Using Decision Trees to Predict the Likelihood of High School Students Enrolling for University Studies

Zdena Dobesova(✉) and Jan Pinos

Department of Geoinformatics, Faculty of Science, Palacky University,
17. listopadu 50, 779 00 Olomouc, Czech Republic
{zdena.dobesova, jan.pinos01}@upol.cz

Abstract. This article presents the use of decision trees to identify the main factors which predict the likelihood of high school students matriculating at the Department of Geoinformatics, Palacky University in Olomouc (Czech Republic). The Department of Geoinformatics has been running a continuous and systematic information campaign about studying the fields of geoinformatics and geography within the department. In order to collect feedback about the information campaign, students who apply to study at the department are then given a questionnaire. Answers received from this questionnaire in two years, (2016 and 2017), were analyzed using decision trees that help us understand what specific type of information positively affects the likelihood of a student actually commencing studies at our department.

Keywords: Data mining · Decision tree · Questionnaire · Geoinformatics GIS · Advertisement · University applicant

1 Introduction

Universities in the Czech Republic offer numerous branches (fields) of study for students desiring a bachelor's degree. Every university has various way of advertising its study programs. One way is through face-to-face meetings engineered between a teacher/other university representative and secondary school students (and possibly their parents and friends). This method is still very important but has its limits when trying to reach higher numbers of possible applicants. Therefore, digital ways of advertising, specifically over the Internet, are used, in addition to classical advertising via printed materials. Some study fields are well known among the secondary school students (such as medicine, engineering, agriculture, law, etc.). On the other hand, newer fields such as geoinformatics have much less name recognition among secondary school students and therefore rely more on advertising in order to gain new students. Simply counting the total number of applicants to the program is not a fine-grained enough way to measure the impact of the department's measures. Therefore, at the department of Geoinformatics we decided to evaluate the information campaign in more detail.

R. Silhavy et al. (Eds.): CoMeSySo 2018, AISC 859, pp. 111–119, 2019.
https://doi.org/10.1007/978-3-030-00211-4_12

Valuable information was obtained by evaluating the information campaign using decision trees as one of the data mining techniques [1, 2]. The advertisement activities and collection of the data were described in a previous article [3]; this article also included information about secondary school students' most frequent sources of information about the department. Our current research extends the previous data set to add information about the commencement of studies, and from that extended data set, we construct a decision tree as a predictive model of whether students matriculate with us. The data set is described in Sect. 2. Section 3 is a short introduction to data mining by the decision tree. The statistical evaluation and decision tree are described in Sect. 4.

2 Public Presentations of the Study Fields and Consequent Data Collection

The delivery of information to secondary school students can be via multiple channels; Palacky University uses many of them. First, the University has extensive information about fields of study and department admission procedures on its web pages in digital form [4]. Additionally, information is presented by a representative of the Department of Geoinformatics at the special study fair Gaudeamus in Brno and Prague [5]. Third, the Open Day event at Palacky University could be considered as an opportunity to inform the students about the programs in person [4]. At this event, interested students can visit the university to speak with the teachers and collect information about the application procedure personally. At both events, the university provides printed leaflets and booklets with a list of fields of study and detailed descriptions.

Additionally, the Department of Geoinformatics at Palacky University organises several special presentations about bachelor's degrees in Geoinformatics and Geography. One presentation occurs as part of the worldwide GIS Day in November [6], then other ones are: GIS week in April, presentations at secondary school and excursions to the department to see laboratories and equipment. For all presentations, the printed booklets about Geoinformatics study are used.

Information about Geoinformatics is also spread indirectly by teachers, friends, schoolmates, siblings and parents. Other information channels like a radio broadcast, TV news, the Teachers' Newspaper (www.ucitelskenoviny.cz) could also be considered. A detailed description of the information channels, events and printed information materials is presented in the article "Discovering association rules of information dissemination about Geoinformatics university study" [3].

This research evaluates the influence of each information activity on the likelihood of the student actually starting the study of Geoinformatics. The research task is to measure the impact of each form of presentation: personal presentation, information on the internet, leaflets and other channels of advertisement, on an applicant's reaching the final decision to study Geoinformatics. Moreover, the event 'GIS weekend' is evaluated in this article. GIS weekend is solely organized for real applicants. The submission period for applications ends in February each year. Subsequently, all applicants are invited in April to visit the Department of Geoinformatics at the university. The GIS weekend event covers the detailed overview of study subjects, presentation of computer classrooms, and opportunities for studying abroad under the ERASMUS+ program.

Moreover, special laboratories are presented. There are eye-tracking laboratory (for cartographic testing), 3D modelling and printing laboratory (creation of earth surface models) and laboratory for the tangible creation of elevation models at the department. Students also create their first digital map on the computer. The students also practice taking an exam in mathematics and geography. Department members have assumed that the personal attendance of the department staff and the presentation of practical examples of Geoinformatics would be very influential for the final decision to actually begin studies at our department.

2.1 Source Data Matrix

The data were collected during admission periods in two years 2016 and 2017. Data from the questionnaire were gathered from applicants. The structure of the questionnaire contained 12 possible responses about the source of information [3]. Applicants check one or more sources of received information about the study:

1. Teacher of Geography at a secondary school *(TeacherG)*
2. Teacher of Information Science at a secondary school *(TeacherI)*
3. Lecture at a secondary school provided by the Department of Geoinformatics *(Lecture)*
4. Attending the GIS Day event at the Department of Geoinformatics *(GIS Day)*
5. The Open Day event at Palacky University (November and January) *(Open Day)*
6. Friends, schoolmates, siblings *(Friend)*
7. Parents and grandparents *(Parent)*
8. Gaudeamus Fair in Prague or Brno *(Gaudeamus)*
9. Leaflets about studying Geoinformatics *(Leaflets)*
10. TV and Newspapers *(TV News)*
11. Internet by your search *(Internet)*
12. Teacher Newspaper *(T News)*

Data were transformed to the matrix with binary values. Each applicant was assigned an identification number (ID) in the matrix. Information about attending the GIS weekend was manually added to the data matrix according to the attendance list.

Moreover, information about the start of the study was added into the matrix as the last column. This column is class (category or predicted) value in the prediction model. An example of input data for the construction of decision trees is shown in Table 1.

Table 1. Input data matrix (short example) for decision tree.

ID	TeacherG	Lecture	Open day	Internet	Friend	…	Gaudeamus	GIS weekend	**Start study**
1	1	1	1	0	0	…	0	1	1
2	0	0	0	1	0	…	1	1	0
3	0	0	1	0	1	…	0	0	1
4	0	0	0	0	0	…	1	1	0
5	0	0	0	1	0	…	0	0	0

3 Decision Tree Method and Prediction

The input data for classification task is a collection of records. Each record is characterised by tuples ($x_1, x_2, \ldots, x_n, \boldsymbol{y}$) where $\mathrm{x_i}$ is input data set, and $\boldsymbol{y}$ is a special attribute, designated as the predicted class (category or target attribute) [1, 2]. In case of the presented situation, the predicted class is 1/0 (yes/no) for the start of the study of Geoinformatics. The data could be used for the design of classification/prediction model that predicts the class for unknown records of data. A decision tree is a typical example of supervised learning [7, 8] where the aim is to predict resulting value.

The construction of optimal decision tree is computationally infeasible because of the exponential size of search space. Efficient algorithms have been developed to induce a reasonably accurate tree in a reasonable amount of time. The Hunt`s algorithm is a greedy strategy that grows a decision tree by making a series of locally optimal decisions of which attribute to use for partitioning data [1, 2]. Hunt's algorithm is a base foundation for algorithm ID3, C4.5 and CART. Decision trees, especially smaller/sized trees are relatively easy to interpret. The interpretation is by simple rule with a condition like "*If conditions Then…. Else*". The presence of redundant attributes does not adversely affect the accuracy of decision trees. The decision tree partitions the data set into the smaller group of records that are more homogenous than the group in a higher level of the tree.

The software WEKA v 3.8 was used for the construction of the decision tree. The algorithm J48 is accessible in WEKA. J48 is a Java implementation for generating a pruned or unpruned C4.5 decision tree. C4.5 was developed by Quinlan [9]. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalised information gain (difference in entropy). The attribute with the highest normalised information gain is chosen as a node in the tree to make the decision. The C4.5 algorithm then recurs on the smaller group of records [7, 8]. The outputs of the tree (leaves) are categorical values.

4 Statistical Evaluation and Data Mining

The survey collects data from the years 2016 and 2017. In 2016, the number of collected applicant questionnaires was 46 and 55 in the year 2017 (total 101). All applicants were admitted in both years. From 48 students who applied, only 22 matriculated in 2016 and 24 did not start the study (more than half of them). The proportion in 2017 was better, 35 students matriculated, and 20 students did not. The overview is displayed in Fig. 1.

Additionally, the influence of attending the GIS weekend event was statistically evaluated. Applicants can freely attend the GIS weekend in April (after submission of their application in February). The expectation is that students with a serious interest in the study of Geoinformatics are more likely to attend this excursion of the department and to prepare for the admission exams. Table 2 shows that only about half of the applicants that attended the event GIS weekend actually started the study. The

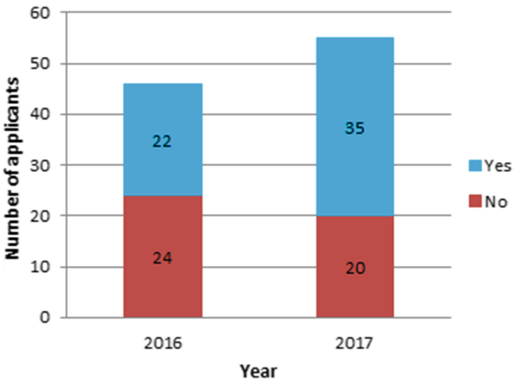

Fig. 1. Count of applicants and count of matriculants (start study Yes/No).

expectation about the positive influence of the GIS weekend event has not been verified. Also, many applicants who did not attend this event, did commence studies in our department (around half of them).

Table 2. Comparison of the number of applicants attending GIS weekend and number of students that start study

Year	Number attending GIS weekend	Number attending GIS weekend and start study	Number not attending GIS week and start study	Total number started study
2016	25	12	10	22
2017	29	17	18	35
Total	54	29	28	57

4.1 Decision Trees

The data matrix described in Sect. 2.1 was used as an input data for constructing a decision tree. Firstly, data for both years were taken as a training data set (totally 101 records). The pruned decision tree is shown in Fig. 2. The confidence factor has a default value 0.25. The confidence factor determines the pruning (smaller values incur more pruning, higher value less pruning). The tree predicts the commencement of studies in our department (rectangle leaf **Yes**) and not commencing studies in our department (rectangle leaf **No**). The total of correctly classified students is 69.3%, and the total of wrongly classified students is 30.6%.

The first node is a condition about attending the event **Open Day** (Yes or No). This condition divided the input data set, and it could be assumed as the most influential. The attribute at the first node produces the highest normalised information gain. When students attend the event Open Day, 38 of them matriculated (also 12 students are wrongly classified, they did not start to study at our department). The next node is a lecture of university staff or university students at a secondary school. The rule is interpreted as:

"When an applicant does not attend the ***Open Day*** *but attends the* ***Lecture****, then 17 applicants matriculate (5 are wrongly classified)."*

The next condition could be constructed according to the tree. The last node presents a **Friend**'s recommendation of the study. When the applicant has no previous information and has only information from a source labeled "Friend" then 9 of them start the study. In case of No information from Open Day, Lecture, GIS week and Friend, then 17 students do not start the study (7 wrongly classified). The first two nodes in the tree (Open Day and Lecture) mean the highest influence to start the successful study. It depicted by positive branches in the tree. In addition, positive branch at the fourth node Friend could be summarize in total number of matriculates: Yes = 38 (OpenDay) + 17 (Lecture) + 9 (Friend) = 64 (totally). The lower position of GIS week in the tree (Fig. 2) corresponds with the previous statistic finding that GIS week has only a partial influence on the motivation to start the study.

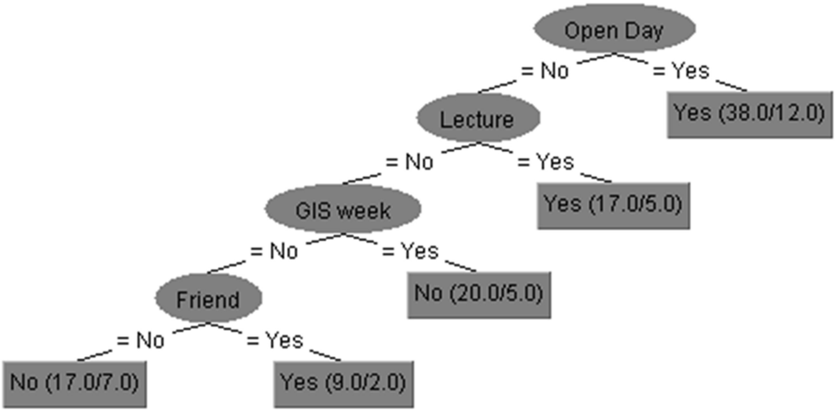

Fig. 2. The pruned decision tree for the whole data set.

Secondly, the construction of unpruned trees were generated for the year 2016 subset and 2017 subset separately. The confidence factor was set to the same value 0.25. The decision trees for these two subsets are showed in Figs. 3 and 4. The correctly classified instances were 35 (76%) and incorrectly classified instances 11 (24%) for the year 2016. The correctly classified instances were 41 (74.5%) and incorrectly classified instances 14 (25.5%) for the year 2017. In both cases, the correctness of decision trees is slightly higher than for the whole data set (above 74% in comparison with 69%). The first node for the whole data set and the data set for the year 2016 is the same; it is Open Day. For 2017 data set the highest node is Friend. The recommendation by Friend directly influences 16 applicants (right short branch in Fig. 4). The node Friend is also placed high in the tree structure (specifically at a second position) in the year 2016. In fact, the high position of nodes Open Day and

Friend in all resulting tree structures confirms the strong influence of a "personal way" of delivering information.

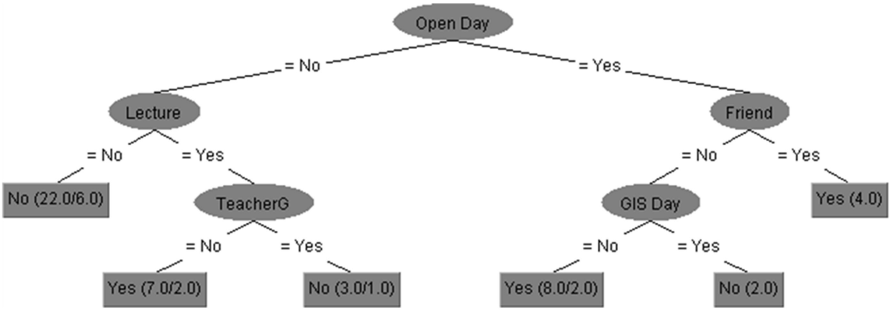

Fig. 3. The unpruned decision tree for data set in 2016.

Negative influence (left branch) can also be found in the decision tree generated for 2016 data set (Fig. 3). This influence is valid for the relatively high number of applicants (22 applicants). The interpretation of the rule is: *"When there is No information from Open Day and No information from Lecture, then 22 students will not start the study (6 are wrongly classified)."*

When all three decision trees are compared, there is low or no occurrence of the node **Internet**. In the 2016 data set and the whole data set, the node Internet is not present in the trees at all. In case of 2017 data set, this node is present but at a low position. The rule could be stated (left branch): "*When there is Friend = No and Lecture = No and Internet = Yes and Open Day = No then the 10 applicants will not start the study (4 are wrongly classified).*" It means that information from the internet does not assure the start of the study. However, the only positive influence is when we combine the Internet with the event Open Day which in this case produces 8 applicants that start the study.

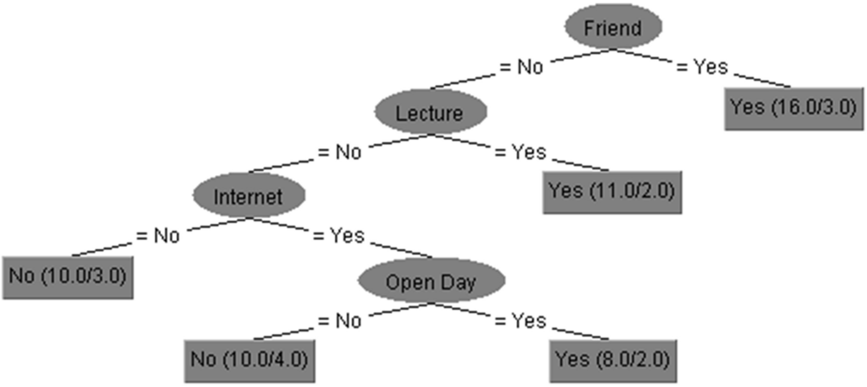

Fig. 4. The unpruned decision tree for data set in 2017.

Researched results could also be compared with the results which were presented in the previous article on this topic: about association rules [3] for the same input data set. Besides the rules with frequent combinations of several sources of information about the study, the rules with information about the start study were deduced in association with the source of information. Three rules were deduced:

- ***OpenDay = 1 and Internet = 1 => Study = 1***

If students attend Open Day and have information on the Internet, they start the study.

- ***Friend = 1 and Internet = 1*** $\geq$ ***Study = 1***

This rule is an exception. Students probably start study if they have the only recommendation of friends and information from the Internet.

- ***Internet = 1 => Study = 0***

This rule expresses that student probably does not start the study if he obtained the information solely from the Internet. The information about the study from Internet has a very low impact on deciding whether to start the study.

Concerning the rules about the positive start of the study corresponds with the information gained from the decision tree mentioned above.

5 Results

In this paper, we found that attendance at the Open Day event and recommendation of a friend (and schoolmates, siblings) have the greatest influence over whether students will commence studies at the Department of Geoinformatics. Both of these are considered a “personal way of delivering the information”. The high position of nodes Open Day and Friend (followed by Lecture, GIS Day, GIS weekend) in the predictive decision trees means that the information gain divides the applicants into two groups. The first group consists of applicants that receive information by these personal ways, and they start the study of Geoinformatics. The second group consists of applicants that do not start the study. It has also been found that the influence of the Internet on the student’s decision is very low. The Internet does not appear as a node in two of three presented decision trees.

Surprisingly, personal attendance at the GIS weekend event does not have a significant influence on the start of the study of Geoinformatics. The GIS weekend appears in the decision trees but in rather a negative way. The low position means that it does not bring minimisation of entropy to divide the input data set.

Comparison of decision trees with the previously presented data mining technique – association rules [3] brings the same main findings. The decision trees clarify a more detailed explanation of the combined elements of information dissemination and advertising. The resulting findings support the intensification of personal ways of delivering the information to potential applicants such as the Open Day, Gaudeamus, GIS Day and lectures at secondary schools in future. The personal presentation of information will assure higher numbers of applicants, and mainly, a higher number of students that actually decide to start the study of geoinformatics.

The construction of a predictive model based on questionnaire data brings a new approach. The use of the decision tree shows, in a straightforward manner, how to predict the likelihood of matriculation based on the way the student received information before enrolment if that data exists.

Acknowledgement. This article has been created with the support of the Operational Program Education for Competitiveness – European Social Fund (Project CZ.1.07/2.3.00/20.0170 Ministry of Education, Youth, and Sports of the Czech Republic) and with the support of the student Project IGA_PrF_2018_028 of the Palacky University Olomouc.

References

1. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining, 1st edn. Addison-Wesley Longman Publishing Co., Inc., Boston (2005)
2. Šarmanová, J.: Methods of Data Analysis (Metody analýzy dat). Technical University of Ostrava, Faculty of Mining and Geology, Ostrava (2012). (In Czech)
3. Dobesova, Z.: Discovering association rules of information dissemination about Geoinformatics university study. In: Silhavy, R. (ed.) Artificial Intelligence and Algorithms in Intelligent Systems, vol. 764, pp. 1–10. Springer, Cham (2019)
4. Palacky University: Why study at Palacky University? In: Proč studovat na Univerzitě Palackého? www.studuj.upol.cz
5. MP-Soft, Gaudeamus Fair. https://gaudeamus.cz/
6. Esri, GIS day. http://www.gisday.com/
7. Witten, I.H.: Data Mining. Morgan Kaufmann, Burlington (2011)
8. Pavel, P.: Methods of data mining (Metody Data Miningu), part two. University of Pardubice, Faculty of Economics and Administration, Pardubice (2014). (In Czech)
9. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., Burlington (1993)