ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study

Nicolae-Bogdan Şara[1], Rasmus Halland[2],
Christian Igel[1], and Stephen Alstrup[1]

1- Department of Computer Science, University of Copenhagen, Denmark

2- MaCom A/S, Denmark

Abstract. Pupils not finishing their secondary education are a big societal problem. Previous studies indicate that machine learning can be used to predict high-school dropout, which allows early interventions. To the best of our knowledge, this paper presents the first large-scale study of that kind. It considers pupils that were at least six months into their Danish high-school education, with the goal to predict dropout in the subsequent three months. We combined information from the MaCom Lectio study administration system, which is used by most Danish high schools, with data from public online sources (name database, travel planner, governmental statistics). In contrast to existing studies that were based on only a few hundred students, we considered a considerably larger sample of 36299 pupils for training and 36299 for testing. We evaluated different machine learning methods. A random forest classifier achieved an accuracy of 93.47 % and an area under the curve of 0.965. Given the large sample, we conclude that machine learning can be used to reliably detect high-school dropout given the information already available to many schools.

1 Introduction

School dropout is a problem for the individual and society. School education is correlated with a person's health and life expectancy, law-abidance, political interest, as well as happiness.[1] It can be argued that school dropouts impose a financial burden on the rest of society. In the USA, it has been estimated that compared to a high school graduate a dropout costs $292,000 on average, because of less tax income, incarceration costs, and other reasons [1]. Around 25 percent of public school students in the USA who entered high school in the fall of 2000 ended up leaving school and failing to earn a diploma within the subsequent four years [2].

In Denmark, about 14% of the pupils who start high school end up dropping out.[2] There are different secondary education programmes in Denmark. In particular, we distinguish between STX (*studentereksamen*) and HF (*højere forberedelseseksamen*). The company MaCom A/S provides online study administration tools to secondary education institutions through their system Lectio, which is used by the majority of Danish schools. Our goal is to use machine

[1] `http://www.oecdbetterlifeindex.org/topics/education`, retrieved November 2014
[2] `http://www.oecdbetterlifeindex.org/countries/denmark`, retrieved November 2014

learning to build a dropout predictor for Lectio, which can bring students at risk of dropping out in the near future to the teacher's attention. This allows the teacher to take countermeasures early.

Related work. The few existing studies on drop-out prediction using machine learning are difficult to compare. They consider different data sets, different levels of education, different prediction goals, different sources of information about the students, and different evaluation procedures. Most of them only build on small populations of some hundreds of students. According to the authors, [3] is probably the first application of machine learning to dropout prediction. The study considers 354 students participating in a distance learning computer science course in Greece. Several machine learning methods were compared, and a naïve Bayes classifier gave the best results. Prediction accuracies of 63 % and 83 % for the beginning of the academic period and for the remaining period, respectively, are reported. The naïve Bayes classifier also performed best in [4] for dropout prediction at a British university reaching an accuracy of 89.5 %.

A Dutch study considering 516 electrical engineering students also compared several algorithms [5]. The best results were obtained using classification and regression trees (CART, [6]) yielding 76 % accuracy, where cost-sensitive learning [7] was found to improve the accuracy. Cost-sensitive learning also increased the performance of the classifiers in a study looking at 670 Mexican middle-school students [8]. It was also applied in the Czech study [9], which considered 775 students and different classifiers and prediction tasks. Adding information from a social network analysis increased the classification performance up to 96.66 % using PART [10] and bagging.

2 Experimental Setup

In the following, we first describe the data and the extracted features, and subsequently discuss the machine learning methods employed.

Data. According to interviews with school inspectors and [11], the most relevant time horizon for predicting dropout is the near future. Therefore, our goal is to build a classifier that can predict whether a student will drop out in the subsequent three months.

We argue that different features describing the students should be used for dropout prediction at the beginning of the education than afterwards, and hence two different classifiers should be used for these two phases. In the present study, we focused only on the students that had already completed the first six months of high school. Thus, our classifier was able to include information about high-school performance during the previous semester.

In Lectio, teachers have the opportunity to specify the reason for the dropout of a student. Advised by school inspectors, we decided to focus only on the dropout reasons "Expelled from school", "Not passed", "The student couldn't be contacted", "The student does not thrive in school environment", "Regretted educational choices", "Not mature enough", "Leave", "Personal circumstances", "Academic level is too high", "Academic level is too low" and filtered the data accordingly (e.g., we excluded dropout due to sudden severe illness, because it cannot be predicted from the input data).

We queried the MaCom Lectio database for students enrolled after 2009 and extracted 72598 pupils, 55259 of which graduated and 17339 dropped out, giving a dropout rate of 23.8%, which is close to the Danish average. This ratio was maintained when randomly splitting the data equally into a training and test set with 36299 samples each.

We augmented the Lectio data with information retrieved from public online sources. After a literature study and interviews with school inspectors, we selected 17 features to describe each student:

- Gender
- Student has Danish name
(using information from `http://www.babyklar.dk`)
- Absences and missing assignments for first months of studies
- Education type (HF or STX)
- Travel time to school (based on querying `http://www.rejseplanen.dk`)
- Average income per postal code
(based on `http://www.statistikbanken.dk/INDKP1`)
- School and class size
- Teacher pupil ratio
- Most recent grade average variation between semesters
- Absences, grades and assignments for one month and one year sample period

All features were normalized to span $[0, 1]$ in the training set.

For every pupil, we picked one assessment date (when the features are computed and the prediction is made) and created a single data point. For a pupil that dropped out, the assessment date was set to three months before s/he left school. In the visualization of the data generating process Fig. 1, this three month period is indicated in red. For a pupil who graduated, the timepoint at which the features were calculated was chosen at random (excluding the first six months). Absences, grades and assignments were measured over two periods, one month and one year, prior to the assessment date (or since school start if the assessment date was in the first study year), indicated in blue and green in Fig. 1, respectively. If the grade variation between consecutive semesters could not be computed because a pupil only received grades once, zero imputation was used (this leaves room for improvement).

Methods. We compared different machine learning algorithms. We selected support vector machines (SVMs, [12]) with Gaussian kernels and random forests

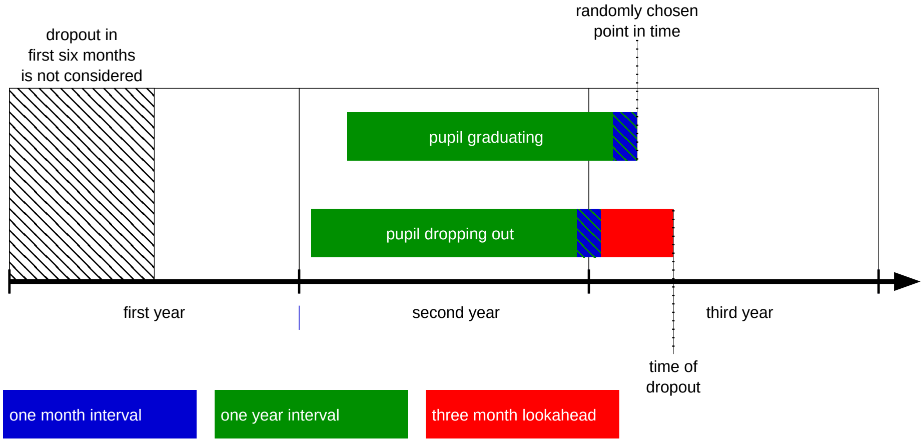

Fig. 1: Visualization of the data generation process.

(RFs, [13]) because of their good performance in general [14]. We added CART because of its interpretability and the good results in [5]. Furthermore, we considered a naïve Bayes classifier, which is easy to implement and worked best in the comparisons in [3, 4].

We used WEKA [15] for the naïve Bayes classifier and the open source machine learning library Shark [16] for all other methods. The naïve Bayes classifier and CART were used with their default parameters. For the SVM and RF we performed model selection. We used grid-search to optimize the 10-fold cross-validation error on the training set. For RF, we varied the number of trees and the number of features considered for choosing a split at each node on a 3×6 grid; 500 trees and 5 features gave the best results. For the SVM, we tuned the regularization parameter and the kernel bandwidth using a 10×11 grid, where the bandwidth was centered around an estimate produced by Jaakkola's heuristic [17].

3 Results

The accuracies of the different methods on the test set are given in Table 1. Figure 2 shows the receiver operating characteristic (ROC) curves visualizing the trade-off between the true positive rate and the false positive rate. The area under the ROC curve (AUC) for each classifier is given in Table 1.

The random forest performed best with an accuracy of 93.5 %, followed by SVM, CART, and finally the naïve Bayes classifier. The four features most frequently used by the RF for splitting were class size, school size, absences last month, and the average income per postal code.

	Random forest	CART	SVM	naïve Bayes
Accuracy (in %)	93.5	89.8	90.4	85.6
AUC (·100)	96.5	86.9	94.8	93.1

Table 1: Prediction accuracy and area under the curve (AUC) on the test data.

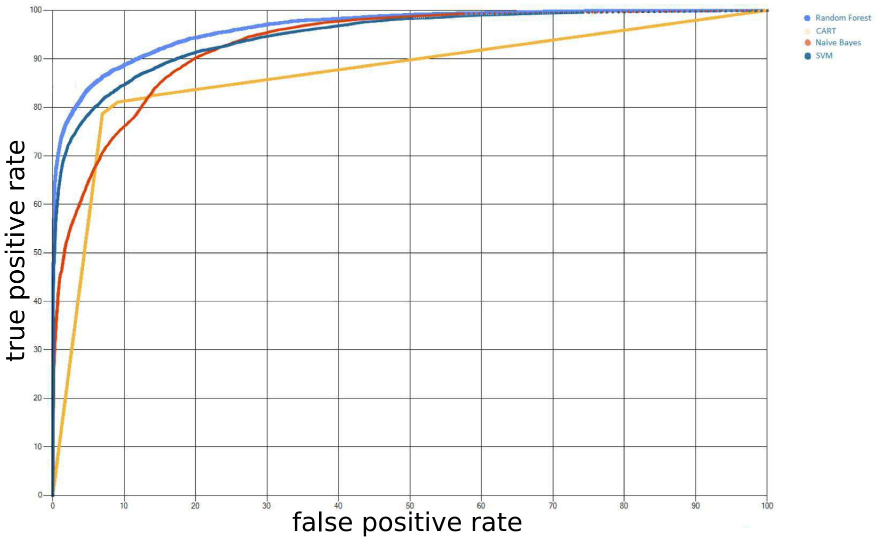

Fig. 2: ROC curves on test set, RF is depicted in light blue, CART in yellow, naïve Bayes in red, and SVM in dark blue.

4 Conclusions

Machine learning techniques can predict high-school dropout with a high accuracy. In our study considering 72598 pupils, a random forest achieved an accuracy of 93.5 % and an AUC of 0.965. Thus, the predictor is accurate enough to be used as a useful support tool for teachers allowing them to take early countermeasures preventing dropout. The ROC analysis showed that by varying the threshold the classifier can be tuned towards a desired false negative rate. Addressing the class imbalance in the training process (e.g., as in [5, 9, 8]) would lead to a different ROC curve, which may suggest an even more desirable trade-off.

In our preliminary investigation, we did not consider dropout in the first six months of high school. Future work will also address—using different input features—the important early dropout scenario. Adding information from social media, as done in [9], is likely to further increase the classification accuracy.

References

[1] A. Sum, I. Khatiwada, J. McLaughlin, and S. Palma. The consequences of dropping out of high school. *Center for Labor Market Studies Publications*, 2009.

[2] R. W. Rumberger and S. A. Lim. Why students drop out of school: A review of 25 years of research. Technical report, University of California, Santa Barbara, 2008.

[3] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas. Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems*, volume 2774 of *LNCS*, pages 267–274. Springer, 2003.

[4] Y. Zhang, S. Oussena, T. Clark, and K. Hyensook. Using data mining to improve student retention in higher education: a case study. In J. Filipe and J. Cordeiro, editors, *12th International Conerence on Enterprise Information Systems (ICEIS)*, pages 190–197. SciTePress, 2010.

[5] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers. Predicting students drop out: A case study. In T. Barnes, M. Desmarais, C. Romero, and S. Ventura, editors, *The 2nd International Conference on Educational Data Mining (EDM 2009)*, pages 41–50, 2009.

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[7] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 973–978. Morgan Kaufmann, 2001.

[8] C. Márquez-Vera, C. Romero, and S. Ventura. Predicting school failure using data mining. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, and J. Stamper, editors, *The 4th International Conference on Educational Data Mining (EDM 2011)*, pages 271–276, 2011.

[9] J. Bayer, H. Bydzovskä, J. Géryk, T. Obsivac, and L. Popelinsky. Predicting drop-out from social behaviour of students. In K. Yacef, O. Zaïane, H. Hershkovitz, M. Yudelson, and J. Stamper, editors, *The 5th International Conference on Educational Data Mining (EDM 2012)*, pages 103–109, 2012.

[10] E. Frank and I. H. Witten. Generating accurate rule sets without global optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, pages 144–151. Morgan Kaufmann, 1998.

[11] ATI Adaptive Technologies, Inc. Using predictive modeling to improve high school dropout prevention, 2008.

[12] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[13] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[14] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.

[15] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2nd edition, 2005.

[16] C. Igel, V. Heidrich-Meisner, and T. Glasmachers. Shark. *Journal of Machine Learning Research*, 9:993–996, 2008.

[17] T. Jaakkola, M. Diekhaus, and D. Haussler. Using the Fisher Kernel Method to Detect Remote Protein Homologies. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 149–158, 1999.