Use of the university’s enrolment campaign database for the development of a computer model to predict student expulsion

A V Zharikov¹, E V Zhuravlev¹, O V Zhurenkov² and D Yu Kozlov¹

¹ Institute of Mathematics and Information Technology, Altai State University, 61 Lenin ave., Barnaul, 656049, Russia
² International Institute of Economics, Management and Informational Systems, Altai State University, 61 Lenin ave., Barnaul, 656049, Russia

E-mail: zharikov@math.asu.ru, evzhuravlev@mail.ru, zhur@pie-aael.ru, dyk.barnaul@gmail.com

Abstract. The article discusses the construction of a computer model to predict the problems occurrence in students in the educational process at the university. The following data sources of Altai State University were used for this purpose: “Admissions Office” (enrollees database) and “Dean’s office” (database of students) for 2013-2018. These data were combined using developed SQL scripts. While analyzing the obtained combined data set, we had to face the difficulties typical for solving data analysis problems. Thus, it turned out that there are incomplete and inconsistent data or cases when one and the same entity is named differently, etc. In order to solve these problems, we wrote a script in the R programming language using regular expressions, the data were unified and standardized, and the missing data were restored using the information from other fields of the data set. Then we discarded the variables with the near-zero dispersion, which could not make a significant contribution to the developed predictive model. After that, the data set under study was divided into 2 parts: the 2013-2017 data were taken to build a predictive model with the use of the logistic regression algorithm, while the data for 2018 were used, in fact, to predict whether a particular student would be expelled. It should be mentioned that the 2013-2017 data were divided into the training and test samples in the proportion of 90% and 10% correspondingly. The test result of the computer model built in the R programming language showed satisfactory accuracy; the most significant factors affecting student expulsion were also identified. The paper substantiates the economic feasibility of using the developed computer model at the university.

1. Introduction
For any university, expelling a student leads to several problems. These are resources wasted on education (especially for state-funded places), various reputational costs that affect the ranking of the university and the number of enrollees, and a reduction in the university funding through various channels which results in the personnel cuts. For example, in the USA the expulsion level is one of the most important indicators of a university competitiveness, demonstrating, on the one hand, the attractiveness of a university (its ability to retain students and to prevent their transfer to another university), and on the other hand, the effectiveness of the university’s educational policy in adapting students to study, helping them in the educational process [1-2]. In the modern system of the Russian higher education, this indicator is also used to evaluate the effectiveness of the university.
The university could have retained some of the expelled students, foreseeing difficulties in their studies and focusing the attention of the university services on such students. In this regard, it is interesting to learn the mechanism for identifying such troubled students.

In the last twenty years, studies on the student expulsion in Russian universities have been actively developing. However, most of the works are descriptive in nature. Mostly these studies are data analyses for separate universities and are not based on any theoretical model [1].

Researchers in this field note that Western models of student expulsion are poorly suited for the Russian education system. Most of the existing concepts of expulsion have been developed for the American system of higher education where voluntary expulsion (a student decides on the expulsion on his/her own) prevails. In Russian universities, forced expulsion is widespread, i.e. students are expelled due to poor academic performance. Therefore, the model for Russian universities should mostly focus on the factors influencing academic performance and possible academic challenges [1]. The reasons for expelling students are associated not only with poor performance, there are quite a few of them. A detailed study of the reasons for expulsion from Russian universities was performed in the work [3].

2. Operational data sources
In the presented study, the data from two operational data sources (ODS) of Altai State University, i.e. “Admissions Office” (enrollees database) and “Dean’s office” (students database), were taken as primary data. The data were taken for 2013–2018, taking into account the entire period of study (4 years for undergraduates), the primary data contained only 2 data sets with a complete time lag of observation.

![Figure 1. Data flows during uploading to the data storage.](image)

Our project is conceived as iterative: in the future we will supplement the data from other ODS. At the next iteration, the “Academic Record Lists” source containing performance data will be added. During subsequent iterations, both internal data (for example, from CCTV cameras) and external data (for example, from social networks, scientific communities, and publication activity data) can be used. Unlike the ODS used in this work, the perspective ODS are dotted in figure 1.

According to the RF Federal Law “On Personal Data”, anonymized data (not containing personal information) get to the intermediate area (figure 1) [4]. For the depersonalization purpose, we used a hash function from the minimum set of personal data, which form a unique set (key) for the vast majority of records. In addition, “Date of Birth” was transformed into “Age”. The information on the
birth date not only refers to personal data, but also can not affect the expulsion, while the age of the student may well be a significant factor.

At the first stage of data origination, excess data were revealed (fields that are not relevant for our studies). First of all, these are numerical characteristics, such as numbers of various documents. Thus, the dimension of the intermediate table was reduced from 124 to 64. The attributes obtained in this way helped to formulate the necessary entities for the designed data storage (figure 2).

**Figure 2.** Conceptual framework of data storage.

Most attributes are dimensions of the designed storage. Besides, some of the attributes form one dimension and can be used for decomposition (granularity) or convolution (generalization) of the corresponding dimension. In the available data, such attributes are: Faculty, Department, Major, Group (form a group with the information from the Curriculum) and Country, Region, Settlement, Address (form a group with the information about the address). Moreover, the last group (address) is used both for the address of the student and for the address of the educational institution that the prospective student finished.

To reduce input errors, it is recommended to use the KLADR directory classifier of the Russian Federation addresses (for the citizens of the Russian Federation) and the directories of the internal information system “Curriculum”, which is reflected in the diagram in figure 2. In the future, we plan to enrich the data with the detailed information on the Unified State Examination and with the academic performance data (from the “Academic Record Lists” ODS).

The analysis of the data quality demonstrated a poor result. A careful study of this aspect revealed the following problems: empty values, incorrectness and occasional inconsistency of data, different data formats. Most of the problems are associated with poor-quality software used to input primary data. The transition to new software, which will take into account inter alia our requirements, will significantly reduce operator errors and improve the quality of the source data.

To solve these problems, we wrote a script in the R programming language using regular expressions, the data were unified and standardized and the missing data were restored using the information from other fields of the data set [5, 6]. Correct records (figure 1) were uploaded into the data storage.

It should be mentioned that a part of the missing values can be restored from other fields of the record. For example, “Type of educational institution” is unmistakably restored by the “Educational institution” value (figure 2). Similarly, the values for such fields as “Country”, “Region”, “Faculty”, “Department”, “Major” are restored. Using the scripts containing regular expressions, additional processing was performed (figure 1): in the rejected entries, the missing values for the listed fields were entered.
3. Development of a computer model to predict student expulsion

Further manipulations were carried out with the data uploaded into the data storage. It was decided to divide the entire data set into two parts. For the 2013–2017 data (with the cardinality of more than 11 thousand records), we decided to build a binary classification model. Then we should identify “troubled” students from the data referring to students admitted to the university in 2018.

At the next stage, variables with near-zero dispersion were discarded [7] on the assumption that they will hardly affect the predictive ability of the developed model. After the above described procedures, 32 variables out of 64 remained in the data set.

Before building a model, an exploratory data analysis was performed. First of all, we tested a hypothesis that the “Mean Unified State Examination Score” affects the expulsion. It should be noted that the reason for expulsion can be not only poor academic performance. As figure 3 shows, the vast majority of the expelled students did not have the worst mean Unified State Examination score on admission.

![Graph characterizing the influence of the mean Unified State Examination score on expulsion: (a) threshold 10; (b) threshold 67.](image)

The factor analysis revealed 12 most significant parameters, among which there are the following attributes: the presence of parents, the number of brothers and sisters, the number of children, the enrollee residence (a village or a city), the residence of a student in a dormitory, the citizenship (the Russian Federation or foreign), the faculty and the learning environment (full-time, part-time, non-resident instruction, individual plan). These factors partially coincide with the ones previously identified by various researchers in this area. Some factors are directly based on the aggregate measures that were identified in the studies (social status, nationality, marital status, financial status) [1].

The graph in figure 4(a) shows that the largest share of the students expelled from Altai State University accounts for the Altai Republic, which is neighboring region to Altai Krai. The graph in figure 4(b) demonstrates that the main type of the educational institution that the enrollee has finished is school. Accordingly, this parameter value accounts for the largest share of the expelled. At the same
time, it can be noted that among the expelled, the share of the ordinary school (not a gymnasium or a lyceum) students is a disproportionately high.

Figure 4. Graphs describing the geography (a) and the types of educational institutions (b) of the Altai State University enrollees.

Next, a predictive model based on the logistic regression algorithm was developed [7], the dependent variable being the binary status of the student (expelled or not). One of the advantages of the chosen method for the current task was its ability to work properly with both numerical and factorial (categorical) variables, i.e. it did not require additional data processing, which would be necessary for a number of other methods. Then, the built logistic regression model was transformed into an optimal logistic regression model containing only 12 predictors by means of discarding predictors based on the Akaike information criterion. Also, the selected algorithm makes it easy to evaluate the significance of predictors, which can be useful in the further construction of other mathematical models. The most significant factors for the current task were (in increasing order of p-value for predictors):

- The type of the entrant’s previous educational institution (school, gymnasium, secondary vocational school, another university, etc.).
- The mean Unified State Examination scores on admission.
- The presence of living parents (especially mothers).
- Whether the entrant is a foreigner or not.
- Whether the entrant is a city or a village resident.

The application of the logistic regression algorithm for a certain set of independent variables values results in the probability of accepting a dependent variable of one out of two options (herein, if the student will be expelled or not). One can classify a student based on the probabilities obtained by choosing a certain threshold probability value using ROC analysis [7–10]. First, a ROC-curve (figure 5) showing the dependence of the correctly classified students number (True positive rate) on the number of incorrectly classified ones (False positive rate) was built.

The area under the curve, the AUC parameter, indicates the predictive power of the model, where AUC = 1 corresponds to the ideal classifier, while AUC = 0.9, which is the value obtained for our data, is considered to provide high classification accuracy [7, 8].
Next, we constructed the curves depending on the cutoff threshold of the following parameters: sensitivity (herein, the proportion of the correctly identified students of one class), specificity (herein, the proportion of the correctly identified students of another class) and accuracy (an indicator of the classification efficiency with the use of the selected model). The value of 0.19 corresponding to the intersection of all the three curves was chosen as the threshold point (figure 6).

Nowadays, it is possible to use the cutoff point to move from the probabilities to the forecast of the student expulsion. The following result was obtained for the test sample (10% of the initial data of students from the 2013-2017 data set): the status predicted by the logistic regression model coincided with the actual status (expelled or not) in 86% of cases. Thus, the test result of the computer model built in the R programming language showed satisfactory accuracy.
Further, the trained model identified potentially “troubled” students who entered the university in 2018 (the students who had already been expelled were previously excluded from this data set). Their share approximated 22%, i.e. 787 students out of the total number of 3,573 people.

Let us suppose that the deans, having the information on such students, provide them with appropriate assistance (in the form of compensating courses, assigning tutors, allocating psychological or financial assistance, etc.). Basically using the existing university tools on an individual basis, they manage to retain 40 students (just over 5% of those at risk). Given that the university receives 98,000 rubles funding per student a year, the economic effect of the corresponding system introduction at the university will reach as much as 4 million rubles per year.

4. Conclusion

Thus, knowing about the factors that influence students’ expulsion, the university can formulate and apply targeted policies aimed at academic, social and psychological support of students. Such policies will favourably affect risk groups as a whole, while the university administration will have the proper idea of particular students and their potential problems in order to take a pointed effort.

We hope that the research in this area will not only help a given university, but also will contribute to the development of a model suitable for Russian universities on the whole. Thus, the inter-university research project “Trajectories and experience of Russian University students” (TERUS) is already being implemented within the network collaboration of the universities participating in the “5-100” program. It is an annual survey of students, starting from the first year and up to the graduation, which is supplemented with administrative data on the student's progress, scholarship, and financial assistance. Data collection started in 2015 and is planned to be continued until at least 2020 [1].

References

[1] Gorbunova E V 2018 Educational Studies Moscow 1 110–131
[2] Des Jardins S L, Ahlburg D A and McCall B P 1999 Economics of Education Review 18(3) 375–390
[3] Ovcharenkov E A 2014 Modern problems of science and education. 6 752–757
[4] Federal Law “On Personal Data” dated 07.27.2006 152-FZ
[5] Friedl J E F 2006 Mastering Regular Expressions 3e (Sebastopol, CA, USA: Publisher “O'Reilly Media”) p 542
[6] Goyaerts J and Levitian S 2012 Regular Expressions Cookbook 2e (Sebastopol, CA, USA: Publisher “O'Reilly Media”) p 612
[7] Shitikov V K and Mastitsky S E 2017 Classification, regression, Data Mining algorithms using R URL: https://github.com/ranalytics/data-mining
[8] Paklin N B and Oreshkov V I 2013 Business Intelligence: From Data to Knowledge 2e (St. Petersburg: Publishing House “Piter”) p 704
[9] Orlova I V 2009 Multidimensional statistical analysis in economic problems: computer modeling in SPSS (Moscow: Publishing House “Vuzovskij uchebnik”) 310 p
[10] Chakraborty D P 2017 Observer Performance Methods for Diagnostic Imaging Foundations, Modeling, and Applications with R-Based Examples (Boca Raton, FL, USA: Publisher “CRC Press”) p 542