Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Assessment of At-Risk Students' Predictions From E-Book Activities Representations in Practical Applications

Erwin D. LOPEZ Z.[a*], Tsubasa MINEMATSU[a], Yuta TANIGUCHI[a], Fumiya OKUBO[a] & Atsushi SHIMADA[a]
[a]*Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan*
*lopez.zapata.erwin.242@s.kyushu-u.ac.jp

Abstract: The use of e-book reading systems such as Bookroll, and their ability to record readers' activities allows the design of predictive models capable of identifying at-risk students from their reading characteristics. Even though previous works have obtained promising results in this task, these results may not evidence the expected prediction performance in practical applications due to their selected assessment methods. Accordingly, in this paper, we assess this performance in two practical scenarios. The first is when we keep stored data from previous years of our course which can be used to train our model, and the second is when we only have data from a different course to use in this training process. In order to obtain a more accurate assessment, we collected 92,574 samples of predictive performances from different models under the above-mentioned conditions. We also considered different feature representations along with variational latent representations, which can leverage our previous data to automatically design general hidden features. From our results, we understand that in the first condition we can expect a relatively good predictive performance, especially when using variational latent representations. However, in the second condition we found that even when using them, the predictive performances are very limited resulting in an impractical solution.

Keywords: e-book data, at-risk students' prediction, model generalizability, VAE

1. Introduction

The adoption of digital textbook reading systems in learning environments has opened the possibility to analyze the students' behaviors and predict their learning outcomes based on the action logs they leave in these systems (Flanagan & Ogata, 2017). An example of these reading systems is Bookroll, an application used in university classrooms for distributing class reading materials. This e-book application also includes features that provide students with a variety of digital interactions, such as adding memos, jumping to an arbitrary page, highlighting text, etc. (Flanagan & Ogata, 2017; Ogata et al., 2017)

From previous works, we know that Bookroll users can be modeled as a set of features extracted from their set of action logs. These features are usually defined as some actions' frequencies (e.g., the number of times that a student uses the highlight function of the application), counted from a certain predefined subset of the total data (e.g., the subset of in-class activities), and selected after conducting a features engineering process. By using their own designed feature representations, different works have attempted to use Machine Learning (ML) supervised models to predict at-risk students (defined as students who get a low score) and obtained promising results (Akçapinar et al., 2019; Hasnine et al., 2018; Huang et al., 2020; Murata et al., 2021; Okubo et al., 2018; Yin et al., 2019). However, apart from Huang et al. (2020), these works limited their evaluation scope to a similar dataset (e.g., subsets of data from the same course) without considering test subsets from different datasets (low variance in the data).

While all these developments are useful in the design of strategies for identifying at-risk students with predictive models, in practice, we are limited to using past data to train our models, which means that our data may preserve some differences. Considering that a usual methodological issue in Learning Analytics (LA) is the presence of biases (Pelanek, 2020), we need to verify whether the reported predictive performance in the literature applies to practical implementation. This assessment is

also in line with the idea of designing a model robust enough to still work with a dataset generated from a different educational environment proposed by Baker (2019) in a LAK 2019 keynote, and by Flanagan et al. (2018) in a workshop from ICCE2018 on predictive models based on Bookroll data.

Consequently, in the present work, we employ an improvement of the best assessment method found in the literature to explore the generalizability and practical performance (predictive performance in a practical condition) of different models and existing sets of features. Additionally, since previously designed feature sets do not necessarily represent the best choices, based on the papers (Le et al., 2018, Lopez et al., 2022) we implement Variational Autoencoders (VAE) and Semi-supervised VAE (Semi-VAE) to generate latent representations and assess the general performance achieved by them.

On that account, our research questions are as follows.

RQ1: In a university course supported by the Bookroll application and in a practical scenario where we use past data from the same course, to what extent can we predict the presence of at-risk students?

RQ2: In a university course supported by the Bookroll application, to what extent can we use past data collected from a course with different characteristics to predict the presence of at-risk students?

RQ3: Can VAE latent representations improve the predictive performance in practical scenarios?

2. Previous Work

2.1 Feature representations of Bookroll data

As mentioned in the previous section, there is found a trend of works that employed ML supervised models to map their designed set of features to the probability of being an at-risk student. We identify a total of 52 possible features through these works with 4 prevalent sets of features, which are designated as Feature Sets 1 to 4 ("FS1", "FS2", "FS3", and "FS4") in the present paper. Specifically, Hasnine et al. (2018) and Akçapinar et al. (2019) make use of similar features related to the engagement (e.g., number of sessions, short events, etc.), included in "FS1". Yin et al. (2019) present a unique set that considers hidden features (e.g., the ratio between the number of PREVs and NEXTs) included in "FS2"; while Huang et al. (2020) and Cheng et al. (2021) employ larger sets of basic features (e.g., number of PREV, NEXT, ADD_BOOKMARK, etc.) considered in "FS3". The "FS4" includes the features used by Okubo et al. (2018) and Murata et al. (2021), which consists of the number of ADD_BOOKMARKs, ADD_MEMOs, read slides, and the total number of events. We should point out that these last works also use data from sources other than Bookroll. However, we include their set of features because only these works employ time-aware models such as RNNs, and therefore, their designed features can be important when assessing these models' performances. Finally, in order to implement VAE models, we also consider all the existing features in the "ALL" Features Set.

2.2 Variational Autoencoders

Since each course has particular characteristics, the same feature may have different meanings in different datasets. Accordingly, the best feature selection may depend on the dataset characteristics.

On that account, we also explore the automatic feature selection with the use of VAE. This model is a special type of autoencoder where the encoder maps the input vector to a probability distribution of latent representations instead of a fixed one. The term "Variational" comes from the Variational Bayesian approach employed to define the optimization equations required to learn the parameters of this probabilistic model without facing intractability problems (Kingma & Welling, 2014). The VAE can be trained just as an autoencoder that learns the mean and standard deviation of their latent probabilistic distribution (Figure 1). The VAE Loss function is shown in equation 1, where D_{KL} is the Kullback-Leibler divergence that decreases when $q_\varphi(z|x)$ is near to be a normal distribution, and β is a hyperparameter introduced by Higgins et al. (2017) to avoid entanglement in the latent space.

$$LOSS = MSE(X, X'') + \beta D_{KL}\left(q_{\varphi}(z|x)||p(z)\right) \quad (1)$$

Although our previous work was the first approximation to the use of VAE models for modeling Bookroll users' data (Lopez et al., 2022), different works in the LA field have used these models earlier. One example is that of Du et al. (2020), who implement a VAE for modeling at-risk students' latent representations and from them generate new samples of at-risk students to address the data imbalance issue presented in the K-12 datasets. Also, Ding et al. (2019) model MOOCs users with the use of a modified LSTM-AE to avoid the challenging process of designing handcrafted features that are effective for their prediction task. In this case, the authors use the model to substitute the feature representations with latent representations when training a predictive model, a method that allows an improvement of its prediction accuracy and reduces overfitting. Similar to this work, Bosch (2017) uses a VAE to extract the latent representations and feed them into a predictive model to detect the students' boredom. Its results also show a performance improvement when compared to predictive models that used manually designed features.

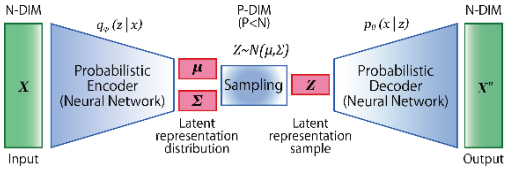

Figure 1. VAE architecture.

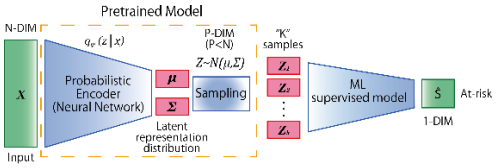

Figure 2. VAE-supervised model integration.

The data pipeline implemented by Ding et al. (2019), Bosch (2017) with educational data, and also by Lopez et al. (2022) with Bookroll data is shown in Figure 2. It can be described as the integration of two models, the first consisting of a trained VAE encoder and the second consisting of an ML-supervised model. Here the trained VAE encoder will apply non-linear transformations to the input vector to obtain its latent representation, and the ML supervised model may learn the relations between this representation and the probability of being an at-risk student. Since the VAE encoder is first pretrained on the task of reconstructing the data, it can encode any input data into general latent features.

2.3 Predictive models and methods

The works from Section 2.1 have examined the predictive performance of different ML supervised models to predict at-risk students from their Bookroll data traces (Akçapinar et al., 2019; Hasnine et al., 2018; Huang et al., 2020; Murata et al., 2021; Okubo et al., 2018; Yin et al., 2019). In Table 1 the reader can observe the models present in the literature and the number of times that have been considered by different works. These frequency values are based not only on a review of the mentioned works but also on the summary presented by Flanagan et al. (2018) from a workshop on learning performance prediction with Bookroll data.

Table 1. *ML supervised models in the literature.*

Model	Frequency
BartMachine	1
CN2Ruler Inducer	1
Generalized Linear Model	1
Rule-Based Classifier	1
AdaBoost	2
Gated Recurrent Unit	2
k-Nearest Neighbors	4
Logistic Regression	5
XGBoost	6
Decision Tree	6
Naïve Bayes	6
Neural Networks	8
Random Forest	8
Support Vector Machine	10

Since these works have adopted different predictive performance metrics and used different datasets, their results cannot be compared. However, from the variety of optimal models found in them, we can infer that even when delimiting our problem as an at-risk students' prediction using Bookroll data, no one model is the best. For this reason, is also difficult to assert if any of the possible set of features in the literature is better at encoding the information about the students' behaviors.

Furthermore, as regarded by the bias-variance trade-off (Kohavi & Wolpert, 1996), there is a chance that some features or models are good at making predictions in datasets with the same characteristics as their training one, but struggle to work with slightly different datasets. This causes assessment methodologies based on two datasets from the same course, and on two datasets from different terms of the same course to lead to different performance results. With this in mind, we should also review and analyze the assessment methodologies present in the literature.

First, Hasnine et al. (2018), Akçapinar et al. (2019), and Murata et al. (2021) only have data from a single course, and consequently, employ a cross-validation method to obtain their results. This methodology avoids a possible bias generated when selecting the train and test splits. However, it is difficult to account for the predictive performance gain due to the lack of variance in data, meaning that their results may be different from those obtained when using past data for training.

Although Huang et al. (2020) do not detail their methodology, we can infer from their results that they assessed an already trained model across 17 different datasets. Some of their performance results suggest that the LA bias problem prevents models to work with different datasets.

Finally, Okubo et al. (2018) and Cheng et al. (2021) implement a methodology based on using data from different sections of the same course. Specifically, the first work uses 15 different datasets and applies cross-validation for each course. In other words, for each validation loop, they define the data from 14 different sections as a train set and the remaining one as a test set. In contrast, the second prepared two datasets, each one corresponding to a course section, and applies 3-fold cross-validation with one of them to generate their train-test subsets, while using the other to assess them. This method is the best designed to avoid possible biases since it considers an additional test set taken from another environment (we will refer to it as assess_test). However, similar to the first kind of work, their results do not address the case when its assess_test contains past data or data with different characteristics.

3. Dataset

We used four datasets previously collected by our laboratory, of which the first one is managed internally, and the other three have been released to the public in the LAK22 Data Challenge Workshop. The datasets details are provided below.

DS1: A dataset containing 5,377,001 logs of unlabeled data from 2054 Kyushu University students enrolled in 16 different courses for the years 2019 and 2020. The considered courses have different characteristics, such as their duration, the professor in charge, their corresponding academic department, etc. The dataset does not preserve these details but contains information about their class schedule.

A-2019: A dataset containing 129,358 logs of labeled data from 50 Kyushu University students enrolled in the course “Programming Theory” in the year 2019.

A-2020: A dataset containing 147,452 logs of labeled data from 62 Kyushu University students enrolled in the course “Programming Theory” in the year 2020.

B-2020: A dataset containing 197,593 logs of labeled data from 90 Kyushu University students enrolled in the course “Cybersecurity Basic Theory” in the year 2020.

We detail some additional characteristics regarding these datasets in Table 2.

Table 2. *Datasets additional characteristics*

Characteristics	A-2019	A-2020	B-2020
# Students	50	62	90
# At-risk students	19	16	15
# Materials	21	12	9
Weeks duration	8	8	7
Professor ID	SA	SA	MT

4. Method

4.1 Experiment Preparations

The labels of our labeled datasets consisted of the students' final grades. We should mention that this grade is specific to its corresponding course. Similar to previous works, we considered the students with low scores (D and F) as at-risk students and prepared our datasets based on these new labels. Then, we prepared feature representations of the data considering FS1, FS2, FS3, FS4, and ALL sets. Also, we trained VAE and Semi-VAE models to generate data latent representations.

4.1.1 VAE latent representations

We first obtained the feature representations of dataset DS1 and then trained a VAE for each set. Each VAE training process was based on two steps. In the first, we made an approximation of the optimal architecture, the number of latent dimensions, and the values of the hyperparameters with an explorative studio with the support of the optimization framework Optuna (Akiba et al., 2019). In the second step, we manually set these values with a visual inspection of the training and validation loss values. For these two steps, we applied z-score normalization across the feature columns and randomly split the 2054 samples into train and validation datasets to avoid overfitting.

4.1.2 Semi-Supervised VAE representations

As the reader may note, the VAE models can eliminate information relevant to the at-risk students' prediction task in their search for better generalizability. The Supervised VAE proposed by Lopez et al. addresses this problem with the use of labels in the whole dataset. However, since we cannot access the labels of DS1 (unlabeled data), we need a model able to use only the existent labels. On that account, we propose a Semi-Supervised VAE model, which has the same architecture as a Supervised VAE (See Figure 3) but defines its prediction loss as shown in Equation 2.

$$LOSS = MSE(X, X'') + \beta D_{KL}\left(q_{\varphi}(z|x)||p(z)\right) + \gamma \sum_{i=1}^{k, s_i \neq -1} s_i \log \widehat{s_i} \qquad (2)$$

Therefore, if we label with a masking value of −1 all the unlabeled samples, the model will be able to only consider the samples with at-risk probability labels to calculate its prediction loss and hence prevent itself to eliminate the information of features related to the prediction performance.

Accordingly, we considered DS1 as our unlabeled dataset and A-2019 as our labeled dataset. We generated the masking labels to DS1 and applied the same methodology described in the previous section to obtain 10 additional Variational models.

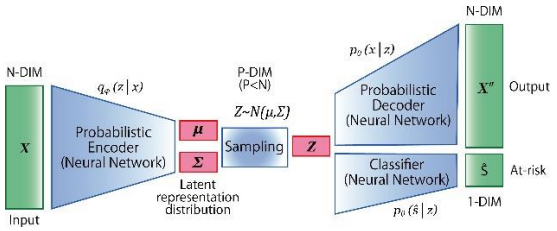

Figure 3. Semi-Supervised VAE architecture.

4.2 Models

Since no one model is the best for our application, we selected 7 models from Table 1. Our selections are the Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), Decision Trees (DT), XGBoost, Neural Networks (DNN), and Gated Recurrent Unit (GRU). We selected the XGBoost over the Random Forest since they share the same architecture but the first usually outperforms the second. Also, we included the GRU because is the only one that considers the time dimension.

4.3 Experiment

We conducted two similar experiments. In the first experiment, we select one Feature Set (e.g., FS1) to represent our different datasets. Then, similar to the best-found assessment method of Cheng et al. (2021), we use 3-fold cross-validation to split the A-2019 dataset into train-test subsets. Next, we train one selected model (e.g., DNN) with these subsets and collect its test performance results. After that, we use the A-2020 dataset as an assess_test subset (Section 2.3) to collect a new predictive performance of the same model. We finally repeat this process across all our selected models, Feature Sets, and latent representations of our Feature Sets (generated by our VAE models, Section 4.1.1 and 4.1.2). In the second experiment, we conduct the same process using the B-2020 dataset as the assessment dataset. The reader may note that the first experiment is intended to answer RQ1 since we use past data to make predictions in a current course, while the second can answer RQ2 since the A-2019 and B-2020 courses exhibit different characteristics (especially, the professor in charge and delivered content).

5. Results and discussion

5.1 Results of models trained on feature representations

From the two described experiments, we collect 92,574 samples of predictive performance results from different models trained on different feature and latent representations. To address RQ1 and RQ2 avoiding a bias due to our proposed latent representations, we consider only the results from models trained on the different Feature Sets. In Figure 4, the reader can see the distribution of the assessment predictive performance (F-score) samples collected in experiment 1 (N = 10,363) and experiment 2 (N = 10,363). We also plot the test predictive performance samples distribution (N = 10,363) to allow verification that our results are in line with the test performances reported in the literature.

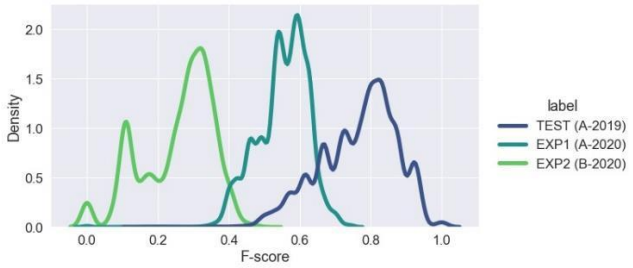

Figure 4. Test and assess_test subsets predictive performances.

From the experiment 1 results, we understand that when using past data to train our models, even if it is from the same course, we should expect a significant decrement (around 0.2 F-score in this case) in the predictive performance in contrast to our k-fold cross-validated test results. However, the mean of the experiment 1 distribution is 0.552 F-score, which is higher than the expected F-score of a non-predictor model (model that always predicts that a student is not at risk, 0.41 F-score for experiment 1), meaning that our designed models are still useful.

On the other hand, the experiment 2 results suggest that using past data from a different course may not be of use for making predictions, since their mean distribution of 0.257 F-score is lower than the expected F-score of a non-predictor model (0.286 for experiment 2).

Finally, although the literature suggests there is no optimal model, we account for the differences between models in our results. To this end, we compared the distributions from different models (N = 300) when using a fixed Feature Set and validated this comparison with a T-Test ($p<0.05$). We then ranked these models (The 1st order means the best model, the largest order the worst, and the same order indicates there is no statistical difference) as shown in Figure 5.

We observe that only the GRU model gets better results in all sets, meaning that we can rely on improving the test performance of our GRU models to also improve their performance in a real application. This result can be explained by the time-awareness of the models, meaning that the data changes in time preserve similar information in any course. Additionally, we also note that the simple models LR and SVM get better performances in the A-2020 assessment set, while there is no best model in the B-2020 assessment set. Finally, we can select only LR, SVM, and GRU to get a more accurate expected performance when using past data from the same course to train our models. Through this process, we get that their corresponding means are 0.582, 0.572, and 0.575 F-scores respectively.

Figure 5. Models' performance rank in the test and assessment subsets.

5.2 Results of models trained on latent representations

To address RQ3, we now use also the VAE latent representations performance results. In this case, we fix a Feature Set and model to compare the results distributions of this model without latent representations (N = 300), the same model with the use of VAE (N = 300), and with the use of Semi-VAE (N = 300). We validated this comparison with a T-Test ($p<0.05$) and ranked the distributions from 1st to 3rd, assigning the same order when there is no statistical difference. The summarized results are shown in Figure 6.

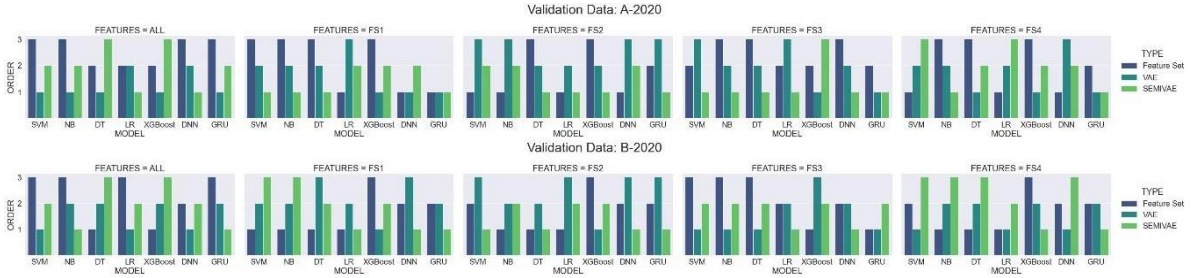

Figure 6. Feature Set, VAE, and Semi-VAE performance rank in the assessment subsets.

We observe that in general the VAE and Semi-VAE latent representations lead to models with better results in the A-2020 assessment data, however, in dataset B-2020 is difficult to determine which representations are better. If we consider each model mean as a sample and compare the distributions (N=7), in the A-2020 assessment data we get that the VAE and Semi-VAE performance are better than the non-latent representation ($p=0.0362$, $p=0.0191$ respectively). However, if we conduct the same process in the B-2020 data we get no statistical difference ($p=0.5506$). In addition, we see in Figure 6 (A-2020 data) that Semi-VAE tends to get better results more often, however, we do not have the statistical validation to claim that it is better than VAE ($p=0.3418$).

Finally, we report in Table 3 some of the better results obtained by VAE and Semi-VAE representations in the A-2020 assessment data. With also these last results in consideration, in response to RQ1, we conclude that in a university course supported by the Bookroll application and in a practical scenario where we use past data from the same course, we can still make predictions about the presence of at-risk students and expect an F-score 0.2 higher than a non-predictor model. In response to RQ2, we conclude that considering the current state of the art, it is not possible to predict the presence of at-risk students when using past data collected from a course with different characteristics (specific differences

in this study: delivered year, course content, and professor in charge). Moreover, in response to RQ3, we conclude that VAE latent representations can improve the predictive performance in the practical scenario when we use past data of the same course. However, we cannot assert whether VAE or Semi-VAE latent representations are better for this task.

Table 3. *Some of the highest performance distributions means (F-score)*

Model	Feature Set	Non-latent	VAE	Semi-VAE
DT	ALL	0.5529	0.6257	0.5334
GRU	ALL	0.5462	0.6215	0.6113
SVM	FS1	0.5898	0.6158	0.6185
NB	FS1	0.5806	0.6197	0.6333
GRU	FS2	0.5689	0.5595	0.6212
NB	FS3	0.4458	0.5740	0.6193
XGBM	FS4	0.5249	0.6294	0.5971
GRU	FS4	0.5872	0.6079	0.6070

5.3 Latent representations

The importance of the VAE models is that they can automatically select some features and how to use them to design their latent representations. In this regard, we additionally conduct a feature importance test (permutation feature importance algorithm proposed by Bisher, Rudin, & Dominici, 2019) for our variational models. The results for the VAE and Semi-VAE models are detailed in Figure 7. Though, these results are specific to our datasets and may not apply to any Bookroll dataset.

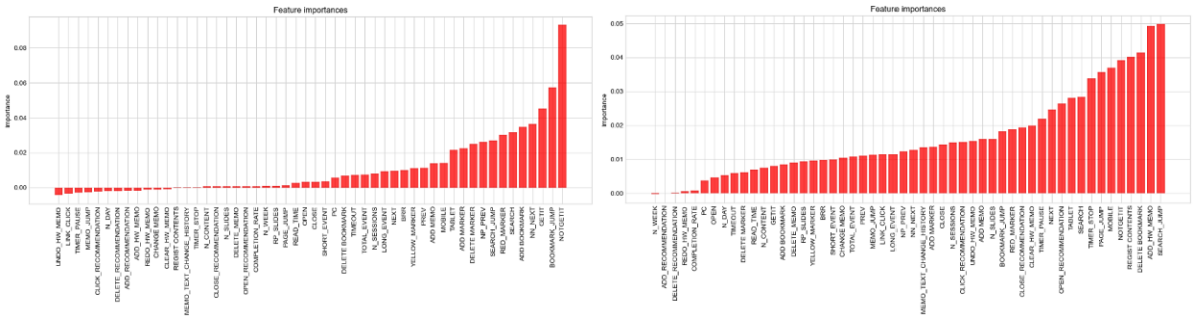

Figure 7. Left. VAE features importance. Right. Semi-VAE features importance.

We observe that the order of importance differs from the VAE to the Semi-VAE model. Since the two models are based on the same architecture, we may attribute these differences to the incorporation of prediction labels in the Semi-VAE model. To better understand these differences, we first should remember that the VAE model attempts to find general latent features that can reconstruct the data. This can be translated into two types of importance attributions. The first is based on the non-linear relationship that a feature has with other features, for example, the number of NEXT events and NN_NEXT (the number of NEXT events longer than 3 seconds) may be strongly related and the model should decide to attribute high importance to only one of these features. The second is based on the frequency of non-zero values of the feature across the students. For example, the feature ADD_HW_MEMO (adding a handwritten note) is not often used by the students and consequently, the VAE model may not attribute high importance to it. Accordingly, we may conclude, for example, that the NOT_GET_IT feature (number of times that a student pressed the button "did not understand this slide") is probably not very related to other features and relatively frequent (relatively to other independent features).

However, in the case of Semi-VAE, the model should also consider how important the feature is to make a prediction. This means that in our previous example the model may not consider that the

NOT_GET_IT feature is really important when estimating the probability of being an at-risk student. Moreover, the reader may note that the second most important feature of our Semi-VAE model is the previously mentioned ADD_HW_MEMO, which received low importance from the VAE model. This probably means that despite its infrequency, when a student uses this feature, our model can make a better prediction.

Besides the differences, we can find some similarities. One example is that redundant features such as N_WEEK and N_DAY have low importance in both models. The reason is probably that other reader's activities features are related to these features in all courses and the variational models do not need to preserve their information since it can be inferred (e.g., a student with a high number of notes, highlights, and search activities probably has a greater number of active days and weeks).

Finally, we would like to note that a Semi-VAE has more difficulty learning its mappings to the latent space since it has to deal with a trade-off between the importance of the features that have unique information and the importance of the ones that are better predictors. In this context, a features engineering process may be required to design better Semi-VAE latent representations, which in turn could lead to a statistically validated improvement of the VAE results. This idea is also supported by the results in Table 3, where the best results obtained by Semi-VAE representations take place when encoding a designed Feature Set (FS1 and FS2), in contrast to VAE, which works better encoding all possible features (ALL Feature Set).

6. Limitations

The experiment results reported in the previous section should be considered in light of some limitations. In this section, we will describe the two which we consider the most important.

The first is the optimality of the hyperparameters selected for our models. Even though we conducted several explorative trials to select the best hyperparameters, ideally, we should sample over different values for each model to avoid any bias from our selections. However, the computation time for our experimentation was more than 100 hours, which means that looping across hyperparameters would require months of computation.

The second is the collection time range of the DS1 unlabeled dataset used for training the variational models. As the reader may note, this dataset was collected in the same years of our datasets A-2019, A-2020, and B-2020. This means that our variational models were able to learn some singular characteristics present only in these years (e.g. The online activities due to the pandemic). While this ability is a good characteristic of the variational models, it also possibly introduces a bias that we cannot account for from our test and validation results. We should note that our first idea to avoid this possible bias was to use a dataset collected from 2018 and 2019. However, the datasets containing 2018 course data did not preserve their class schedule information.

7. Conclusions

From the inclusion of digital textbook reading systems in classrooms, several works have contributed to the idea of using the readers' data to identify at-risk students. The present work complements these contributions. Specifically, our results reinforce the idea that readers' interactions data preserve information related to their academic performance, but also evidence that these relationships are to a certain extent dependent on the course's characteristics. For this reason, it is difficult to use a model to make predictions across different courses. Nevertheless, we proved that the VAE latent representations and the Time-aware predictive models tend to manifest better generalizability characteristics, allowing an improvement of the predictive performance in a practical implementation.

Additionally, this paper introduces the Semi-supervised VAE to generate latent representations able to encode not only the feature relations present in different courses but also the course-specific features and their relationship with the academic performance. Since results do not provide evidence to prove or disprove that the Semi-VAE representations are better than the VAE ones, we conclude that further research is required to understand their differences. Considering that the Semi-VAE training is more complicated and easier to find a local minimum, a starting point can be to study techniques to improve their training. Some ways are to conduct a features engineering process or to design "warm-up" functions and apply them to certain loss components.

Acknowledgments

This work was supported by JST AIP Grant Number JPMJCR19U1, JSPS KAKENHI Grant Number JP18H04125, and JP22H00551, Japan.

References

Akçapinar, G., Hasnine, M. N., Majumdar, R., Flanagan, B., & Ogata, H. (2019). Developing an early-warning system for spotting at-risk students by using ebook interaction logs. *Smart Learning Environments*, 6(1):4.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019) Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25rd ACM SIGKDD*, doi:10.1145/3292500.3330701.

Baker, R. S. (2019) Challenges for the future of educational data mining: The baker learning analytics prizes. *Journal of Educational Data Mining*, 11(1):1–17.

Bosch, N. (2017) Unsupervised deep autoencoders for feature extraction with educational data. *Deep Learning with Educational Data Workshop at the 10th International Conference Educational Data Mining*.

Chen, C.H., Yang, S. J. H., Weng, J.X., Ogata, H., & Su, C.Y. (2021) Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers. *Australasian Journal of Educational Technology*, 37(4):130–144.

Ding, M., Yang, K., Yeung, D.Y., & Pong, T.C. (2019) Effective feature learning with unsupervised learning for improving the predictive models in massive open online courses. *Proceedings of the 9th International Conference on Learning Analytics Knowledge*.

Du, X., Yang, J., & Hung, J.L. (2020) An integrated framework based on latent variational autoencoder for providing early warning of at-risk students. *IEEE Access*, 8:10110–10122.

Fisher, A.J., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of machine learning research : JMLR, 20.

Flanagan, B., Chen, W., & Ogata, H. (2018) Joint activity on learner performance prediction using the bookroll dataset. *Workshop Proceedings 26th International Conference on Computers in Education*, 487–492.

Flanagan, B., & Ogata, H. (2017) Integration of learning analytics research and production systems while protecting privacy. *Workshop Proceedings 25th International Conference on Computers in Education*, 355–360.

Hasnine, M., Akçapınar, G., Flanagan, B., Majumdar, R., Mori, K., & Ogata, H. (2018) Towards final scores prediction over clickstream using machine learning methods. *Proceedings of the 26th International Conference on Computers in Education*, 11.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., … Lerchner, A. (2017) beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*.

Huang, A.Y.Q., Lu, O.H.T., & S. J. H. Yang. (2020) Evaluation of classification algorithms for predicting students' learning performance based on bookroll reading logs. *Cognitive Cities*, 262–272.

Kingma, D. P., & Welling, M. (2014) Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Kohavi, R., & Wolpert, D.H. (1996) Bias plus variance decomposition for zero-one loss functions. *ICML*.

Le, L., Patterson, A., & White, M. (2018) Supervised autoencoders: Improving generalization performance with unsupervised regularizers. Advances in Neural Information Processing Systems, volume 31.

Lopez, E. D., Minematsu, T., Taniguchi, Y., Okubo, F., & Shimada, A. (2022) Exploring the use of probabilistic latent representations to encode the students' reading characteristics. *CEUR Workshop Proceedings*, 3120.

Murata, R., Minematsu, T., & Shimada, A. (2021) Early detection of at-risk students based on knowledge distillation rnn models. *Proceedings of the 14th International Conference on Educational Data Mining*, 699–703.

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. *Smart Sensors at the IoT Frontier*, 327–350.

Okubo, F., Yamashita, T., Shimada, A., Taniguchi, Y., and Shin'ichi, K. (2018) On the prediction of students' quiz score by recurrent neural network. CEUR Workshop Proceedings, 2163.

Pelanek, R. (2020) Learning analytics challenges: trade-offs, methodology, scalability. Proceedings of the 10th International Conference on Learning Analytics & Knowledge.

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.