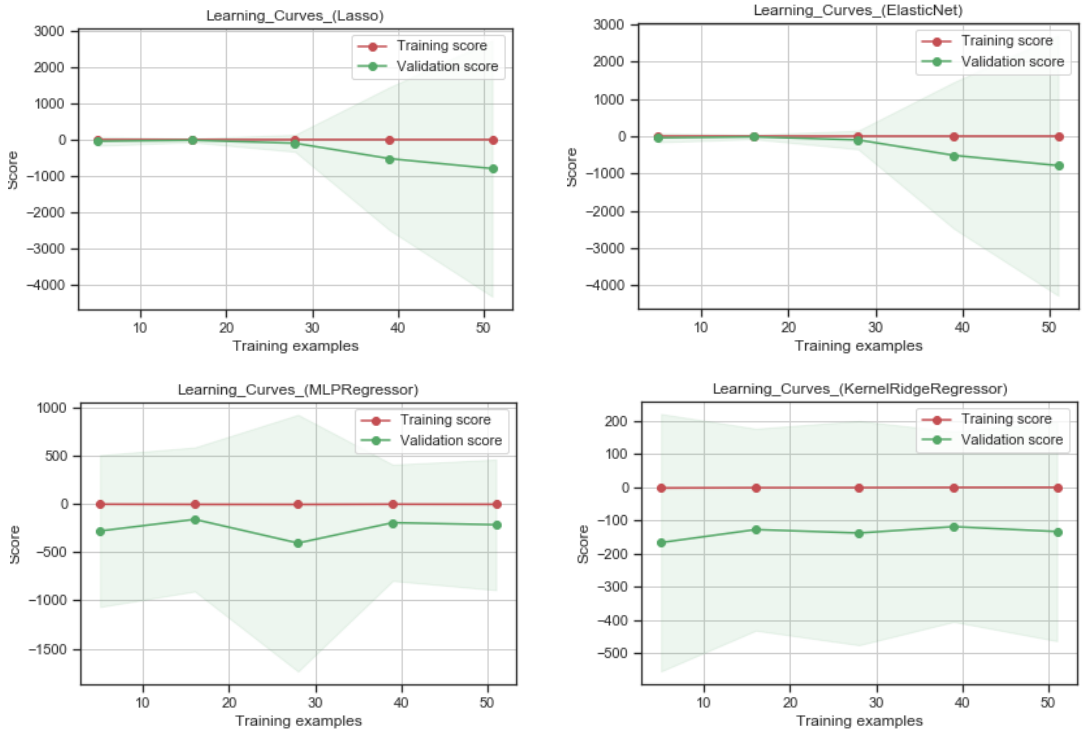


Figure 2. Score distribution (histogram) and fitting by a Gaussian distribution. (for dataset 1)

Learning curves for different machine learning models are illustrated in Figure 3. Validation scores significantly lower than training scores indicate overfitting occurring in some methods. Since validation scores approach training scores as the number of training sample increases, it suggests that more complex models such as AutoML may perform even better as the number of samples is increased.



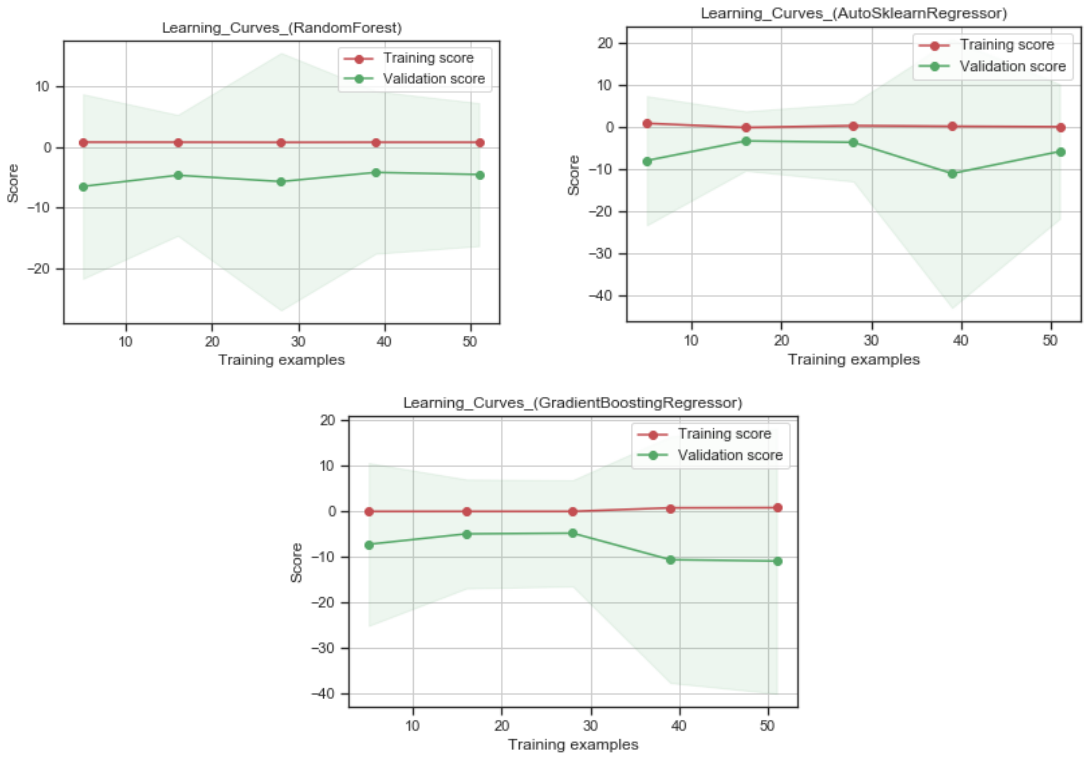


Figure 3. Learning curves of compared methods for training and test data sets. (for dataset 1)

Distributions of RMSE (root mean squared error) for dataset 1 is indicated in Table 1 and Figure 4. They are obtained by 30 validations, resulting from performing 3-fold cross-validation 10 times.

Table 1: RMSE mean and standard deviation for methods compared

method	**mean**	**std.dev.**
LASSO	141.7914	200.6841
Elastic Net	137.4087	192.9349
MLP regression	70.7624	50.5348
kernel ridge regression	54.7972	26.6537
random forest regression	23.4546	15.7610
AutoML	23.4572	15.4890
gradient boosting	25.7972	16.3673

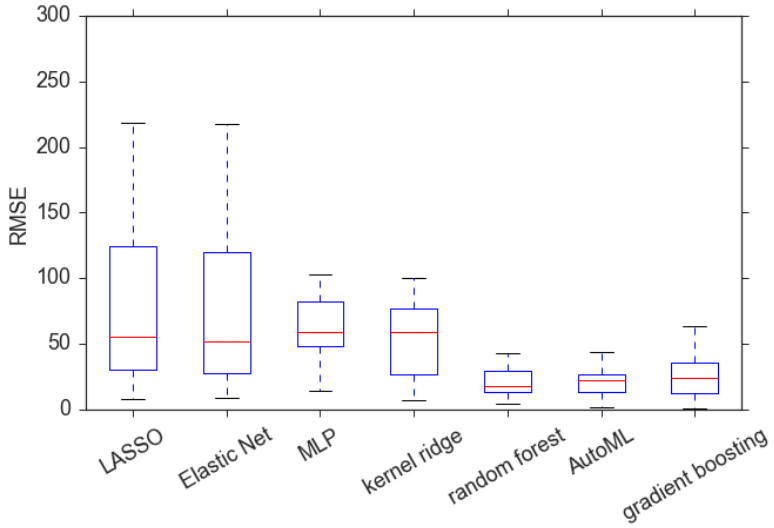


Figure 4. Distributions of RMSE over 3-fold cross-validation using 10 different partitions. Boxes represents the lower to upper quartile values of the data, with a red line at the median. The whiskers show the range of data. (for dataset 1)

After training random forest, features that contributed more were ranked using Gini-importance values. The result is indicated in Figure 5. It shows that "hour_17" feature contributed significantly, which indicates that the number of accesses that a student makes at 5 pm greatly affects how well the student perform in terms of the final grade score. The random sequence of characters ranked third in Gini-importance is an ID for a book that students could access from the e-learning system.

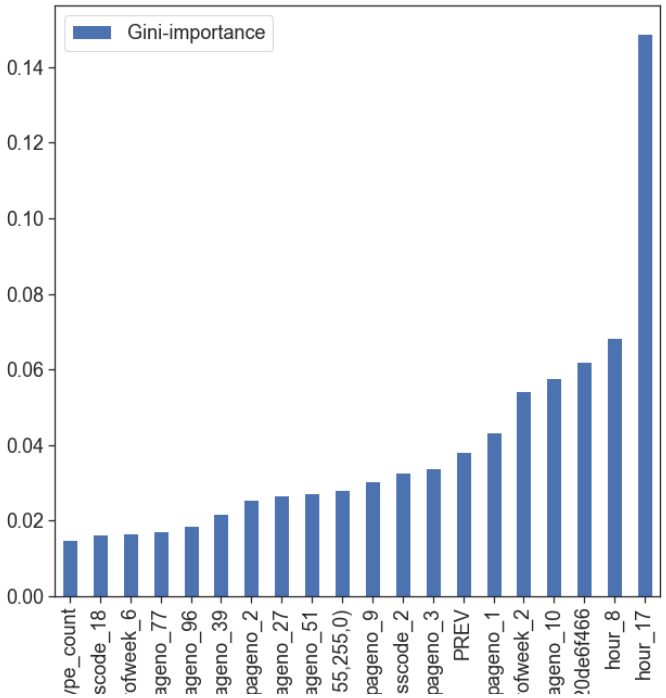


Figure 5. Features sorted by Gini-importance, indicating how much each feature contributed to making predictions in random forest. (for dataset 1)

3.2 Dataset 2

Figure 6 illustrates the distribution of final grade scores for dataset 2. It is less concentrated around score 100 when compared to dataset 1. However, it is still not well fitted to a Gaussian distribution, indicating simple models like linear regression would not be appropriate.

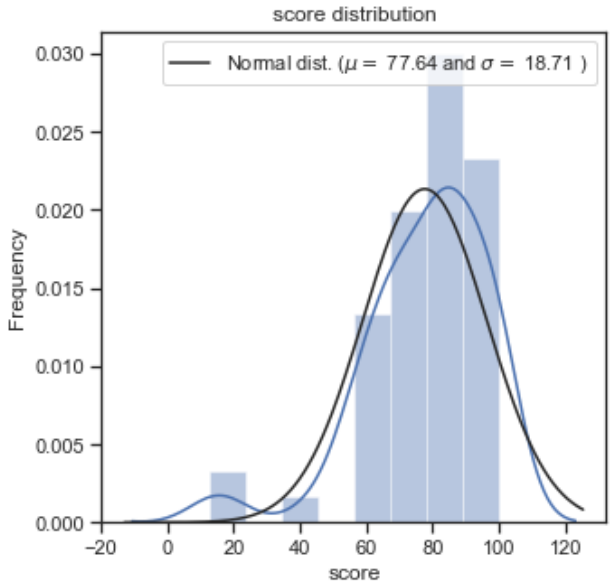
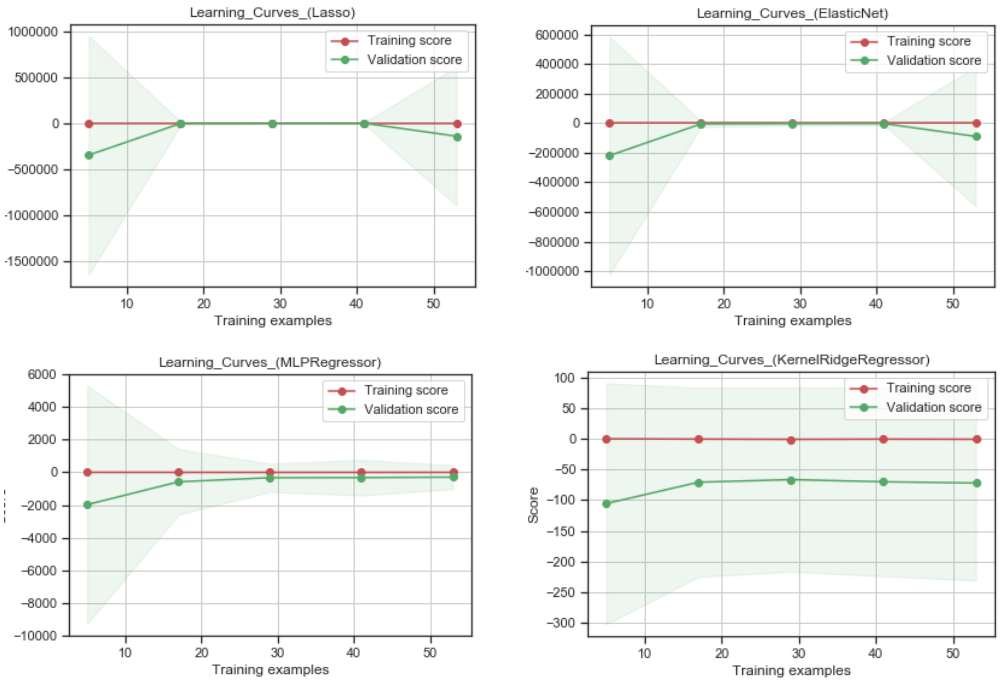


Figure 6. Score distribution (histogram) and fitting by a Gaussian distribution. (for dataset 2)

Figure 7 illustrates learning curves for dataset 2. Like in dataset 1, validation scores are sometimes much lower than training scores in some models. It suggests that with more data, complex models may have validation scores closer to training scores, resulting in better predictions.



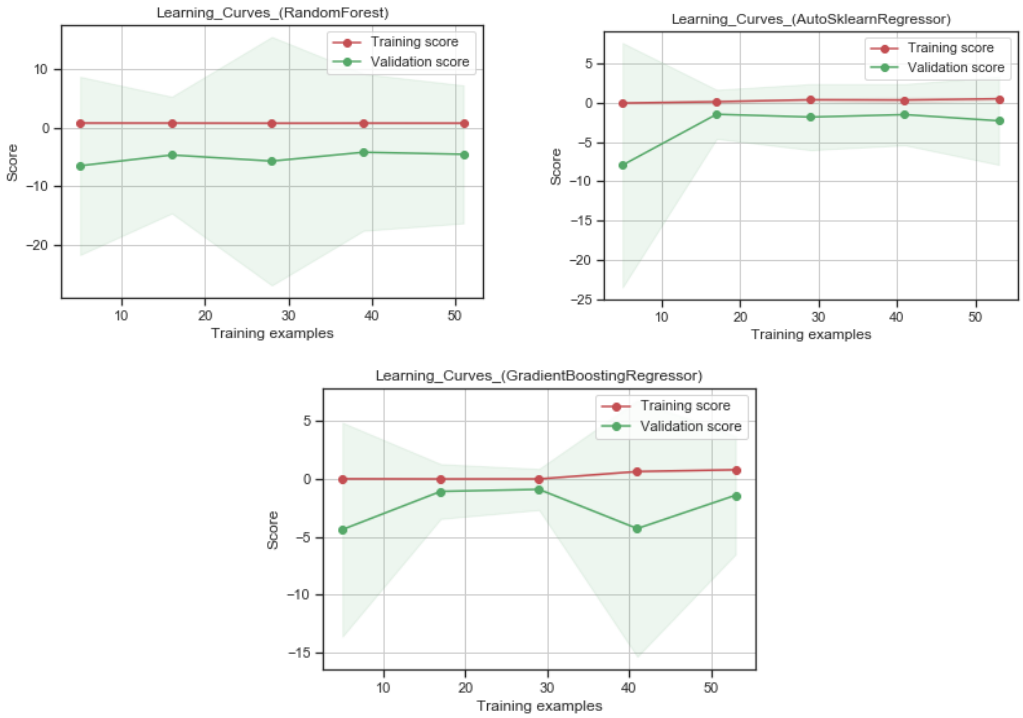


Figure 7. Learning curves of compared methods for training and test data sets. (for dataset 2)

Distributions of RMSE (root mean squared error) for dataset 2 is indicated in Table 2 and Figure 8. They are obtained by 30 validations, resulting from performing 3-fold cross-validation 10 times.

Table 2: RMSE mean and standard deviation for methods compared (for dataset 2)

method	**mean**	**std.dev.**
LASSO	1238.1671	5716.8297
Elastic Net	995.1362	4514.8347
MLP regression	95.9370	130.1382
kernel ridge regression	50.0973	26.7959
random forest regression	14.9165	10.0090
AutoML	14.3660	9.4390
gradient boosting	13.3342	9.9282

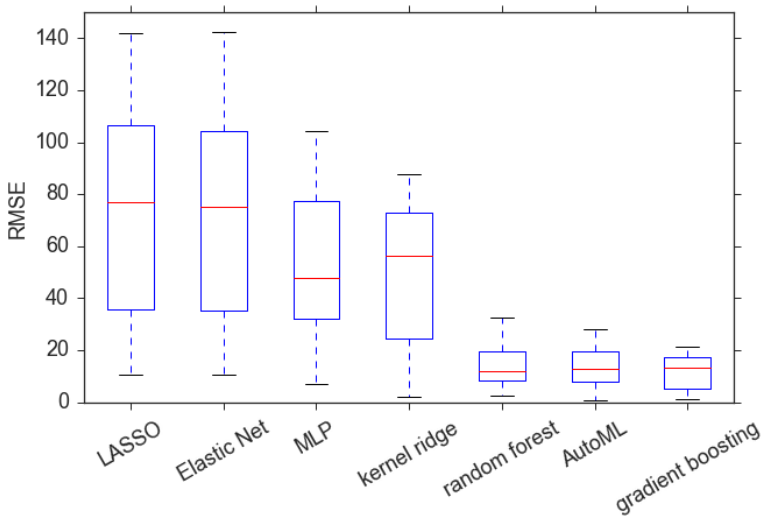


Figure 8. Distributions of RMSE over 3-fold cross-validation using 10 different partitions. Boxes represents the lower to upper quartile values of the data, with a red line at the median. The whiskers show the range of data. (for dataset 2)

Features that contributed according to random forest are indicated in Figure 9. This time, the ID of a book contributed to most, suggesting referring more or less to a specific book in the e-learning system greatly affected how well the student performs in terms of the final grade score. A time feature "hour_5" was also important, suggesting accessing at a certain time of the day also affects the student's performance.

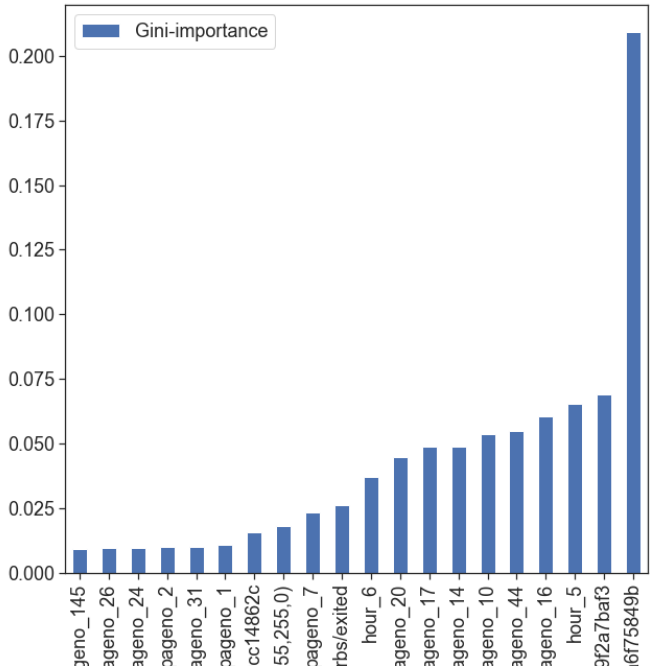


Figure 9. Features sorted by Gini-importance, indicating how much each feature contributed to making predictions in random forest. (for dataset 2)

4. Discussion

Possibly due to the size of datasets, AutoML didn't perform any better than random forest or gradient boosting. Using more data might make these complex models more competitive. Random forest performed best for dataset 1, and gradient boosting did so for dataset 2. These are ensemble methods, showing their effectiveness with datasets of this size. The difference in the performance of two models between datasets may result from distributions of data. In order to explain it, we need to conduct further analysis on the distributions.

Features that contributed most to making predictions included ones regarding times, representing at which time of the students were accessing the e-learning system. Features representing which part of the teaching material were also important.

RMSE obtained for dataset 2 is as low as 13.34 where the grade ranges between 0 and 100, indicating using log data from an e-learning system can be an effective way of predicting students' performance. One must note, however, that in the datasets used for analysis, scores are mostly above 70, predictors at this moment may not be useful for predicting how well a moderately performing student would perform among others. On the other hand, it could be useful for detecting students that will perform significantly worse than others.

5. Conclusion

One interesting observation obtained from this analysis is that at what time (hour) of the day students access the e-learning system greatly contributes to predicting the final grade score of students. It may suggest that this feature represents a student's attitude and motivation toward the course. The results suggest that analyzing log data can contribute to making e-learning better.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP16K00228, JP16H02904.

References

Brendan Flanagan, Hiroaki Ogata, Integration of Learning Analytics Research and Production Systems While Protecting Privacy, Proceedings of the 25th International Conference on Computers in Education (ICCE2017), pp.333-338, 2017.

Hiroaki Ogata, Chengjiu Yin, Misato Oi, Fumiya Okubo, Atsushi Shimada, Kentaro Kojima, and Masanori Yamada, E-Book-based learning analytics in university education, Proceedings of the 23rd International Conference on Computer in Education (ICCE 2015) pp.401-406, 2015.

Hiroaki Ogata, Misato Oi, Kousuke Mohri, Fumiya Okubo, Atsushi Shimada, Masanori Yamada, Jingyun Wang, and Sachio Hirokawa, Learning Analytics for E-Book-Based Educational Big Data in Higher Education, In Smart Sensors at the IoT Frontier, pp.327-350, Springer, Cham, 2017.