Anna Kijanka Ph.D.
The School of Banking and Management in Krakow
a.kijanka@wszib.edu.pl
Tomasz Molęcki
The School of Banking and Management in Krakow
t.molecki@wszib.edu.pl

LOGIT MODEL TO PREDICT THE NUMBER OF STUDENTS ENROLLING IN A SECOND-CYCLE PROGRAMME AT THEIR HOME NON-PUBLIC SCHOOL OF HIGHER EDUCATION

Introduction

The basic source of revenue for non-public schools of higher education are fees, mainly in the form of tuition fees paid by students. It is a major challenge for the management to ensure an adequate level of income so that the operating costs are covered and a positive financial result is generated. Therefore, it is highly desirable that a sufficient number of candidates should apply for studies, which will ensure an adequate level of the school finance. It seems to be of crucial importance that the school management is aware of the future career paths plans of the graduates of the first-cycle degree studies as this knowledge will make it possible to take actions resulting in the largest possible number of applicants willing to undertake Master's degree studies (hereinafter referred to as SUM) at their home school.

The query of the literature on the subject indicates some interest in this issue but the number of accessible publications is insignificant[1]. The authors of the existing articles focus mainly on the issues of quality determinants[2] and education costs[3].

The aim of this study is to estimate the probability of the continuation of education (SUM) at the same school of higher education, and thus to approximate the number of candidates for the second-cycle studies in the group of the current graduates of the first-cycle degree studies.

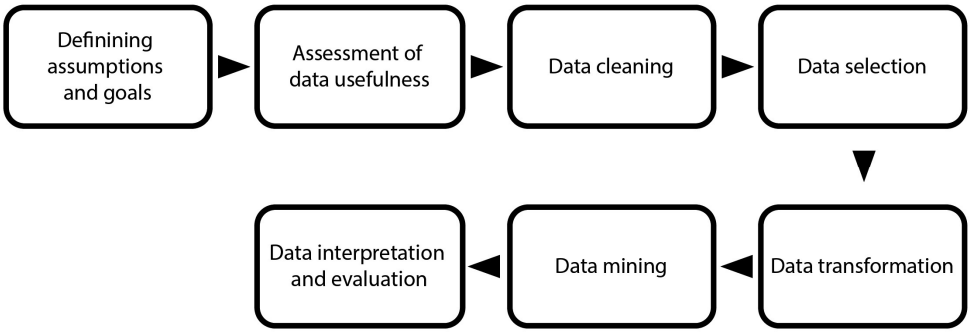
[1] A. Nandeshwar S. Chaudhari *Enrollment Prediction Models Using Data Mining,* http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf (accessed: 22.05.2022).
[2] G. Zieliński, K. Lewandowski, *Determinanty percepcji jakości usług edukacyjnych w perspektywie grup interesariuszy*, http://zif.wzr.pl/pim/2012_3_3_4.pdf (accessed: 25.05.2022).
[3] I. Gawryś, P. Trippner, *Analiza poziomu rentowności przedsiębiorstwa na przykładzie niepublicznej uczelni wyższej w roku akademickim 2015/2016*, „Annales Universitatis Mariae Curie-Skłodowska", 2017, No. LI, 5.

1. Knowledge discovery

The research data was obtained from the management system of one of the Krakow non-public schools of higher education (hereinafter referred to as the School). The choice of the data was dictated mainly by their availability. It also depended on the authors' subjective opinion on its usefulness in achieving the purpose of the research.

The determination of the information (the variables) that was applied in the analyses required numerous actions such as, in particular, data cleaning (correction of mistakes, supplementing the missing data) and data transformation (encoding or deleting student's personal details, data enrichment by calculating new attributes and aggregation). The above listed actions were an introduction to the data *mining process* (Fig.1) which makes it possible to discover associations, hidden patterns, rules and trends in the data and, as a result, to discover knowledge[4].

Figure 1. Steps in knowledge discovery process



Source: Authors' own study based on: M. Szeliga, *Data Science i uczenie maszynowe*, Wydawnictwo Naukowe PWN, Warszawa 2017, pp. 2-3.

A detailed presentation of variables subject to analysis is given in Table 1.

Dependent variable „SUM" is an encoded zero-one information specifying whether the first-cycle graduate undertook Master's degree studies at the School. The following explanatory variables were distinguished: Gender, Graduation-year, Mode of studies, Grade point average, Scholarships received, Age – graduation, Distance. For the sake of the requirements of the software that was used in the modelling process, the variables were simplified to the following forms: *gender*, *graduation_year*, *mode*, *average*, *scholarship*, *age*, *km* respectively.

[4] T. Morzy, *Eksploracja danych*, Wydawnictwo Naukowe PWN, Warszawa 2013, pp. 3-9.

Table 1. Variables to be mined

Specification	Description
SUM (*sum*)	Quality variable indicating whether the student continued his/her education and enrolled in the Master's degree studies (SUM). Information encoded as 0 means *no continuation* while 1 indicates *continuation*.
Gender (*gender*)	Quality variable describing gender: 0 – male, 1 - female.
Graduation year (graduation_ year)	The year of the undergraduate thesis defense (quantitative variable)
Mode of study (*mode*)	Quality variable describing the mode of studies encoded as 0 for full-time study and 1for part-time study.
Grade point average (*average*)	Grade point average for studies in a scale of 2 to 5, calculated on the basis of semester grades.
Scholarship (scholarship)	Information about received scholarships. Quality variable encoded as 0 (no scholarships) or 1 (scholarship granted).
Graduate's age (*age*)	Enum variable representing student's age in the year of the thesis defense. The calculation of the variable is based on the thesis defense year and the student's date of birth.
Distance (*km*)	Qualitative, enum variable including the distance from the student's place of residence to Kraków (in kilometers)

Source: Auhtors' own research.

In order to achieve the aim of the research, a data mining method was used with a logistic regression model. The choice of the research method was dictated by the qualitative character of the dependent variable. It was expected that the constructed model would make it possible to indicate factors that determine the decisions of the first-cycle graduates to continue education in master's degree studies at their home school.

Four econometric models were formed for particular fields of study that are available at the investigated School. The article presents a model developed on the basis of the personal details of students and graduates of the Management field of study. The data concerned graduates who received their diplomas in 2013-2018.

2. The essence of the logistic regression model

The logistic regression model, also referred to as a logit model, makes it possible to determine the probability of an object belonging to one of the two classes, depending on the vector of independent variables that characterizes the object[5]. The binary, qualitative dependent variable is characteristic for the logistic regression. The dependent variable most often represents the

[5] T. Kufel, *Ekonometria. Rozwiązywanie problemów z wykorzystaniem programu GRETL,* Wydawnictwo Naukowe PWN, Warszawa 2011, p. 142.

occurrence or the lack of occurrence of the event to be predicted, e.g. company's bankruptcy (1 – yes, 0 – no) or undertaking education at a given level (1 – yes, 0 – no). Due to the fact that the linear regression model, when additional conditions are not imposed, is not able to ensure that the value of the dependent variable always falls within the range corresponding to the possible probabilities, the following model is used to model the probability:

$$P = P(Y = 1/X_1, X_2, \ldots, X_k) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k)}} \qquad (1)$$

where:

- Y is a binary dependent variable,
- $X_1, X_2, \ldots, X_k$ are k explanatory (or independent) variables that can be measurable or qualitative,
- P stands for the conditional probability of success (of the object belonging to category coded as 1),
- α_0, α_1, …, α_k are model structural parameters (coefficients) [6].

Following a simple transformation of relationship (1), an equivalent form of the logistic regression model is obtained:

$$\frac{P}{1 - P} = e^{\alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k} \qquad (2)$$

where the left side of the equation indicates the probability of success, i.e. the ratio of the probability of success to the probability of failure.

After taking the logarithm of both sides of equation (2), the third, equivalent logit form of the model is obtained:

$$logit\ P = \ln \frac{P}{1 - P} = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k \qquad (3)$$

The most commonly used assessment method of the logistic regression model parameters is the maximum likelihood method due to the fact that the estimators obtained with it are compatible, asymptotically unbiased, asymptotically effective and have an asymptotic normal distribution[7].

The interpretation of the obtained assessments of logistic regression model parameters are as follows:

[6] A. Stanisz, *Modele regresji logistycznej,* wyd. StatSoft Polska, Kraków 2016, p. 166.
[7] Ibidem, p. 63.

- if $\alpha_i > 0$, then it should be considered that the increase in the value of feature X_i, with the remaining variables unchanged, leads to an increase in the probability of success, i.e. of the object belonging to the class coded as 1,
- $\alpha_i < 0$ indicates that the increase in the value of variable X_i, *ceteris paribus*, leads to the decrease in the probability of success.

Conclusions regarding the phenomenon being modeled can also be expressed in terms of *odds*. The relative change in the probability of an event occurring under the influence described by variable X_i is given by e^{α_i}:

- if $e^{\alpha_i} > 1$, then the factor described by variable X_i has a stimulating effect on the occurrence of the phenomenon under investigation,
- if $e^{\alpha_i} < 1$, then factor X_i has an inhibiting effect on the occurrence of the phenomenon,

while a unit increase of variable X_i results in the change by $(e^{\alpha_i} - 1) \cdot 100\%$ of the probability of the object belonging to the class coded as 1, with the unchanging influence of the remaining independent variables.

The estimated model coefficients α_i *which* are interpreted in terms of logit inform about the change in the logarithm of the odds associated with a 1-unit increase in predictor. When the probabilities are higher or lower than 0.5, logit values are higher or lower than 0, respectively.

The advantage of the logistic regression is that it does not require some of the assumptions that are necessary for linear regression. The vectors of explanatory variables and residuals do not have to be normally distributed and heteroskedasticity is acceptable. However, the following conditions have to be met:

- The relationship between the logarithm of odds and the vector of explanatory variables must be linear, in line with equation (3).
- The observations must be independent.
- The model should consider all significant variables.
- Independent variables cannot be colinear.
- Logistic regression is sensitive to the occurrence of outliers. They should be removed before the analysis starts.

When summarizing the characteristics of the logistic regression model, one should emphasize its basic feature which is related to the possibility of quantifying the probability of a particular

observation belonging to the class coded as 1. This makes it also possible to classify the object to one of the two classes. The classification rule[8] usually adopts $p = 0.5$ as the probability cutoff value.

3. Probability model of further education at SUM level

The logit model to assess the probability of further education at the SUM level at the home school of the graduates of Management was developed as a result of several formalized activities[9]. The values of regression parameters were calculated on the basis of the data concerning 1 205 graduates (learning sample) in 2013-2017. For year 2018 (verification sample – 144 graduates) an *ex post* prediction was conducted to compare the regression results with the empirical data and to check the predictive ability of the model in this way.

The estimation stage of the model parameters was preceded by a basic analysis of the data structure.

Basic measures of the variables statistical description

Almost 38% graduates of the first-cycle degree program in Management who completed their studies in 2013-2017 decided to undertake further education at the SUM level.

Table 2 presents basic measures of the statistical description of quantitative explanatory variables: average, age, km.

Table 2. Location and variation measures of variables (learning sample)

Measures	Variables		
	average	*age*	*km*
Average	4.07	28.12	44.84
Median	4.04	25	26.00
Standard deviation	0.373	6.883	88.265
Minimum	3.16	21	0
Maximum	4.98	62	1000
Number	1205	1205	1205
Coefficient of variation	9%	24%	197%

Source: Authors' own research and calculations.

[8] G. G. Judge, C. Hill, W. E. Griffiths, H.Lütkepohl, T. Lee, *The Theory and Practice of Econometrics*, John Wiley&Sons, New York 1985.
[9] G. James, D. Witten, T. Hastie, R. Tibshira, *An Introduction to Statistical Learning with Applications in R*, Springer, New York 2013, pp. 131-135.

The grade point average of half of the graduates of Management who were granted diplomas in 2013-2017 was not less than 4.04. An insignificant variation of average grades is noticeable (coefficient of variation of 9%). The average age of the graduates under investigation exceeded slightly 28 years and the oldest graduate's age was 62. Half of the graduates did not live further than 26 km from Krakow, while the average distance from their place of residence to the School was 44 km. Variable *km* has a high cv (197%) [10].

Percentage shares of individual categories were determined in the qualitative variables: *gender, graduation_year, mode, scholarship* and *sum.* Gender ratios in the sample under investigation were fairly even: females constituted 55.8% of the sample. Over 83% of graduates studied in the part-time mode. Scholarships were received by slightly over 39% of students.

The determined Pearson linear correlation coefficients (Table 3) indicate an average correlation between variables *scholarship* and *average* (R=0.51), *average* and *gender* (R=0.34), and *age* and *average* (R=0.36). The strength of the remaining linear relationships is insignificant ($|R|$<0.3).

Table 3. Correlations between explanatory variables

Variable	*gender*	*Graduation-year*	*mode*	*average*	*scholarship*	*age*	*km*
gender	1	0.00	0.05	0.34	0.25	0.05	0.01
graduation-year	0.00	1	0.01	0.07	-0.02	0.04	0.10
tryb	0.05	0.01	1	0.12	0.05	0.29	-0.22
average	0.34	0.07	0.12	1	0.51	0.36	-0.03
scholarship	0.25	-0.02	0.05	0.51	1	0.14	-0.04
age	0.05	0.04	0.29	0.36	0.14	1	-0.16
km	0.01	0.10	-0.22	-0.03	-0.04	-0.16	1

Source: Authors' own research and calculations.

Estimation of the model parameters

The parameters of the logistic regression model were estimated with the *Gretl* program (Table 4).

[10] The outliers have an insignificant impact on the model predicative capacity, which was verified through the elimination of outliers and the comparison of the model prediction results.

Table 4. Logit estimation, observations used 1–1205, Dependent (Y): sum

	Coefficient	*Standard error*	*z*	*value of p*	
const	117.906	90.9627	1.296	0.1949	
gender	−0.0937891	0.131848	−0.7113	0.4769	
graduation_year	−0.0600870	0.0451917	−1.330	0.1836	
mode	−0.287385	0.173286	−1.658	0.0972	*
average	0.550652	0.209129	2.633	0.0085	***
scholarship	0.503834	0.142989	3.524	0.0004	***
age	0.0161146	0.00970125	1.661	0.0967	*
km	0.000299949	0.000701698	0.4275	0.6690	

Arithmetic mean of the dependent variable	0.375104	Standard deviation of the dependent variable	0.484351
McFadden's R-squared	0.331648	Corrected R-squared	0.221613
Log-likelihood	−772.0163	AIC	1560.033
BIC	1600.786	HQ	1575.381

Source: Authors' own research and calculations with the use of the Gretl application

The assumed significance level α equaled 10%. After the sequential elimination of statistically insignificant variables (the value of $p > 0.1$), a model was obtained with parameters given in Table 5.

Table 5. Model 2: Logit estimation, observations used 1–1205, Dependent (Y): sum

	Coefficient	*Standard error*	*z*	*Values of p*	
const	−2.92799	0.746180	−3.924	<0.0001	***
mode	−0.304450	0.169854	−1.792	0.0731	*
average	0.492461	0.200156	2.460	0.0139	**
scholarship	0.502971	0.141746	3.548	0.0004	***
age	0.0161220	0.00959630	1.680	0.0930	*

Arithmetic mean of the dependent variable	0.375104	Standard deviation of the dependent variable	0.484351
McFadden's R-squared	0.330201	Corrected R-squared	0.223930
Log-likelihood	−773.1697	AIC	1556.339
BIC	1581.811	HQ	1565.932

Source: Authors' own research and calculations with the use of the Gretl application.

Thus, the regression model for the field of study of *Management* is given by:

$$P(Y = 1/X_1, X_2, X_3, X_4) = \frac{1}{1 + e^{-(-2.92799 - 0.304450 \cdot X_1 + 0.492461 \cdot X_2 + 0.502971 \cdot X_3 + 0.0161220 \cdot X_4)}} \quad (5)$$

where: $X_1 - mode, X_2 - average, X_3 - scholarship, X_4 - age$

Verification of the model consistency with empirical data

In the case of binary models there are various ways measure to assess the fit of the model to empirical data. The simplest measure is the coefficient of determination R^2 (the corrected *R-squared* which considers the number of variables in the model). *McFadden's R-squared* coefficient can also be used. It should be emphasized, however, that low variability explanation level is typical for all logit models[11].

Numerous practitioners believe that "good" quality of the model is determined by the accuracy of predictions that it provides[12]. *Count R-squared* is used in the process. The number of correct predictions was determined (the number of cases for which the model's classification was correct), i.e. the cases where the model described belonging to the coded class as 1 (continuation of studies at SUM) when the estimated value of probability was higher than the value of the probability limit. The cutoff value was adopted at the level of 0.5. The corrected R-squared determines the share of the correctly classified cases in the total number of cases. The basic method to assess the logistic regression model is to present classification (prediction) results in the form of a hit table (Table 6).

Table 6. Hit table (learning sample)

Empirical (actual)	Predicted		Total
	0	1	
0	**524**	**229**	753
1	**229**	**223**	452
Total	753	452	1205

Source: Authors' own research and calculations with the use of the Gretl application.

The count R-squared in the model was (524 + 223) / 1205 = 62.0%, which means that almost 62% of the classification cases turned out to be correct.

The accuracy of predictions can also be assessed by the odds ratio, calculated as:

$(223 \cdot 524) / (229 \cdot 229) = 2.23.$

It should be pointed out that the odds ratio values higher than 1 indicate that predicting with the use of the model is better than random classification[13].

Prediction capacities of the model were verified by the data from 2018 in which 110 students graduated from the first-cycle degree program in Management (Table7).

[11] T. Kufel, *Ekonometria*…, op. cit., p. 146.
[12] P. Cichosz, *Data Mining Algorithms: Explained Using R*, Willey, Chichester 2015, pp. 134-157.
[13] T. Kufel, *Ekonometria*…, op. cit., p. 146.

Table 2. Hit table (verification sample)

Empirical (actual)	Predicted		Total
	0	1	
0	**28**	**28**	56
1	**21**	**67**	88
Total	49	95	144

Source: Authors' own research and calculations with the use of MS Excel.

It should be noted that the share of the correctly classified cases in the total number of cases, i.e. the Count R-squared increased to (28+67)/144 = 66%, while the odds ratio reached the value of 3.19.

4. Interpretation of the results and the application of the acquired knowledge to predict the number of students continuing their education at the home school

The model of logistic regression for the field of study of *Management* showed 4 determinants of further studies at second-cycle level (at the adopted significance level of 10%):

- *mode* of studies (full-time/part-time),
- grade point average in studies (*average*),
- scholarship received (*scholarship*),
- age of the graduate at the moment of diploma thesis defense (*age*).

On the basis of the parameter sign next to the variable (formula 5), it is possible to determine the influence of this explanatory variable on the probability of further studies at the home school. A positive sign of the estimated regression coefficient next to variables *average*, *scholarship* and *age* indicates that the factors that influence the decision on undertaking the second-cycle degree studies are good learning performance, scholarship and the increasing maturity of the graduate. The increase in the grade point average by one point indicates an increase by 63.6% of the probability that the graduate will continue education at the home school (with an unchanged influence of the remaining independent variables). Obtaining a scholarship results in the increase in the „success" probability by 65.4% (*ceteris paribus*). An additional year in the graduate's age involves an increase in the probability that the student will continue education at the School by 1.6% (with the assumption that other variables in the model remain unchanged). A negative sign next to variable *mode* indicates that the change of the mode from full-time to part-time decreases the probability of the student belonging to the class of SUM studies (ceteris paribus) by 26.2%.

The logistic model was applied to the processed data concerning students of the last semester of the first-cycle degree studies (and whose defense of thesis was set for 2019). On the basis of the estimated model parameters the probability of their continuation of studies at the second- cycle degree studies was determined. Figure2 presents a part of the spreadsheet used for the students of the investigated field of study. A cutoff value of 0.5 was adopted.

Figure 2. Part of the spreadsheet used to predict the number of the students of Management continuing education at SUM

mode	*average*	*scholarship*	*age*	*logit*	*P*	*prediction*
0	3.66	0	21	-0.78702074	0.312808732	0
1	4.29	1	47	0.14092269	0.535172484	1
0	3.99	0	21	-0.62450861	0.348756734	0
1	4.05	1	25	-0.33195195	0.41776576	0
1	4.16	0	26	-0.76463024	0.317641836	0
1	4.29	0	29	-0.65224431	0.342483966	0
1	4.74	1	25	0.00784614	0.501961525	1
1	4.63	1	41	0.21162743	0.55271028	1
1	4.08	0	29	-0.75566112	0.319589024	0
0	3.08	1	25	-0.50518912	0.376321983	0

Source: Authors' own research and calculations.

The result obtained indicates that every third student of the last semester of the first-cycle degree studies in Management will undertake the second-cycle degree program at the home school.

Conclusions

The study is an attempt to assess the number of students who will undertake the second-cycle degree program in a non-public school of higher education after they graduated from the first-cycle degree studies at this school. The data used was stored in the school IT systems and the mining was conducted with the use of a logistic regression model. Four determinants were identified that could have an impact on students' decisions on undertaking the second-cycle studies at the SUM level at the home school. The interpretation and assessment of the model parameters lead to the conclusion that a high grade point average, a scholarship and graduate's maturity increase the probability of the continuation of education at the home school. Studying in the part-time mode decreases this probability.

The number of students who will start the second-cycle degree studies was approximated. The estimation may support the School management in planning organizational activities.

Despite the fact that the developed model proved its usefulness, some of its predictive weaknesses should be pointed out which result from an insufficient range of information and, consequently, an insignificant number of the variables under analysis. The authors are aware that the presented model needs improvement. It is advisable that the set of variables should be extended so that they better present the modeled reality. It seems that in particular the model lacks the information on the student's financial situation, which is important when making a decision on paid studies at a non-public school of higher education. It also seems important to measure student satisfaction with the quality of services provided by the School. Information on further careers of the graduates who did not continue education in the School is also crucial as well as the information on the fields of study that are available in competitive institutions. It should be remembered that in 2020, when the global pandemic started, the schools of higher education were forced to implement new teaching methods. As a result, the existing models should consider new conditions and variables, which may be a difficult challenge determining a further direction of research.

Bibliography

[1] Cichosz P., *Data Mining Algorithms: Explained Using R*, Willey, Chichester 2015,
[2] Gawryś I., Trippner P., *Analiza poziomu rentowności przedsiębiorstwa na przykładzie niepublicznej uczelni wyższej w roku akademickim 2015/2016*, „Annales Universitatis Mariae Curie-Skłodowska", 2017, No, LI, 5,
[3] James G., Witten D., Hastie T., Tibshira R., *An Introduction to Statistical Learning: with Applications in R*, Springer Texts in Statistics, Springer Science+Business Media, New York 2013,
[4] Judge G, G., Hill C., Griffiths W, E., Lütkepohl H., Lee T., *The Theory and Practice of Econometrics*, John Wiley&Sons, New York 1985,
[5] Kufel T., *Ekonometria*, *Rozwiązywanie problemów z wykorzystaniem programu GRETL*, Wydawnictwo Naukowe PWN, Warszawa 2011,
[6] Morzy T., *Eksploracja danych*, Wydawnictwo Naukowe PWN, Warszawa 2013,
[7] Nandeshwar A., Chaudhari S., *Enrollment Prediction Models Using Data Mining* http://nandeshwar,info/wp-content/uploads/2008/11/DMWVU_Project,pdf,
[8] Stanisz A., *Modele regresji logistycznej,* wyd, StatSoft Polska, Kraków 2016,
[9] Szeliga M,, *Data Science i uczenie maszynowe*, Wydawnictwo Naukowe PWN, Warszawa 2017,
[10] Zieliński G., Lewandowski K., *Determinanty percepcji jakości usług edukacyjnych w perspektywie grup interesariuszy*, http://zif.wzr.pl/pim/2012_3_3_4.pdf

Abstract

The article is an attempt to use a logistic regression model to estimate the number of the first-cycle degree program graduates who will continue education on supplementary master's degree studies at the home school.

Data available in the IT systems of the school under analysis was mined and on this basis factors were identified that can influence the graduates' decisions regarding the continuation of studies at the level of a supplementary master's degree program (SUM). The interpretation and assessment of the developed logit model parameters made it possible to draw applicable conclusions.

Key words

Logit, data mining, enrolment to school of higher education, logistic regression, decision-making process, company,