Advances in Social Science, Education and Humanities Research, volume 437
International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020)

Analysis of Students' Academic Performance by Using Machine Learning Tools

Gafarov F.M.* Rudneva Ya.B. Sharifov U.Yu. Trofimova A.V. Bormotov P.M.

Kazan Federal University, Kazan 420000, Russia
**Corresponding author. Email: fgafarov@yandex.ru*

ABSTRACT
In higher education, considerable experience has been gained in applying analytics using multidimensional databases (including retrospective ones). One of the promising areas in this area is data mining. Data mining as an interdisciplinary field of research allows creating predictive models of students' academic success. However, questions remain in the scientific community about the types and sources of data relevant for building prognostic models, about the methods of processing this data, and about the variables that determine students' academic success. The purpose of the study is to analyze, using machine learning methods and artificial neural networks, which variables affect the academic success of students. SPSS Statistics and data mining methods using the Python programming language were used to process and analyze data. The study analyzed data on student performance at Kazan Federal University from 2012 to 2019. Preliminary results showed that data mining methods have good potential for creating information-analytical systems that allow not only modeling or visualizing data, but also predicting stable trends.
Keywords*: academic (educational) analytics, data mining, Python, predictors, academic success, forecasting, neural networks*

1. INTRODUCTION

Active integration into the digital environment and the use of digital technologies provides Russian universities with a unique opportunity to solve a number of strategic problems: (1) through the use of Data Science methods, to modernize or design algorithms for filling and subsequent analytical processing of databases in the main areas of activity, (2) to quickly receive and analyze information on the current state of educational activity (descriptive analytics), (3) using data mining methods, to develop as a separate direction predictive analytics. In the current situation, the academic (educational) analytics of Russian universities is based on traditional static data processing algorithms.

However, international practice over the past decade has gained considerable experience in applying the results of academic (educational) analytics using multidimensional databases (including retrospective) [1]. It is worth noting that a number of researchers distinguish between academic analytics and data mining. The essence of the differences is that academic analytics is evaluated as a descriptive approach to data processing that allows solving operational problems of managing the educational system, and data mining is aimed at revealing hidden patterns that need to be taken into account when making strategic decisions based on data [2-3].

A significant number of research works in the field of data mining are concentrated on the use of clustering, classification, and visualization methods. One of the promising, actively developing areas is the development of predictive models of students' academic success based on machine learning [4-5]. Predictive models predict the student's educational route long before he graduates. An important role in the formation of such models is played by predictors, that is, key forecasting parameters. Recently, research interest in the development of predictive models based on the analysis of "digital traces" has intensified. quantitative data on the activity of students in computer educational environments [6].

The purpose of our study is to analyze, using machine learning methods and artificial neural networks, which variables affect the academic success of students. The study included three stages:

1. data structuring,
2. substantiation of criteria for assessing academic success,
3. designing a predictive model of academic success based on machine learning.

2. METHODOLOGY

2.1. Used data and methods of their structuring

At the first stage of research, software tools using the Python programming language were developed to structure and aggregate data. The Python programming

language has an extensive set of libraries that allow you to access databases (cx_Oracle, pyodbc), upload and download data to files of different formats (pandas, xml), perform statistical data analysis (scipy, numpy), and also perform intelligent analysis using modern machine learning methods (tensorflow, kers, sklearn). Software modules were created to clear data from incomplete, inaccurate or erroneous information, to select students with similar characteristics, to convert text data to numeric and to encode bare grade points in accordance with a particular cluster. Further, the data were aggregated in such a way that the information on the given parameters, which is contained in the university data warehouse, including retrospective information about the student's performance for the entire period of study, was summarized in a unique record.

As the variables in the study, we used the socio-demographic characteristics of students (gender, age, country of residence, place of birth, family status), their socio-economic status (living conditions while studying at the university), information about the educational background before entering the university (type of graduated institution, the results of the Unified state exam in three disciplines), data on the features of admission to the university (year of admission, type and category of study, educational direction) and the results of academic performance at the university (by year, by semester, academic disciplines).

The general selection of the studied data was formed in accordance with the following restrictions: (1) the data should contain information only about graduates of the Kazan (Volga) Federal University, (2) who received basic higher education (bachelor's degree), (3) accepted to the university according to the results of the Unified State Examination and (4) having data on academic performance at the university for a full four-year study cycle. Based on these limitations, 14,724 students were included in the selection. The selection covered four full time periods of study at the university: 2012 - 2016, 2013 - 2017, 2014 - 2018, 2015 - 2019.

2.2. Academic Success Assessment Methods

The second stage of the study was to determine the criteria for assessing the academic success of students. In the scientific literature, this issue continues to be debatable over a long period of time. Predictors of learning success are complex determinant relationships, so one of the relevant topics for the educational community (Data Mining Research) is the integration of data mining with the social and human sciences (psychology, pedagogy, sociology) [2]. One of the promising areas of research in this area is the study of the relationship between the socio-psychological characteristics of the student, his academic performance and the organizational conditions that the university provides (infrastructure, personnel, educational policy) [7-11]. However, the majority of researchers use the results of students' academic performance as a basic criterion for assessing academic success (by year, by semester, by academic subject, and by area of study).

In our study, the dependent variable of the model is also equivalent to the aggregate indicator of graduate achievement for the entire period of study at the university, normalized to a 100 point grading system (in the study, we did not consider students expelled for any reason).

It is worth noting that our task was not to establish the exact value of each student's performance, but to segment his academic results for the entire period of study with respect to the general aggregated indicator of academic performance at the university. For this purpose, the K-means method was used.

K-Means is an uncontrolled machine learning technique for segmenting a dataset into clusters. K-Means seeks to distribute the data set into separate groups, usually without prior knowledge of where to start. The K-Means algorithm includes:

1. determination of K-centroids (the number of necessary clusters), one for each group,
2. the definition of each data element to a group, the group is determined by the nearest centroid relative to the data element,
3. after distributing all the data into groups, the centroids are recalculated,
4. repeating steps 2 and 3 until either the centroids change during recalculation or the specified number of iterations is reached.

2.3. Design methods for the predictive model of academic success (Machine Learning)

At the third stage, to construct the predictive model, we used machine learning methods from the Scikit-learn library [12] and neural network methods implemented in the Keras library [13] of Python. Scikit-learn is an open source machine learning library that supports learning with and without a teacher. This library also contains various tools for data preprocessing, selection and evaluation of models, and many other useful utilities [12]. Keras is a high-level neural network API that can run on top of Google's low-level TensorFlow [14] library. To select the optimal architecture of neural networks, we used the keras-tuner library, which allows you to configure hyper parameters (number of layers, number of neurons, types of activation functions, etc.) of a neural network in automatic mode.

3. RESEARCH RESULTS

Previously, as a hypothesis, variables were determined (Table 1), which can affect the academic success of students.

In accordance with socio-demographic characteristics, the sample included 70% of girls and 30% of boys in the age group from 18 to 23, unmarried / unmarried, who are citizens of the Russian Federation (the share of foreign students in the sample was less than 6%). 78% of them indicated the city as their place of residence (the cities in the sample were not ranked by population, only administrative status was recorded). 67% graduated from high school, 31% graduated from high schools and lyceums, 2% - vocational schools. 23% of students lived in a dormitory during their studies. Based on the characteristics of university enrollment, the sample was distributed as follows: 97% of students studied in full-time, 50% of them on the terms of concluding a contract (paid), 50% on budgetary financing. In the areas of training: (1) mathematics, engineering, information technology - 14%, (2) medicine - 4.5%, (3) natural science disciplines - 13%, (4) humanitarian disciplines - 34.5%, (5) economics, management, jurisprudence - 34%. For the full four-year training cycles, the sample was distributed almost evenly in the range from 21 to 26%. The exception was the 2013 admission year, which accounted for 30% of the data.

Using the K-Means, five groups of students corresponding to different levels of achievement were differentiated. Students with an average grade point for the entire study period ranging from 86 to 98 points were assigned to cluster 5 as academically successful (22%), students with an average score of 55 to 65 were assigned to cluster 1 as academically unsuccessful (6%). Students in clusters 2-4 were identified as a group with average academic performance. The linear Pearson's correlation coefficient (r) between two variables: the average grade point of each student and his grade level by the K-means method was $r = 0.96$ at the level of statistical significance $p \leq 0.01$. For machine learning, bare grade points were coded according to the established cluster.

Table 1 Description of variables used in the study

Variable	Description
Gender	Male/Female
Age at admission (Yearstud)	17 years and younger; from 18 to 23 years old; 24 years and more
Country of birth (Isrussia)	Citizenship of the Russian Federation / Foreign citizens
Type of training (Typetraining)	Full-time / Part-time
Category (Categoryid)	Budget/Contract
The year of admission to university	2012, 2013, 2014, 2015
Type of university before entering university (Typeofschool)	School, lyceum, gymnasium, university, college / technical school
Birthplace	City, Village
Living conditions during studies at the university	Dormitory, renting an apartment, accommodation with parents
Marital status (Martialstatusid)	Married / Not married
Direction of training (specialization)	In five enlarged areas: (1) mathematics, engineering, it; (2) medicine; (3) chemistry, physics, geography, geology, ecology (natural sciences); (4) philology, pedagogy, psychology, religious studies, cultural studies, journalism, social and political sciences (humanitarian disciplines); (5) economics, management, jurisprudence
Academic success before university	Unified State Exam (USE) in three disciplines
Academic Performance Results	Points for all disciplines provided for by the curriculum of the educational program (for a 100-point grading system)

Of the studied variables that could potentially affect the academic success of students (table 1), the average USE score in three disciplines turned out to be statistically significant at $p \leq 0.01$ (Pearson's linear correlation coefficient was $r = 0.42$). Using the K-means method, five groups of students were determined in accordance with

different levels of academic success (USE) before entering the university: cluster 1 - low results, cluster 5 - high, clusters 2-4 - medium. Further, the frequency of transitions of students from a group of clusters according to the USE results to group clusters by academic performance at the university was established.

The choice of the number of performance levels was due to the need for sufficient data detalization. In addition, five clusters is the optimal number of groups for subsequent management decisions [11].

To design a predictive model of academic success at the third stage of the study, we used the data of students assigned by the K-Means method to the first and fifth cluster in terms of academic performance at the university. In the design of the model, the most popular methods of supervised and unsupervised machine learning were used. As input data, the encoded data presented in Table 1 was used; the value of the column “Results of academic performance” was predicted.

In order to identify the impact of the exam results on the forecasting of academic success of students, classifiers and

neural networks were trained taking into account and without taking into account the column "Educational success before entering university". The dataset was divided into training (80%) and test selection (20%). Prediction accuracy was estimated using the metrics.accuracy_score function of the sklearn library (Table 2).

In addition to the naive Bayesian classifier, all the other machine learning methods used gave quite good forecasting results.

To improve the accuracy of forecasting, a neural network with a three-layer architecture, consisting of an input, hidden and output layer, was used. The output neural layer consisted of only one neuron with the logistic softmax activation function, giving an output of 0 or 1, which means that the student entered the cluster of successful or unsuccessful students. The number of neurons and activation functions of the input and hidden layers were selected automatically based on the keras-tuner library. The architectures of 3 neural networks that provide the highest forecasting accuracy are presented in Tables 3 and 4. The following notation for column names is used in these tables:

Table 2 The accuracy of the forecast of student success with and without taking into account the results of the USE

Machine learning method	Forecast accuracy	
	USE taken into account	USE not taken into account
Support vector method (kernel-radial basis function)	86.9	82.5
Support vector method (kernel polynomial)	86.4	81.9
Gradient boosting decision tree algorithm (XGBoost)	86.1	82.4
Logistic Regression	86.7	83.1
Random forest algorithm	85.8	82.4
Decision Tree	83.0	81.2
K-Neighbors Classifier, k=10	86.1	81.7
Gaussian Naive Bayes	70.2	56.2

1. Input layer activation function is f_input
2. Number of neurons in the input layer is n_input
3. Hidden layer activation function is f_hidden
4. The number of neurons in the hidden layer is n_hidden

The following notation is used for activation functions: Rectified Linear Unit-relu, Sigmoid, Scaled Exponential Linear Units-selu, Exponential Linear Unit-elu and hyperbolic tangent-tanh.

Table 4 Architectures of the 3 best neural networks for predicting success taking into account USE

f_input	n_input	f_hidden	n_hidden	Forecast accuracy (%)
relu	20	sigmoid	90	90.1
relu	40	tanh	80	90.1
relu	70	sigmoid	80	90.09

Table 5 Architectures of the 3 best neural networks for predicting success without taking into account USE

f_input	n_input	f_hidden	n_hidden	Forecast accuracy (%)
selu	100	selu	10	86.01
selu	95	elu	45	85.7
elu	40	relu	70	85.6

4. DISCUSSION OF RESULTS

Clusterization of performance levels allowed us to restore the trajectory of each student moving from cluster to cluster depending on the year of university entry, the type and category of education, and the direction of training. In figure 1, transitions of a student contractor who studied at the university from 2015 to 2019 in the integrated direction of mathematics, engineering, it (the Y axis - clusters of achievement, the X axis - semesters) are

recorded. The graph shows that the student's academic success was higher in the first and third years, in the second year, academic performance dropped sharply.

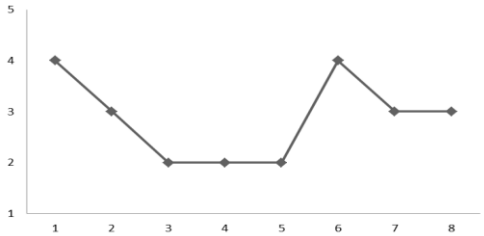

Figure 1 Graph of student-contractor transitions by clusters

The study of the frequency of student transitions between two groups of clusters - academic success before entering the university (USE) and academic performance at the university - showed that (1) only 26% of students had high USE results at the university entrance, (2) a significant proportion of students with low results of the Unified State Examination during the study at the university got the opportunity to take higher positions in academic performance. This suggests that the USE is a significant variable, but not the only one (table 5).

Table 5 Transfer of students between clusters for the entire period of study at the university (% of students)

Clusters according to the USE	University performance clusters				
	1	2	3	4	5
1	18.40	31.91	29.07	15.74	4.88
2	9.61	27.66	29.84	24.51	8.38
3	3.83	18.95	27.78	31.56	17.88
4	2.47	10.83	20.78	33.25	32.66
5	1.44	5.89	12.72	32.48	47.47

The second conclusion is that students who have shown good results at the level of secondary general education can potentially be at risk.

The results show that the use of artificial neural network methods allows to achieve forecasting accuracy of 90 percent (including the USE) or 86 percent (excluding the USE).

Thus, the use of machine learning methods and neural networks confirms the hypothesis that, along with the results of academic success before the university (USE), additional variables must be taken into account as predictors of academic success.

5. CONCLUSION

Currently, Russian universities have developed long-term digital strategies that require significant financial investments in information systems, which are the basic component of the digital university model and allow universities to cope with the modern challenges of higher education. However, the data that are not always used, as well as the tools for collecting, summarizing and analyzing them, provide the expected benefits and results. In turn, the digitalization of management entails a certain digital strategy for organizing the educational process, without which it is impossible to implement a predictive (as opposed to situational) approach in making managerial decisions. One of the goals of our work is to design, based on the results, an easy-to-use software module that will allow management decision-makers to independently analyze data and identify potential "risk" groups among students.

The results of our study showed that machine learning methods have good potential for predicting student performance. The development of this area of academic (educational) analytics will contribute to the development of differentiated actions aimed at different groups of students in accordance with their potential, as well as a more efficient allocation of resources of institutions.

ACKNOWLEDGMENT

The study (all theoretical and empirical tasks of the research presented in this paper) was supported by a grant from the Russian Science Foundation (Project No. 19-18-00253, "Neural network psychometric model of cognitive-behavioral predictors of life activity of a person on the basis of social networks").

REFERENCES

[1] H. Aldowah, H. Al-Samarraie, W. M. Fauzy, Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics, 37, 2019, pp.13–49. DOI: https://doi.org/10.1016/j.tele.2019.01.007

[2] Romero S.V. Cristobal «Data mining in education» WIREs Data Mining Knowl Discovery. 2013. T. 3. № 1. pp. 12–27. DOI: https://doi.org/10.1002/widm.1075

[3] P. Baepler, CJ.Murdoch, Academic analytics and data mining in higher education. Int J Schoolarship Teach Learn 2010, 4:1–9. DOI https://doi.org/10.20429/ijsotl.2010.040217

[4] A. Hellas, P. Ihantola, A. Petersen, V. V.Ajanovski, M. Gutica, T. Hynninen, S. N. Liao, Predicting Academic Performance: A Systematic Literature Review. In Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, New York, USA: ACM, 2018, pp. 175-199. DOI: https://doi.org/10.1145/3293881.3295783

[5] K. Y. Diaz Pedroza, B. Y. Chindoy Chasoy, A. A. Rosado Gómez, Review of techniques, tools, algorithms and attributes for data mining used in student desertion. Journal of Physics: Conference Series, Volume 1409, Issue 1, article id. 012003 (2019). DOI: https://doi:10.1088/1742-6596/1409/1/012003.

[6] M.M. Tamada, J.F. Netto, D.P. Lima, Predicting and Reducing Dropout in Virtual Learning using Machine Learning Techniques: A Systematic Review. 2019 IEEE Frontiers in Education Conference (FIE), pp. 1-9. DOI: 10.1109/FIE43999.2019.9028545

[7] F.Araque, C. Roldán, A. Salguero, Factors influencing university dropout rates. Computers & Education, 53 (3), 2009, pp. 563–574. DOI: https://doi:10.1016/j.compedu.2009.03.013

[8] G. Gray, C. McGuinness, P. Owende, M. Hofmann, Learning Factor Models of Students at Risk of Failing in the Early Stage of Tertiary Education. Journal of Learning Analytics, 3(2), 2016, pp. 330-372. DOI: https://doi.org/10.18608/jla.2016.32.20

[9] R. Asif, A. Merceron, S.A. Ali, N.G. Haider, Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113(1), 2017, pp. 177-194. https://doi.org/10.1016/j.compedu.2017.05.007

[10] S. Lonn, B. Koester, Rearchitecting Data for Researchers: A Collaborative Model for Enabling Institutional Learning Analytics in Higher Education.Journal of Learning Analytics, 6(2), 2019, pp. 107–119. DOI: https://doi.org/10.18608/jla.2019.62.8

[11] V. Migueis, A. Freitas, P. J.Garciab, A.Silva, Early segmentation of students according to theiracademic performance: A predictive modelling approach. Decision Support Systems, Volume 115, November 2018, Pages 36-51 DOI: https://doi.org/10.1016/j.dss.2018.09.001

[12] F. Pedregosa, G. Varoquaux, A.Gramfort, V. Michel , B.Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, Volume 12, 2011, Pages. 2825–2830.

[13] Chollet, François (2015). Keras. Available at: https://keras.io

[14] Abadi, Martin et al. (2016) Tensorflow: A system for large-scale machine learning. In 12th Symposium on Operating Systems Design and Implementation, USENIX Association, pp. 265–283.