Ensemble learning for early identification of students at risk from online learning platforms

Li Yu and Tongan Cai*

The Pennsylvania State University, University Park PA 16803, USA,
cta@psu.edu

Abstract. Online learning platforms has made knowledge easily and readily accessible for people, yet the ratio of students withdrawing or failing a course is relatively high comparing to in-class learning as students do not get enough attention from the instructors. We propose an ensemble learning framework for the early identification of students who are at risk of dropping or failing a course. The framework fuses student demographics, assessment results and daily activities as the total learning statistics and considers the slicing of data with regard to timestamp. A stacking ensemble classifier is then built upon eight base machine learning classification algorithms. Results show that the proposed model outperforms the base classifiers. The framework enables the early identification of possible failures at the half of a course with 85% accuracy; with full data incorporated an accuracy of 94.5% is achieved. The framework shows great promise for instructors and online platforms to design interventions before it is too late to help students to pass their courses.

Keywords: Online Learning Environment, Student Dropout, Ensemble learning

1 Introduction

Online learning platforms are changing the way people acquiring knowledge. Instead of going to a classroom, students nowadays can get access to online study materials any time and anywhere by taking massive open online courses (MOOCs). Take Coursera as an example, Coursera is now the world's largest online learning platform for higher education, with more than 200 of the world's top universities and industry educators partners offering courses, certificates, and degrees. It has had over 45 million learners around so far[1]. Another online learning platform, the Open University, is one of the largest universities in Europe with 174,898 students. More than 2 million people worldwide have been learning with the Open University[2].

As [14] suggested decades ago, online learning is superior for distributing course materials efficiently and grant easy access to students without space or

* Contact Author

[1] https://about.coursera.org/press

[2] http://www.open.ac.uk/about/main/strategy-and-policies/facts-and-figures

time constraints, yet the online learning platform can cause students lack of support from the instructor—the instructors may overlook students in need. Since the online courses often have hundreds or thousands of students in one session, it is of great importance for instructors to identify students who are at risk of failing and withdrawing timely and accurately.

We would like to leverage some novel statistical and machine learning ideas to identify students who are at risk of failing and withdrawing in a online study platform. With achieving accurate identification of such students, we would also like to identify them at an early stage of the course. Specifically, we validate our framework on the Open University Learning Analytics Dataset (OULAD) [6]. Comparative results shows that the proposed method can accurately identify students at risk at an early stage.

The following of the paper is organized as follows: Section 2 reviews both early and recent related works in this domain, Section 3 introduces in detail about the proposed method, Section 4 gives a comprehensive presentation and analysis of our experiment & result. A conclusion is then drawn and future possibilities are discussed.

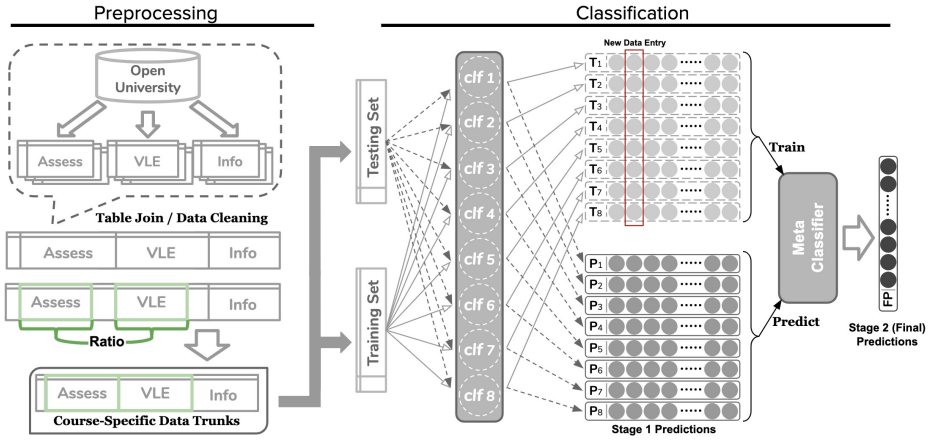


Fig. 1. Flow of proposed framework.

2 Related Works

Although online learning platforms have been running since 1970s, learning analytics have not shed much light on it until recent years. Traditional learning analytic works like [11] focus on cognitive aspect of learning, including learning patterns and decision approaches, but online learning analytics favors a data-driven way. In 2015, the Knowledge discovery in databases (KDD) Cup was held on a dataset collected on a massive open online course (MOOC) in order to predict the dropout (withdrawal) of students. The data contains students'

daily activity logs. [9] adopts a multi-view semi-supervised learning on behavior features, [10] utilizes a clustering method on learning events to stratify student with different levels of activity, and [15] introduces an optimization of a joint embedding function to represent both students and course elements into a single shared space. While some of them achieves nearly 90% of classification accuracy, further improvement was limited due to the size of the dataset.

In 2017, a complete version of learning analytics dataset was published by the Open University [6] with richer features. This dataset is soon studied by various works. [13] uses a decision tree method to study the relation of demographics information to student performance. [3] considers the temporal continuity in the data, models it as a time series and deploys a Long-short-term-memory (LSTM) deep learning model for the task. Both of them reported around 95% classification accuracy. [8] and [12] adopt clustering idea to stratify students into groups with different risk levels. [4] considers a binary feature of student daily activities and first assessment result to classify the student of pass/fail, which achieves 93% precision rate. There are also works using Gaussian Mixture Model [1] and Markov Chain [7]. However, these works use the full dataset. Considering that withdrawn students have clearly missing activity logs during a later time period, the result of them are arguable. Some works start to model an "early identification" process by slicing the features into weeks, intervals or portions. [5] developed a self-learning framework for incremental training on the data to achieve early identification, which achieves similar result to those with legacy data. [2] adopts a time series forest on the activity data for student withdrawal prediction using different percentages of data.

We argue that previous works either missed the significance of course progress modelling or didn't use all data available (activity, assessment, demographics). Specifically, we would like to present a way to fully use the rich data available and also formulate the course progress to achieve early identification. We also propose a novel ensemble classification model for the task.

3 Method

The proposed framework can be stratified into two stages. As shown in Fig. 1, there is a processing stage for data engineering and a classification stage for decision making. The following sections will elaborate on the details of data acquisition & analysis, information fusion & ratio formulation, choice of base classifiers and the stacking ensemble method.

3.1 Data Acquisition & Analysis

As aforementioned, the dataset we use is the OULAD released by Open University in 2017. It is a comprehensive dataset that includes 22 courses and the learning analytics data of 32,593 students. The courses on OU are organized as modules and modules can be presented multiple times in a year. There are a total of 7 courses and 22 presentations of them. As can be seen in Table. 1,

Table 1. Summary of presentations and students performance

module	presentation	Distinction	Pass	Withdrawn	Fail	Total
AAA	2013J	20	258	60	45	383
	2014J	24	229	66	46	365
BBB	2013B	155	648	505	459	1767
	2013J	176	896	644	521	2237
	2014B	166	561	490	396	1613
	2014J	180	972	749	391	2292
CCC	2014B	192	471	898	375	1936
	2014J	306	709	1077	406	2498
DDD	2013B	54	456	432	361	1303
	2013J	98	731	681	428	1938
	2014B	119	360	490	259	1228
	2014J	112	680	647	364	1803
EEE	2013J	127	482	243	200	1052
	2014B	72	285	173	164	694
	2014J	157	527	306	198	1188
FFF	2013B	118	664	411	421	1614
	2013J	187	908	675	513	2283
	2014B	107	547	462	384	1500
	2014J	258	859	855	393	2365
GGG	2013J	141	451	66	294	952
	2014B	128	350	100	255	833
	2014J	127	317	126	179	749

each presentation is named as the year and the order of month it was presented. For example, 2013A means that the module was presented in January of 2013 whereas 2013J indicates the starting month to be October. The number of enrolled students for each module presentation ranges from several hundreds to a few thousands. Students in each module has final results as either "Distinction", "Pass", or "Withdrawn" and "Fail" and the number of students ending with each of the four results are 3024, 12361, 7052, and 10156 respectively. It can be seen that the number of Distinction/Pass almost equals to that of Fail/Withdrawn.

The learning analytics data of students in OULAD can be categorized into three parts. The first part is the demographics of students which include the basic information such as gender, region, education levels, age etc. The second part is their assessment results on tests, exams, and finals over the course time. Their assessment is evaluated as scores from 0 to 100. The third part is the clickstream recordings (10,655,280 entries) of students on each day of a given module presentation. It is the logs of interactions students made with the Virtual Learning Environment (VLE). We average the number of clicks on each day of the module AAA 2013J for the four different final results and plot them in Fig. 2. It is clear that students who managed to pass (Distinction/Pass) the course

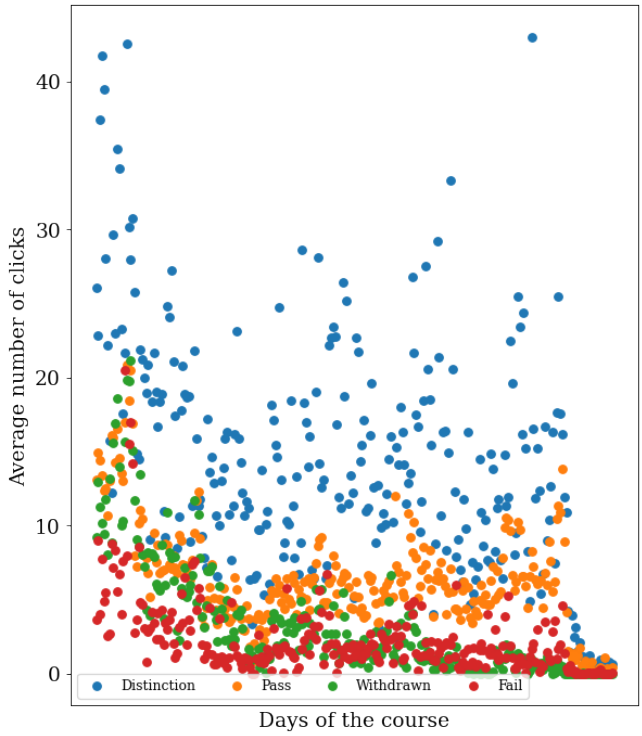


Fig. 2. Average per-day clicks.

tended to be more active than those who failed or withdrew, in terms of daily clicks on VLE. Those who finished with distinction also had more clicks than those who just passed. One interesting observation is that the number of clicks from the withdrawn category gradually diminished and reduced to zero towards the end of the course, which agrees with our expectation that students would not access the course materials after they dropped out.

3.2 Information Fusion & Ratio Formulation

OULAD is presented as a collection of tabular data linked with identifiers like student ids and presentation codes. Fig. 3 shows a typical structure of one module presentation. We can see that the VLE data of students can be retrieved from the tables studentVle and vle on site id, which is the id of sites belonging to one of the twenty activity types such as homepage, resource, and forum on VLE. Similarly, assessment results of students are the combination of table studentAssessment and assessments on assess id, which is the id of assessments such as tests, exams or finals. We have to join the tables in order to get a full representation of student's learning statistics on one of the three categories: Info, VLE, and Assess. During the joining, we find that some assessment and VLE data are missing for

students who dropped even before the class started, so their data are filled with zeros. For the demographics, we perform one-hot encoding on each columns of students' information since they are basically categorical data. We also exclude the info of imd band due to the large portion of missing values.

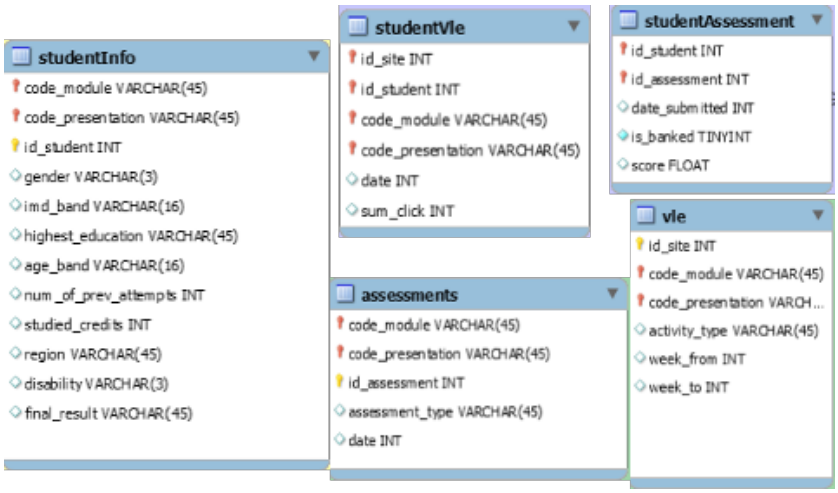


Fig. 3. Data schematic of a typical course.

As mentioned previously, our goal is to identify students at risk of failing or withdrawing a course at an early stage. To accommodate for this, we extract the VLE and Assess data of students based on a ratio of the module presentation length. The ratio starts at 5% and increments every 5% till it reaches 100%. Take AAA 2013J as an example, it spans a total of 268 days and we choose the first 134 days of VLE and Assess data if the ratio is 50%. The demographics data is used as a whole because it won't change during the course presentation. This gives us a ratio dependent student learning statistics so that we can perform classification at different stages of a course and identify students at risk before the course ends.

3.3 Base Classifiers

We choose eight base classifiers with different characteristics and decision boundaries and tune the hyperparamters individually on our extracted dataset. They are:

- K-Nearest-Neighbor Classifier (KNN) with 185 neighbors
- Random Forest Classifier (RF) with 170 estimators
- Support Vector Machine (SVM) with RBF kernel
- Multi-layer Perceptron Classifier (MLP)

- Logistic Regression (LR) with L2 penalty
- Bernoulli Naive Bayes Classifier (BNB)
- Gradient Boosting Decision Tree Classifier (GBDT) with maximum depth of 4
- XGBoost Classifier (XGB) with default hyperparameters

3.4 Stacking Ensemble

Stacking (meta) ensemble method was proposed by David H. Wolpert in 1992 [16] as "stacked generalization", but it is until recent years that stacking ensemble method is widely used in machine learning tasks. The idea of stacking ensemble method is intuitive: determine the final decision based on the results given by multiple base classifiers. It is expected that the stacking ensemble result should be better (at least no worse) than any one of the base classifiers. First, each base classifier is trained in a cross validated fashion so that every data entry gets a preliminary classification result as probability ranging from 0 to 1. The results from multiple classifiers are then regarded as new features and stacked as a new data entry. The new data entries are then feed forward to a second stage "meta" classifier for a final classification decision.

In this proposed frameworks, the stacking ensemble model consists of the eight base classifiers mentioned and has one logistic regression classifier as the "meta" classifier. Each of the hyper-parameters of the base classifiers are tuned on the full dataset using 5-fold cross-validation. For the validation of the stacking ensemble model, we split the data into test and train sets.

4 Experiment & Result

4.1 Deployment & Experimental Settings

We deploy our proposed ensemble model with Python Scikit-Learn's implementation of Nearest-Neighbor, Random Forest, SVM, Neural Network (MLP), Logistic Regression and Bernoulli NB classifiers and XGBoost package. When performing hyper-parameter tuning, we construct 5-fold validation with StratifiedK-Fold cross-validation method to make each fold to resemble the data distribution of the whole dataset. During validation phase, the training and testing data is split randomly with ratio 80% to 20%. We run our proposed method using a quad-core computer

4.2 Performance vs. Percentage of Data Used

The performance metrics we selected are accuracy, precision and recall. In this binary classification problem, a confusion matrix can be formulated as Table 2.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Table 2. Confusion Matrix

Where we note

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \text{Recall} = \frac{\text{TN}}{\text{FN+TN}}$$

We treat each module presentation separately because they have different length of days. The curves of prediction accuracy, precision, and recall with respect to the ratio of data being used are plotted as follows, with Fig. 4(a) shows the prediction accuracy, Fig. 4(b) the prediction precision, and Fig. 4(c) the prediction recall rate.

The proposed framework can achieve very high classification accuracy, since the accuracy rate for the courses can converge to about 90% to 95%. The framework is also able to identify the students at risk accurately in a relatively early stage. Around the middle of the course progress, the accuracy rate rises to higher that 85%, which means it can identify students who are at risk of failing or withdraw (or have withdrawn) prematurely. For validation, we also included the precision and recall curves to support our observation.

We also plot the ROC curves with respect to different stages of the courses. Three ratios of data are chosen as milestones of a typical course presentations. They are respectively 40% (Fig. 4(d)), 70% (Fig. 4(e)), and 100% (Fig. 4(f)). With 40% data incorporated, the AUC score for each of the course are already reasonable, with most curves show higher than 0.8 scores. About half of the cases achieve 0.95 AUC score with 70% data used, indicating a convincing prediction is fulfilled at 70% time of course period. It is not surprising that perfect left-cornered ROC curves are spotted if all data are used, when the course has finished and results are released. From the analysis of accuracy and ROC over different time stamp of the course presentations, we validate that the proposed framework is effective to identify students at risk, at a relatively early stage.

4.3 Base Classifiers vs. Stacking Ensemble

We compare the stacking ensemble model with the base classifiers. The comparison are based on the full dataset with each base classifier tuned on separate

classification task. Top values are marked and top 3 for each metric are blackened and shown in Table 3. There are very strong classifiers like Random Forest,

Table 3. Comparison of ensemble with base classifiers.

Model	Accuracy	Precision	Recall
Nearest-Neighbor	0.8605	0.8006	0.8929
Gradient Boosting	**0.9253**	0.9397	**0.9056**
XGBoost	0.9237	0.933	**0.9078***
Random Forest	**0.9293***	**0.9653***	0.8935
SVM	0.9035	**0.9437**	0.8643
Neural Network	0.908	0.9079	0.8982
Logistic Regression	0.8991	0.8736	0.9092
Bernoulli NB	0.894	0.8928	0.8838
Stacking Ensemble	**0.9293***	**0.9494**	**0.9056**

XGBoost and Gradient Boosting that achieves very good result on one or more metrics. The stacking ensemble seems to be seeking a balance between accuracy, precision and recall, achieving the "generally" best result for the classification.

4.4 Comparison with Related Works

A comparison of our proposed method with related works on the same dataset is shown in Table 4. For the first two works in the table [2,3], they only considered the VLE files and excluded the fail cases. One thing worth noting here is the way we calculate accuracy, precision and recall. We calculate and average them over all the 22 module presentations. Whereas the accuracy reported in [2,3] is the highest score among the 22 presentations. Considering the long sequences of inactivity in later period of a class for the withdrawn cases, their result can be trivial to get. The third work worked on only the demographic data and the fourth one only worked on binarized daily VLE features. Both of them excluded the withdrawn cases. The fifth work validates the method only on a specific course and may not be suitable for all the courses. Comparing to the related works, we can conclude that the proposed method shows improvements over the previous works in terms of 1) a better coverage of cases and features 2) a globally good result on all the course data.

5 Conclusion & Discussion

In this project, we present an ensemble learning method for the early identification of students at risk of failing and withdrawal in an online learning environment. The proposed information fusion utilizes as much available information

Table 4. Comparison with related works.

Work	Method	Accuracy	Precision	Recall	Note
[2]	Time Series Forest	0.9399	/	/	VLE, Excl. "fail"
[3]	LSTM	0.9725	0.9279	0.8592	VLE, Excl. "fail"
[13]	Decision Tree	<0.85	<0.87	<0.35	Demographics, Excl. "withdrawn"
[4]	DT/RF/LR/SVM	0.8797	0.9333	0.8950	Binary VLE, Excl. "withdrawn"
[1]	Mixture Model	0.9200	/	0.9350	Specific course
Ours	Stacking	0.9293	0.9494	0.9056	/

as possible by formulating the student activity data as daily numerical features and combining with the assessment results for the corresponding time period. The student demographics information is also considered as categorical features and included in training. The classification module utilizes various classifiers of different types in order to improve the robustness and adaptability of the model.

The proposed method is validated on the Open University Learning Analytics Data and achieves great accuracy in identifying the students at risk. It achieves 94.94% classification precision and maintained 90.56% recall for the identification of students at risk using full available data. The proposed method also achieves the goal of "early identification", that achieves higher than 85% accuracy with only half of the data incorporated, indicating that the proposed framework can correctly identify students who are at risk around the mid-term of the course.

We also recognize that there are some weaknesses in this project. First, the proposed stacking ensemble is not presenting a drastic improvement from the strong XGBoost and Random Forest classifiers. With limited data, the base classifiers can be good enough to do the job. A trial is made ruling out the XGBoost, Gradient Boosting and Random Forest classifier. The overall result of classification decreased, but the ensemble model shows the improvement from the baselines.

Table 5. Comparison of ensemble with base classifiers.

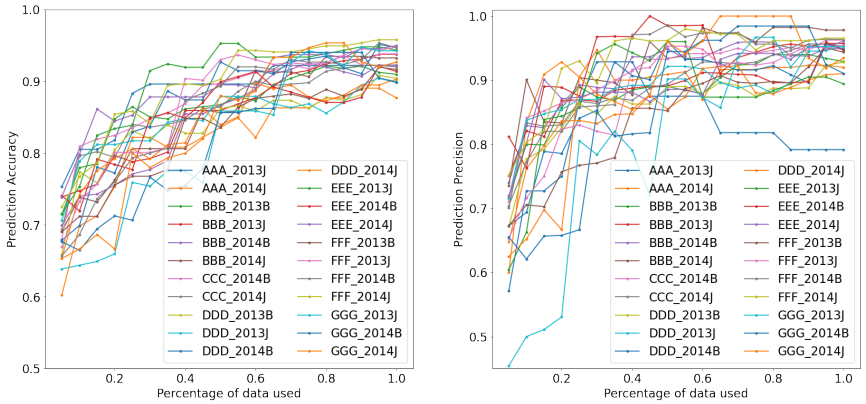
Model	Accuracy	Precision	Recall
Nearest-Neighbor	0.8605	0.8006	0.8929
SVM	0.9035	**0.9437***	0.8643
Neural Network	0.908	0.9079	**0.8982**
Logistic Regression	0.8991	0.8736	**0.9092***
Bernoulli NB	0.894	0.8928	0.8838
Reduced Model	**0.9143***	**0.9249**	**0.8967**

Another problem is that the proposed method, like other previous works, can't distinguish between the students who are likely to withdraw or who are likely to fail. Considering that the withdrawal cases have long sequences of inactive states, the identification of withdrawal is trivial if it is not required to also predict the actual withdraw timestamp of the student. Future works may try to work on early identification of the withdrawal and failing cases separately before it is too late and study the factors that result in the withdrawal/fail.

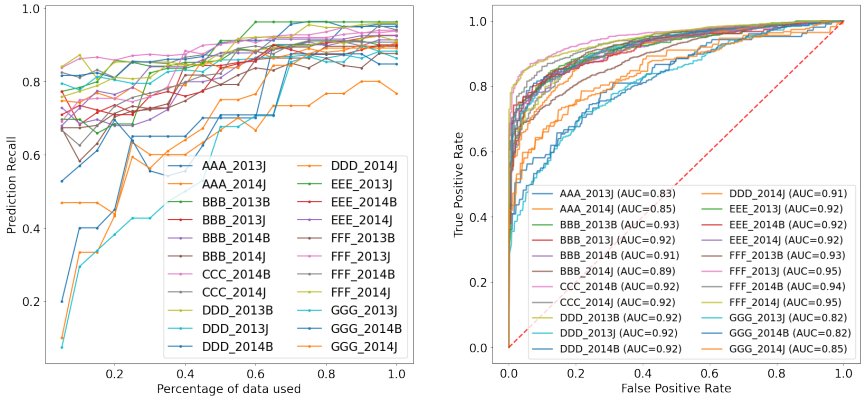
References

1. Raghad Alshabandar, Abir Hussain, Robert Keight, Andy Laws, and Thar Baker. The application of gaussian mixture models for the identification of at-risk learners in massive open online courses. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2018.
2. Liu Haiyang, Zhihai Wang, Phillip Benachour, and Philip Tubman. A time series classification method for behaviour-based dropout prediction. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, pages 191–195. IEEE, 2018.
3. Saeed-Ul Hassan, Hajra Waheed, Naif R. Aljohani, Mohsen Ali, Sebastián Ventura, and Francisco Herrera. Virtual learning environment to predict withdrawal by leveraging deep learning. *International Journal of Intelligent Systems*, 34(8):1935–1952, 2019.
4. Hendrik Heuer and Andreas Breiter. Student success prediction and the trade-off between big data and data minimization. *DeLFI 2018-Die 16.E-Learning Fachtagung Informatik*, 2018.
5. Martin Hlosta, Zdenek Zdrahal, and Jaroslav Zendulka. Ouroboros: early identification of at-risk students without models based on legacy data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 6–15. ACM, 2017.
6. Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific data*, 4:170171, 2017.
7. Jakub Kuzilek, Jonas Vaclavek, Viktor Fuglik, and Zdenek Zdrahal. Student dropout modelling using virtual learning environment behaviour data. In *European Conference on Technology Enhanced Learning*, pages 166–171. Springer, 2018.
8. Jakub Kuzilek, Jonas Vaclavek, Zdenek Zdrahal, and Viktor Fuglik. Analysing student vle behaviour intensity and performance. In *European Conference on Technology Enhanced Learning*, pages 587–590. Springer, 2019.
9. Wentao Li, Min Gao, Hua Li, Qingyu Xiong, Junhao Wen, and Zhongfu Wu. Dropout prediction in moocs using behavior features and multi-view semi-supervised learning. In *2016 international joint conference on neural networks (IJCNN)*, pages 3130–3137. IEEE, 2016.
10. Zhenhui Liu, Jingjing He, Yufei Xue, Zhenzhong Huang, Manli Li, and Zhihui Du. Modeling the learning behaviors of massive open online courses. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2883–2885. IEEE, 2015.
11. Theodore J. Mock, Teviah L. Estrin, and Miklos A. Vasarhelyi. Learning patterns, decision approach, and value of information. *Journal of Accounting Research*, 10(1):129, Apr 1, 1972.

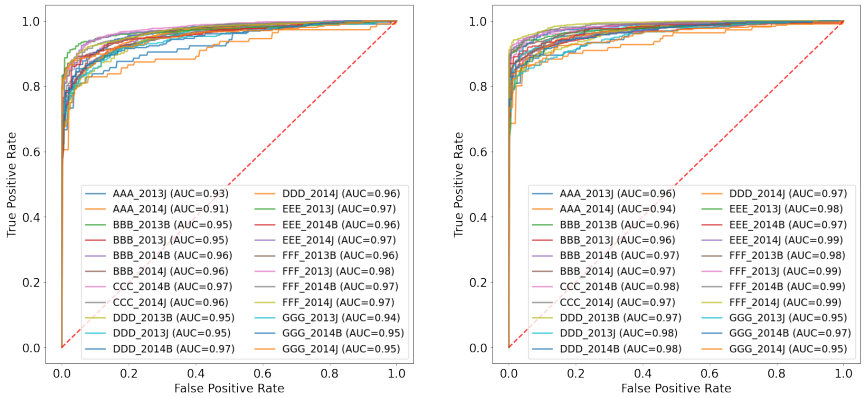
12. Robert L. Peach, Sophia N. Yaliraki, David Lefevre, and Mauricio Barahona. Data-driven unsupervised clustering of online learner behaviour. *arXiv preprint arXiv:1902.04047*, 2019.
13. Saman Rizvi, Bart Rienties, and Shakeel A. Khoja. The role of demographics in online learning; a decision tree based approach. *Computers & Education*, 137:32–47, 2019.
14. Robert W. Taylor. Pros and cons of online learning - a faculty perspective. *Journal of European Industrial Training*, 26(1):24–37, Feb 1, 2002.
15. Milagro Teruel and Laura Alonso Alemany. Co-embeddings for student modeling in virtual learning environments. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 73–80. ACM, 2018.
16. David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.



(a) Accuracy vs. percentage of data used (b) Precision vs. percentage of data used



(c) Recall vs. percentage of data used (d) ROC curve using 40% of data



(e) ROC curve using 70% of data (f) ROC curve using all of data

Fig. 4. Identification result. Plot a-c shows the prediction accuracy, precision, and recall rate over incremental amount of data. Plot d-f shows the ROC curve using 40%, 70%, and 100% data.