MODELING STUDENT DROPOUT USING STATISTICAL AND DATA MINING METHODS

PETR BERKA [a,b,*], LUBOŠ MAREK[c], MICHAL VRABEC[c]

berka@vse.cz, marek@vse.cz, vrabec@vse.cz

[a] University of Economics in Prague, Faculty of Informatics and Statistics, Department of Information and Knowledge Engineering, W. Churchill Sq. 4, Prague, Czech Republic

[b] University of Finance and Administration, Department of Computer Science and Mathematics, Estonská 500, Prague 10, Czech Republic

[c] University of Economics in Prague, Faculty of Informatics and Statistics, Department of Statistics and Probability, W. Churchill Sq. 4, Prague, Czech Republic

Abstract

Not completing the study by a large portion of students is a serious problem at universities worldwide. Regardless of the country, numbers are very similar: about one-half of students who enrolled for the bachelor study leave university before obtaining the degree. To deal with this problem, we create models to distinguish between students who successfully completed their study and students who dropped out of university. Models created using traditional statistical analysis techniques (logistic regression) are compared with models created using data mining methods (decision trees, association rules). We use data about students who enrolled for their bachelor study at the University of Economics in Prague in the academic year 2013/2014 in our analysis.

Key words

student dropout, logistic regression, decision trees, association rules

JEL classification

C38, D83

1. Introduction

Not completing the study by a large portion of students is a serious problem at the universities worldwide. Regardless on the countries, the numbers are very similar: about one-half of students who enrolled for the bachelor study leave the university before obtaining the degree. In-depth studies based on detailed questionnaires report various factors that influence the student's dropout (Sagenmuller, 2018; Norton and Cherastidtham, 2018; Vossensteyn et al., 2015, Fischer et al., 2016). We are approaching this problem from the data analysis point-of-view. Using the data collected about the students in the university information system we try to identify factors that influence the student dropout. We created several classification and descriptive models that relate together the information available about the students when entering the university with the information on how the students finished their studies. Our work thus fits into the area of educational data mining, in particular into dropout and retention analysis (Dekker et al., 2009; Kotsiantis et. al., 2003; Lykourentzou et. al., 2009).

The rest of the paper is organized as follows: section 2 describes the used data, section 3 presents the tasks and used methods, section 4 shows the results of the analysis and section 5 concludes the paper.

2. Student dropout data

We used data about all bachelor students who enrolled for their study at the University of Economics, Prague (VSE) in the academic year 2013/2014. The available dataset contains 3465 students. We removed data students who did not study in Czech study programs and/or who are still in the bachelor program (e.g. due to interruption of study). The dataset contains data about 3339 students after this filtering; the number of female students (1681) was slightly higher than the number of male students (1658), about one half of the students come from Prague and Central Bohemian region, followed by students from South Bohemian region, most of the students study in the full-time form of study (3196 out of 3339).

Two types of variables can be found in the data: socio-demographic characteristics (e.g. age, sex, secondary education, citizenship) and study progress at the university (e.g. credits spent and lost, study results, date of defense, date of dismission). As we focus on models based on the information about the students that is available when starting their study we used only variables of the first type (Table 1 lists the used variables of the first type that are originally recorded about each student).

Table 1: List of used original variables

Code	Explanation	Type	Values or range
BirthDate	Birthdate	date	date (e.g. 25.6.1991)
Sex	Sex	nominal	"male", "female"
Citizenship	Citizenship	nominal	code of country (e.g. CZ)
AddCountry	Country of residence	nominal	code of country (e.g. CZ)
SSCountry	Country of the secondary school	nominal	code of country (e.g. CZ)
YearSS	Year ending secondary school	numeric	year (e.g. 2010)
SSProgram	Type of secondary education	nominal	code of education (e.g. 7941K41)
AdmissForm	Form of admission	nominal	"no-exam", "exam", "repeal"
AdmissDate	Date of admission	date	date (e.g. 11.7.2013)
StudyForm	Form of study (full-time, part-time)	nominal	"pres", "comb"
RecCred	Number of recognized credits	numeric	0 – 26
DismissDate	Dismission date	date	date (e.g. 29.9.2015)
Defense	Date of defense	date	date (e.g. 15.6.2016)

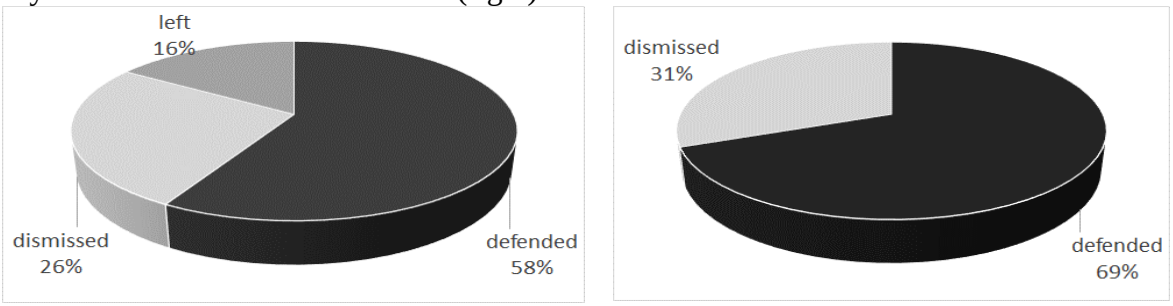
Source: the authors.

We extended the data by adding following new (input) variables:

- YearsSSVSE - Number of years between leaving the secondary school and entering the university, derived from YearSS and AdmissDate (integer number, e.g. 0),
- AdmissAge - Age when entering the university, derived from Birthdate and AdmissDate (decimal number, e.g. 19.83),
- SSType3 - Type of the secondary school, derived from SSProgram; we distinguish between gymnasium (Gym), lyceum (Lyc) and other secondary school (SS).

Based on the stored data we also created a new variable that reflects the way how a student left the university. The considered possibilities are "defended" (if a date of defense was found in the data base), "dismissed" (if a date of dismission letter was found in the database) and "left" by their own (if no previous condition applies). This variable (denoted *Target*) will be the target for the classification tasks (Figure 1 shows relative frequencies of the value of *Target*).

Figure 1: Relative frequencies for *Target* considering all three values (left) and considering only values defended and dismissed (right).



Source: the authors.

3. Data analysis tasks and methods

3.1. Task description

We defined two types of tasks of our analysis: binary classification and dependency analysis. When doing classification we tried to create models that will predict the value of *Target* based on the variables known early after entering the University. We considered only two values of *Target* for binary classification: defended and dismissed. This decision resulted in further reduction of the dataset; we filtered out students who left the University on their own. So we finally worked with the data about 2808 students. We tried to found interesting differences in the groups defined by the values “defended”, “dismissed” of *Target* during the dependency analysis. We performed our analysis on the university level not considering faculties or study programs.

3.2. On used methods

We used logistic regression and decision trees for the classification tasks and association rules mining for the dependency analysis. Logistic regression is a standard statistical technique used to create classification or prediction models, decision trees are used both in statistical data analysis and data mining (machine learning) for classification and prediction tasks as well as for concept description, and association rules mining is a typical data mining technique used to find interesting relationships between conjunctions of categories (values of categorical variables).

Logistic regression is a statistical technique that uses a logistic function to model a binary dependent variable having e.g. values 0 and 1 (Hilbe, 2009). Suppose $\boldsymbol{x}$ is a vector of explanatory variables and $P(y=1|\boldsymbol{x})$ is the response probability to be modeled. In the logistic model the logarithm of the odds for the value "1" of the binary dependent variable y is a linear combination of the independent variables x_1, $x_{2, \ldots,}$ x_m, and the model obeys the formula

$$\ln \frac{P(y=1|x_1, x_2,..., x_m)}{1-P(y=1|x_1, x_2,..., x_m)} = q_0 + q_1x_1 + q_2x_2 + \cdots + q_mx_m \,. \qquad (1)$$

It follows from (1) that

$$P(y=1|\, x_1, x_2, \cdots, x_m) = \frac{e^{q_0+\sum_j q_j x_j}}{1+e^{q_0+\sum_j q_j x_j}} = \frac{1}{1+e^{-q_0-\sum_j q_j x_j}} \,. \qquad (2)$$

The value of $P(y=1|\boldsymbol{x})$ is then used for classification. It is worth to mention, that the logistic function $1/(1+e^{-\mathbf{z}})$, also called sigmoid function, is widely used to model the activity of a single neuron in artificial neural networks.

Decision trees belong to the most popular models for solving classification tasks but, due to their interpretability, can also be used for concept description. Algorithms for building decision trees recursively partition the attribute space in a top-down manner (therefore the general name for these algorithms is TDIDT – top-down induction of decision trees) into regions homogeneous with respect to the target. The tree-learning algorithm repeatedly performs the following 3 steps:

1. Select the best splitting attribute as a root of the current (sub)tree.
2. Divide data in this node into subsets according to the values of the selected attribute and add a new node for each this subset.
3. If there is an added node, for which the data do not belong to the same class, go to 1.

This method, also known as "divide and conquer" has been implemented in various algorithms. A standard algorithm widely used by statisticians is CART (Breiman et. al., 1984) a typical tree learning algorithm used in data mining and machine learning community is C4.5 (Quinlan, 1993).

Association rules have been proposed by R. Agrawal in the early 90th as a tool for so-called market basket analysis (Agrawal et. al., 1993). An association rule has the form of an implication

$$X \Rightarrow Y, \tag{3}$$

where X and Y are sets of items (itemsets) and $X \cap Y = \emptyset$. An association rule expresses that transactions containing items of set X tend to contain items of set Y. This idea of association rules was later generalized to any data in the tabular, attribute-value form. In this generalization, association rules have the form of relationships between conjunctions of attribute-value pairs (categories) called antecedent (*Ant*) and succedent (or consequent) (*Suc*)

$$Ant \Rightarrow Suc. \tag{4}$$

The two basic characteristics of an association rule are *support* and *confidence*. Let a (for the analyzed data) be the number of examples (rows in the data table) that fulfill both *Ant* and *Suc*, b the number of examples that fulfill *Ant* but do not fulfill *Suc*, c the number of examples that fulfill *Suc* but do not fulfill *Ant*, and d the number of examples that fulfill neither *Ant* nor *Suc*.

Support of an association rule is then defined as

$$sup = \frac{a}{a+b+c+d}, \tag{5}$$

and confidence is defined as

$$conf = \frac{a}{a+b}. \tag{6}$$

In association rule discovery, the task is to find all rules $Ant \Rightarrow Suc$ that have their support and confidence above the user-defined thresholds *minconf* and *minsup*. There is a number of algorithms and systems that perform this task. We are using the LISp-Miner, developed at the University of Economics, Prague (Rauch and Šimůnek, 2014) in the reported work. LISp-Miner offer a wide variety of different forms of association rules that can be found in the data (the system can be freely downloaded from http://lispminer.vse.cz). We choose two procedures to analyze the student's data, 4ft-Miner and KL-Miner.

The 4ft-Miner procedure offers a more general form of the above described association rules. 4ft-Miner mines for patterns (4ft-rules) of the form

$$\varphi \approx \psi / \gamma, \tag{7}$$

where φ, ψ and γ are cedents (more complex formulas that just conjunctions of attribute-value pairs as in "standard" association rules) and ≈ (called quantifier) denotes a relationship between φ and ψ for the examples from the analyzed data table, that fulfill γ. If γ is empty then the procedure analyzes the whole data table. The relationships between φ and ψ are defined (and evaluated) using frequencies *a*, *b*, *c*, *d*, from the four-fold contingency table. One example relationship, used in our analysis is the so-called *founded implication* evaluated using support and confidence given in formulas (5, 6). For more details about 4ft-Miner, refer e.g. to (Rauch, 2013).

The KL-Miner procedure mines for KL-patterns R ~ C/γ. The KL-pattern R ~ C/γ means that the attributes *R* and *C* are in a relation given by the symbol ~ when the condition γ is satisfied. The relation ~ is thus evaluated on a *K* x *L* contingency table.

4. Data analysis results

4.1.Binary classification

As already stated in Section 3.1., we created models that are able, based on variables known about the students when starting their study, to classify whether a student will successfully end the study by defending the thesis or whether he will be dismissed. The dependent (target) variable is now binary, and the independent (input) variables are: *Sex*, *Citizenship*, *AddCountry*, *SSCountry*, *AdmissForm*, *StudyForm*, *RecCred*, *SSType3*, *YearsSSVSE*, *AdmissAge*. We used the SAS Enterprise Miner system version 7.1. developed by SAS (https://www.sas.com/en_us/software/enterprise-miner.html) for all the computations.

To check the strength of the relationship of these variables with the target, we first compute the R-Squares (squared correlation) value for the whole data set. The results are shown in Table 2. To create and evaluate the models (logistic regression, decision tree), we randomly split the data into training and testing part by the ratio 60:40.

The regression coefficients are estimated using maximum likelihood estimation. To find the optimal parameters of the model that maximize the product of $P(y_i \mid \boldsymbol{x}_i)$ over the training data $[\boldsymbol{x}_1, y_1], \ldots, [\boldsymbol{x}_n, y_n]$, the numerical Newton-Raphson method is used. Table 3 shows the estimated values of the parameters and their standard errors for the value "dismissed" of the *Target*.

Table 2: R-Squares for input variables

Effect	DF	R-Square
AdmissAge	1	0.031166
Sex	1	0.028087
YearsSSVSE	1	0.026437
Citizenship	22	0.018397
AdmissForm	3	0.017135
SSType3	3	0.008533
StudyForm	1	0.007965
AddCountry	11	0.005416
SSCountry	7	0.002484
RecCred	1	0.000825

Source: the authors.

To create the decision trees we used the following parameters to control the tree growing:

1. Number of splits is set to 2, so a binary tree is created.
2. Maximal depth of the tree is set to 6.
3. Minimal number of examples in a leaf is set to 5.

Chi-square criterion χ^2 is used to find a splitting attribute (see formula 8). Here χ^2_{node} refers to the value computed for the node to be split and χ^2_{left} and χ^2_{right} refer to the values computed for the two children of the split node.

$$\chi^2 = \chi^2_{\text{node}} - (\chi^2_{\text{left}} + \chi^2_{\text{right}}) \tag{8}$$

The χ^2 value itself is computed according to the formula 9, here o_{ij} are observed and e_{ij} are expected values.

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{9}$$

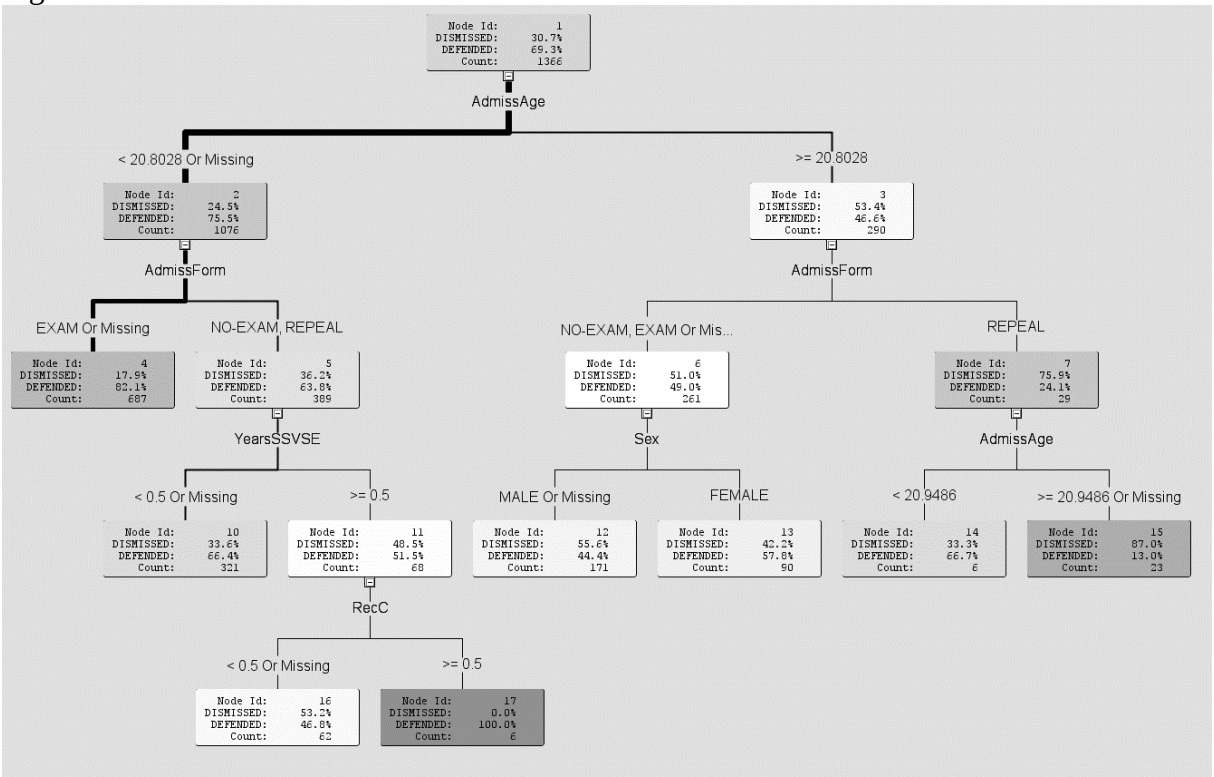
Table 3: Maximum likelihood estimates for the value "dismissed" of the Target

Parameter		Estimate	Standard Error
Intercept		-7.8153	2.9543
AddCountry	CZ	8.2840	51.4622
AdmissAge		0.0037	0.0013
AdmissForm	exam	-0.5731	0.0914
AdmissForm	no-exam	0.2410	0.0995
Citizenship	CZ	-8.6997	51.4473
RecCred		-0.0721	0.0498
SSType3	Gym	-0.1157	0.1055
SSType3	Lyc	-0.1667	0.1697
Sex	female	-0.2546	0.0667
StudyForm	comb	-0.0951	0.1897
YearsSSVSE		-0.1843	0.1329

Source: the authors.

Figure 2 shows the decision tree. As we can see the root variable is that with the highest value of R-square as shown in Table 2.

Figure 2: Decision tree



Source: the authors.

We evaluated the quality of both models using values from the confusion matrix. This matrix has four entries for binary classification: *TP* is the number of correctly classified positive

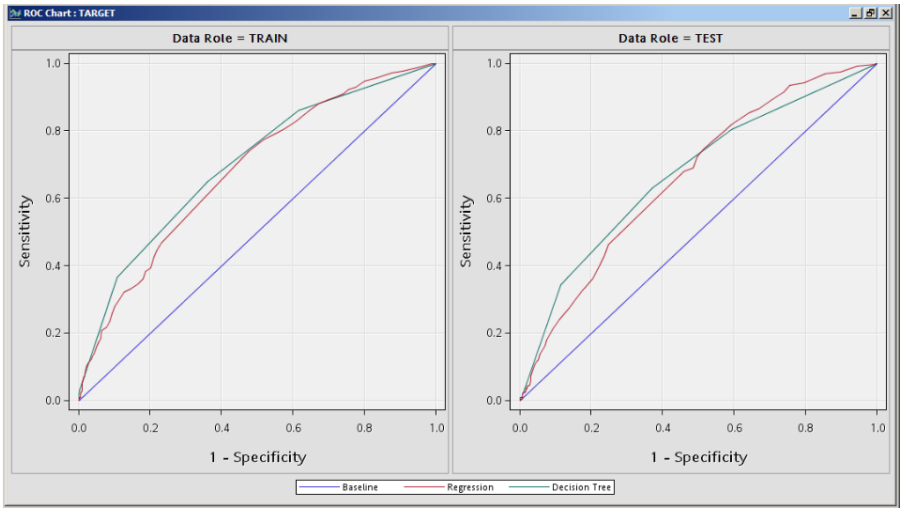
examples, *TN* is the number of correctly classified negative examples, *FP* is the number of wrongly classified positive examples and *FN* is the number of wrongly classified negative examples. Table 4 shows the classification accuracies of both models, Figure 3 shows the ROC curve. While classification accuracy is the most simple criterion that shows the proportion of correctly classified examples (*TP*+*TN*)/(*TP*+*TN*+*FP*+*FN*) for a default threshold of the score, ROC curve relates together the true positive rate *TP*/(*TP*+*FN*) and the false positive rate *FP*/(*TN*+*FP*) for various settings of the threshold of the score. Here score is a quantity related to the class predicted by the model (so score can be the probability of the class) and the threshold is used in the decision strategy in such a way that if the score is above the threshold then the example is classified into class "positive" otherwise it is classified into class "positive". While classification accuracy evaluates the model with respect to all classes, ROC is constructed only for one class (in our case for the class "dismissed" which is the "positive" one).

Table 4: Classification accuracies of the models

Model	Training set	Testing set
Decision tree	0.7226	0.7185
Logistic regression	0.7086	0.7021

Source: the authors.

Figure 3: ROC Curve



Source: the authors.

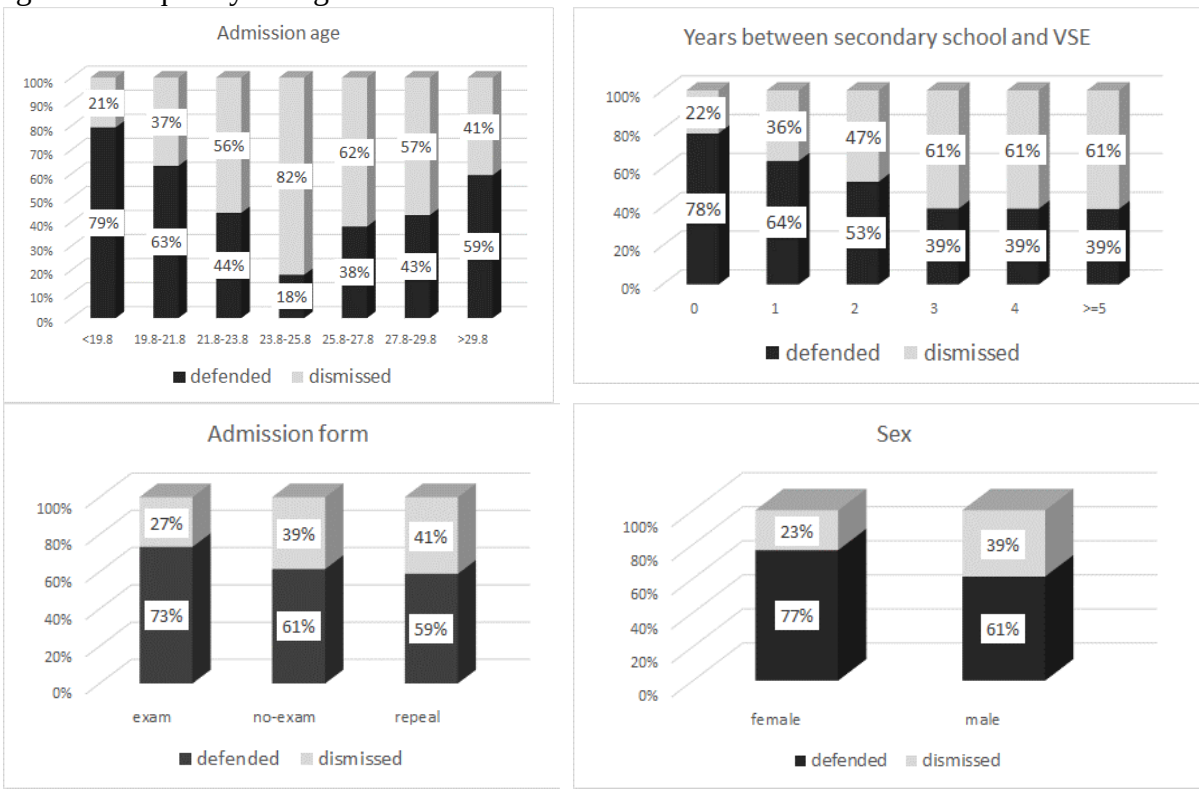
As can be seen from the evaluation of both models, decision tree slightly outperforms logistic regression in terms of overall accuracy. The ROC curve gives an ambiguous picture when analyzing the model performance for the class "dismissed"; for larger values of the threshold the decision tree again performs better (larger values of the threshold correspond to the part of ROC closer to the point [0, 0]), while for smaller values of the threshold logistic regression performs better.

4.2. Dependency analysis

Association rules mining can be performed only for categorical variables, numeric variables must be discretized in advance. This transformation can be done directly in the LISp-Miner system. We discretized the values of the variable *YearsSSVSE* into intervals, 0, 1, 2, 3, 4, 5_and_more, we discretized the variable *AdmissAge* into intervals of 2 years starting with the value 17.8 (this corresponds to equidistant discretization of the range of values found in data), and we discretized the variable *RecCred* into intervals. After this necessary preprocessing we run different procedures implemented in LISp-Miner.

We used KL-Miner to check the predictive power of individual characteristics of students known when starting their study to differentiate between students who completed their study and students who were dismissed. Figure 4 shows some interesting results of this analysis in the form of relative frequency histograms.

Figure 4: Frequency histograms for selected variables and the class variable defended/dismissed



Source: the authors.

We used 4ft-Miner for several tasks. The aim of the first task was to find strong implications, i.e. association rules having high values of confidence as defined in formula (6), that relate values of input variables (we used the same set of variables as for logistic regression) to the values "dismissed" or "defended" of the variable *Target*. The aim of the second task was to find some characteristics of students who left the university on their own. The results of these tasks are reported in Tables 5 – 7. Table 5 shows some interesting associations for the students who were dismissed, Table 6 shows some interesting associations for students who defended, Table 7 shows some interesting characteristics of students who left.

Table 5: Example 4ft patterns in the form '$\varphi \Rightarrow$ Target(dismissed)'

4ft pattern	a	b	a/(a+b)
AdmissAge(21.8-23.8) & AdmissForm(exam) ⇒ Target(dismissed)	59	38	0.608
YearsSSVSE(>5) ⇒ Target(dismissed)	59	38	0.608
AdmissAge(21.8-23.8) & AdmissForm(exam) & RecCred(0) ⇒ Target(dismissed)	52	28	0.650
SSType3(SS) ⇒ Target(dismissed)	34	22	0.607
AdmissAge(23.8 – 25.8) ⇒ Target(dismissed)	29	6	0.829
StudyForm(pres) & YearsSSVSE(>5) ⇒ Target(dismissed)	27	5	0.844
AdmissAge(23.8 – 25.8) & AdmisForm(exam) ⇒ Target(dismissed)	15	3	0.833

Source: the authors.

Table 6: Example 4ft patterns in the form '$\varphi \Rightarrow$ Target(defended)'

4ft pattern	a	b	a/(a+b)
YearsSSVSE(0) $\Rightarrow$ Target(defended)	1170	334	0.778
StudyForm(pres) & YearsSSVSE(0) $\Rightarrow$ Target(defended)	1161	325	0.781
AdmissForm(exam) $\Rightarrow$ Target(defended)	1060	359	0.747
Sex(female) $\Rightarrow$ Target(defended)	874	259	0.771
SSType3(gym) & StudyForm(pres) & YearsSSVSE(0) $\Rightarrow$ Target(defended)	*844*	*212*	*0.799*
Sex(female) & SSType3(gym) $\Rightarrow$ Target(defended)	618	152	0.803
. . .			
Sex(female) & SSType3(SS) & StudyForm(pres) & YearsSSVSE(0) $\Rightarrow$ Target(defended)	11	0	1

Source: the authors.

Table 7: Example 4ft patterns in the form 'Target(left) $\Rightarrow$ ψ',

4ft pattern	a	b	a/(a+b)
Target(left) $\Rightarrow$ SSType3(gym)	280	141	0.665
Target(left) $\Rightarrow$ SSType3(gym) & StudyForm(pres)	269	152	0.639
Target(left) $\Rightarrow$ YearsSSVSE(0)	257	164	0.610
Target(left) $\Rightarrow$ StudyForm(pres) & YearsSSVSE(0)	254	167	0.603
Target(left) $\Rightarrow$ Sex(male)	250	171	0.594

Source: the authors.

4.3. Interpretation and discussion of the results

The results of classification analysis are not very impressive. The accuracies of both models, 0.72 for decision tree and 0.70 for logistic regression, are only slightly above the baseline derived from the relative frequency of the majority class (0.69). So the data collected about the students when enrolling for the study is not too good predictors. It is obvious that the study progress in the early stages of the study plays a key role. This corresponds to the conclusions of Dekker et. al. (2009) who used pre-university data and data from the first year of study to predict dropout of freshmen at a particular university department. The authors used logistic regression, decision trees, decision rules, Bayesian networks and random forest with very similar results. When applying the different algorithms only to the pre-university data, they achieved classification accuracy between 0.68 (decision tree) and 0.71 (Bayesian network), when enriching these data by the information about study achievement during the first year, their classification accuracy improved to values between 0.75 (Bayesian network) and 0.8 (decision tree). Logistic regression model achieved the accuracy of 0.69 on the pre-university data and 0.79 on the full data. The dropout rate in their data was 0.49, so their baseline accuracy was 0.51.

The results of the dependency analysis can be used to differentiate between students who have the potential to successfully finish their studies and students who have not. We can thus identify promising students worth to retain at the University. Consider e.g. the 4ft pattern in italics from Table 6

SSType3(gym) & StudyForm(pres) & YearsSSVSE(0) $\Rightarrow$ Target(defended) (0.799). (20)

The analyzed data contain 1438 students of this group (which is almost 1/2 out of all students in the data). 844 students from this group completed the study, 212 were dismissed, 184 left by their own and 198 did not study. So a significant number of students that have a high chance to complete the study (the confidence of this pattern is 0.799) left the University earlier than necessary. This finding is also confirmed by the patterns shown in Tables 6 and 7.

Table 6 also illustrates the well-known trade of between support and confidence of association rules; rules with high support have usually lower confidence and rules with high confidence have usually lower support.

5. Conclusion

We report some initial analysis of the data about all students who enrolled for their bachelor study at the University of Economics, Prague in the academic year 2013/2014. We created both classification models to classify students into groups "dismissed" and "defended" and description models that relate together different characteristics of students. Our previous work shows that to distinguish between successful and unsuccessful students, the key variables are related to the study progress. The variables known about the students when they enter the University are less important. Anyway, in the work reported in this paper, we rely only on these variables as our aim was to create an "early warning" system capable to identify students who might run into troubles during their study. Following the idea of interpretability, we intentionally narrow our classification analysis to methods that provide interpretable models (decision trees, logistic regression). So we omit methods creating more complex and harder to understand models like SVM, neural networks or gradient boosting approaches. The performance of our classification models is comparable with the results published for a similar task by other authors. We also identified some "patterns" of students who have a good chance to successfully finish their studies. To do this, we used the association rules mining approach, again with the aim to obtain interpretable and understandable results. The main drawback of this approach is the (usually) large number of found associations. So we present only some illustrative examples rather than a detailed analysis of all created rules. It is of no surprise, that some of the shown rules correspond to partial paths in the decision tree created for classification (Figure 2).

We will extend our work in several directions. We plan to use additional data about the students, in particular, more details about study achievements collected on a yearly basis, to improve the classification accuracy. We also plan to perform similar analyses on time-shifted data, i.e. on the data about students who enrolled for their study in the years following 2013/2014 to check the stability of the found patterns.

Acknowledgments

This paper was institutionally supported by the long-term research scheme of the Faculty of Informatics and Statistics of University of Economics in Prague.

References

[1] Agrawal, R., Imielinski, T. Sawami, A. 1993. Mining associations between sets of items in massive databases. In Proceedings of the ACM-SIGMOD International Conference on Management of Data, Washington D.C., 1993, pp. 207-216.

[2] Breiman, L., et al. 1984. Classification and regression trees. Boca Raton : Chapman and Hall/ CRC, 1984. ISBN 978-0412048418.

[3] Dekker, G. W., Pechenizkiy, M., Vleeshouwers, J. M. 2009. Predicting students drop out: A case study. In International Conference on Educational Data Mining. Cordoba, Spain, 2009, pp. 41-50.

[4] Fischer, J. et al. 2016. Eurostudent VI. Základní výsledky šetření postojů a životních podmínek studentů vysokých škol v České republice. Research Report. Prague : Ministry of Education, Youth and Sports of the Czech Republic, 2016.

[5] Hilbe, J. M. 2009. Logistic regression models. London : Chapman & Hall/CRC, 2009. ISBN 978-1138106710.

[6] Kotsiantis, S., Pierrakeas, C., Pintelas, P. 2003. Preventing student dropout in distance learning systems using machine learning techniques. In International Conference on Knowledge-Based Intelligent Information & Engineering Systems. Oxford, 2003. pp. 3-5.

[7] Lykourentzou, I., Giannoukos, I., Nikolopoulos, V. 2009. Dropout prediction in e-learning courses through the combination of machine learning techniques. In Computer & Education Journal, 2009, vol. 53, iss. 3, pp. 950-965.

[8] Norton, A., Cherastidtham, I. 2018. Dropping out. The benefits and costs of trying university. Grattan Institute. [cit. 2019-03-10] https://grattan.edu.au.

[9] Quinlan, J. R. 1993. C4.5: Programs for machine learning. Amsterdam : Morgan Kaufman, 1993. ISBN 978-1558602380.

[10] Rauch, J. 2013. Observational calculi and association rules. New York : Springer. ISBN 978-3-642-11736-7.

[11] Rauch, J., Šimůnek, M. 2014. Dobývání znalostí z databází, LISp-Miner a GUHA. Oeconomia, Praha, 2014. ISBN 978-80-245-2033-9.

[12] Sagenmuller, I. 2018. Student retention: 8 reasons people drop out of higher education. [cit. 2019-03-10] http://www.u-planner.com.

[13] Vossensteyn, H. et al. 2015. Dropout and completion in higher education in Europe. Publications Office of the European Union, 2015. [cit. 2019-03-10] http://publications .europa.eu.