Transdisciplinarity and the Future of Engineering
B.R. Moser et al. (Eds.)
© 2022 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/ATDE220700

Dropout Prediction by Interpretable Machine Learning Model Towards Preventing Student Dropout

Miki KATSURAGI[1] and Kenji TANAKA
Dept of Technology Management for Innovation, School of Engineering
The University of Tokyo, Japan

Abstract. In the education industry, the needs of online learning are significantly increasing. However, the web-based courses demonstrate higher dropout rates than traditional education courses. As a result, engaging students with data analysis is getting more crucial especially for distance learning. In this study, we analyze data on the daily learning status of students in order to predict the student's dropout in online schools. Specifically, we trained a dropout prediction machine leaning model with 1) Basic attributes of students, 2) Progress of learning materials, and 3) Slack conversation data between students and teachers. The experimental results show that the accuracy rate of the machine learning model has reached 96.4%. As a result, the model was able to predict 78% of the students who actually dropped out of school. We also looked into feature importance by SHAP value to gain ML model interpretability.

Keywords. distance learning, dropout prevention, churn prediction, deep neural network, explainable AI

Introduction

The needs of online learning are significantly increasing in education industry since the COVID-19 outbreak forced many schools to massively push their education system towards an online environment. That said, dropout rates in distance learning are higher than those in conventional schools, thus limiting dropout is essential in distance learning [1].

In the previous research, according to the student survey results we find students often feel more isolated in online courses, and the students who dropout were not comfortable with the communication with other students / teachers. Other study also indicates isolation may be a central factor in the high rates of dropout [2].

To predict student dropout, wide range of Machine Learning approaches like Logistic Regression, Deep Neural Networks (DNN) etc. has been proposed by researchers for predicting student dropout or churn. Original approaches to model student dropout in higher education reach 40 years back to the famous survival model [3]. One of the widely used are decision tree algorithms that outperform the others due to its ability to interpretability [4]. However, the traditional parametric model or a single artificial

[1] Correspondig author, Mail: kmiki@g.ecc.u-tokyo.ac.jp, tanaka@tmi.t.u-tokyo.ac.jp

intelligence-based method cannot achieve relatively high-precision prediction, so the establishment of a combined prediction model to improve the prediction accuracy is an inevitable trend to solve the problem of dropout prediction [5. Based on this, Wang, Wei, used the neural network to achieve higher accuracy without feature engineering [6].

That said, there are less interpretation why such prediction results were obtained in past churn analysis and researches especially in complex machine learning models like DNN. Therefore, in this study, in order to improve student satisfaction, we tried to extract the features importance using Shapley value in addition to the prediction results. This has the advantage that the driving factors which lead to student dropout can be inferred from the feature importance.

1. Data-set of this study

1.1. Data of Japanese online-high school case

This study uses the data from a Japanese leading online school which was collected from November 2019 to July 2020 for 5790 students. We tried to predict the dropout status on July 2020 by past 3 months data. The data consists of following 3 tables and we joined those data using the student ID as a key.

- Student demographic data
- Student learning status (course material progress)
- Slack conversation data

We used the student's daily conversation frequency data of Slack since we find communication is the important factor for student's satisfaction in the previous study. 80% of the data was randomly assigned for training, 10% or validation, and 10% for testing.

1.2. Dropout Definition

In our dataset a student's career can result in four types: enrolled, graduation, leave of absence, and withdrawal. In this paper, we defined "Dropout" as "Leave of Absence" and "Withdrawal". The dropout status was also one-hot encoded with the value 0(non-dropout) and 1(dropout). Therefore, the problem can be stated as a supervised, binary classification model.

1.3. Dataset pre-processing/Exploration

Before train the ML models, we tried to see the relationship between the dropout status and each other feature by calculating the Clermont correlation coefficient because the data includes qualitative variable (the number of rows in the cross-table is r, and the number of columns is c) shown in Equation (1).

$$V = \sqrt{\frac{\chi^2}{N \times \min(r-1, c-1)}} \tag{1}$$

The top 8 clamor correlation coefficients for each feature are shown in Table 1.

Table 1. Relationship between feature and target correlation
(The association is closest whenthe coefficient is 1).

Feature	*Cramer's V*
Homeroom teacher	0.107
Number of material count	0.053
Daily number of remarks at Slack	0.052
Course understanding rate	0.052
Understanding count of material	0.051
Course progress rate	0.051
Not understanding count of material	0.050
Test score	0.031

As shown in Table 1, homeroom teacher has the highest correlation with the student dropout status and other features are came from student learning progress.

2. Method

As shown in Figure 1, three types of machine learning models were trained to minimize the ROC[2] AUC[3]: AdaNet, Gradient Boosting Decision Tree, and a model of a forward propagation neural network (Deep Neural Network, DNN). The training data were used to aggregate the values for 3 months. The reason for setting the training data aggregation period to 3 months is to facilitate data collection.

Each model's ROC AUC is shown in Table 2. As a result, Gradient Boosting Decision Tree performed the best with the ROC AUC score 0.969. We adopted this model because a higher ROC AUC means more stable accuracy which can lead to more avoidance of missed student dropouts.

Table 2. ROC AUC by Each Model.

Model name / Duration	**ROC AUC**
Adanet	0.967
Gradient Boosting Decision Tree	**0.969**
Deep Neural Network	0.943

[2] Receiver Operating Characteristic Curve

[3] Area within the ROC curve, the closer to 1, the better

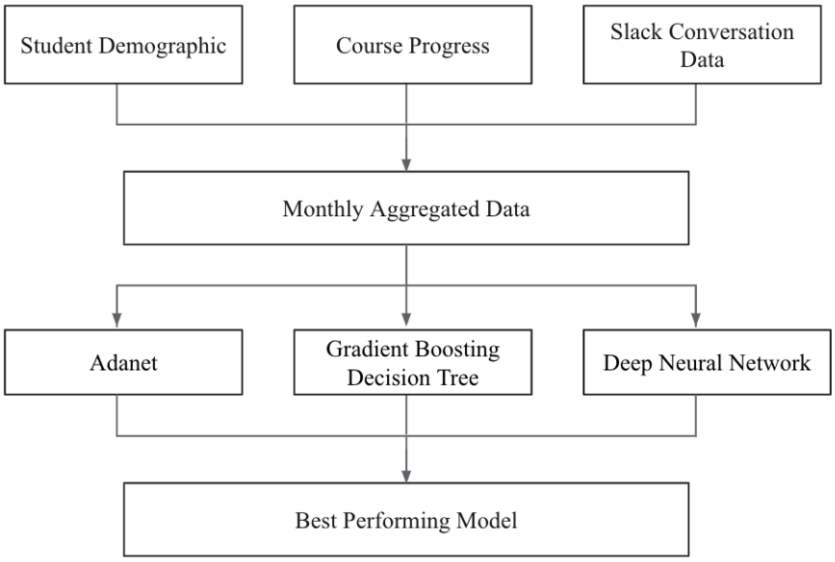

Figure 1. Data and model structure

Table 3. shows the model evaluation score of the Gradient Boosting Decision Tree. The Accuracy was 96.4%, and the F1 Score (harmonic mean of the fit rate and the recall rate, the closer to 1, the better) was 0.773, which is sufficient accuracy.

Table 3. Model Evaluation Scores

Model Evaluation Metrics	**Score**
Accuracy	96.4%
F1 Score	0.773
Precision	0.768
Recall	0.779

For the Gradient Boosting Decision Tree, feature attribution value[4] (normalized to sum 1), was calculated, which are shown in Figure 2 and Table 4 below.

[4] A score that indicates the relative importance of each input feature calculated by Shapley value, which predicts when the column values change. It is determined by whether the degree changes, and the larger the change in the prediction, the higher the importance.

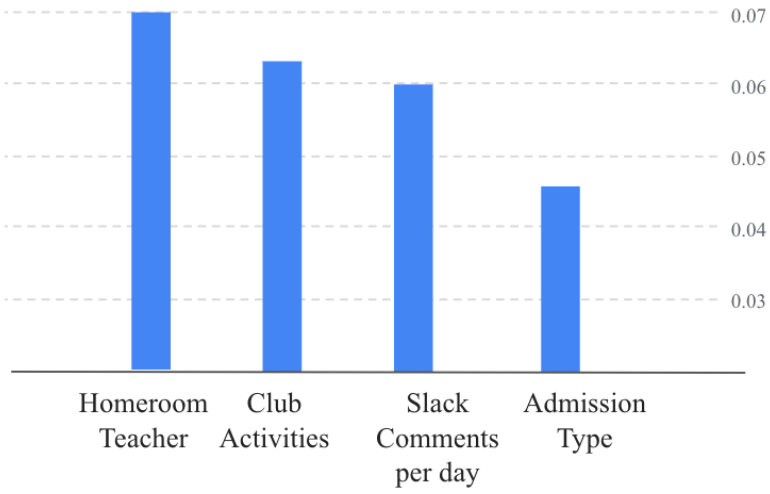

Figure 2. Each Feature's Attribution Values which has over 0.01 Score

Table 4. Feature attribution values.

Feature	Attribution Score
Homeroom Teacher	0.073
Club activities	0.066
Slack comments per day	0.059
Admission Type	0.047
Enrollment month	0.006
Region	0.006
Gender	0.004
Live with parents	0.002

As seen in the Figure 2 and Table 4, Homeroom teacher has the highest feature importance. That said, since each homeroom teacher is in charge of a specific course and region, it won't be the only factor affecting the results and some covariates might be exist. In future research, it is necessary to clarify these covariates.

3. Conclusion

In this study, we trained ML models that predict student's dropout or absence from school based on student attributes, class enrollment, and communication status in Slack at a Japanese leading online school. By analyzing the feature importance, the factors that may have influenced the students' career paths were revealed. We found that items such as homeroom teacher, Slack conversation frequency, and student's club affiliation had an impact on the student's career path in predicting the student's absence from school and withdrawal from school. On the other hand, information related to course completion, such as the number of classes viewed and the level of understanding of course materials, did not contribute as much to the prediction results as expected. Therefore, future

research should use more granular information, such as the test score and grades, in addition to students' course completion and progress. With those result, the online school can detect which factors attribute the most to dropout and they can think of a solution.

This study had several strengths but also some limitations that should be considered. A key strength was the interpretability of ML model by Sharpley values. Specifically, it turned oud the slack conversation data is effective to know how actively the student interacted with other students, which attributes to the prediction result. However, as for the homeroom teacher, it was the most attributed feature but only their names are used, which is considered too little information. Since some factors such like the number of students, and careers vary from teacher to teacher, it is necessary to collect those information and analyze these factors in more detail.

References

[1] C. De la Varre, et al., Reasons for student dropout in an online course in a rural K–12 setting, *Distance Education*, 2014, 35.3, pp. 324-344.

[2] A. P. Rovai, Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks. *The Internet and Higher Education*, (2002). 5, pp. 319–332. doi:10.1016/S1096-7516(02)00130-6.

[3] M. Fei, D.Y. Yeung, Temporal models for Predicting Student Dropout in Massive Open Online Courses, *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, DOI: 10.1109/ICDMW.2015.174.

[4] B. Máté, M. Nagy, and R. Molontay. Interpretable Deep Learning for University Dropout Prediction, Proceedings of the 21st Annual Conference on Information Technology Education, 2020, https://dl.acm.org/doi/pdf/10.1145/3368308.3415382, accessed June 20, 2022.

[5] P.M. da Silva, M.N.C. Lima et al., Ensemble regression models applied to dropout in higher education, *2019 IEEE 8th Brazilian Conference on Intelligent Systems (BRACIS)*, 2019, DOI: 10.1109/BRACIS.2019.00030.

[6] W. Wang, Y. Han and C. Miao, Deep model for dropout prediction in MOOCs. *Proceedings of the 2nd International Conference on Crowd Science and Engineering*. 2017, pp. 26-32, https://doi.org/10.1145/3126973.3126990.

[7] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, Predicting Student Dropout in Higher Education, *ICML Workshop on Data4Good: Machine Learning in Social Good Applications*, 2016, pp. 411–421, https://arxiv.org/pdf/1606.06364.pdf.

[8] D. Delen, A Comparative Analysis of Machine Learning Techniques for Student Retention Management. *Decision Support Systems*, 2010, Vol. 49 (4), pp. 498–506.

[9] Z. Kovacic, Early Prediction of Student Success: Mining Students' Enrolment Data. *Proceedings of Informing Science and IT Education Conference*, 2010, pp. 647–665, https://proceedings.informingscience.org/InSITE2010/InSITE10p647-665Kovacic873.pdf , accessed June 20, 2022.

[10] Y. Zhang, S. Oussena, T. Clark, and H. Kim, Using Data Mining to Improve Student Retention in Higher Education: A Case Study. *ICEIS 2010 - Proceedings of the 12th International Conference on Enterprise Information Systems*, Volume 1, DISI, Funchal, Madeira, Portugal, June 8 - 12, 2010, https://shura.shu.ac.uk/11970/1/%255BCam_Ready%255DICEIS2010%2520Use%2520Data%2520Mining_Ying.pdf, accessed June 20, 2022.

[11] D. Delen, A Comparative Analysis of Machine Learning Techniques for Student Retention Management, *Decision Support Systems*, 2010, Vol. 49 (4), pp. 498–506.