Milica Gerasimović
Advisor/Coordinator
Institute for Improvement of Education,
Belgrade

Uglješa Bugarić
Professor
University of Belgrade Faculty of
Mechanical
Engineering

Marija Božić
MScEE
PhD Program in Environmental Education,
Autonomous University of Barcelona
Spain

Enrollment Management: Development of Prediction Model Based on Logistic Regression

This paper presents the development of a model for forecasting and decision-making by applying logistic regression. The prediction of professional choices for graduates has been verified on the sample of 159 graduates. Predictor variables are grouped in nine input variables and the data collected from the Unique Education Information System database. Professional choices of graduates after finishing vocational school are grouped into three output variables and the data collected in the survey. The obtained results show that the application of logistic regression can predict the number and structure of students enrolled in higher education.

Keywords: *prediction, decision-support systems, models, enrollment management, graduate's professional choice, logistic regression*

1. INTRODUCTION

Higher education has an important role in economic and social development, and therefore all the changes and reform processes in that domain should lead to the improvement of socioeconomic climate in one country. Higher education reform underlines the need for changing and specifying the role of universities and faculties in the enrollment management process. Previous reform activities of the world's leading universities and faculties (Aberdeen, 2003), as well as of the Serbian higher education system, indicate the importance of the reform of the management process as a key factor in improving both the quality of education and the scientific research in universities (Gerasimović, 2012). In that context, the existence of the Unique Education Information System (EIS) under the jurisdiction of the Ministry of Education and Science of the Republic of Serbia, is of great importance in improving the process of enrollment management in higher education institutions in Serbia.

The constant development, implementation and improvement of the EIS provides a more effective resource planning, resource management and better monitoring of the activities in the educational system on all its management levels. The EIS enables the data exchange from schools to municipalities, from schools to Regional School Ministry Administration, as well as from schools administration to the central level of the Ministry. Furthermore, the maximum possible data transfer automatization has been provided in the opposite direction, from the central level of the Ministry to the schools.

The implementation of the EIS increases efficiency and the effectiveness of the education system by managing the following activities: planning schools'

network, planning necessary teaching and non-teaching staff, financial planning, monitoring and controlling students' achievements and schools' performances.

The EIS is designed through the following program modules:

- Students' database
- School employees' database
- Inventory and schools premises database
- Financial module

Analytical software for simulation of the implementation of different education funding formulas (version available in the Ministry).

In the focus of the research presented in this paper is the "students' database", which is as an integral program module of the EIS. During the analysis of the EIS database, it has been noticed that it provides an important quantity of information regarding statistic students data, and furthermore, information about the students performance in different courses during their education (quarters' reports, school year end reports), as well as the parents information (education level, forms of employment, etc.). In order to select the required information from the mentioned EIS database, it is essential to determine the key factors of the graduate students' professional choice. Based on those factors a research on the sample of graduate student population has been conducted in order to gather baseline data needed to develop the predictive model. Implementation of this model of forecasting and decision making will practically mean to use the input data from the EIS database and generate the output data that can help higher education institutions predict the number of enrolled students and make enrollment decisions.

This paper proposes a model for forecasting and decision making based on logistic regression and it represents an integral part of the support system in decision making (Figure 1).

The support system enables the accomplishment of the objectives defined by the strategic orientation of

higher education institutions. The software support provides the efficiency of decision making, ease of use and adaptability.
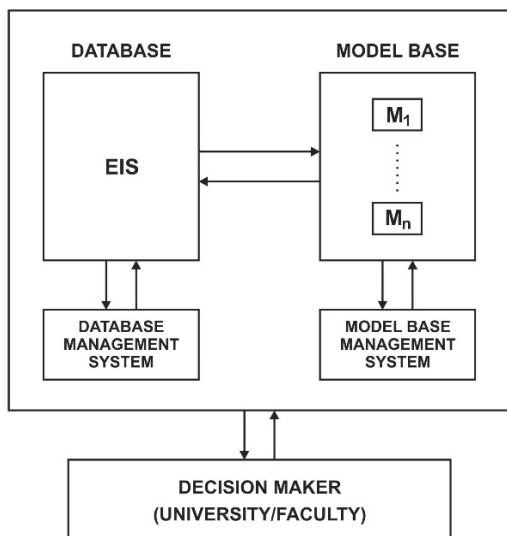

Figure 1. Support system for decision making process

2. LITERATURE REVIEW

Forecasting is predicting the outcome of an event in the future. It is used in predicting future event consequences, future event occurrence time, but also time series like future values of periodic data series in certain point of time.

Forecasting can be used on time series for predicting the value of the periodic data series over a period of time. There are various estimation and forecasting methods. The statistical forecasting methods are widely used especially the method of regression analysis.

The characteristic of the method of regression analysis is that the dependent variable, namely the forecasting value, is expressed as a mathematical function of one or more variables, predictive values, known in the time of forecasting (Hillier & Lieberman, 2001). There are a great number of articles, monographs and books in the field of operational research that deal with predicting and decision making in various domains. Among them are those that refer to higher education and predicting the professional choice of graduate students, (Gerasimovic et al., 2011; Miljkovic et al., 2011), selection and retention of students (Hossler, 1984), the role of financial support for students (Spaulding & Olswang, 2005), enrollment management strategies (Antons & Maltz, 2006; Bilous, 2009), prediction student retention (Duke, 2006), the impact of the tuition fees on student enrollment decision (Edward, 1990), etc. Different methodologies are applied on those researches. However, a review and systematization of the existing papers, research operations models and methods in the field of prediction and decision making in the process of enrollment management, indicates that developing different forecasting, optimization and decision making models, among other is prepared by application of regression analysis, i.e. logistic regression (Chang, 2006; Edward, 1990).

3. METHODOLOGY

The aim of this paper is to develop a model for predicting the number and structure of students enrolled in higher education institutions of technical sciences by use of regression analysis, more precisely logistic regression, as a part of methodology which should be used in process of enrollment management in higher education institutions in Serbia.

3.1 Logistic Regression

Logistic regression is a statistical method used to test models for prediction of categorical outcomes (dependent variable size (y)), with two or more categories, while the predictor i.e. the independent variables (x) are categorical (discrete), continuous, or a combination of both in the same model (Moore & McCabe, 2004).

One of the mathematical functions that best describe the connection between the independent variables and the probability of finding the variable in a particular category is represented by the following formula (Moore & McCabe, 2004):

$$\ln\left(\frac{p}{1-p}\right) = b_0 + bx \qquad (1)$$

Where: p – outcome probability of dependent variable y, b_0, b – logistic regression weights, x – value of the independent variable.

The odds ratio of the possibility or impossibility of occurrence of a certain event usually is presented as (Moore & McCabe, 2004):

$$odds = \frac{p}{1-p} \qquad (2)$$

The logistic regression is used for statistical modeling of the categorical dependent variable. Binary data are the most common form of categorical data.

The dependent variable takes values 0 and 1, depending on whether the observed event occurred or not. If the logistic model is consisted of the combination of continuous and categorical predictor values, which is the case analyzed in this paper, the functional dependence between the occurrence probability of the observed event and the mentioned independent variables is given in the form of:

$$\ln it(P) = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{j=k+1}^{n} b_j \cdot \sum_{i=1}^{m} c_{j,i} \qquad (3)$$

The coefficients b_j are derived from the categorical independent variables which have a total of n-k+1, the factors $c_{j,i}$ have values 0 or 1, and their number (m) depends on the number of categories for each variable.

3.2 The sample description

In order to develop the prediction model, a research has been conducted in two vocational schools (VET schools) in Belgrade on the sample of 159 students. The

Table 1. Input and output variables

No	Variables	Mark	Code and frequencies
	Input variables (EIS database)		
1	Gender	G	1- male (88.24%) 2- female (11.76%)
2	Second grade success	S2	2- sufficient (9.24%) 3- good (47.06%) 4- very good (31.94%) 5- excellent (11.76%)
3	Third grade success	S3	2- sufficient (8.40%) 3- good (45.38%) 4- very good (31.09%) 5- excellent (15.13%)
4	Fourth grade success	S4	2- sufficient (7.57%) 3- good (42.69%) 4- very good (24.29%) 5- excellent (25.45%)
5	Mother's education level	ME	1- primary school (3.36%) 2- high school (5.04%) 3- VET school (61.34%) 4- higher school(15.13%) 5- university (11.76%) 6- no answer (3.36%)
6	Father's education level	FE	1- primary school (21.85%) 2- high school (63.87%) 3- VET school (6.72%) 4- higher school (7.56%) 5- university (0.00%) 6- no answer (0.00%)
7	Mother's work status	MW	1- unemployed (21.85%) 2- full-time employed (63.87%) 3- occasionally employed (6.72%) 4- no answer (7.56%)
8	Father's work status	FW	1- unemployed (9.24%) 2- full-time employed (67.23%) 3- occasionally employed (13.45%) 4- no answer (10.08%)
9	Parents schooling financial support	FS	1- yes (67.23%) 2- no (15.97%) 3- no answer (16.81%)
	Output variables (survey)		
1	Job	J	1- yes (29%) 0- no (71%)
2	Continuation of schooling at FME in Belgrade	FME	1- yes (31%) 0- no (69%)
3	Continuation of schooling at some other faculty	OF	1- yes (40%) 0- no (60%)

research was conducted targeting the higher education institutions of engineering, more specifically, the Faculty of Mechanical Engineering in Belgrade (FME). The data collected from the EIS database of the 159 students is used as starting values (nine input variables). The data collected in the survey on the same sample of students (experiment) is used as control values (three output variables) in the developing phase of the predictive models. The encoding mode of input and output variables and their descriptive statistics are presented in Table 1.

Using the input data from the EIS database and the results obtained by the experiment, a model has been developed based on the logistic regression, which can provide support in the decision making in the process of managing and forecasting the enrollment in higher education institutions of technical sciences.

Before performing the regression analysis the validity check of the assumptions of logistic regression was performed.

Table 2. Collinearity analysis

Input variables	TOL	VIF
Gender (G)	0.806	1.241
Second grade success (S2)	0.406	2.466
Third grade success (S3)	0.301	3.321
Fourth grade success (S4)	0.506	1.975
Mother's education level (ME)	0.848	1.179
Father's education level (FE)	0.852	1.173
Mother's work status (MW)	0.856	1.169
Father's work status (FW)	0.914	1.094
Parents schooling financial support (FS)	0.864	1.157

Correlation analysis proved the non existence of the assumption of multi-collinearity. The results given in table 2 show that the values of the Tolerance (TOL) coefficient for each independent variable are larger than the critical value of 0.20, therefore, the assumption of no multi–collinearity is not undermined. This conclusion is supported by the VIF coefficient value that ranges from 1,094 to 3,321, which is below the critical value of 5.

4. RESULTS AND DISCUSSION

The tested model of professional choices prediction was carried out on the sample of 159 graduates. The factors influencing the enrollment on FME are grouped into 9 predictors. Three of them are continuous and six are categorical independent variables. The categorical independent variables are the gender, the employment status of the mother, the employment status of the father, the schooling financial support, the level of education of the mother and the level of education of the father which can be found in EIS. The encoding of these variables is shown in Table 3.

Gender is a dichotomous variable. The female gender is coded as 1, while the male is 0. Other categorical variables have three or five categories (Table 3), and one of them is always a reference while the rest are valuated based on it. In that way the probability of the observed event will equal one for each categorical size. For example, the categorical variable of the employment status of the father is coded as following: unemployed (1) employed (2), while the reference group is coded as - other.

The continuous independent variables (the success in second grade (x_1), the success in the third grade (x_2), and success in fourth grade (x_3)) were measured so that higher numbers indicate better results.

Table 3. Coding of categorical variables

Variable	Category	Frequency	Coding (1)	(2)	(3)	(4)	(5)
Level of education of the father (FE)	Primary school	7	1	0	0	0	0
	High school	3	0	1	0	0	0
	Vocational school	88	0	0	1	0	0
	Higher school	26	0	0	0	1	0
	University	31	0	0	0	0	1
	Other	4	0	0	0	0	0
Level of education of the mother (ME)	Primary school	5	1	0	0	0	0
	High school	7	0	1	0	0	0
	Vocational school	98	0	0	1	0	0
	Higher school	27	0	0	0	1	0
	University	18	0	0	0	0	1
	Other	4	0	0	0	0	0
Employment status of the mother (MW)	Unemployed	36	1	0	0	0	0
	Employed	112	0	1	0	0	0
	Other	11	0	0	0	0	0
Employment status of the father (FW)	Unemployed	16	1	0	0	0	0
	Employed	126	0	1	0	0	0
	Other	17	0	0	0	0	0
Parents schooling financial support (FS)	No	25	1	0	0	0	0
	Yes	106	0	1	0	0	0
	Other	28	0	0	0	0	0
Gender (G)	Female	20	1	0	0	0	0
	Male	139	0	0	0	0	0

Table 4. Classification

Observed		Predicting Faculty of Mechanical Engineering (FME)		Exact mach (%)
		No	Yes	
Faculty of Mechanical Engineering (FME)	No	108	7	93.9
	Yes	16	28	63.6

Table 5. Predicting the probability of enrollment on the FME

Variable	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
(G)_1	-1.897	1.078	3.098	1	0.078	0.150	0.018	1.240
(S2)	0.060	0.533	0.013	1	0.910	1.062	0.374	3.019
(S3)	0.193	0.646	0.090	1	0.765	1.213	0.342	4.304
(S4)	1.641	0.429	1.,632	1	0.000	5.159	2.226	11.957
(ME)			3.224	5	0.666			
(ME)_1	-22.927	12911.176	0.000	1	0.999	0.000	0.000	.
(ME)_2	-24.916	12911.176	0.000	1	0.998	0.000	0.000	.
(ME)_3	-24.411	12911.176	0.000	1	0.998	0.000	0.000	.
(ME)_4	-23.648	12911.176	0.000	1	0.999	0.000	0.000	.
(ME)_5	-23.220	12911.176	0,000	1	0.999	0.000	0.000	.
(MW)			0,016	2	0.992			
(MW)_1	0.070	1.242	0,003	1	0.955	1.073	0.094	12.239
(MW)_2	-0,012	1.218	0,000	1	0.992	0.988	0.091	10,747
(FW)			1,885	2	0.390			
(FW)_1	-1.955	1.424	1,884	1	0.170	0.1142	0.009	2.308
(FW)_2	-0.884	0.976	0,820	1	0.365	0.413	0.061	2.799
(FS)			0,661	2	0.719			
(FS)_1	0.813	1.008	0,651	1	0.420	2.255	0.313	16.278
(FS)_2	0.311	0.719	0,188	1	0.665	1.365	0.334	5.586
(FE)			5,970	5	0.309			
(FE)_1	39.706	20043.248	0,000	1	0.998	1.755E17	0.000	.
(FE)_2	21.191	29682.089	0,000	1	0.999	1.597E9	0.000	.
(FE)_3	41.737	20043.247	0,000	1	0.998	1.337E18	0.000	.
(FE)_4	42.223	20043.247	0,000	1	0.998	2.175E18	0.000	.
(FE)_5	39.700	20043.247	0,000	1	0.998	1.744E17	0.000	.
Invariable	-23.629	15330.800	0,000	1	0.999	0.000		

The dependent variable - further education at FME is dichotomous and can only have two values: Yes or No, which are coded as following: 1-Yes, 0-No.

The statistical analysis is performed using the SPSS 18.0 software.

4.1 Model testing

The goodness of fit of the model has been performed by using the Hosmer-Lemeshow test for logistic regression.

In case of this model results of Hosmer-Lemeshow test are: χ^2= 7,990 with statistical significance of 0,434, which means that the given model is fitting.

In further analysis, it is determined how accurately the model predicts the category of *intends to enroll on FME* (Table 4).

Valid data for further analysis is accuracy of predicting the enrollment on FME, i.e. the number of students who will opt to enroll. The obtained results lead to conclusion that the examined regression model correctly classifies 63,6% of the cases of enrollment on the observed Faculty. That means the developed regression model correctly predicts 136 of 159 cases in sample, i.e. 28 of 44 students which decided to enroll FME.

The contribution or the importance of each predictor variable in the model is shown in Table 5. It was established that only one independent variable gave a statistically significant contribution to the predictive capabilities of the model (Sig<0,05). It is the predictor: *Success in the fourth year* whose odds ratio Exp(B),

$\dfrac{p}{1-p}$ is 5,159 (Exp(B)=5,159). In other words, the

improvement in performance in the final year for one unit (sufficient (2), good (3), very good (4), excellent (5)) increases the probability of enrollment in FME about 5 times.

According to data shown in Table 5, it is possible to determine an analytical expression for calculating the probability that the observed case belongs in the certain category.

The logistic regression equation (3) of the analyzed model is:

$$\ln it(P) = -23,629 + 0,060 \cdot x_1 + 0,193 \cdot x_2$$
$$+1,641 \cdot x_3 - 1,897 c_{4,1} - 22,927 \cdot c_{5,1}$$
$$-24,916 \cdot c_{5,2} - 24,411 \cdot c_{5,3} - 23,648 \cdot c_{5,4}$$
$$-23,220 \cdot c_{5,5} + 0,070 \cdot c_{6,1} - 0,012 \cdot c_{6,2} \qquad (4)$$
$$-1,955 \cdot c_{7,1} - 0,884 \cdot c_{7,2} + 0,813 \cdot c_{8,1}$$
$$+0,311 \cdot c_{8,2} + 39,706 \cdot c_{9,1} + 21,191 \cdot c_{9,2}$$
$$+41,737 \cdot c_{9,3} + 42,233 \cdot c_{9,4} + 39,700 \cdot c_{9,5}$$

Where:

$c_{4,1}$ = 1 for female sex, in other cases 0,

$c_{5,1}$ = 1 for "mother–primary school", in other cases 0,

$c_{5,2}$ = 1 for "mother–high school", in other cases 0,

$c_{5,3}$ = 1 for "mother–vocation. school",in other cases 0,

$c_{5,4}$ = 1 for "mother–higher school", in other cases 0,

$c_{5,5}$ = 1 for "mother–university", in other cases 0,

$c_{6,1}$ = 1 for "mother–unemployed", in other cases 0,

$c_{6,2}$ = 1 for "mother–employed", in other cases 0,

$c_{7,1}$ = 1 for "father–unemployed", in other cases 0,

$c_{7,2}$ = 1 for "father–employed", in other cases 0,

$c_{8,1}$ = 1 for "no financial support", in other cases 0,

$c_{8,2}$ = 1 for "financial support", in other cases 0,

$c_{9,1}$ = 1 for "father–primary school", in other cases 0,

$c_{9,2}$ = 1 for "father–high school", in other cases 0,

$c_{9,3}$ = 1 for "father–vocation. school", in other cases 0,

$c_{9,4}$ = 1 for "father – higher school", in other cases 0,

$c_{9,5}$ = 1 for "father–university", in other cases 0.

The coefficients b_0, b_i, b_j, from the logistic regression equations correspond to the values of the coefficients B (Table 5) and determine the strength and direction of the independent variable and the probability of the analyzed event. For example, the variable sex of pupils has a negative coefficient B (= −1,897). That is, the female students rarely choose to enroll on FME. More specifically, the data Exp(B)=$p/(1–p)$ = 0,150, (Table 5) shows that girls opt to enroll on FME 13 times less than the boys.

5. CONCLUSIONS

The aim of this study was to develop a predictive and decision-making model using logistic regression, which is confirmed on the example of the enrollment management in higher education institutions. For reasons of gathering the relevant input data (professional orientation of the graduates), the survey was conducted in two VET schools in Belgrade, on a sample of 159 graduates. The survey was conducted targeting the higher education institutions of engineering, more specifically, the Faculty of Mechanical Engineering. The data collected in the survey (experiment) was used in the development and verification of the predictive model, as baseline (nine input variables) and as control values (three output variables). The processing and the systematization of data were performed using the regression analysis – the logistic regression.

The model testing, i.e. determining how well the logistic model predicts the results was conducted using the Pearson's χ^2 test. The analysis showed that the entire model with all predictors is statistically significant, χ^2(20, N=159)=81,772, p<0,001, which means that the model distinguishes respondents who intend to enroll in the Faculty of Mechanical Engineering in Belgrade from those who do not.

The contribution of the model is confirmed by classification rate. The results show that the tested model correctly classifies 63.6% of the cases of *enrollment on Faculty of Mechanical Engineering with experimentally obtained results (the survey)*.

The model proposed in this paper is just one of the models that can be found among other models with different performances and prediction techniques used to solve the problem of enrollment management in faculties/universities of different scientific fields (technical, social, etc.), legal status (state, private), size, etc.

It can be concluded that the proposed model can provide decision support in the process of enrollment management in higher education institutions. Also, by experimental verification of the developed model for prediction of enrollment, it can be concluded that Unique Educational Information System (EIS) should be used as a support in the process of enrollment management in higher education institutions.

REFERENCES

[1] Antons, C. M., Maltz, E. N., (2006). Expanding the Role of Institutional Research at Small Private Universities: A Case Study in Enrollment Management Using Data Mining. *New directions for institutional research*, 2006 (131), 69-81.

[2] Bilous, O. *Enrollment management of higher education: admission offices and recruitment activities. A case study of Montclair State University NJ.* USA, 2009.

[3] Chang, L., (2006). Applying Data Mining to Predict College Admissions Yield: A Case Study. *New Directions for Institutional Research*, 2006 (131), 53-68.

[4] Edward, P. St. J., (1990). Price response in enrollment decisions: An Analysis of the High School and Beyond Sophomore Cohort. *Research in Higher Education*, 31(2), 161-176.

[5] Gerasimovic, M. *Contribution to development of methodology for prediction and decision-making by applying artificial neural networks.* Faculty of Mechanical Engineering Belgrade University, 2012.

[6] Gerasimovic, M., Stanojevic, Lj., Bugaric, U., Miljkovic, Z., Veljovic, A., (2011). Using artificial neural networks for predictive modeling of graduate's professional choice. *The New Educational Review*, 23 (1), 175-189.

[7] Herzog, S., (2006). Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks Vis-à-Vis Regression. *New directions for institutional research*, 2006 (131), 17-33.

[8] Hillier, F.S., Lieberman, G.J. *Introduction to Operations Research.* McGraw Hill, 2001.

[9] Hossler, D. *Enrollment Management: An Integrated Approach.* College Board Publications, New York, 1984.

[10] Miljkovic, Z., Gerasimovic M., Stanojevic, Lj., Bugaric, U., (2011). Using artificial neural networks to predict professional movements of gra–duates. *Croatian Journal of Education,* 13(3), 117-141.

[11] Spaulding, R., Olswang, S., (2005). Maximizing Enrollment Yield Through Financial Aid Packaging Policies. *Nasfaa journal of student financial aid*, 35 (1), 27-39.

[12] SPSS 18.0 http://en.softonic.com/s/spss-18-down–load Accessed 20 January 2011.

[13] The "Robert Gordon" University of Aberdeen: *Principles of Teaching, Learning and Assessment,*

Module 1, Centre for Enhancement of Learning and Teaching, Aberdeen, 2003.

[14] Moore, D.S., McCabe, G.P. Introduction to the practice of statistic. WH Freeman & Co.: New York, 2004.

УПРАВЉАЊЕ И ПРЕДВИЂАЊЕ УПИСА: РАЗВОЈ МОДЕЛА ПРЕДВИЂАЊА ЗАСНОВАНОГ НА ЛОГИСТИЧКОЈ РЕГРЕСИЈИ

Милица Герасимовић, Угљеша Бугарић, Марија Божић

У раду је приказан развој модела предвиђања заснованог на логистичкој регресији. Предвиђање професионалног опредељења матураната верифи–ковано је на узорку од 159 ученика. Предикторске величине су груписане у 9 улазних променљивих, а потребни подаци су добијени из базе Јединственог информационог система у образовању. Професионално опредељење матураната по завршетку средње школе је груписано у три излазне променљиве, а вредности су добијене експери–менталним истраживањем. Добијени резултати показују да се применом логистичке регресије може предвидети број и структура ученика који се уписују на факултете.