Accuracy of a Cross-Program Model for Dropout Prediction in Higher Education

Kerstin Wagner, Agathe Merceron, Petra Sauer
Beuth University of Applied Sciences Berlin
{kerstin.wagner, sauer, merceron}@beuth-hochschule.de

ABSTRACT: Reducing dropout rates in higher education would allow increasing the number of graduates. If one can predict early enough whether a student might drop out, targeted counseling could be put in place. This work replicates the approach of Berens et al. (2019) to predict whether students might dropout using academic performance data from their first semester. Further, the approach is extended by comparing the results of the cross-program model on specific programs of study with the results of the models trained for each specific program. The findings support the generalization of the approach of Berens et al. (2019) to the German context, which could serve to establish best practices for dropout prediction in higher education.

Keywords: dropout prediction, replication study, machine learning, models' comparison

1 INTRODUCTION AND RELATED WORK

A goal of the European Commission in 2010 was to increase the number of 30-34-year-olds with higher educational attainment from 31% to at least 40% in 2020 (European Commission, 2010). A way to achieve this is to reduce dropout rates. This requires early knowledge of students who run the risk of not completing their studies. The subsequent supervision of first-year students and the support provided by targeted study counseling must be supported by an effective dropout forecast and the analysis of possible causes (Dekker, Pechenizkiy, & Vleeshouwers, 2009).

Looking for a general approach based only on academic performance data, this work replicates to a large extent the work of Berens, Schneider, Görtz, Oster, and Burghoff (2019) – hereinafter also referred to as original study – regarding the chosen algorithms and academic performance features. The purpose is to use the data the university has on the academic performances of their current and former students to identify students at risk at the end of their first bachelor semester. As an extension of the original study, two additional modeling ways are compared: 1) a cross-program model built with the data of three bachelor programs together and evaluated separately on each specific program and 2) three models built for each specific degree program. The comparison of the results of the original study with all our results indicates that the approach of Berens et al. (2019) is generalizable to other higher education institutions in Germany.

Predicting dropout with machine learning algorithms in higher education institutions is an important task and has been investigated in many works (Ochoa & Merceron, 2018). Researchers use sociodemographic data, performance data or a mixture of both to solve this task. Sociodemographic data might include gender, ethnicity, income, date of birth. Performance data might include pre-university grade, major or degree program declared, enrollment in university courses, university grades. Dekker et al.

(2009), Aulck, Nambi, Velagapudi, Blumenstock, and West (2019) or Berens et al. (2019), as the original study, have obtained good prediction results with performance data only; adding sociodemographic data hardly improves the results. The work of Berens et al. (2019) is particularly relevant to us as it predicts students' dropout in a German context which is also the case of the present investigation. Building on these findings, this work uses only performance data. Different kinds of features can be extracted from performance data, in particular, global features and local features as introduced by Manrique, Nunes, Marino, Casanova, and Nurmikko-Fuller (2019). Local features are specific to a particular program of study like grades in the courses of this program. By contrast, global features can be extracted for any program of study like the number of passed exams, the average grade in passed exams and so on. Note that models built with machine learning algorithms that use local features can be trained only with the data of that particular program while models that use global features only, can be trained with data coming from all programs of an institution; we call these last models cross-program models. Dekker et al. (2009) have investigated one degree program only and use local features while Aulck et al. (2019) and Berens et al. (2019) have used global features and build one cross-program model. Manrique et al. (2019) have investigated two programs of study; they have built two models using local features and a cross-program model. Interestingly, the performance of the two models using local features tends to be better than the performance of the cross-program model. This finding leads us not only to replicate the work of Berens et al. (2019) but also to extend it by investigating whether individual models built separately for each degree program using global features give better results than the cross-program model. From a machine learning perspective, more training data is better. This would speak for a model integrating data from different programs of study. However, data from another program of study could also add noise.

No known algorithm works better in all contexts. Dekker et al. (2009) have obtained the best results with decision trees and Aulck et al. (2019) with logistic regression. Three algorithms – logistic regression, random forests, and neural networks – have given very similar results in the work of Berens et al. (2018); the addition of the ensemble method AdaBoost slightly improved the results.

Models are evaluated differently. Dekker et al. (2009), Aulck et al. (2019) and Manrique et al. (2019) have used k-fold cross-validation. Berens et al. (2019) have picked out a single cohort to evaluate their model. This work has used a time-aware validation in the spirit of Krauss, Merceron, and Arbanowski (2019) to reflect the intended use of the model: it is to build with data of passed students to predict whether new-comers might drop out. This approach is also used in Asif, Merceron, Ali, and Haider (2017) or Baneres, Rodriguez, and Serra (2019).

2 METHOD

This study uses data from six-semester bachelor's degree programs, which include 4,312 students from 2005 until summer 2019. The original study takes the data of two German universities: the entire bachelor courses of a state university (SU) with 14,496 records and a private university of applied science (PUAS) with 7,600 records while our work is based on three bachelor programs of a German state university of applied sciences. Our records include for each student the enrollment date in the degree, every single course they enrolled in, the respective enrollment semester and the grade earned, the graduation date and the result for students who completed the degree.

For the further processing of the data some preprocessing was necessary: to take account of changes to the curricula over the years, all data had to be converted to the present curricula and pseudonymization of the records were carried out by aggregation of grade from grades ({1.0, 1.3}, {1.7}, {2.0, 2.3, 2.7}, {3.0, 3.3, 3.7}, {4.0}) to (1.3, 1.7, 2.3, 3.3, 4.0) with 1.3 is the best and 4.0 the worst. Two other possible outcomes of an exam are "not participated" and "failed". To earn a degree, a student has to successfully pass every single course and has three attempts to do so. To attempt an exam, a student has to enroll in the corresponding course.

Our use-case is to predict students who drop out of the degree. A student switching from one degree to another degree within the same university or to another university is therefore considered as a dropout as well. To uncover students still enrolled in the university but not enrolled in any course of a specific program, we find dropouts from the data: a student that has not enrolled in any course of the degree during more than two consecutive semesters has dropped out of the degree. This threshold results from the longest interruption that we have identified in graduates.

A preliminary exploration of the data on a smaller dataset has shown that students dropping out and students completing the degree strongly differ in the courses of the first semester. The frequency of occurrence of "not participated" and "failed" is much higher for students who drop out than for students who complete their studies. This observation suggests that the use of appropriate algorithms on the data of the courses of the first semester can predict students' dropout with good results. That's why we focus on the first semester in this study.

The global features chosen for this study are given in Table 1 right. Our work differs from the original regarding the features in the following points: 1) we don't distinguish between important and other successfully completed exams because all courses of our first semester are mandatory and considered as core courses of the programs, 2) we don't distinguish between exams not participated in and no-show exams because this distinction does not exist in our data.

Table 1: Academic performance features – Comparison with Berens et al. (2019)

Berens et al.		**This work**	
Variable	**Values**	**Variable**	**Values**
No. of important successfully completed exams	1 to 9	No. of successfully completed exams	0 to max
No. of other successfully completed exams	0 to max		
Average grade per semester	1.00 to 4.00	Average grade per semester	1.3 to 4.0
No. of failed exams per semester	0 to max	No. of failed exams per semester	0 to max
No. of exams per semester not participated in	0 to max	No. of exams per semester not participated in	0 to max
No. of no-show exams per semester	0 to max		
Class label	1 = dropout 0 = graduate	Class label	1 = dropout 0 = graduate

In this study we have used the five different algorithms: decision tree, logistic regression, neural network, random forest, and AdaBoost; the implementation was done in the Python scikit-learn library.

Concerning the algorithms in the original study, the replication differs in three points: we have 1) added decision trees due to their good interpretability like logistic regression, 2) used random forest instead of bagged random forest, because the benefit of the bagged version was not clear to us, and 3) added the decision tree to AdaBoost due to good results as a single classifier. The metrics used to evaluate the models are precision, recall, accuracy and area under the ROC curve. The original study uses a classification threshold, which cannot be repeated in the present research because of the time-aware evaluation. Instead, this work has optimized the hyper-parameters of each model by 10-fold cross-validated grid search tuned for the recall metric.

Figure 1 shows the different training/test sets we have used and their numbers of records: [a] cross-program model: the dataset of the three degree programs is split into 80% training data corresponding to students with the oldest matriculation date and 20% test data (students with the newest matriculation date), [b] program-specific models: is similarly split, but for each degree program because we have trained program-specific models, and [c] cross-program model with program-specific test: the training set is the union of the training sets of [b] and the test sets are the same as in [b] (training set [a] is not necessarily disjunct from test set I, test set II and test set III). Variant [a] corresponds to the replication study while variants [b] and [c] are carried out for its extension.

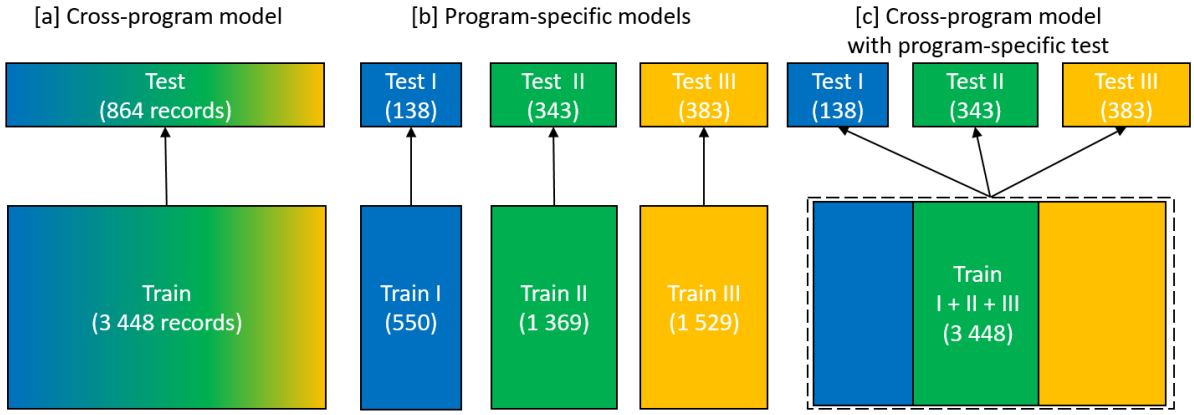

Figure 1: Schematic illustration of the training and test splits

3 RESULTS

Figure 2 presents the scores for each model and each variant [a/b/c]. First, we compare the results of the variant [a] to the results of the original study: recall 71.49% (SU) and 69.89 (PUAS); accuracy 76.60% (SU) and 83.64% (PUAS) – best results obtained with AdaBoost. Our AdaBoost achieves better results: recall 78.19% and accuracy 83.55%; only accuracy (PUAS) is marginally better. AdaBoost outperforms slightly the other models in the original study, which is not the case in our replication. Our other models for variant [a] achieve similar results: recall between 74.49% and 78.56% and accuracy between 81.46% and 83.89%. The best model for variant [a] in terms of recall is the decision tree (78.56%), closely followed by the neural network (78.37%) and in terms of accuracy the neural network (83.89%), closely followed by the decision tree (83.78%). The most important decision feature of the decision tree is the number of successfully completed exams and this is also confirmed by the result of the logistic regression coefficients. It stresses the observation that students who are not successful in their first semester tend to drop out faster.

The comparison of the cross-program models [a/c] with the program-specific models [b] as the extension of the replication study does not show a clear picture. In some cases, the cross-program model outperforms the specific models, for example, the decision tree and degree program III with a recall for [a] of 78.56%, for [b] of 70.64% and for [c] of 78.44%. In other cases, the specific model outperforms the cross-program models like for the AdaBoost and degree program II: recall [b] = 83.98% opposed to recall [a] = 78.19% and recall [c] = 79.84%.

The ROC AUC reaches the highest value for variant [a] with the logistic regression (92.21%). The best AUC score of 95.35% is achieved by the logistic regression in the specific model [b] for program II. AdaBoost performs worse for this metric. Similar trends can be observed for precision.

The differences between the results of the different models tend to be marginal, although the performance for program II tends to be better than for the other programs. Overall, the results show the appropriateness of a cross-program model and confirm the approach of Berens et al. (2019).

Models / Programs		Precision			Recall			Accuracy			ROC AUC		
		a)	b)	c)	a)	b)	c)	a)	b)	c)	a)	b)	c)
AdaBoost	I	94.63%	93.22%	**98.11%**	78.19%	**70.51%**	66.67%	83.55%	80.43%	80.43%	85.37%	81.92%	**82.50%**
	II		96.23%	**97.49%**		**83.95%**	79.84%		**86.30%**	84.26%		**87.98%**	87.42%
	III		**88.78%**	84.53%		**79.82%**	79.36%		82.77%	**83.81%**		83.24%	**84.53%**
Decision Tree	I	94.65%	92.73%	**98.15%**	**78.56%**	65.38%	**67.95%**	83.78%	77.54%	**81.16%**	90.93%	76.60%	**91.42%**
	II		**98.01%**	97.06%		81.07%	**81.48%**		**85.42%**	85.13%		**94.29%**	93.58%
	III		89.02%	**90.00%**		70.64%	**78.44%**		78.33%	**82.77%**		86.80%	**88.20%**
Logistic Regression	I	**95.80%**	**100.00%**	**100.00%**	75.97%	**67.95%**	64.10%	82.85%	**81.88%**	79.71%	**92.21%**	93.23%	**93.93%**
	II		**98.02%**	97.47%		**81.48%**	79.42%		**85.71%**	83.97%		**95.35%**	95.02%
	III		**95.18%**	91.76%		72.48%	**76.61%**		82.25%	**82.77%**		**89.56%**	88.92%
Neural Network	I	95.07%	98.21%	**100.00%**	78.37%	70.51%	**67.95%**	**83.89%**	82.61%	**81.88%**	85.77%	**84.42%**	83.97%
	II		**97.55%**	97.51%		**81.89%**	80.66%		**85.71%**	84.84%		**88.45%**	87.83%
	III		**92.47%**	91.49%		78.90%	**78.90%**		**84.33%**	83.81%		**85.21%**	84.60%
Random Forest	I	94.82%	**100.00%**	98.11%	74.49%	61.54%	**66.67%**	81.46%	78.26%	**80.43%**	90.95%	90.49%	**91.99%**
	II		96.53%	**97.47%**		**80.25%**	79.42%		**83.97%**	**83.97%**		91.98%	**93.35%**
	III		87.28%	**91.26%**		69.27%	**76.61%**		76.76%	**82.51%**		83.66%	**89.17%**

Figure 2: Heatmap of achieved metric scores – highest values are in bold

4 CONCLUSION AND FUTURE WORK

In this study, we have investigated five different algorithms using a global feature set to predict students' dropouts using only data from the first semester. Further, we have built cross-program models, models specific to each of the three programs and tested a cross-program model on each study program separately. Overall, the results show that a cross-program model as proposed in the original study is generalizable. Further research is needed to understand why prediction tends to work better for the study program II.

Despite the differences from the original study, our models get comparable results and even better for recall. AdaBoost works best in the original study, which is not the case here. Further investigation is needed to understand why. The obvious next step concerns the dataset: we have used only three degree programs with 4,312 records. So, subsequent activity is to consider more study programs up until a university-wide analysis and prediction system as in Berens et al. (2019). Especially interesting could also be the consideration of different online degree programs as well as the inclusion of master's degree programs.

Further work also consists of taking higher semesters into account. This data might reveal the self-regulating skills of the students better. A preliminary study has shown that about 1/3 of the students who drop out do so during or immediately after the first semester. This means 2/3 of the students drop out later. A follow-up is to predict dropouts related to the semester as in Berens et al. (2019). We will consider how adding more data from higher semesters will impact the performance of the classifiers.

ACKNOWLEDGEMENT

We thank Lennart Egbers and Stephan Wagner for the substantial data preprocessing and their preliminary study.

REFERENCES

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Aulck, L. S., Nambi, D., Velagapudi, N., Blumenstock, J. E., & West, J. (2019). Mining University Registrar Records to Predict First-Year Undergraduate Attrition. *EDM*.

Baneres, D., Rodriguez, M. E., & Serra, M. (2019). An Early Feedback Prediction System for Learners At-Risk Within a First-Year Higher Education Course. *IEEE Transactions on Learning Technologies*, *12*(2), 249–263. https://doi.org/10.1109/TLT.2019.2912167

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). *Early Detection of Students at Risk—Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods*. JEDM | Journal of Educational Data Mining, 11(3), 1–41. https://doi.org/10.5281/zenodo.3594771

Dekker, G., Pechenizkiy, M., & Vleeshouwers, J. (2009). Predicting Students Drop Out: A Case Study. *Proceeding of the 2nd International Conference On Educational Data Mining*, 41–50. http://www.educationaldatamining.org/EDM2009/uploads/proceedings/edm-proceedings-2009.pdf

European Commission. (2010, March 3). *Europe 2020—A European strategy for smart, sustainable and inclusive growth*. Retrieved from https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A52010DC2020

Krauss, C., Merceron, A., & Arbanowski, S. (2019). The Timeliness Deviation: A novel Approach to Evaluate Educational Recommender Systems for Closed-Courses. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, 195–204. https://doi.org/10.1145/3303772.3303774

Manrique, R., Nunes, B. P., Marino, O., Casanova, M. A., & Nurmikko-Fuller, T. (2019). An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, 401–410. https://doi.org/10.1145/3303772.3303800

Ochoa, X., & Merceron, A. (2018). Quantitative and Qualitative Analysis of the Learning Analytics and Knowledge Conference 2018. *Journal of Learning Analytics*, *5*(3), 154–166. https://doi.org/10.18608/jla.2018.53.10