PRODUCT Management & Development
An International Journal for Transdisciplinary Research in Innovation

ISSN 2237-5228

Journal homepage
www.pmd.igdp.org.br

Artigo original

Exploring student course transition prediction in online schools using positive and unlabeled data

KATSURAGI, Miki[2]*, TANAKA, Kenji[1]

[1]*The University of Tokyo (Dept of Technology Management for Innovation, School of Engineering), Tokyo, Japão*
[2]*The University of Tokyo (Dept of Technology Management for Innovation, School of Engineering), 154-0005 2-27-14-503 Setagaya-ku, Mishuku, Japão*

Abstract

In this study, we analyzed the daily learning data of students enrolled in an online school to predict school withdrawal using machine learning techniques. Specifically, we focused on predicting student withdrawal one month in advance based on their attribute data, class enrollment data, and communication records between teachers and students over the preceding three months. Unlike traditional binary classification methods that simply categorize outcomes as Positive or Negative, we employed a Positive and Unlabeled (PU) Learning approach to consider not only the timing of withdrawal but also the potential for future withdrawals. This approach resulted in an improved recall rate, increasing the accuracy of our predictions of student course changes from 63% to 72%.

Keywords: distance learning, dropout prevention, PU Learning.

1. Introduction

While the need for and prevalence of distance education continue to rise, the rates of withdrawal and leave of absence among students in online schools have become a significant concern (De la Varre et al., 2014). Particularly in online courses, it has been noted that students are prone to feelings of isolation, and the lack of communication with other students and teachers often leads to withdrawals, with this isolation being a potential major cause (Rovai, 2002).

To prevent such withdrawals and leaves of absence, various predictive analyses utilizing machine learning models have been conducted domestically and internationally, mostly treating withdrawal as a simple binary classification problem. However, this approach, while predicting withdrawals at a specific point in time, fails to consider future risks. For instance, predicting withdrawals and leaves of absence on October 1st, a month in advance (September 1st), does not account for the potential risks of withdrawals occurring after October 1st (refer to Figure 1).

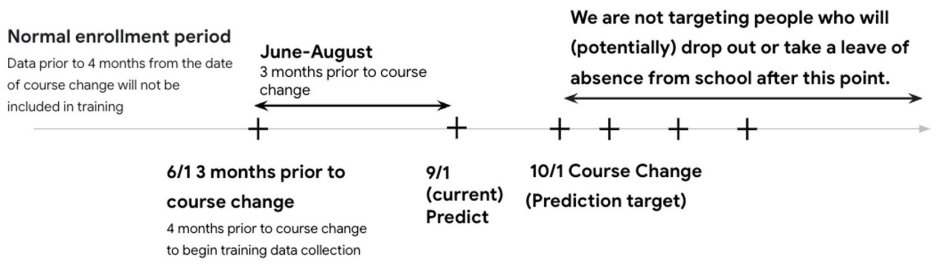

Figure 1. Current forecast timeframe example and issues

Therefore, this study went beyond the mere binary classification of Positive and Negative, and trained a predictive model using Positive and "Unlabeled" data. This model enabled us not only to predict withdrawals but also to consider future risks, improving the recall rate of withdrawal predictions from 63% to 72%. Furthermore, it was found that students who withdrew or took leaves of absence from our school often tended to change courses beforehand (for example, switching from a course including in-person attendance to an online-only course), significantly impacting the school. Hence, this study focuses on predicting such course changes among students.

Received: January 6, 2024. **Accepted:** August 31, 2024.
***Corresponding author:** KATSURAGI, Miki. E-mail: miki.katsuragi1@gmail.com

2. Literature review

The efficacy of machine learning in educational settings, particularly for predicting student behavior, has been a growing area of interest. This section reviews the existing literature on machine learning approaches to predict student dropout and the application of Positive and Unlabeled data analysis. We begin by exploring conventional methods and then delve into more advanced techniques, highlighting their contributions and limitations in the context of our study.

2.1. Machine learning for predicting dropout and leave of absence

In the field of dropout prediction, machine learning approaches have been widely adopted. One of the most classical approaches to predict student dropout involves modeling dropout using survival models (Fei & Yeung, 2015). Additionally, decision tree algorithms are widely used due to their ease of interpretation (Baranyi et al., 2020). Besides these, various machine learning approaches such as logistic regression (Takahashi & Komatsugawa, 2018) and random forests (Chung & Lee, 2019) have also been proposed. However, these traditional parametric models and single AI-based methods often struggle to achieve high prediction accuracy, leading researchers to try approaches that combine multiple models to enhance predictive accuracy (Da Silva et al., 2019). Notably, Wang et al. (2017) demonstrated the effective learning of high-precision models using neural networks without complex feature engineering.

2.2. Predictive analysis using positive and unlabeled data

Positive and Unlabeled Data (also known as PU Learning) is a specialized approach to binary classification in machine learning. Unlike conventional binary classification, where data is presumed to belong to either Positive or Negative classes, the dataset used in this method comprises Positive labeled data and an unlabeled set containing both Positive and Negative samples in unknown proportions. This method is distinct from traditional semi-supervised learning as it does not require prior assumptions about the distribution of data. This approach emerged in the early 2000s (Liu et al., 2002) and has been utilized in applications like personalized advertising and diabetes prediction (Claesen et al., 2015).

2.3. Definition of “Positive” and “Unlabeled” data

In PU Learning, the "unlabeled" dataset refers to a collection where both Positive and Negative instances are mixed together. The advantage of this method is its ability to learn effectively from a limited amount of Positive data and a large amount of unlabeled data, particularly in situations where it is difficult or costly to accurately label all Positive instances. On the other hand, "Positive" data in PU learning has the same meaning as Positive data in traditional binary classification, referring to data that is definitely known to be Positive at the time of analysis.

This study employed PU Learning, treating student course changes as the target variable, with course changes labeled as Positive data and all other cases (students who had not yet changed courses at the time of analysis) as Unlabeled data. While machine learning has been widely applied to predict student dropout, PU data analysis in this context is still underexplored. By leveraging both positive and unlabeled data, our approach offers a more nuanced understanding of dropout behavior, leading to potentially more accurate predictive models. This method advances research in educational data analysis and has practical implications for improving student retention.

3. Analysis method

3.1. Data used

In this study, we utilized a dataset of 2,500 students from a domestic online school. This dataset, collected over three months from June to August 2021, comprises the following elements:

1. **Student Basic Information:**
- **Grade Levels**: The specific grade levels students were enrolled in (e.g., primary, secondary, high school).
- **Courses**: The specific courses or subjects in which students were enrolled (e.g., 3 days a week course, 5 days a week course)
- **Affiliations**: Any affiliations with particular school programs or extracurricular activities (e.g., Computer Programming Courses).
- **Types of Enrolment**: Details on whether the student was enrolled full-time, part-time, or in a special program.
- **Club Activities**: Participation in various club activities, indicating the level of engagement outside of standard coursework (e.g., eSports Club)

2. **Course Enrollment Status:**
- **Report Submission Records**: Data on the frequency, timeliness, and completeness of students' report submissions for their courses.

• **Attendance**: Detailed records of student attendance, including the number of classes attended, missed, and the overall attendance rate (e.g., percentage of classes attended over the three-month period).

3. **Communication Records from Teachers to Students**:

This data includes logs of communication between teachers and students, such as emails, messages, and other forms of interaction. These records reflect the frequency and type of communication, providing insight into the level of support and intervention each student received.

3.2. The definition of "student course changes"

In this study, our primary goal was to predict "negative course changes", which are strongly linked to a higher risk of future dropout and significant revenue loss for the school. Specifically, these changes involve a student reducing their course load, leading to a decrease in tuition (e.g., Changing the course 5day-per-week course to a 3day-per-week course leads to the tuition decrease from 500,000 yen to 300,000 yen), and are labeled as Positive in our model. This approach is based on the school's experience that students who lower their tuition are more likely to drop out. Given the rarity of actual dropouts (less than 5% of students), predicting these course changes is a practical strategy for preventing dropout and minimizing revenue loss.

3.3. Data integration and processing

The data sets, which include student attributes, course enrollment status, and teacher-student communication records, were integrated using unique student IDs to form a comprehensive dataset. This integrated dataset was then used to train a machine learning model designed to predict these negative course changes.

• **Positive Data**: Defined as cases where a student reduced their course load and tuition by October 1st, 2021.

• **Unlabeled Data**: All other cases where the student maintained their current course load or did not make any change in their enrollment status as of the analysis date.

Data preprocessing:

• **Handling of Imbalanced Data**: Given the low incidence of actual dropouts, techniques such as oversampling the Positive class or applying weighted loss functions were considered to address the imbalance and improve model performance.

• **Normalization and Encoding**: Continuous variables, such as attendance rate, were normalized to ensure comparability, while categorical variables, such as grade level and course type, were appropriately encoded.

Dataset characteristic:

• The dataset includes both categorical (e.g., grade level, course type) and continuous variables (e.g., attendance rate).

• The data was anonymized to protect student privacy, with all personally identifiable information (PII) removed or encoded.

3.4. Experimental method

In this research, we compared the results of predicting student outcomes using two different methods: Random Forest and the Elkanoto Classifier, a PU Learning method proposed by Elkan & Noto (2008).

Random Forest:

We implemented the Random Forest algorithm using the scikit-learn library in Python (Pedregosa et al., 2011). The model was configured with the following parameters:

• Number of Trees (n_estimators): 100

• Maximum Depth (max_depth): None (allowing the nodes to expand until all leaves are pure or until they contain fewer than the minimum number of samples required to split)

• Minimum Samples per Leaf (min_samples_leaf): 1

• Bootstrap Sampling: True, to enable random sampling with replacement.

These parameters were chosen based on standard practices in the literature (Breiman, 2001), and we conducted a grid search with cross-validation to fine-tune the hyperparameters for optimal performance.

Elkanoto Classifier:

We also applied the Elkanoto Classifier, a method specifically designed for Positive and Unlabeled (PU) learning scenarios, as introduced by Elkan & Noto (2008). This classifier estimates the likelihood of data points being positive based on the assumption that the labeling process is unbiased and random. The implementation details followed the guidelines from the original paper to ensure accuracy.

By comparing these methods, we were able to assess the effectiveness of traditional supervised learning (Random Forest) against a specialized PU learning approach in predicting student course changes.

4. Results

In this section, we present the results of our analysis comparing the performance of the Random Forest model and the Elkanoto Classifier in predicting student course changes. Our goal was to assess which method better identifies students at risk of making negative course changes, as defined in our earlier discussion.

4.1. Model evaluation metrics

We evaluated the performance of both models using several metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of the model's effectiveness, particularly in the context of our imbalanced dataset where recall is a critical measure due to the low incidence of actual negative course changes. Table 1 summarizes the evaluation metrics for both the Random Forest and Elkanoto Classifier models.

Table 1. Evaluation metrics of each machine learning model

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	94.5%	0.62	0.63	0.62
PU Learning	91.4%	0.50	0.72	0.55

- **Accuracy** refers to the overall correctness of the model, representing the proportion of correctly predicted instances (both positive and unlabeled) out of the total predictions.
- **Precision** indicates the proportion of true positive predictions out of all positive predictions made by the model. In our context, it measures how many of the predicted negative course changes were actually correct.
- **Recall** (also known as sensitivity) measures the proportion of actual positive cases (students who made negative course changes) that were correctly identified by the model. Given the focus on identifying students at risk, a higher recall is preferred.
- **F1 Score** is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives.

As seen in Table 1, the Random Forest model demonstrated higher accuracy and precision, but the Elkanoto Classifier exhibited a significantly better recall. This suggests that while the Random Forest model was more accurate overall, the Elkanoto Classifier was more effective at identifying at-risk students, making it more suitable for our goal of early intervention.

4.2. Data feature examination through clustering

To further understand the distinctions between students who made negative course changes and those who did not, we applied clustering techniques, specifically t-SNE (t-distributed stochastic neighbor embedding), to visualize the data distribution. Figure 2 presents the comparison of numerical distribution between students who changed courses and those who did not, focusing on key features like attendance rate and number of attendances.

- **Attendance Rate** is a continuous variable representing the percentage of classes a student attended during the study period. It is a crucial factor as previous analysis showed its strong correlation with course changes.
- **Number of Attendances** counts the total classes attended by each student during the same period.

Figure 2 shows a bimodal distribution for non-changers (indicated by 'n') and a unimodal distribution for course changers (indicated by 'y'). The data reveals that students who made course changes generally had lower attendance rates, suggesting that declining attendance might be an early indicator of potential course changes

Figure 3 displays the clustering results using t-SNE, highlighting the separation between students who made negative course changes and those who did not. The clear boundary between the clusters indicates that the selected features (e.g., attendance metrics) were effective in distinguishing between these two groups.

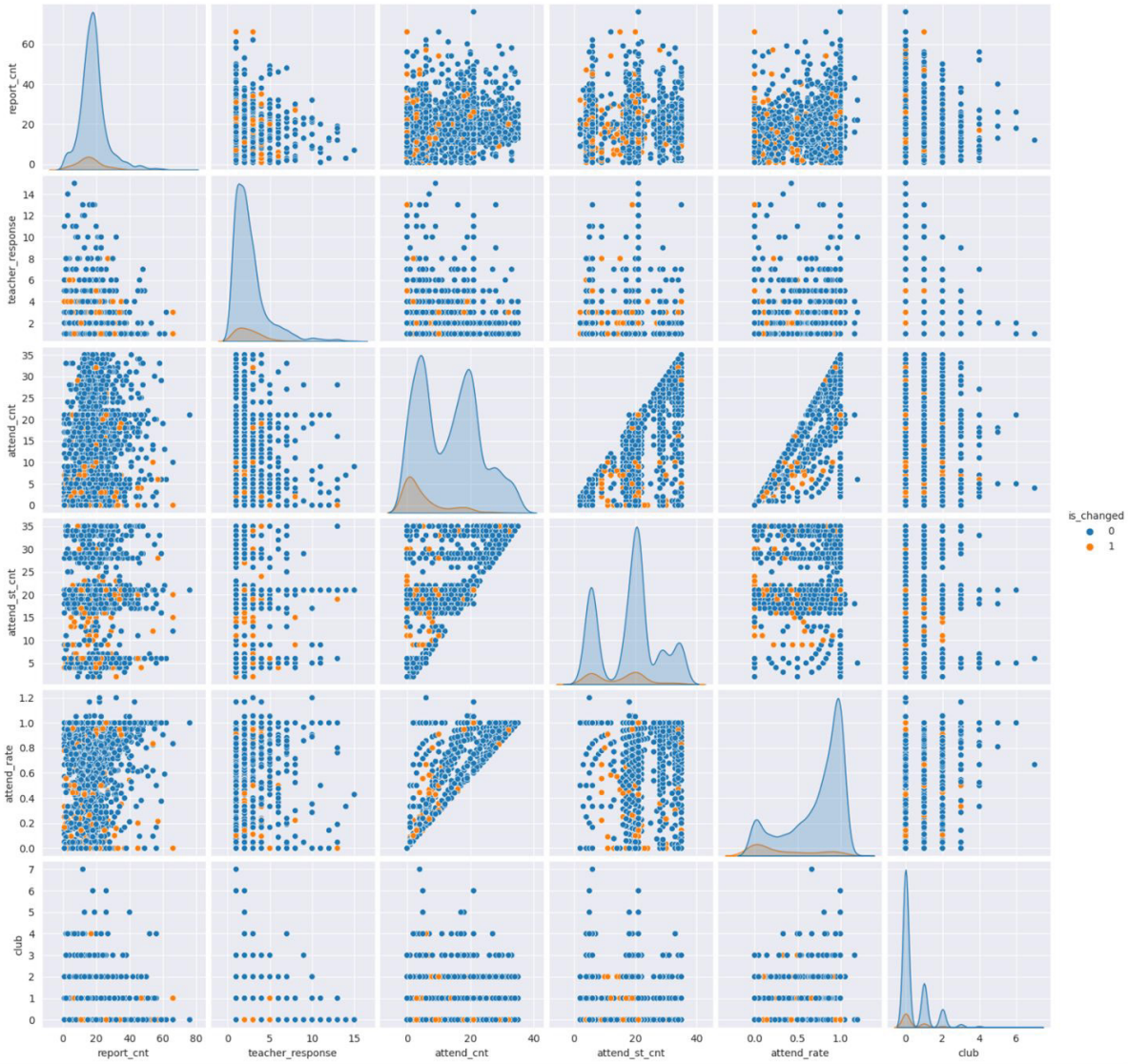

Figure 2. Comparison of numerical distribution between course changers and non-changers

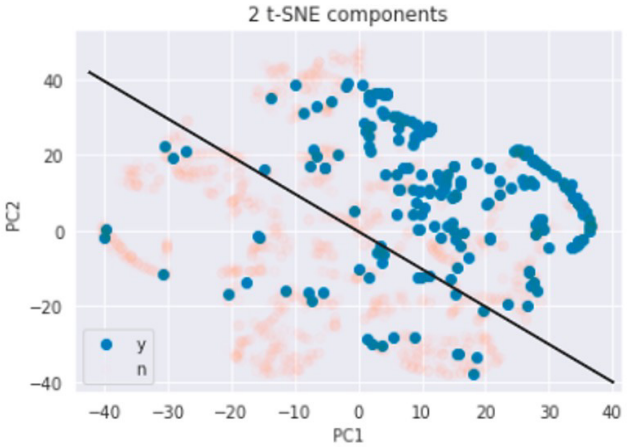

Figure 3. Clustering results by t-SNE

Table 2 provides the average attendance and attendance rate according to the boundary lines identified in Figure 3.

The clustering analysis shows that students who did not change courses had a significantly higher attendance rate and number of attendances. These findings reinforce the importance of these variables in predicting course changes.

Table 2. Average attendance and attendance rate according to t-SNE boundaries

Feature Value	Upper right (Non-changers)	Lower left (Changers)
Number of attendees	5.96	18.74
Attendance rate	0.47	0.89

4.3. Model interpretation and implications

The differences observed in the performance metrics and the clustering analysis suggest that while the Random Forest model offers good overall accuracy, the Elkanoto Classifier's strength lies in its ability to detect at-risk students earlier and more reliably. For educators and administrators, this implies that adopting a PU Learning approach may be more beneficial when the goal is to identify and intervene with students who are likely to make decisions that could lead to dropout.

5. Conclusion

In this study, we developed a machine learning model to predict course changes among students at a domestic online school by analyzing data on student attributes, course enrollment status, and teacher-student communication. By employing PU Learning, we successfully identified students at risk of potential course changes, providing predictions not only at a specific point in time but also capturing the ongoing risk. Our analysis revealed that student attendance rates are a significant factor influencing the likelihood of course changes, dropout, and leave of absence.

Given these findings, we identified specific factors that contribute to student course changes, such as low attendance rates. These insights can be directly translated into actionable strategies to mitigate the identified risks:

1. **Enhanced Individualized Instruction by Teachers**: Since low attendance rates were associated with a higher likelihood of course changes, teachers can use this information to provide more personalized support to students who show signs of disengagement. By tailoring guidance to the individual needs of these students, it is possible to increase their motivation and reduce the risk of them changing courses or dropping out.
2. **Promotion of Student-to-Student Communication**: Our findings suggest that a lack of engagement may be a contributing factor to course changes. Schools can address this by creating opportunities for students to interact with each other, such as through online interactive events and discussion groups. This increased interaction can help build a stronger sense of community, reducing feelings of isolation and promoting student retention.
3. **Organizing Offline Events**: Given the importance of community, particularly during times when face-to-face interaction is limited, organizing offline events can provide students with much-needed personal connections. These events can encourage active participation and strengthen the bond between students and the school, further reducing the likelihood of course changes.

These strategies are grounded in our analysis and are intended to enhance student engagement and satisfaction, especially in remote learning environments. By utilizing the predictive insights generated by our machine learning model, schools can identify students who are at risk early and implement targeted interventions that address their specific needs.

Ultimately, the approach proposed in this study is expected to enhance the effectiveness of online education by reducing dropout and leave of absence rates. Moving forward, such data-driven strategies will likely play a critical role in improving the quality of online education and enriching the learning experiences of students.

However, it is important to note that the reliance on specific datasets in this study may limit the generalizability of our findings to other educational contexts or student populations. Furthermore, the effectiveness of the predictive model may vary depending on the quality and comprehensiveness of the data used. Future research should consider incorporating a broader range of variables, including psychological factors and external influences, to further refine the model's accuracy and applicability.

6. References

Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable deep learning for university dropout prediction. *Proceedings of the 21st Annual Conference on Information Technology Education*, 13-19.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, *96*, 346-353.

Claesen, M., De Smet, F., Gillard, P., Mathieu, C., & De Moor, B. (2015). Building classifiers to predict the start of glucose-lowering pharmacotherapy using Belgian health expenditure data. *Clinical Orthopaedics and Related Research*, 1-23.

Da Silva, P. M., Lima, M. N. C. A., Soares, W. L., Silva, I. R. R., Fagundes, R. A. A., & de Souza, F. F. (2019). Ensemble regression models applied to dropout in higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)* (pp. 120-125). IEEE.

De la Varre, C., Irvin, M. J., Jordan, A. W., Hannum, W. H., & Farmer, T. W. (2014). Reasons for student dropout in an online course in a rural K–12 setting. *Distance Education*, *35*(3), 324-344.

Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pp. 213-220.

Fei, M., & Yeung, D. Y. (2015). Temporal models for Predicting Student Dropout in Massive Open Online Courses. In *IEEE International Conference on Data Mining Workshop* (pp. 256-263). IEEE.

Liu, B., Lee, W. S., Yu, P. S., & Li, X. (2002). Partially supervised classification of text documents. *Proceedings of the Nineteenth International Conference on Machine Learning*, *2*, 387-394.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiis, R., Dubourg, V. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Rovai, A. P. (2002). Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks. *The Internet and Higher Education*, *5*(4), 319-332.

Takahashi, H., & Komatsugawa, H. (2018). Analysis method of dropout students using student ICT-based data. In *The 43rd Annual Conference of JSiSE* (pp. 17-18). JSISE.

Wang, W., Yu, H., & Miao, C. (2017). Deep model for dropout prediction in MOOCs. *Proceedings of the 2nd International Conference on Crowd Science and Engineering*, 26-32.

Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). Using data mining to improve student retention in higher education: a case study. In *International Conference on Enterprise Information Systems* (pp. 190-197). SciTePress.