Predicting Students Drop Out: A Case Study

Gerben W. Dekker[1], Mykola Pechenizkiy[2] and Jan M. Vleeshouwers[1]
g.w.dekker@student.tue.nl, {m.pechenizkiy, j.m.vleeshouwers}@tue.nl
[1]Department of Electrical Engineering, Eindhoven University of Technology, the Netherlands
[2]Department of Computer Science, Eindhoven University of Technology, the Netherlands

Abstract. The monitoring and support of university freshmen is considered very important at many educational institutions. In this paper we describe the results of the educational data mining case study aimed at predicting the Electrical Engineering (EE) students drop out after the first semester of their studies or even before they enter the study program as well as identifying success-factors specific to the EE program. Our experimental results show that rather simple and intuitive classifiers (decision trees) give a useful result with accuracies between 75 and 80%. Besides, we demonstrate the usefulness of cost-sensitive learning and thorough analysis of misclassifications, and show a few ways of further prediction improvement without having to collect additional data about the students.

1 Introduction

The monitoring and support of the first year students is a topic that is considered very important at many educational institutions. At some of the faculties yearly student enrollment for a bachelor program can be lower than desired, and when coupled with a high drop out rate of freshmen the need in effective approaches for predicting student drop out as well as identifying the factors affecting it speaks for itself.

At the Electrical Engineering (EE) department of Eindhoven University of Technology (TU/e), the drop out rate of freshmen is about 40%. Apart from the department's aim to enforce an upper bound to the drop-out rate, there are other reasons to want to identify successful and unsuccessful students in an early stage. In the Netherlands, there is the legal obligation that universities have to provide students with the necessary support to evaluate their study choice. In general, students who choose to pursue their study career at another institution, should do this at an early stage. For EE students there is a very concrete reason to evaluate before the end of the first semester: the EE program of the nearby Fontys University of Applied Science accepts TU/e drop outs in their curriculum until the beginning of January, without any time losses involved. Besides, there is always a subset of students which the department considers a "risk group", i.e. students who may be successful but who need extra attention or specific individual care in order to succeed. Detecting this risk group in an early stage is essential for keeping these students from dropping out. It enables the department to direct its resources to the students who need it most.

Current approach at EE department. To support students in making this decision, every enrolled student receives a study advice in December. This advice tells the student whether or not he or she is encouraged to proceed his study career at the faculty. It is based upon the grades and other results of the student so far and upon information obtained from 1st-semester-teachers and student-mentors, examined and interpreted by

the department's student counselor. The final semester examinations are not taken into account, because they are in January; postponing the advice until after the results are known would preclude students from switching to Fontys. The advices seem to be quite accurate in practice: students who are assessed as potentially successful are in general the same students that are successful after a year. Moreover, the students who are not encouraged to proceed their current study program, generally do not continue into the second year.

The objectives. Despite the success, the assessment remains unsatisfactory because of its rather subjective character. Therefore, a more robust and objective founding of the process may lead to advices which are more consistently followed up by students. Besides, a closer analysis is likely to lead to an improved selection process.

First of all, the department is interested in which of the currently available student data are the strongest predictors of success, and in the performance of this predictor. Obviously, the lower the predictor's quality, the more the department is curious to know what information makes the current assessment work. If the predictor quality is high, the department's interests are directed towards: (1) using the predictor as a back-up of the current assessment process; (2) identifying success-factors specific to the EE program; (3) identifying what data might result in a further increase of the predictor quality, and as a consequence, collect these data; (4) considering a more differentiated view on the risk group; (5) modifying the assessment process time-line, resulting in an earlier prediction, ideally even before entering the study. Furthermore, if strong predictors for academic success can be found, these will also be used to gain understanding of success and risk factors regarding the curriculum. Awareness of these factors by teachers, education personnel and management will help to select appropriate measures to support the risk group, eventually resulting in a decrease of the drop-out rate.

In this paper we present the results of the educational data mining case study aimed to address these identified issues. First, we discuss related work on addressing the problem of student dropout (Section 2). Then, we consider the settings of our EDM case study and present the analysis of classification results (Section 3). In Section 4 we present the further evaluation of one of the models. We conclude this paper with a summary of the results and discussions of further work in Section 5.

2 Background and Related Work

The topic of explanation and prediction of academic performance is widely researched. In the earlier studies, the model of Tinto [12] was the predominant theoretical framework for considering factors in academic success. Tinto considers the process of student attrition as a socio-psychological interplay between the characteristics of the student entering university and the experience at the institute. This interaction between the student's past and the academic environment leads to a degree of integration of the student into this new environment. According to this model, a higher degree of integration is directly related to a higher commitment to the educational institute and to the goal of study completion. Later studies tried to operationalize this model identifying the factors like peer group interactions, interactions with faculty, faculty concern for

student development and teaching, academic and intellectual development, and institutional and goal commitments that affect the student's integration [10]. These factors proved to have a predictive capacity across different institutions, and showed therefore to be a potential tool in identifying students who might drop out. Other studies tried to identify the significant factors in a more detailed way. Many studies included a wide range of potential predictors, including personality factors, intelligence and aptitude tests, academic achievement, previous college achievements, and demographic data and some of these factors seemed to be stronger than others, however there is no consistent agreement among different studies ([1], [3], [5], [13]). One of the recent European studies [3] has confirmed that sex (only in technical schools), age at enrollment, score on pre-university examination, type of pre-university education, type of financial support, father's level of education and whether or not living at the university town may all have an impact on the drop out. All studies show that academic success is dependent on many factors, where grades and achievements, personality and expectations, as well as sociological background all play a role.

The use of data-mining techniques in this field, known as educational data mining (EDM), is relatively new. The methodology is not yet transparent and it is not clear which data mining algorithms are preferable in this context. Clustering as means of data exploration and classification for building predictors have been tried in [4]. Association analysis has become also a popular approach in EDM [7], while one of the recent EDM case studies indicates that it is easy to underestimate the required efforts and overestimate the usefulness of this technology for small datasets [6]. The results of the case study presented in [2] indicate that Bayesian networks and neural networks are consistently outperformed by decision tree algorithms on relatively small educational datasets. However, the related work is still too scarce and in general it is hard to conclude from the recent studies (e.g. [2], [4], [8], [11]) which approach should be favored or even to measure whether learnt models outperform more traditional ways of predicting academic success.

3 Prediction of student drop out

In this case study we consider data collected over the period 2000 – 2009 that contains information about all the students being involved in the EE program. We selected a target dataset of 648 students who were in their first year phase at the department and came either from VWO (which is pre-university secondary education) or from polytechnical education (finishing at least a year of education at a polytechnical school grants access to university too). The latter group is a minority of about 10% of the considered students in the dataset.[1]

In order to get labels for the supervised learning of predicting models the students are classified in the following way: if a student was able to get his *propedeuse* (in the

[1] The further discussion of background knowledge and different issues related to the data preprocessing, data cleaning and transformation processes goes beyond the scope of this paper. An interested reader can find this information in the online technical report at http://www.win.tue.nl/~mpechen/research/edu.html.

Netherlands, a diploma which a student acquires after having successfully completed the first year at a university) in three years, he is classified as successful, and otherwise as unsuccessful.

We considered three datasets: a dataset with pre-university data only containing 495 instances (242 instances classified as unsuccessful, 253 instances classified as successful), each described with 13 attributes (Appendix A), a dataset with university grades only containing 516 instances (253 instances classified as unsuccessful, 263 instances classified as successful), each described with 74 attributes (for each of the 37 available courses we have two attributes saying how many attempts were taken, and what the highest grade was), and dataset with both sets of attributes containing also 516 students (missing values for pre-university data were replaced with zeros).

In our experimental study we used several popular Weka [14] classifiers (with their default settings unless specified otherwise). We compared the two decision tree algorithms CART (*SimpleCart*) and C4.5 (*J48*), a Bayesian classifier (*BayesNet*), a logistic model (*SimpleLogistic*), a rule-based learner (*JRip*) and the Random Forest (*RandomForest*). We also considered the *OneR* classifier as a baseline and as an indicator of the predictive power of particular attributes.

These classifiers are run on the dataset containing the pre-university data. We used 10-fold cross validation for estimating generalization performance. The statistical significance of differences in performance of *OneR* and other learners is tested with the two-sided paired t-tester in Weka's Experimenter, using a significance level of 5%.

3.1 Classification with pre-university or university data only

The classification accuracies for the dataset containing only the pre-university related data are shown in Table 1. The OneRule classifier reached the accuracy of 68% taking the *VWO Science mean* as a predictor. None of the other classification algorithms was able to learn a model which would outperform it (statistically) significantly.

Attribute ranking (with respect to the class attribute) according to the information gain criterion showed that the *VWO Science mean*, *VWO main* and *VWO Math mean* were by far the best attributes in information gain (information gains 0.16, 0.13, 0.12 respectively), with the next "closest" attribute *VWO Year* lagging behind (0.05). Furthermore, these three attributes are highly correlated and therefore it is logical to expect it would be hard to learn a more complex and yet generalizable classifier with a relatively small dataset. Learning a classifier with feature selection also does not improve the results a lot. Learning a *J48* tree using only the three mentioned attributes gives an average accuracy of 71%.

Table 1. Classification accuracy on pre-university dataset

Classifiers	*OneR*	*CART*	*J48 -M 2*	*J48 -M 10*	*BayesNet*	*Logit*	*JRip*	*RF*
Accuracy	0.68	0.68	0.70	0.69	0.71	0.69	0.70	0.65

The same classification techniques were applied to the dataset with the university grades (Table 2). The OneRule algorithm results in the classifier which checks the grade for Linear Algebra (*LinAlgAB*), and decides positive if this grade is bigger than 5.5 (that is exactly the minimum for passing a course). Again we can see that more sophisticated classification techniques do not improve accuracy very much. However, it is worth noticing that the CART classifier is statistically significantly better than the base line with a classification accuracy that is 4.8% higher on average.

Table 2. Classification accuracy on university grades dataset

Classifiers	*OneR*	*CART*	*J48 -M 2*	*J48 -M 10*	*BayesNet*	*Logit*	*JRip*	*RF*
Accuracy	0.76	0.81 o	0.79	0.79	0.75	0.79	0.78	0.80

o – statistically significant improvement

The CART classifier learnt a compact tree with five leaves and uses *LinAlgAB* as root of the tree, and *CalcA*, *Calc1* and *Project nAttempts* as further discriminators. It is worth noticing that the grades of the Networks course are not used at all, while some of its attributes have higher information gains. Correlation analysis however does show that correlation between Linear Algebra and Networks attributes is rather strong, but weak between Linear Algebra and Calculus attributes.

3.2 Classification with complete data

Classification accuracies for the dataset containing both pre-university and university related data are shown in Table 3 (column indexes correspond to those in Tables 1 and 2).

Table 3. Accuracy and rates of total dataset

Classifiers	*OneR*	*CART*	*J48 -M 2*	*J48 -M 10*	*BayesNet*	*Logit*	*JRip*	*RF*
Accuracy	0.75	0.79	0.80 ○	0.80	0.75	0.79	0.77	0.79
True positives	0.64	0.79 ○	0.80 ○	0.75 ○	0.72 ○	0.79 ○	0.73 ○	0.82 ○
False negatives	0.36	0.21 ○	0.20 ○	0.25 ○	0.28 ○	0.21 ○	0.27 ○	0.18 ○
True negatives	0.86	0.80 ●	0.80 ●	0.84	0.79	0.80 ●	0.82	0.77 ●
False positives	0.14	0.20 ●	0.20 ●	0.16	0.21 ●	0.20 ●	0.18	0.23 ●

○, ● – statistically significant improvement or degradation

It can be seen that these accuracies are comparable with those achieved on the dataset with university related data only. Apparently, the pre-university data does not add much independent information that can improve classification accuracy. However, we can see that the trees learnt with *J48* are now statistically significantly better than the base line model. The other tree-based classifiers also achieve reasonable accuracy, while the *Bayes Net* and *JRip* algorithms slightly fall behind.

To get a better insight on the performance of classifiers, the scoring of the algorithms is shown in more detail now. A remarkable fact is that the base line model has a higher false negative rate than all other models. This is an interesting finding, because according to the student counselor it is better to give an erroneous positive advice to a student who should actually be classified as negative, than to give a erroneous negative advice to a

student who should be classified as positive. Cost-sensitive learning can be used to balance classification accuracies or boost the accuracy for a particular type of prediction.

3.3 Boosting accuracy with cost-sensitive learning

In order to "advice" a classification algorithm to prefer one type of misclassification to another a cost matrix (that has a direct mapping to the confusion matrix) is commonly used as an input to a meta classifier:

	classified as negative	classified as positive
actual negative	$C(-,-)$	$C(-,+)$
actual positive	$C(+,-)$	$C(+,+)$

By choosing the weights $C(i, j)$ in a certain way we can achieve a more balanced classification in case of severe class imbalances (using the diagonal entries), or a more cost-effective classification (using the off-diagonal entries).

Since cost matrices are equivalent under scaling, and we only want to increase the cost of false negatives over false positives, it suffices to build a matrix with only one free coefficient and structure [[0 1] [C 0]], with C > 1.

Since our experiments favored tree-based learners we used *J48*, *J48graft* and *CART* as base classifiers in Weka's *CostSensitiveClassifier*. To prevent the tree from growing too big, we used the *CfsSubsetEval* feature subset selection algorithm that tries to select the most predictive attributes with low intercorrelation. The *J48* and *J48graft* classifiers were forced to have at least 10 instances for each node in order to prevent overfitting and unnecessarily complex models. Combining these *CART*, *J48* and *J48graft* with the two ways of using the cost matrix in cost-sensitive approach (*data weighing* and *model cost*), six experiments were conducted using *F* measure for defining the precision-recall tradeoff (we used $\beta = 1.5$). For each combination, the settings giving the highest *F* measure is presented in Table 4. The tree learnt with the "plain" *J48* is presented in the first data column.

The results indicate that it is necessary to sacrifice some of the achieved accuracy to be able to shape the misclassification. Only model 5 achieves a high accuracy and a high *F* measure, all other models lose in accuracy if *F* is increased. During the experiment, it became clear that there is not much room for enhancement: if recall increased to values higher than 85%, the overall accuracy results were unacceptable. The only exception is model 7 (notice the size of this tree being much larger comparing to other models and also seem to be too detailed to be meaningful for decision making).

In some cases, small trade-offs could be made changing C. Compare for instance model 5 with model 6: a three percent point drop in accuracy gives a three percent rise in recall.

The created decision trees are remarkably similar: in every tree the *LinAlgAB* attribute is dominant, with *CalcA* as first node in most of the cases. When *NetwB* is chosen as the first node, the recall is lower, although the difference is too small to draw decisive conclusions.

Table 4. Accuracy results with cost-sensitive learning

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Type	J48	J48	J48	CART	CART	CART	J48graft	J48graft
Learner option	-	Data weighting	Model cost	Data weighting	Model cost	Model cost	Data weighting	Model cost
$C(+,-)$	-	2	3	2	3	4	4	3.2
Confusion matrix	212 41 651 98	175 78 49 214	206 47 62 201	169 84 50 213	201 52 57 206	181 72 51 212	160 93 31 232	161 92 56 207
Accuracy	0.79	0.75	0.79	0.74	0.79	0.76	0.76	0.71
Precision	0.83	0.73	0.81	0.72	0.80	0.75	0.71	0.69
Recall	0.75	0.81	0.76	0.81	0.78	0.81	0.88	0.79
F_β	0.77	0.79	0.78	0.78	0.79	0.79	0.82	0.76
nLeaves	5	11	5	10	7	7	21	8
TreeDepth	3	6	3	5	4	4	8	5
Root node	LinAlgAB <= 5	LinAlgAB <= 5	LinAlgAB <= 5	LinAlgAB < 5.5	LinAlgAB <= 5.5	LinAlgAB <= 5.5	LinAlgAB <= 5	LinAlgAB <= 5
First node	NetwB <= 5.7	CalcA <= 5	NetwB <= 5.7	VWO-Science-mean	CalcA < 5.15	CalcA < 5.15	CalcA <= 5	CalcA <= 5
Second node	CompB-nAttempts	CompB-nAttempts	CompB-nAttempts	LinAlgA, CalcA	VWO-Science-mean	VWO-Science-mean	VWO-Science-mean	LinAlgB, NetwA2

4 Further evaluation of the obtained results

As the final step, we examined one of the models (model 7 from Table 4) in more detail to see if we can gain better understanding of the classifier errors. The student counselor compared all the wrongly classified instances of model 7 with his own given advices to check for interesting patterns. One of the first assessed things was the question whether the learned model is incorrect or the classification criterion is chosen incorrect. To examine this, two methods were used. Firstly, the false negative and false positive sets have been checked manually by the student counselor. His conclusions were that about 25% of the false negatives should be true negatives instead. This finding might indicate a wrong classification measure. Concerning the false positive set a conclusion is less obvious: about 45% of this set was classified as positive by the student counselor as well as by the tree, but did not meet the classification criterion. A substantial subset of these students have chosen not to continue their bachelor program in Electrical Engineering although all indications for a successful continuation were present. Qualifying these students as *false* positive does not seem to be appropriate. So from this evaluation based on domain expertise we can conclude that some of the mistakes might be due to the classification measure, and some of them raise suspicion on behalf of the learned model.

The second way to check the viability of the model is to compare the results obtained with this classifier with respect to the three class classification problem, i.e. identifying first manually the third so-called *risk* group and then checking whether wrongly classified students will be in the risk class (that would indicate that the learned model is actually more accurate and also that it has difficulties in predicting the students who are difficult to classify into success or failure categories per se). However, we observe that only 25% of the misclassified instances are in this category. It should be noted that this is still twice as much as the risk students ratio in the total dataset. Therefore, this also indicates that the learned model should be improved. Furthermore, 25% of the instances in the false

positive class would be classified as *good* using the three-class classification thus indicating a real difference between two classifiers. So from this test we can also conclude that the model as well as the classification criterion should be revised.

After the analysis of errors, the misclassified sets are looked up in the database to search for meaningful patterns manually. A very clear pattern popped up immediately: almost all misclassified students did not have a database entry concerning *LinAlgAB* (and therefore were mapped to zero). Checking out different students showed that there are many possible reasons now to have a zero value in the *LinAlgAB* record: a) a student might be of a cohort in which the *LinAlgAB* exam was in January or later; b) a student might have not shown up during the exam; and c) a student might have taken another way to get its *LinAlgAB* grade: in some years it was possible to bypass the regular exam by doing the subexams *LinAlg1*, *LinAlg2*, *LinAlg3*, *LinAlg4* and *LinAlg5*. A student succeeding in taking this path can well be an excellent student, but gets a zero mark for the *LinAlgAB* attribute. Due to this effect, 216 of the 516 students do have a zero entry in their *LinAlgAB* record (of which 155 instances were classified as unsuccessful and 61 instances as successful). Moreover, the same effect will play a role for the other courses too. Given the dominant position of the *LinAlgAB* attribute in the decision trees generated in section 3.3, attempts in completing the data-set should be considered worthwhile.

5 Conclusions and Future work

Student drop out prediction is an important and challenging task. In this paper we presented a data mining case study demonstrating the effectiveness of several classification techniques and the cost-sensitive learning approach on the dataset from the Electrical Engineering department of Eindhoven University of Technology.

Our experimental results show that rather simple classifiers give a useful result with accuracies between 75 and 80% that is hard to beat with other more sophisticated models. We demonstrated that cost-sensitive learning does help to bias classification errors towards preferring false positives to false negatives.

Surprisingly (according to the student counselor) the strongest predictor of success is the grade for the Linear Algebra course, which has in general not been seen as the decisive course. Other strong predictors are grades for Calculus, Networks and the mean grade for VWO Science courses. The most relevant information is collected at the university itself: the pre-university data can be summarized into a few attributes.

The in depth model evaluation pointed to three major improvements that can be assessed. Firstly, a key improvement in this dataset would be to find a solution for the changing course organization over the set. Aggregating the available information about student performance for a course in a way that can be used for all students in the dataset might prevent the type of misclassifications that is now strongly prevalent. A second, related improvement would be a better way to encode grades in general. Mapping all unknown or not available information to zero showed to be not effective. Specifically, Linear Algebra

grades should be available. A more advanced solution dealing with missing values also can be considered in this respect.

The quality of the classification criterion is the third improvement that might be considered. The simple binary classification as used in this study has some disadvantages: a negative classification can only be given after three years, and there is no guarantee that a student who does not get his propedeuse after three years will be not successful in the long run. Also, students who do not receive a propedeutical diploma, should not necessarily be "disqualified": they may have had different motives to discontinue their studies. This touches on a more fundamental topic: it is not easy to find an objective way of classifying students. In this paper we experimented with the so-called 0/1 loss and cost-sensitive classification. AUC optimization is also one of the directions of further work.

As a final remark we would like to point out that this study shows that learning a model on less rich datasets (i.e. having only pre-university and/or first-semester data) can be also useful, provided the data preparatory steps are carried out carefully.

6 References

[1] Herzog, S. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. In *Proc. of 44th Annual Forum of the Association for Institutional Research* (AIR), 2004.

[2] Herzog, S. Estimating student retention and degree-completion time: Decision trees and neural networks vis-a-vis regression, *New Directions for Institutional Research*, p. 17-33, 2006.

[3] Lassibille, G., Gomez, L. N. Why do higher education students drop out? Evidence from Spain, *Education Economics* 16(1), p. 89-105, 2007.

[4] Luan, J. Data mining and its applications in higher education. *New Directions For Institutional Research*, p. 17-36, Spring 2002.

[5] Parmentier, P. La reussite des etudes universitaires: facteurs structurels et processuels de la performance academique en premiere annee en medecine. PhD thesis, Catholic University of Louvain, 1994.

[6] Pechenizkiy, M., Calders, T., Vasilyeva, E., De Bra, P. Mining the student assessment data: Lessons drawn from a small scale case study. In Proc. of the 1st Int. Conf. on Educational Data Mining (EDM'08), p. 187-191, 2008.

[7] Romero, C., Ventura, S. Educational data mining: a survey from 1995 to 2005, *Expert Systems with Applications* 33, p. 135-146, 2007.

[8] Romero, C., Ventura, S., Espejo, P. G., Hervas, C. Data mining algorithms to classify students. In *Proc. of the 1st Int. Conf. on Educational Data Mining* (EDM'08), p. 187-191, 2008.

[9] Superby, J., Vandamme, J.-P., Meskens, N. Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Proc. of the Workshop on Educational Data Mining* at ITS'06, p. 37-44, 2006.

[10] Terenzini, P. T., Lorang, W. G., Pascarella, E. T. Predicting freshman persistence and voluntary dropout decisions: a replication, *Research in Higher Education* 15(2), p. 109-127, 1981.

[11] Thai Nge, N., Janecek, P., Haddawy, P. A comparative analysis of techniques for predicting academic performance. In *Proc. of 37th Conf. on ASEE/IEEE Frontiers in Education*, 2007.

[12] Tinto, V. Limits of theory and practice in student attrition, *Journal of Higher Education* 53, p. 687-700, 1982.

[13] Touron, J. The determination of factors related to academic achievement in the university: implications for the selection and counseling of students, *Higher Education* 12, p. 399-410, 1983.

[14] Witten, I. H., Frank, E. *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2 ed., 2005.

Appendix A. Attributes in the pre-university dataset.

Attributes	Type	Remarks
IDNR	numerical	Used only to check data sanity
VWO Year	nominal	Major changes in Dutch education system, {1..4, 'n/a'}
VWO Profile	nominal	The pre-university education curriculum, {1..5, 'n/a'}
VWO nCourses	numerical	The number of courses taken.
VWO mean	nominal	{ n/a, poor, average, above average, good, excellent }
VWO Science nCourses	nominal	{ n/a, < 3, 3, >3 }
VWO Science mean	nominal	As VWO mean
VWO Math nCourses	nominal	{n/a, 0,1,2}
VWO Math mean	nominal	As VWO mean
HO Education	nominal	{n/a, electrical, technical, other}
HO Year	nominal	Same categories as VWO Year
HO Grade	nominal	As VWO mean
GapYear	nominal	{n/a, < -1, -1, 0, 1, >1 }
Classification	nominal	{-1, 1}