Will I Graduate? Determinants of Students' Dropout and Success Rates

Magdalena Kowalewska (412860) and Szymon Socha (411462)

Advanced Econometrics project, 30.05.2022

ABSTRACT

The aim of the project was to develop a model with binomial dependent variables explaining determinants of students' graduate rate. In order to decide which model is the most appropriate, three models were developed (probit, logit and linear probability). Using the general to specific approach all insignificant variables were removed. Afterwards all models were subjected to tests such as Linktest, Likelihood test and Hosmer-Lemeshow test. Furthermore, R square statistics were calculated to understand each model and compare how well the dependent variable is explained in each of them. Based on the AIC, the logit model was selected. Within the key findings of the model three were identified: administrative studies had the largest dropout probability out of all the course categories, prior academic achievements of the student seem not to be a factor of high value when assessing the probability of the student to graduate and if the father drops out in 12th grade (last year in high school) then the student is less likely to drop in comparison to a student whose father left after completing any of grades in secondary or high school. The decision to dropout has many consequences. Identifying students who can potentially drop out at an early stage can help the educational institutions develop strategies, which would provide the necessary aid. Furthermore, the topic of education is also important as it is one of the factors that drive economic growth.

INTRODUCTION

Students' dropout phenomenon is a complex topic, which is a concern for both individuals and the society. For the purpose of this project dropouts are defined as students who started their academic journey but for some reason decided not to complete the course they started. Education is one of the key contributors to economic growth. Knowledge shared within a country (or region) drives productivity, creativeness and fosters entrepreneurship and technological progress[1]. Therefore, multiple researchers have explored this topic in hope of finding the root cause, which would enable the development of an appropriate action plan.

According to multiple studies dropping out has various causes, within which we can recognise internal and external factors. Barbara M. Kehm et al. identifies nine factors contributing to students leaving school: study conditions at university, academic integration at university, social integration at university, personal efforts and motivations for studying, information and

[1]Márquez-Ramos, Laura & Mourelle, Estefanía. (2018). Education and economic growth: an empirical analysis of nonlinearities. Revista de Economia Aplicada. 26. 1-28. 10.1108/AEA-06-2019-0005.

admission requirements, prior academic achievement in school, personal characteristics of the student and sociodemographic background of the student, external conditions[2].

Dropout rates can also be influenced by socio-economic factors such as a country's GDP, inflation rate, unemployment rate or GDP per capita. These factors determine students' approach to education and willingness to stay in school. Usually the poorer the country the more emphasis is put on helping the family by starting earning money as soon as possible. In such a culture attending educational institutions is a lower priority, hence, a larger number of students decide to leave school without formal education.

Regardless of the country the students live in, each student is influenced by their personal life, whether it is the social circle, family situation, actual and psychological age, or health (both physical and mental). Most of these elements can rarely be targeted using facility's aids, as in most cases the institution is unaware about such circumstances. However, teachers can influence students, especially younger students. Thus, it is crucial for the schools to thoroughly assess candidates for teacher positions.

Even Though, multiple studies tackle the topics of certain factors contributing to students dropping out of schools, very few include socio-economic factors, family background, course arrangements and admission process. Hence, this project aims at determining, which factors are the main contributors, when contributors of various types (personal, socio-economic, family background, current academic course, academic history) are examined simultaneously. The research question is which type of contributing factors has the largest importance when determining if a student will graduate or dropout.

LITERATURE REVIEW

Samuli Laato, Emilia Lipponen, Heidi Salmento, Henna Vilppu and Mari Murtonen (2019) performed a study to understand the reasons why adult students dropout of online courses. The first hypothesis was that dropout rates were higher at the beginning of the course and lowered towards the end of the course. The second hypothesis was that the majority of students leave the course during the individual task phase of course (after registration and first logging in to the online course). In order to test the hypotheses, the researchers performed a case study and a qualitative research. The findings support the first hypothesis, however, the researchers did not find enough evidence to support the second hypothesis. Additionally, it was found that lecturers play an significant role throughout the whole study period, and lower dropouts rates were observed when the student was taking law courses, social studies or independant units[3]. Hence, one of the hypotheses tested in this study is H1: there is a negative correlation between students attending social study courses and dropout rate.

[2] Kehm, Barbara & Larsen, Malene & Sommersel, Hanna. (2019). Student dropout from universities in Europe: A review of empirical literature. Hungarian Educational Research Journal. 9. 147-164. 10.1556/063.9.2019.1.18.

[3]Laato, S.; Lipponen, E.; Salmento, H.; Vilppu, H. and Murtonen, M. (2019). Minimizing the Number of Dropouts in University Pedagogy Online Courses. In Proceedings of the 11th International Conference on Computer Supported Education - Volume 1: CSEDU, ISBN 978-989-758-367-4; ISSN 2184-5026, pages 587-596. DOI: 10.5220/0007686005870596

Meta-analysis of 44 empirical studies performed by Barbara M. Kehm, Malene Rode Larsen and Hanna Bjørnøy Sommersel (2019), explored 9 factors affecting the students' dropout rates. These factors are: study conditions at university, academic integration at university, social integration at university, personal efforts and motivations for studying, information and admission requirements, prior academic achievement in school, personal characteristics of the student, sociodemographic background of the student, external conditions. The study aimed at answering three questions: "*What is dropout? Why does it occur? What can be done to reduce or prevent it?*" The researchers concluded that the higher number of resources available at the institution the lower the risk of students dropping out. Furthermore, the more activity is demanded from students, the higher their motivation and thus, lower dropping out rate. The environment also is one of the factors correlated with dropout rates, however, the studies are not clear whether this is a direct or indirect relationship. Not surprisingly, the researchers also concluded that the better students' academic performance the lower the dropout rate. However surprisingly, the correlation between high school grades and students dropping out in college was not statistically significant. The researchers also found that dropout rates are higher among male students and that parents' occupation can potentially be a significant factor when determining the future dropout rate, though some studies have found this factor to be insignificant. Hence, other hypotheses tested in this study are H2: variables corresponding to prior academic achievements will not be significantly associated with students' dropout rate.[4] H3: there is a negative correlation between parents' academic achievements and children being school dropouts.

Mayra Alban and David Mauricio performed a systematic review of literature using data mining techniques to predict if university students will graduate or dropout. The researchers analysed a number of studies performed from 2006-2018. They found that students leavins universities is a common problem around the world, which causes revenue reduction for universities, financial losses from the government (if the government finances the institution) and future problems for students who dropped out and their families. Thus, it is essential to further examine the issue and develop a way to predict students' success rate.[5]

DATA

The dataset was downloaded from the UC Irvine Machine Learning Repository website (https://archive-beta.ics.uci.edu/ml/datasets/predict+students+dropout+and+academic+success). It is a combination of datasets from various higher education institutions. It contains information about students in various fields of study such as agronomy, design, education, nursing, journalism, management, social service, and technologies. The dataset includes information about students such as their academic path, demographics, and social-economic factors. The explanatory variable is the status of the student at the end of the normal study time.

[4] Kehm, Barbara & Larsen, Malene & Sommersel, Hanna. (2019)... op. cit.

[5] Alban, Mayra & Mauricio, David. (2019). Predicting University Dropout through Data Mining: A Systematic Literature. 10.17485/ijst/2019/v12i4/139729.

The dataset is composed of 4424 observations and 37 variables. We observe that some variables have many levels. In order to make data more consistent and results more robust we combine them into more numerous groups. Detailed description of the variables:

- Marital status
    - type: Numeric/discrete
    - description: 1 – single, 2 – married, 3 – widower, 4 – divorced, 5 – facto union, 6 – legally separated. We combine widower with divorced and legally separated and married with factor union.
- Application mode
    - type: Numeric/discrete
    - description: 1 - 1st phase - general contingent, 2 - Ordinance No. 612/93, 5 - 1st phase - special contingent (Azores Island), 7 - Holders of other higher courses, 10 - Ordinance No. 854-B/99, 15 - International student (bachelor), 16 - 1st phase - special contingent (Madeira Island), 17 - 2nd phase - general contingent, 18 - 3rd phase - general contingent, 26 - Ordinance No. 533-A/99, item b2) (Different Plan), 27 - Ordinance No. 533-A/99, item b3 (Other Institution), 39 - Over 23 years old, 42 - Transfer, 43 - Change of course, 44 - Technological specialization diploma holders, 51 - Change of institution/course, 53 - Short cycle diploma holders, 57 - Change of institution/course (International). We combine all ‘phases’, ‘changes of institutions’, ‘ordinances’ into respective groups.
- Application order
    - type: Numeric/discrete
    - description: Application order (between 0 - first choice; and 9 last choice). We observe one observation for 0 and one for 9. We consider them as outliers and join them to the closest group (0 to 1 and 9 to 6).
- Course
    - type: Numeric/discrete
    - description: 33 - Biofuel Production Technologies, 171 - Animation and Multimedia Design, 8014 - Social Service (evening attendance), 9003 - Agronomy, 9070 - Communication Design, 9085 - Veterinary Nursing, 9119 - Informatics Engineering, 9130 - Equinculture, 9147 - Management, 9238 - Social Service, 9254 - Tourism, 9500 - Nursing, 9556 - Oral Hygiene, 9670 - Advertising and Marketing Management, 9773 - Journalism and Communication, 9853 - Basic Education, 9991 - Management (evening attendance). We combine the same groups of majors into similar more general groups. We treat Social Service (evening attendance), Tourism, Social Service as Administration; Communication Design , Journalism and Communication, Basic Education, Animation and Multimedia Design as Arts/Social studies; Agronomy, Equinculture, Informatics Engineering, Biofuel Production Technologies as Biology/Nature; Management, Management (evening attendance), Advertising and Marketing Management as Economics; Veterinary Nursing, Nursing, Oral Hygiene as Medicine.

- Daytime/evening attendance
    - Numeric/binary
    - description: 1 – daytime, 0 - evening
- Previous qualification
    - Numeric/discrete
    - description: 1 - Secondary education, 2 - Higher education - bachelor's degree, 3 - Higher education - degree, 4 - Higher education - master's, 5 - Higher education - doctorate, 6 - Frequency of higher education, 9 - 12th year of schooling - not completed, 10 - 11th year of schooling - not completed, 12 - Other - 11th year of schooling, 14 - 10th year of schooling, 15 - 10th year of schooling - not completed, 19 - Basic education 3rd cycle (9th/10th/11th year) or equiv., 38 - Basic education 2nd cycle (6th/7th/8th year) or equiv., 39 - Technological specialization course, 40 - Higher education - degree (1st cycle), 42 - Professional higher technical course, 43 - Higher education - master (2nd cycle). We consider one person with Doctorate as an outlier and remove it. We combine all ‘ primary and secondary’, ‘higher education’, ‘special courses’ into respective groups.
- Previous qualification (grade)
    - type: Numeric/continous
    - description: Grade of previous qualification (between 0 and 200)
- Nationality
    - type: Numeric/discrete
    - description: 1 - Portuguese, 2 - German, 6 - Spanish, 11 - Italian, 13 - Dutch, 14 - English, 17 - Lithuanian, 21 - Angolan, 22 - Cape Verdean, 24 - Guinean, 25 - Mozambican, 26 - Santomean, 32 - Turkish, 41 - Brazilian, 62 - Romanian, 100 - Moldova (Republic of), 101 - Mexican, 103 - Ukrainian, 105 - Russian, 108 - Cuban, 109 - Colombian. We observe that all non-Portuguese students are a significant minority. We decided to group them together as 'non-portugese'.
- Mother’s qualification
    - type: Numeric/discrete
    - description: 1 - Secondary Education - 12th Year of Schooling or Eq., 2 - Higher Education - Bachelor's Degree, 3 - Higher Education - Degree, 4 - Higher Education - Master's, 5 - Higher Education - Doctorate, 6 - Frequency of Higher Education, 9 - 12th Year of Schooling - Not Completed, 10 - 11th Year of Schooling - Not Completed, 11 - 7th Year (Old), 12 - Other - 11th Year of Schooling, 14 - 10th Year of Schooling, 18 - General commerce course, 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv., 22 - Technical-professional course, 26 - 7th year of schooling, 27 - 2nd cycle of the general high school course, 29 - 9th Year of Schooling - Not Completed, 30 - 8th year of schooling, 34 - Unknown, 35 - Can't read or write, 36 - Can read without having a 4th year of schooling, 37 - Basic education 1st cycle (4th/5th year) or equiv., 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv., 39 - Technological specialization course, 40 - Higher education - degree (1st cycle),

41 - Specialized higher studies course, 42 - Professional higher technical course, 43 - Higher Education - Master (2nd cycle), 44 - Higher Education - Doctorate (3rd cycle). We combine all high school and lower not completed, higher education, special courses into respective groups. We consider ‘Can't read or write’ and ‘Can read without having a 4th year of schooling’ and remove them.

- Father’s qualification
    - type: Numeric/discrete
    - description: 1 - Secondary Education - 12th Year of Schooling or Eq., 2 - Higher Education - Bachelor's Degree, 3 - Higher Education - Degree, 4 - Higher Education - Master's, 5 - Higher Education - Doctorate, 6 - Frequency of Higher Education, 9 - 12th Year of Schooling - Not Completed, 10 - 11th Year of Schooling - Not Completed, 11 - 7th Year (Old), 12 - Other - 11th Year of Schooling, 13 - 2nd year complementary high school course, 14 - 10th Year of Schooling, 18 - General commerce course, 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv., 20 - Complementary High School Course, 22 - Technical-professional course, 25 - Complementary High School Course - not concluded, 26 - 7th year of schooling, 27 - 2nd cycle of the general high school course, 29 - 9th Year of Schooling - Not Completed, 30 - 8th year of schooling, 31 - General Course of Administration and Commerce, 33 - Supplementary Accounting and Administration, 34 - Unknown, 35 - Can't read or write, 36 - Can read without having a 4th year of schooling, 37 - Basic education 1st cycle (4th/5th year) or equiv., 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv., 39 - Technological specialization course, 40 - Higher education - degree (1st cycle), 41 - Specialized higher studies course, 42 - Professional higher technical course, 43 - Higher Education - Master (2nd cycle), 44 - Higher Education - Doctorate (3rd cycle). For Father’s qualification we make the same changes as for Mother’s qualification.
- Mother’s occupation
    - type: Numeric/discrete
    - description: 0 - Student, 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers, 2 - Specialists in Intellectual and Scientific Activities, 3 - Intermediate Level Technicians and Professions, 4 - Administrative staff, 5 - Personal Services, Security and Safety Workers and Sellers, 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry, 7 - Skilled Workers in Industry, Construction and Craftsmen, 8 - Installation and Machine Operators and Assembly Workers, 9 - Unskilled Workers, 10 - Armed Forces Professions, 90 - Other Situation, 99 - (blank), 122 - Health professionals, 123 - teachers, 125 - Specialists in information and communication technologies (ICT), 131 - Intermediate level science and engineering technicians and professions, 132 - Technicians and professionals, of intermediate level of health, 134 - Intermediate level technicians from legal, social, sports, cultural and similar services, 141 - Office workers, secretaries in general and data processing operators, 143 - Data, accounting, statistical,

financial services and registry-related operators, 144 - Other administrative support staff, 151 - personal service workers, 152 - sellers, 153 - Personal care workers and the like, 171 - Skilled construction workers and the like, except electricians, 173 - Skilled workers in printing, precision instrument manufacturing, jewelers, artisans and the like, 175 - Workers in food processing, woodworking, clothing and other industries and crafts, 191 - cleaning workers, 192 - Unskilled workers in agriculture, animal production, fisheries and forestry, 193 - Unskilled workers in extractive industry, construction, manufacturing and transport, 194 - Meal preparation assistants. We reduce the number of levels by joining them into more general categories: 0 as Student, from 110 to 119 as Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers, from 120 to 129 as Specialists in Intellectual and Scientific Activities, from 130 to 139 as Intermediate Level Technicians and Professions, from 140 to 149 as Administrative staff, from 150 to 159 as Personal Services Security and Safety Workers and Sellers, from 160 to 169 as Farmers and Skilled Workers in Agriculture, Fisheries and Forestry, from 170 to 179 to Skilled Workers in Industry, Construction and Craftsmen, from 180 to 189 to Installation and Machine Operators and Assembly Workers, from 190 to 199 as Unskilled Workers, from 100 to 109 as Armed Forces Professions.

- Father's occupation
    - type: Numeric/discrete
    - description: 0 - Student, 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers, 2 - Specialists in Intellectual and Scientific Activities, 3 - Intermediate Level Technicians and Professions, 4 - Administrative staff, 5 - Personal Services, Security and Safety Workers and Sellers, 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry, 7 - Skilled Workers in Industry, Construction and Craftsmen, 8 - Installation and Machine Operators and Assembly Workers, 9 - Unskilled Workers, 10 - Armed Forces Professions, 90 - Other Situation, 99 - (blank), 101 - Armed Forces Officers, 102 - Armed Forces Sergeants, 103 - Other Armed Forces personnel, 112 - Directors of administrative and commercial services, 114 - Hotel, catering, trade and other services directors, 121 - Specialists in the physical sciences, mathematics, engineering and related techniques, 122 - Health professionals, 123 - teachers, 124 - Specialists in finance, accounting, administrative organization, public and commercial relations, 131 - Intermediate level science and engineering technicians and professions, 132 - Technicians and professionals, of intermediate level of health, 134 - Intermediate level technicians from legal, social, sports, cultural and similar services, 135 - Information and communication technology technicians, 141 - Office workers, secretaries in general and data processing operators, 143 - Data, accounting, statistical, financial services and registry-related operators, 144 - Other administrative support staff, 151 - personal service workers, 152 - sellers,

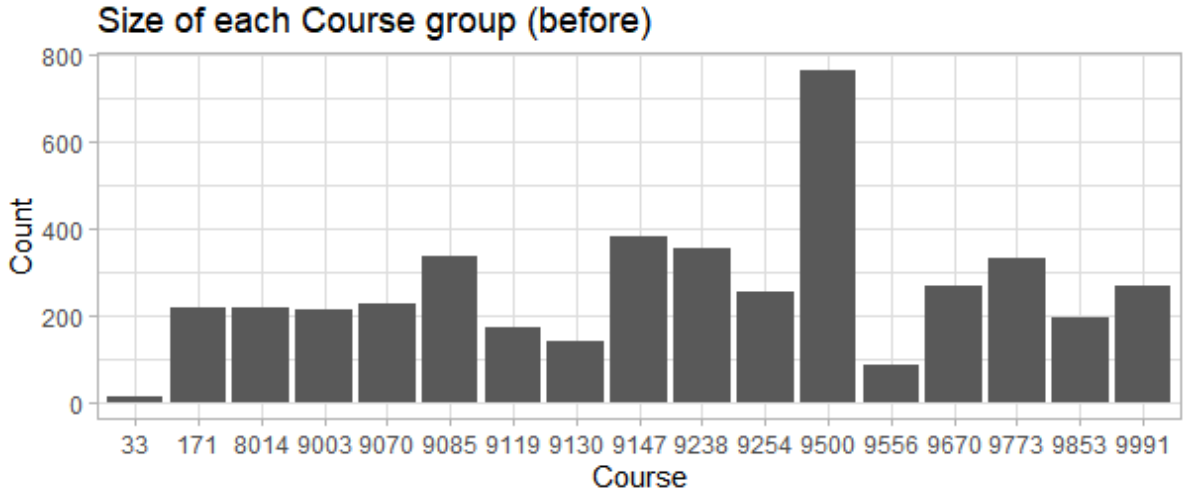
153 - Personal care workers and the like, 154 - Protection and security services personnel, 161 - Market-oriented farmers and skilled agricultural and animal production workers, 163 - Farmers, livestock keepers, fishermen, hunters and gatherers, subsistence, 171 - Skilled construction workers and the like, except electricians, 172 - Skilled workers in metallurgy, metalworking and similar, 174 - Skilled workers in electricity and electronics, 175 - Workers in food processing, woodworking, clothing and other industries and crafts, 181 - Fixed plant and machine operators, 182 - assembly workers, 183 - Vehicle drivers and mobile equipment operators, 192 - Unskilled workers in agriculture, animal production, fisheries and forestry, 193 - Unskilled workers in extractive industry, construction, manufacturing and transport, 194 - Meal preparation assistants, 195 - Street vendors (except food) and street service providers. For Father's occupation we make the same changes as for Mother's occupation.

- Admission grade
    - type: Numeric/continuous
    - description: Admission grade (between 0 and 200)
- Displaced
    - type: Numeric/binary
    - description: 1 – yes, 0 – no
- Educational special needs
    - type: Numeric/binary
    - description: 1 – yes, 0 – no
- Debtor
    - type: Numeric/binary
    - description: 1 – yes, 0 – no
- Tuition fees up to date
    - type: Numeric/binary
    - description: 1 – yes, 0 – no
- Gender
    - type: Numeric/binary
    - description: 1 – male, 0 – female
- Scholarship holder
    - type: Numeric/binary
    - description: 1 – yes, 0 – no
- Age at enrollment
    - type: Numeric/discrete
    - description: Age of student at enrollment. We reduce the number of levels by treating all students at the age of 17, 18 and 19 as one '17-19' group, all between 20 and 30 as '20-29 and all above 30 as '30+'.
- International
    - type: Numeric/binary

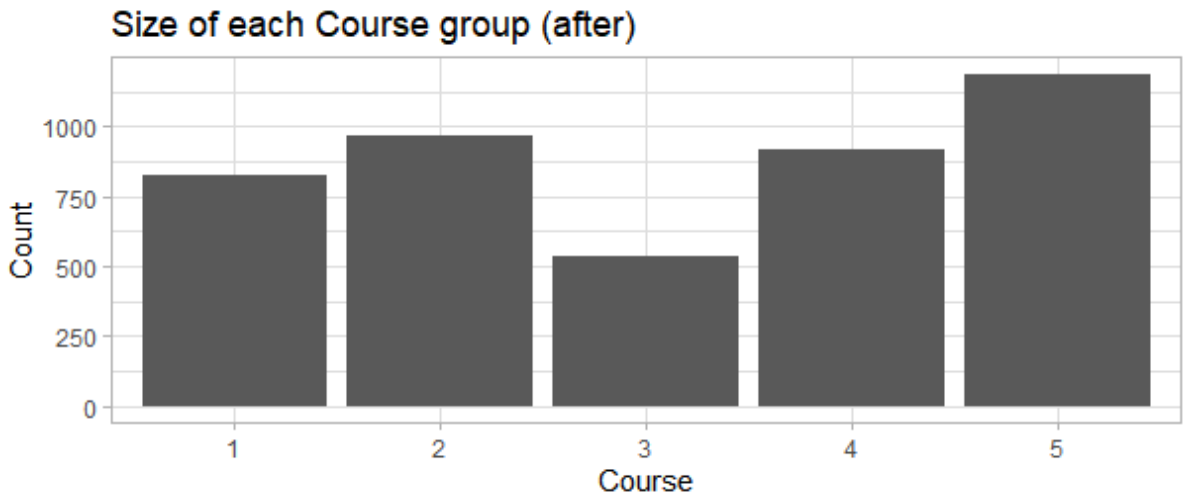
    - description: 1 – yes, 0 – no. After reducing the number of levels in the 'Nationality' variable these two variables are identical. We remove the 'International' variable.
- Curricular units 1st sem (credited)
    - type: Numeric/discrete
    - description: Number of curricular units credited in the 1st semester. We observe many 0 values and many less numerous levels. We simplify it to two 'non-zero' and 'zero' levels.
- Curricular units 1st sem (enrolled)
    - type: Numeric/discrete
    - description: Number of curricular units enrolled in the 1st semester. We treat all values higher than 7 as 8 and all smaller than 5 as 0.
- Curricular units 1st sem (evaluations)
    - type: Numeric/discrete
    - description: Number of evaluations to curricular units in the 1st semester. We treat all values higher than 16 as 17 and all smaller than 5 as 0.
- Curricular units 1st sem (approved)
    - type: Numeric/discrete
    - description: Number of curricular units approved in the 1st semester. We treat all values higher than 8 as 9.
- Curricular units 1st sem (grade)
    - type: Numeric/discrete
    - description: Grade average in the 1st semester (between 0 and 20)
- Curricular units 1st sem (without evaluations)
    - type: Numeric/discrete
    - description: Number of curricular units without evaluations in the 1st semester. We treat all values higher than 0 as 1.
- Curricular units 2nd sem (credited)
    - type: Numeric/discrete
    - description: Number of curricular units credited in the 2nd semester. We apply the same modification as for the 1st semester.
- Curricular units 2nd sem (enrolled)
    - type: Numeric/discrete
    - description: Number of curricular units enrolled in the 2nd semester. We apply the same modification as for the 1st semester.
- Curricular units 2nd sem (evaluations)
    - type: Numeric/discrete
    - description: Number of evaluations to curricular units in the 2nd semester. We apply the same modification as for the 1st semester.
- Curricular units 2nd sem (approved)
    - type: Numeric/discrete
    - description: Number of curricular units approved in the 2nd semester. We apply the same modification as for the 1st semester.

- Curricular units 2nd sem (grade)
    - type: Numeric/discrete
    - description: Grade average in the 2nd semester (between 0 and 20)
- Curricular units 2nd sem (without evaluations)
    - type: Numeric/discrete
    - description: Number of curricular units without evaluations in the 1st semester. We apply the same modification as for the 1st semester.
- Unemployment rate
    - type: Numeric/continuous
    - description: Unemployment rate (%)
- Inflation rate
    - type: Numeric/continuous
    - description: Inflation rate (%)
- GDP
    - type: Numeric/continuous
    - description: GDP
- Target
    - type: Categorical
    - description: Target. The problem is formulated as a three category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course

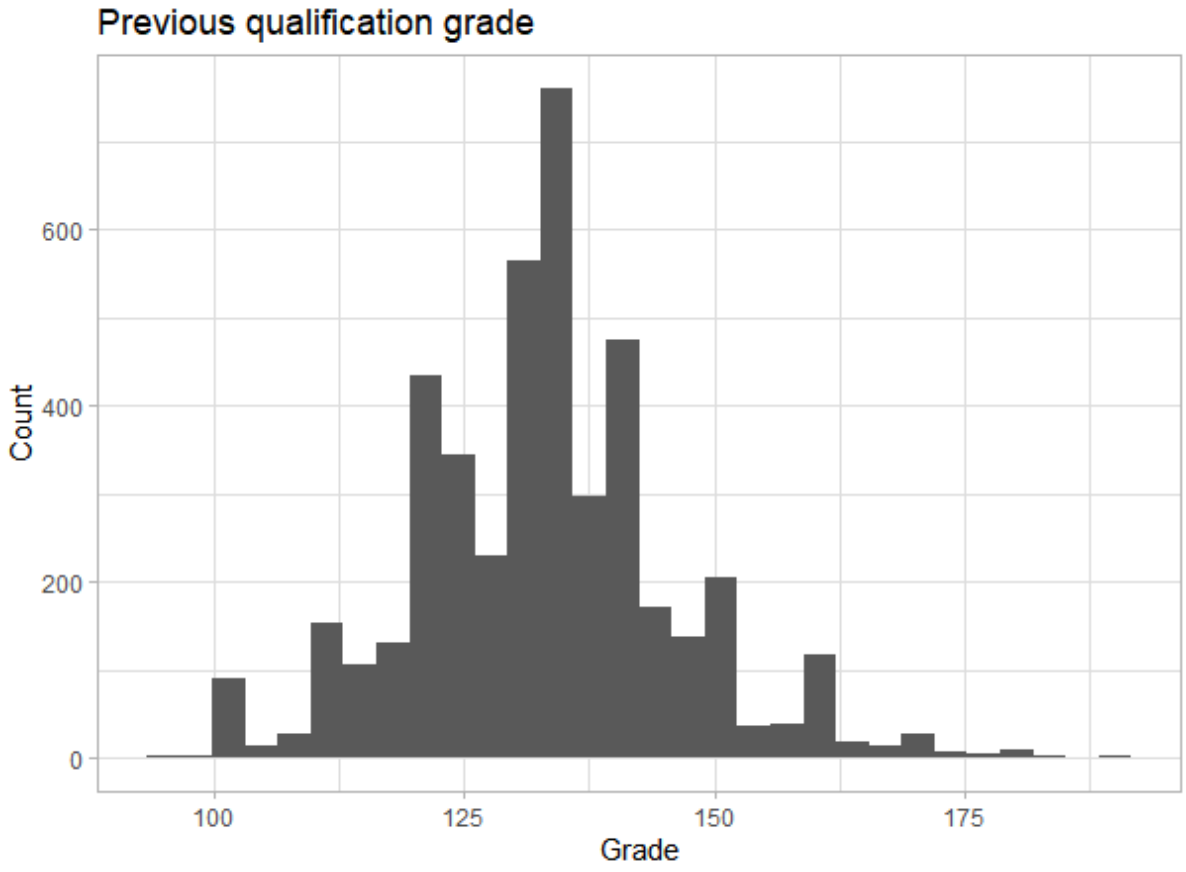
Below we show, using the variable 'Courses' as an example, the effects of reducing the number of levels. In the first graph we observe many different groups. Such large differences as between level 33 and 9500 could negatively affect the performance of the final model.



After the reduction of the number of levels, we observe a balancing of the numbers of individual groups.



Similar results are obtained for the other variables. However, they do not differ in their form from the graphs presented above, so we do not show them all here. In addition to categorical variables, we also analyse continuous variables. However, we do not find any irregularities for them. Below is an example of a histogram for the variable 'Previous qualification grade'.



METHOD/MODEL

We start the modelling with estimation of three models; linear probability model with White's robust matrix, probit and logit. All three models are estimated on the full set of variables. Linear

probability model does not pass the RESET test. We have to reject the null hypothesis (p-value=0 with the confidence level of 0.05), the model has the inappropriate form. We also check the homoscedasticity assumption with Breusch-Pagan test. We have to reject the null hypothesis that residuals of the linear probability model are heteroscedastic. Because both tests failed (RESET and Breusch-Pagan) the form of the model is incorrect and homoscedasticity assumption is not fulfilled. Because of that, the estimates of the model and standard errors of the estimates are biased and inconsistent. In order to fix the heteroskedasticity issue we employ White's estimator of the variance-covariance matrix (1). However, it fixes only one of the issues. The form of the model will remain incorrect. What's more, predictions of the linear probability model are likely to appear outside the 0-1 interval. That is why using linear probability models in binary choice class models is an incorrect approach.

The correct approach is employing a probit or logit model. We run them on the same set of variables as the linear probability model.

Table 1. Stargazer output for linear probability model (with White's robust matrix), logit and probit

	Dependent variable:		
	coefficient test	probit	logit
	(1)	(2)	(3)
ď.żMarital.status.21	-0.002	0.082	0.089
	(0.021)	(0.173)	(0.321)
ď.żMarital.status.31	0.006	0.367	0.590
	(0.033)	(0.310)	(0.578)
Application.mode.21	-0.178*	-0.708	-1.347
	(0.097)	(0.470)	(0.860)
Application.mode.71	0.053	0.127	0.389
	(0.052)	(0.313)	(0.571)
Application.mode.151	0.126*	1.085*	2.034*
	(0.068)	(0.568)	(1.102)
Application.mode.171	-0.040***	-0.237**	-0.451**
	(0.014)	(0.097)	(0.185)
Application.mode.181	-0.020	-0.069	-0.121
	(0.028)	(0.230)	(0.443)
Application.mode.391	-0.020	-0.229	-0.351
	(0.022)	(0.162)	(0.300)
Application.mode.431	-0.036	-0.180	-0.304
	(0.026)	(0.171)	(0.316)

Application.mode.441	0.133**	0.799	1.557*
	(0.061)	(0.490)	(0.919)
Application.mode.511	-0.059*	-0.346	-0.593
	(0.035)	(0.221)	(0.410)
Application.mode.531	0.091	0.734	1.288
	(0.075)	(0.667)	(1.228)
Application.order.21	0.010	0.055	0.222
	(0.016)	(0.119)	(0.228)
Application.order.31	-0.020	-0.122	-0.129
	(0.022)	(0.144)	(0.274)
Application.order.41	-0.004	0.056	0.100
	(0.021)	(0.159)	(0.299)
Application.order.51	0.004	0.060	0.082
	(0.027)	(0.214)	(0.402)
Application.order.61	-0.024	-0.174	-0.236
	(0.031)	(0.198)	(0.376)
Course.21	-0.119***	-0.802***	-1.438***
	(0.016)	(0.132)	(0.253)
Course.31	-0.094***	-0.803***	-1.467***
	(0.021)	(0.169)	(0.318)
Course.41	-0.074***	-0.615***	-1.121***
	(0.016)	(0.128)	(0.244)
Course.51	-0.045***	-0.240*	-0.375
	(0.016)	(0.140)	(0.272)
Daytime.evening.attendance..11	0.024	0.219	0.380
	(0.020)	(0.152)	(0.284)
Previous.qualification.31	-0.092**	-0.552**	-1.065**
	(0.042)	(0.229)	(0.418)
Previous.qualification.91	0.034	0.080	0.085
	(0.024)	(0.208)	(0.385)
Previous.qualification.391	-0.068	-0.410	-0.854
	(0.058)	(0.463)	(0.868)
Previous.qualification..grade.	-0.0001	-0.002	-0.002
	(0.0005)	(0.003)	(0.007)
Nacionality.Portuguese1	-0.055	-0.290	-0.528
	(0.039)	(0.294)	(0.558)
Mother.s.qualification.31	0.015	0.143	0.234
	(0.023)	(0.157)	(0.296)
Mother.s.qualification.91	-0.012	-0.064	-0.160

	(0.014)	(0.105)	(0.197)
Mother.s.qualification.341	-0.082*	-0.802	-1.415
	(0.046)	(0.512)	(0.959)
Mother.s.qualification.371	-0.030*	-0.174	-0.396
	(0.018)	(0.137)	(0.258)
Mother.s.qualification.391	-0.001	0.294	0.725
	(0.088)	(0.506)	(0.908)
Father.s.qualification.31	-0.002	-0.093	-0.071
	(0.029)	(0.166)	(0.311)
Father.s.qualification.91	0.029**	0.210**	0.384*
	(0.015)	(0.105)	(0.199)
Father.s.qualification.341	-0.034	-0.630	-1.126
	(0.054)	(0.557)	(1.045)
Father.s.qualification.371	0.022	0.118	0.241
	(0.017)	(0.130)	(0.247)
Father.s.qualification.391	-0.034	-0.645	-1.111
	(0.075)	(0.518)	(0.963)
Mother.s.occupation.11	-0.052	-0.476	-0.882
	(0.051)	(0.474)	(0.883)
Mother.s.occupation.21	-0.038	-0.540	-0.931
	(0.046)	(0.445)	(0.828)
Mother.s.occupation.31	-0.039	-0.436	-0.759
	(0.041)	(0.427)	(0.797)
Mother.s.occupation.41	-0.018	-0.358	-0.693
	(0.038)	(0.411)	(0.767)
Mother.s.occupation.51	-0.041	-0.350	-0.661
	(0.039)	(0.413)	(0.769)
Mother.s.occupation.61	0.008	-0.028	0.020
	(0.049)	(0.497)	(0.938)
Mother.s.occupation.71	-0.041	-0.457	-0.849
	(0.041)	(0.431)	(0.806)
Mother.s.occupation.81	-0.104	-0.796	-1.364
	(0.071)	(0.553)	(1.019)
Mother.s.occupation.91	-0.007	-0.219	-0.405
	(0.038)	(0.411)	(0.767)
Mother.s.occupation.101	0.123	1.405	2.765
	(0.193)	(2.097)	(5.144)
Father.s.occupation.11	0.018	-0.192	-0.393
	(0.051)	(0.460)	(0.864)

Father.s.occupation.21	-0.013	-0.173	-0.464
	(0.053)	(0.466)	(0.872)
Father.s.occupation.31	-0.007	-0.335	-0.652
	(0.044)	(0.428)	(0.803)
Father.s.occupation.41	-0.049	-0.539	-0.964
	(0.043)	(0.429)	(0.805)
Father.s.occupation.51	0.008	-0.225	-0.449
	(0.042)	(0.422)	(0.793)
Father.s.occupation.61	0.043	0.100	0.105
	(0.043)	(0.459)	(0.868)
Father.s.occupation.71	-0.017	-0.351	-0.668
	(0.042)	(0.424)	(0.797)
Father.s.occupation.81	-0.011	-0.342	-0.640
	(0.043)	(0.433)	(0.815)
Father.s.occupation.91	-0.039	-0.499	-0.916
	(0.041)	(0.421)	(0.790)
Father.s.occupation.101	0.010	-0.151	-0.324
	(0.045)	(0.438)	(0.823)
Admission.grade	0.001*	0.004	0.007
	(0.0004)	(0.003)	(0.006)
Displaced.11	-0.011	-0.166*	-0.283*
	(0.011)	(0.086)	(0.162)
Educational.special.needs.11	-0.031	0.039	-0.001
	(0.051)	(0.293)	(0.514)
Debtor.11	-0.093***	-0.545***	-1.033***
	(0.019)	(0.131)	(0.243)
Tuition.fees.up.to.date.11	0.196***	1.551***	2.732***
	(0.018)	(0.158)	(0.294)
Gender.11	-0.033***	-0.195**	-0.382***
	(0.012)	(0.080)	(0.148)
Scholarship.holder.11	0.079***	0.447***	0.870***
	(0.012)	(0.091)	(0.176)
Age.at.enrollment.20.291	-0.004	-0.016	-0.072
	(0.013)	(0.093)	(0.174)
Age.at.enrollment.30.1	-0.007	-0.133	-0.242
	(0.024)	(0.190)	(0.354)
Curricular.units.1st.sem..credited..zero1	0.095**	0.388	0.719
	(0.043)	(0.257)	(0.471)
Curricular.units.1st.sem..enrolled.	-0.001	-0.206**	-0.447***

	(0.008)	(0.089)	(0.170)
Curricular.units.1st.sem..evaluations.	-0.006***	-0.025	-0.035
	(0.002)	(0.021)	(0.039)
Curricular.units.1st.sem..approved.	0.042***	0.310***	0.574***
	(0.007)	(0.045)	(0.083)
Curricular.units.1st.sem..grade.	-0.0003	-0.008	-0.025
	(0.002)	(0.027)	(0.053)
Curricular.units.1st.sem..without.evaluations.1	-0.003	-0.154	-0.327
	(0.027)	(0.172)	(0.317)
Curricular.units.2nd.sem..credited..zero1	0.075*	0.112	0.226
	(0.043)	(0.269)	(0.494)
Curricular.units.2nd.sem..enrolled.	-0.076***	-0.445***	-0.792***
	(0.009)	(0.084)	(0.158)
Curricular.units.2nd.sem..evaluations.	-0.011***	-0.051***	-0.091**
	(0.002)	(0.019)	(0.036)
Curricular.units.2nd.sem..approved.	0.115***	0.494***	0.903***
	(0.006)	(0.038)	(0.072)
Curricular.units.2nd.sem..grade.	-0.003	0.081***	0.174***
	(0.002)	(0.028)	(0.054)
Curricular.units.2nd.sem..without.evaluations.zero1	-0.019	-0.157	-0.220
	(0.028)	(0.170)	(0.311)
Unemployment.rate	-0.003	-0.043***	-0.075**
	(0.002)	(0.016)	(0.031)
Inflation.rate	0.005	0.040	0.053
	(0.004)	(0.027)	(0.051)
GDP	0.001	-0.030	-0.052
	(0.003)	(0.019)	(0.035)
Constant	1.345***	0.136	0.097
	(0.108)	(0.800)	(1.513)
Observations		3,490	3,490
Log Likelihood		-791.161	-785.683
Akaike Inf. Crit.		1,746.322	1,735.365
Note:		*p<0.1; **p<0.05; ***p<0.01	

The results for all models are similar. Coefficients' are comparable in signs and significance levels. However, we have to remember that none of them fulfil necessary assumptions so that interpreting the results would be inappropriate.

We choose the model with the lowest Akaike Information Criterion (logit) and we proceed with General-to-Specific procedure. The procedure required 20 iterations to obtain a model with all significant variables. After each iteration we use Likelihood ratio test to check if variables that were selected in each iteration are jointly insignificant. As an example, we show the Likelihood ratio test output for the transition from iteration 5 to iteration 15.

```
Likelihood ratio test

Model 1: Target ~ Application.mode.2 + Application.mode.15 + Application.mode.17 +
    Application.mode.39 + Application.mode.44 + Application.mode.51 +
    Course.2 + Course.3 + Course.4 + Course.5 + Daytime.evening.attendance..1 +
    Previous.qualification.3 + Father.s.qualification.9 + Father.s.qualification.39
+
    Mother.s.occupation.7 + Mother.s.occupation.8 + Father.s.occupation.3 +
    Father.s.occupation.4 + Father.s.occupation.7 + Father.s.occupation.8 +
    Father.s.occupation.9 + Admission.grade + Displaced.1 + Debtor.1 +
    Tuition.fees.up.to.date.1 + Gender.1 + Scholarship.holder.1 +
    Curricular.units.1st.sem..credited..zero + Curricular.units.1st.sem..enrolled. +
    Curricular.units.1st.sem..approved.                                          +
Curricular.units.1st.sem..without.evaluations. +
    Curricular.units.2nd.sem..enrolled. + Curricular.units.2nd.sem..evaluations. +
    Curricular.units.2nd.sem..approved. + Curricular.units.2nd.sem..grade. +
    Unemployment.rate + GDP
Model 2: Target ~ Application.mode.15 + Application.mode.17 + Application.mode.39 +
    Application.mode.44 + Application.mode.51 + Course.2 + Course.3 +
    Course.4 + Previous.qualification.3 + Mother.s.qualification.34 +
    Father.s.qualification.9 + Father.s.qualification.39 + Mother.s.occupation.8 +
    Father.s.occupation.4 + Father.s.occupation.9 + Displaced.1 +
    Debtor.1 + Tuition.fees.up.to.date.1 + Gender.1 + Scholarship.holder.1 +
    Curricular.units.1st.sem..credited..zero + Curricular.units.1st.sem..enrolled. +
    Curricular.units.1st.sem..approved. + Curricular.units.2nd.sem..enrolled. +
    Curricular.units.2nd.sem..evaluations. + Curricular.units.2nd.sem..approved. +
    Curricular.units.2nd.sem..grade. + Unemployment.rate
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  38 -803.09
2  29 -805.20 -9 4.2117     0.8969
```

We cannot reject the null hypothesis for the test above (p-value=0.8969 with the confidence level of 0.05) that these two models are the same (coefficients for selected variables are equal

to zero). We are allowed to restrict model 5 to model 15. We show output for interim iterations (5th and 15th iteration) along with the final model below. All variables for the final model are statistically significant.

	Dependent variable:		
	Iter. 5	Iter. 15	Final
Application.mode.21	-1.563**		
	(0.759)		
Application.mode.151	2.366**	2.552**	2.689***
	(1.030)	(1.023)	(1.031)
Application.mode.171	-0.434**	-0.442***	-0.349**
	(0.171)	(0.170)	(0.163)
Application.mode.391	-0.411*	-0.448**	
	(0.213)	(0.205)	
Application.mode.441	0.712**	0.795**	1.002***
	(0.352)	(0.345)	(0.339)
Application.mode.511	-0.676*	-0.722**	
	(0.368)	(0.367)	
Course.21	-1.317***	-1.111***	-1.117***
	(0.242)	(0.192)	(0.189)
Course.31	-1.372***	-1.162***	-1.197***
	(0.301)	(0.251)	(0.249)
Course.41	-1.037***	-1.009***	-1.000***
	(0.232)	(0.195)	(0.194)
Course.51	-0.240		
	(0.255)		
Daytime.evening.attendance..11	0.456*		
	(0.265)		
Previous.qualification.31	-0.894***	-0.860***	-0.709**
	(0.295)	(0.291)	(0.289)
Mother.s.qualification.341		-1.398***	-1.359***
		(0.450)	(0.439)
Father.s.qualification.91	0.323**	0.275**	0.307**
	(0.140)	(0.137)	(0.136)
Father.s.qualification.391	-1.380	-1.362	
	(0.963)	(0.953)	

Mother.s.occupation.71	-0.312		
	(0.289)		
Mother.s.occupation.81	-0.867	-0.988	
	(0.682)	(0.672)	
Father.s.occupation.31	-0.237		
	(0.235)		
Father.s.occupation.41	-0.575**	-0.500**	-0.494**
	(0.240)	(0.224)	(0.222)
Father.s.occupation.71	-0.214		
	(0.220)		
Father.s.occupation.81	-0.238		
	(0.277)		
Father.s.occupation.91	-0.405**	-0.356**	-0.318**
	(0.174)	(0.153)	(0.152)
Admission.grade	0.007		
	(0.005)		
Displaced.11	-0.259*	-0.243	
	(0.152)	(0.148)	
Debtor.11	-0.997***	-0.964***	-0.954***
	(0.235)	(0.232)	(0.231)
Tuition.fees.up.to.date.11	2.662***	2.743***	2.722***
	(0.286)	(0.284)	(0.282)
Gender.11	-0.391***	-0.377***	-0.374***
	(0.144)	(0.142)	(0.141)
Scholarship.holder.11	0.903***	0.881***	0.905***
	(0.170)	(0.168)	(0.166)
Curricular.units.1st.sem..credited..zero1	1.000***	1.047***	1.157***
	(0.251)	(0.246)	(0.239)
Curricular.units.1st.sem..enrolled.	-0.483***	-0.475***	-0.462***
	(0.153)	(0.150)	(0.150)
Curricular.units.1st.sem..approved.	0.555***	0.569***	0.569***
	(0.077)	(0.075)	(0.075)
Curricular.units.1st.sem..without.evaluations.1	-0.347		
	(0.288)		
Curricular.units.2nd.sem..enrolled.	-0.767***	-0.766***	-0.791***
	(0.152)	(0.147)	(0.148)
Curricular.units.2nd.sem..evaluations.	-0.112***	-0.108***	-0.103***
	(0.028)	(0.028)	(0.028)
Curricular.units.2nd.sem..approved.	0.880***	0.856***	0.871***

	(0.068)	(0.066)	(0.066)
Curricular.units.2nd.sem..grade.	0.166***	0.160***	0.155***
	(0.036)	(0.035)	(0.035)
Unemployment.rate	-0.055**	-0.061**	-0.059**
	(0.028)	(0.026)	(0.026)
GDP	-0.034		
	(0.033)		
Constant	-2.416***	-1.366**	-1.771***
	(0.892)	(0.552)	(0.529)
Observations	3,490	3,490	3,490
Log Likelihood	-803.094	-805.200	-811.871
Akaike Inf. Crit.	1,682.188	1,668.400	1,671.741
Note:		*p<0.1; **p<0.05; ***p<0.01	

With a help of General-to-Specific procedure we obtain the model with all variables statistically significant. Now we have to test if the model fulfils the logit model assumptions. First, to check the form of the model we use an alternative of the RESET test for binary choice model, the Linktest. P-value for the yhat is close to zero, p-value for yhat2 equals 0.00745 with the confidence level of 0.05. Because yhat2 is significant we conclude that the form of the model is incorrect.

First form of the model with all significant variables does not pass the Linktest. It means that the form of the model is incorrect and we probably have omitted some relevant variables. We can fix this issue by creating new variables or transforming the existing ones. To correct the form of the model we add two interactions: *Curricular.units.1st.sem..approved.*Curricular.units.2nd.sem..approved.* and *Curricular.units.1st.sem..credited..zero*Scholarship.holder.1*, remove one correlated variable *Debtor* and one variable that in the meantime becomes insignificant *Previous.qualification.3*. Then we again check if the form of the model is appropriate. This time, corrected formula of the model do pass the the Linktest. P-value for the yhat is close to zero, p-value for yhat2 equals 0.0616 with the confidence level of 0.05. We can conclude that the form of the model is correct.

Below we show the model obtained with General-to-Specific procedure (model 1) and its modification with the correct form (final model, model 2).

Model obtained from GTS procedure and its correct form modification

	Dependent variable:
Curricular.units.1st.sem..approved.	0.166***
	(0.036)
Unemployment.rate	-0.055**
	(0.028)
GDP	-0.034
	(0.033)
Constant	-2.416***
	(0.892)

	(1)	(2)
Application.mode.151	2.689***	2.526**
	(1.031)	(1.004)
Application.mode.171	-0.349**	-0.328**
	(0.163)	(0.163)
Application.mode.441	1.002***	0.888**
	(0.339)	(0.349)
Course.21	-1.117***	-1.150***
	(0.189)	(0.190)
Course.31	-1.197***	-1.125***
	(0.249)	(0.245)
Course.41	-1.000***	-1.030***
	(0.194)	(0.195)
Previous.qualification.31	-0.709**	
	(0.289)	
Mother.s.qualification.341	-1.359***	-1.323***
	(0.439)	(0.434)
Father.s.qualification.91	0.307**	0.301**
	(0.136)	(0.136)
Father.s.occupation.41	-0.494**	-0.483**
	(0.222)	(0.223)
Father.s.occupation.91	-0.318**	-0.329**
	(0.152)	(0.151)
Debtor.11	-0.954***	
	(0.231)	
Tuition.fees.up.to.date.11	2.722***	3.111***
	(0.282)	(0.274)
Gender.11	-0.374***	-0.332**
	(0.141)	(0.140)
Scholarship.holder.11	0.905***	2.330***
	(0.166)	(0.669)
Curricular.units.1st.sem..credited..zero1	1.157***	1.258***
	(0.239)	(0.263)
Curricular.units.1st.sem..enrolled.	-0.462***	-0.529***
	(0.150)	(0.159)
Curricular.units.1st.sem..approved.	0.569***	0.958***
	(0.075)	(0.108)
Curricular.units.2nd.sem..enrolled.	-0.791***	-0.825***
	(0.148)	(0.149)
Curricular.units.2nd.sem..evaluations.	-0.103***	-0.093***
	(0.028)	(0.028)
Curricular.units.2nd.sem..approved.	0.871***	1.339***
	(0.066)	(0.112)
Curricular.units.2nd.sem..grade.	0.155***	
	(0.035)	

Unemployment.rate	-0.059**	-0.073***
	(0.026)	(0.026)
Curricular.units.1st.sem..approved.:Curricular.units.2nd.sem..approved.		-0.068***
		(0.014)
Scholarship.holder.11:Curricular.units.1st.sem..credited..zero1		-1.560**
		(0.690)
Constant	-1.771***	-2.191***
	(0.529)	(0.539)
Observations	3,490	3,490
Log Likelihood	-811.871	-815.581
Akaike Inf. Crit.	1,671.741	1,677.162
Note:	*p<0.1; **p<0.05; ***p<0.01	

Next, we check if the model has an appropriate form. In order to do this, we employ Hosmer-Lemeshow and Osius-Rojek tests. P-value for the Hosmer-Lemeshow test is close to zero. It suggests that our model is inappropriate, however because the Hosmer-Lemeshow test has limitations we employ the Osius-Rojek goodness of fit test. P-value for the Osius-Rojek equals 0.95 with the confidence level of 0.05. We cannot reject the null hypothesis that the model is appropriate. We conclude that the model is appropriate.

We also use pseudo R^2 statistics to measure the performance of the final model. McKelvey and Zavoina's R^2 equals to 0.8619838, which means that if one observes a latent variable, our model explains 86% of its total variance. Count equals 0.9131805, which means that 91% of observations are predicted correctly. McFadden's R^2 is equal to 0.6479259, we cannot interpret its value.

With the correct model we can move to interpreting its results. For logit and probit models we are not allowed to interpret coefficients qualitatively, we can only interpret their signs.

By analysing the output above we can say that:

- if one student applied an international student (Application.mode.15), then probability of a student being a Graduate increases in comparison to base level (applied in 1st phase)
- if one student applied in the 2nd phase - general contingent (Application.mode.17), then probability of a student being a Graduate decreases in comparison to base level (applied in 1st phase)
- if one student applied with technological specialisation diploma (Application.mode.44) , then probability of a student being a Graduate increases in comparison to base level (applied in 1st phase)
- if one student studies Arts/Social studies (Course.2), then probability of a student being a Graduate decreases in comparison to base level (studying Administration)

- if one student studies Biology/Nature (Course.3), then probability of a student being a Graduate decreases in comparison to base level (studying Administration)
- if one student studies Economics (Course.4), then probability of a student being a Graduate decreases in comparison to base level (studying Administration)
- if one student's mother's qualification is unknown (Mother.s.qualification.34), then probability of a student being a Graduate decreases in comparison to base level (mother with Secondary Education - 12th Year of Schooling or Eq. (finished))
- if one student's father's qualification is 12th Year of Schooling - Not Completed (Father.s.qualification.9)), then probability of a student being a Graduate increases in comparison to base level (father with Secondary Education - 12th Year of Schooling or Eq. (finished))
- if one student's father's work in administrative staff (Father.s.occupation.4), then probability of a student being a Graduate decreases in comparison to base level (father being a Student)
- if one student's father is an unskilled worker (Father.s.occupation.9), then probability of a student being a Graduate decreases in comparison to base level (father being a Student)
- if one student pays tuition on time (Tuition.fees.up.to.date.1), then probability of a student being a Graduate increases in comparison to base level (student do not pay in time)
- if one student is a male (Gender.1), then probability of a student being a Graduate decreases by in comparison to base level (is a female)
- if one student holds a scholarship (Scholarship.holder.1), then probability of a student being a Graduate increases in comparison to base level (does not hold scholarship)
- if one student's number of curricular units credited in 1st semester is zero (Curricular.units.1st.sem..credited..zero), then probability of a student being a Graduate increases in comparison to base level (student's number of curricular units credited in the 1st semester is not zero)
- increasing the number of curricular units enrolled in the 1st semester negatively influences the probability of student being a Graduate
- increasing the number of curricular units approved in the 1st semester positively influences the probability of student being a Graduate
- increasing the number of curricular units enrolled in the 2nd semester negatively influences the probability of a student being a Graduate
- increasing the number of curricular evaluations in the 2nd semester negatively influences the probability of a student being a Graduate
- increasing the number of curricular units approved in the 2nd semester positively influences the probability of a student being a Graduate
- increasing the unemployment rate negatively influences the probability of a student being a Graduate

As a bonus, we go through exactly the same procedure for the probit model as we did for logit (General-to-Specific and correcting form of the model). We obtain a valid probit model and we compare it with the logit.

Comparison of probit and logit models

	Dependent variable:	
	Target	
	probit	logistic
	(1)	(2)
Application.mode.151	1.407***	2.526**
	(0.494)	(1.004)
Application.mode.171		-0.328**
		(0.163)
Application.mode.441	0.605***	0.888**
	(0.182)	(0.349)
Course.21	-0.638***	-1.150***
	(0.098)	(0.190)
Course.31	-0.719***	-1.125***
	(0.130)	(0.245)
Course.41	-0.556***	-1.030***
	(0.101)	(0.195)
Mother.s.qualification.341	-0.808***	-1.323***
	(0.244)	(0.434)
Father.s.qualification.91	0.194***	0.301**
	(0.072)	(0.136)
Father.s.occupation.41	-0.279**	-0.483**
	(0.118)	(0.223)
Father.s.occupation.91	-0.179**	-0.329**
	(0.081)	(0.151)
Tuition.fees.up.to.date.11	1.722***	3.111***
	(0.146)	(0.274)
Gender.11		-0.332**
		(0.140)
Scholarship.holder.11	1.376***	2.330***
	(0.342)	(0.669)
Curricular.units.1st.sem..credited..zero1	0.800***	1.258***
	(0.133)	(0.263)
Curricular.units.1st.sem..enrolled.	-0.168**	-0.529***
	(0.084)	(0.159)
Curricular.units.1st.sem..approved.	0.328***	0.958***
	(0.041)	(0.108)
Curricular.units.2nd.sem..enrolled.	-0.491***	-0.825***

	(0.084)	(0.149)
Curricular.units.2nd.sem..evaluations.	-0.060***	-0.093***
	(0.015)	(0.028)
Curricular.units.2nd.sem..approved.	0.476***	1.339***
	(0.035)	(0.112)
Curricular.units.2nd.sem..grade.	0.068***	
	(0.018)	
Unemployment.rate	-0.301**	-0.073***
	(0.124)	(0.026)
Unemployment.rate_2	0.011**	
	(0.005)	
Curricular.units.1st.sem..approved.:Curricular.units.2nd.sem..approved.		-0.068***
		(0.014)
Scholarship.holder.11:Curricular.units.1st.sem..credited..zero1	-0.936***	-1.560**
	(0.353)	(0.690)
Constant	0.051	-2.191***
	(0.751)	(0.539)
Observations	3,490	3,490
Log Likelihood	-827.706	-815.581
Akaike Inf. Crit.	1,699.413	1,677.162
Note:	*p<0.1; **p<0.05; ***p<0.01	

We can observe that the results are very similar. All variables have the same sign and close estimations. One of the issues of the probit model could be the fact that its Constant variable is insignificant. It can also be noticed that Akaike Information Criterion indicates that a better model is the logit model, which is in line with what we have obtained on the raw models with all variables included, at the very beginning of the analysis.

RESULTS

Out of the 37 variables, 15 of the old variables (prior to encoding the data into dummy variables) were identified as significant. The variables used in the final logit model are as follows:

- Application.mode.15 (International student)
- Application.mode.17 (2nd phase - general contingent)
- Application.mode.44 (Technological specialization diploma holders)
- Course.2 (Arts/Social studies)
- Course.3 (Biology/Nature)
- Course.4 (Economics),
- Mother.s.qualification.34 (Unknown)
- Father.s.qualification.9 (12th Year of Schooling - Not Completed)
- Father.s.occupation.4 (Administrative staff)

- Father.s.occupation.9 (Unskilled Worker)
- Tuition.fees.up.to.date.1
- Gender.1
- Scholarship.holder.1
- Curricular.units.1st.sem..credited..zero
- Curricular.units.1st.sem..enrolled
- Curricular.units.1st.sem..approved
- Curricular.units.2nd.sem..enrolled
- Curricular.units.2nd.sem..evaluations
- Curricular.units.2nd.sem..approved
- Unemployment.rate.

To evaluate the results quantitatively the marginal effects were calculated.

```
Call:
logitmfx(formula = logit_main_model, data = data)

Marginal Effects:
                                                        dF/dx  Std. Err.       z                 P>|z|
Application.mode.151                                 0.4401471  0.0779050  5.6498      0.00000001606380 ***
Application.mode.171                                -0.0812318  0.0398471 -2.0386             0.0414912 *
Application.mode.441                                 0.2125583  0.0756693  2.8090             0.0049689 **
Course.21                                           -0.2707912  0.0405158 -6.6836      0.00000000002331 ***
Course.31                                           -0.2585030  0.0478761 -5.3994      0.00000006685606 ***
Course.41                                           -0.2436247  0.0412936 -5.8998      0.00000000363891 ***
Mother.s.qualification.341                           -0.2874065  0.0723380 -3.9731      0.00007094211261 ***
Father.s.qualification.91                            0.0751054  0.0337346  2.2264             0.0259901 *
Father.s.occupation.41                              -0.1181179  0.0525776 -2.2465             0.0246692 *
Father.s.occupation.91                              -0.0815004  0.0371939 -2.1912             0.0284349 *
Tuition.fees.up.to.date.11                           0.5245445  0.0277940 18.8726 < 0.00000000000000022 ***
Gender.11                                           -0.0825074  0.0346496 -2.3812             0.0172567 *
Scholarship.holder.11                                0.5047214  0.1068828  4.7222      0.00000233315408 ***
Curricular.units.1st.sem..credited..zero1            0.2859744  0.0519885  5.5007      0.00000003782228 ***
Curricular.units.1st.sem..enrolled.                 -0.1320555  0.0396597 -3.3297             0.0008693 ***
Curricular.units.1st.sem..approved.                  0.2391756  0.0263487  9.0773 < 0.00000000000000022 ***
Curricular.units.2nd.sem..enrolled.                 -0.2061071  0.0371283 -5.5512      0.00000002836926 ***
Curricular.units.2nd.sem..evaluations.              -0.0232032  0.0069647 -3.3315             0.0008637 ***
Curricular.units.2nd.sem..approved.                  0.3343032  0.0270873 12.3417 < 0.00000000000000022 ***
Unemployment.rate                                   -0.0182300  0.0065065 -2.8018             0.0050818 **
Curricular.units.1st.sem..approved.:Curricular.units.2nd.sem..approved. -0.0170209  0.0034346 -4.9557      0.00000072055869 ***
Scholarship.holder.11:Curricular.units.1st.sem..credited..zero1         -0.3555359  0.1334209 -2.6648             0.0077041 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

dF/dx is for discrete change for the following variables:

 [1] "Application.mode.151"                                  "Application.mode.171"
 [3] "Application.mode.441"                                  "Course.21"
 [5] "Course.31"                                             "Course.41"
 [7] "Mother.s.qualification.341"                            "Father.s.qualification.91"
```

```
 [9] "Father.s.occupation.41"                                          "Father.s.occupation.91"
[11] "Tuition.fees.up.to.date.11"                                      "Gender.11"
[13] "Scholarship.holder.11"                                           "Curricular.units.1st.sem..credited..zero1"
[15] "Scholarship.holder.11:Curricular.units.1st.sem..credited..zero1"
```

The marginal effects can be interpreted in the following way:

- if one student applied as an international student (Application.mode.15), then the probability of a student being a Graduate increases by 44 percentage points in comparison to base level (applied in 1st phase)
- if one student applied in the 2nd phase - general contingent (Application.mode.17), then probability of a student being a Graduate decreases by 8 p.p. in comparison to base level (applied in 1st phase)
- if one student applied with technological specialisation diploma (Application.mode.44) , then probability of a student being a Graduate increases by 21 p.p. in comparison to base level (applied in 1st phase)
- if one student studies Arts/Social studies (Course.2), then probability of a student being a Graduate decreases by 27 p.p. in comparison to base level (studying Administration)
- if one student studies Biology/Nature (Course.3), then probability of a student being a Graduate decreases by 26 p.p. in comparison to base level (studying Administration)
- if one student studies Economics (Course.4), then probability of a student being a Graduate decreases by 24 p.p. in comparison to base level (studying Administration)
- if one student's mother's qualification is unknown (Mother.s.qualification.34), then probability of a student being a Graduate decreases by 29 p.p. in comparison to base level (mother with Secondary Education - 12th Year of Schooling or Eq. (finished))
- if one student's father's qualification is 12th Year of Schooling - Not Completed (Father.s.qualification.9)), then probability of a student being a Graduate increases by 8 p.p. in comparison to base level (father with Secondary Education - 12th Year of Schooling or Eq. (finished))
- if one student's father's work in administrative staff (Father.s.occupation.4), then probability of a student being a Graduate decreases by 12 p.p. in comparison to base level (father being a Student)
- if one student's father is an unskilled worker (Father.s.occupation.9), then probability of a student being a Graduate decreases by 8 p.p. in comparison to base level (father being a Student)
- if one student pays tuition on time (Tuition.fees.up.to.date.1), then probability of a student being a Graduate increases by 52 p.p. in comparison to base level (student do not pay in time)
- if one student is a male (Gender.1), then probability of a student being a Graduate decreases by 8 p.p. in comparison to base level (is a female)
- if one student holds a scholarship (Scholarship.holder.1), then probability of a student being a Graduate increases by 50 p.p. in comparison to base level (does not hold scholarship)

- if one student's number of curricular units credited in 1st semester is zero (Curricular.units.1st.sem..credited..zero), then probability of a student being a Graduate increases by 29 p.p. in comparison to base level (student's number of curricular units credited in the 1st semester is not zero)
- if one student's number of curricular units enrolled in the 1st semester increases by 1 unit (Curricular.units.1st.sem..enrolled.), then probability of a student being a Graduate decreases by 13 p.p. in comparison to average level (average level: 5.946705)
- if one student's number of curricular units approved in the 1st semester increases by 1 unit (Curricular.units.1st.sem..approved.), then probability of a student being a Graduate increases by 24 p.p. in comparison to average level (average level: 4.639542)
- if one student's number of curricular units enrolled in the 2nd semester increases by 1 unit (Curricular.units.2nd.sem..enrolled.), then probability of a student being a Graduate decreases by 21 p.p. in comparison to average level (average level: 6.039542)
- if one student's number of curricular units evaluations in the 2nd semester increases by 1 unit (Curricular.units.2nd.sem..evaluations.), then probability of a student being a Graduate decreases by 2 p.p. in comparison to average level (average level: 7.733238)
- if one student's number of curricular units approved in the 2nd semester increases by 1 unit (Curricular.units.2nd.sem..approved.), then probability of a student being a Graduate increases by 33 p.p. in comparison to average level (average level: 4.461032)
- if the unemployment rate increases by 1 percentage point (Unemployment.rate.), then probability of a student being a Graduate decreases by 2 p.p. in comparison to average level (average level: 11.71544)

We cannot interpret estimations for interactions.

H1: there is a negative correlation between students attending social study courses and dropout rate.

H2: variables corresponding to prior academic achievements will not be significantly associated with students' dropout rate.

H3: there is a negative correlation between parents' academic achievements and children being school dropouts.

Based on the findings we can conclude that H1 cannot be supported. The student, who studies Arts/Social studies (Course.2), has the probability of being a Graduate decreases by 27 p.p. in comparison to base level (studying Administration). Moreover, students who take courses assigned to other initiatives also have a negative probability of dropping out vs administration studies. This indicates that administrative studies students have the higher probability of dropping out and the difference between other course categories seems to be insignificant.

The second hypothesis H2 was confirmed by the data. None of the dummy variables corresponding to prior academic achievements had smaller p-value < 0.05, indicating that this factor is insignificant when determining future dropping out rates.

Based on the results we can only conclude that if the father drops out in 12th grade (last year in high school) then the student is less likely to drop in comparison to a student whose father graduated any of grades in secondary or high school. Which could potentially support the hypothesis if we would exclude the fathers who graduated year 12. On the other hand, when analysing the mother's qualification, it can be observed that only the “unknown” variable was significant. If unknown means less educational experience than the available options then the H3 would be supported for mothers’ academic achievements. Unfortunately, there was no explanation of the data of the unknown option provided by the creators.

It can also be stated that students who identify as males have a higher probability of dropping out. On the contrary, students who are on scholarships or pay their tuition on time have a higher probability of becoming graduates. Among significant variables, there are more variables connected to the first semester than the second one, which might indicate that the first semester plays a larger role in determining whether a student drops out. This could potentially be connected to the psychological phenomenon called goal gradient argued by Kurt Lewin. The longer the students attend the course the higher their commitment level. The only significant socio-economic variable was unemployment rate.

FINDINGS AND FUTURE EXTENSION

It was found that 15 variables have a significant effect on students’ dropout rate, which are: Application.mode, Course category, Mother’s qualification, Father.s.qualification, Father’s occupation, Tuition.fees.up.to.date, Gender, Scholarship holder, Curricular.units.1st.sem (credited, enrolled, approved), Curricular.units.2nd.sem (enrolled, evaluations, approved) and Unemployment.rate. It was concluded that the administrative studies had the largest dropout probability out of all the course categories. Prior academic achievements of the student seem not to be a factor of high value when assessing the probability of the student to graduate. However, the results also indicate that if the father drops out in 12th grade (last year in high school) then the student is less likely to drop in comparison to a student whose father graduated any of grades in secondary or high school. On top, students who identify as males have a higher probability of dropping out, and students who are on scholarships or pay their tuition on time have a higher probability of becoming graduates. Lastly, it is possible that the number of student dropouts is more dependent on the first semester than the second one, but more research is needed. As an extension of the research it would also be interesting to examine studies performed in different regions as most of the studies have been conducted in the USA or Europe.

BIBLIOGRAPHY

1. Alban, Mayra & Mauricio, David. (2019). Predicting University Dropout through Data Mining: A Systematic Literature. 10.17485/ijst/2019/v12i4/139729.
2. Kehm, Barbara & Larsen, Malene & Sommersel, Hanna. (2019). Student dropout from universities in Europe: A review of empirical literature. Hungarian Educational Research Journal. 9. 147-164. 10.1556/063.9.2019.1.18.
3. Laato, S.; Lipponen, E.; Salmento, H.; Vilppu, H. and Murtonen, M. (2019). Minimizing the Number of Dropouts in University Pedagogy Online Courses. In Proceedings of the 11th International Conference on Computer Supported Education - Volume 1: CSEDU, ISBN 978-989-758-367-4; ISSN 2184-5026, pages 587-596. DOI: 10.5220/0007686005870596
4. Márquez-Ramos, Laura & Mourelle, Estefanía. (2018). Education and economic growth: an empirical analysis of nonlinearities. Revista de Economia Aplicada. 26. 1-28. 10.1108/AEA-06-2019-0005.

APPENDIX

Linear probability model RESET test

```
        RESET test


data:  linear_probity_raw
RESET = 288.63, df1 = 2, df2 = 3406, p-value < 0.00000000000000022
```

Linear probability model Breush-Pagan test

```
        studentized Breusch-Pagan test


data:  linear_probity_raw.residuals ~ ď.żMarital.status.2 + ď.żMarital.status.3 +
Application.mode.2 + Application.mode.7 + Application.mode.15 +
Application.mode.17 + Application.mode.18 + Application.mode.39 +
Application.mode.43 + Application.mode.44 + Application.mode.51 +
Application.mode.53 + Application.order.2 + Application.order.3 +
Application.order.4 + Application.order.5 + Application.order.6 +     Course.2 +
Course.3 + Course.4 + Course.5 + Daytime.evening.attendance..1 +
Previous.qualification.3 + Previous.qualification.9 + Previous.qualification.39 +
Previous.qualification..grade. + Nacionality.Portuguese +
Mother.s.qualification.3 + Mother.s.qualification.9 + Mother.s.qualification.34 +
Mother.s.qualification.37 + Mother.s.qualification.39 + Father.s.qualification.3 +
Father.s.qualification.9 + Father.s.qualification.34 + Father.s.qualification.37 +
Father.s.qualification.39 + Mother.s.occupation.1 + Mother.s.occupation.2 +
Mother.s.occupation.3 + Mother.s.occupation.4 + Mother.s.occupation.5 +
Mother.s.occupation.6 + Mother.s.occupation.7 + Mother.s.occupation.8 +
Mother.s.occupation.9 + Mother.s.occupation.10 + Father.s.occupation.1 +
Father.s.occupation.2 + Father.s.occupation.3 + Father.s.occupation.4 +
Father.s.occupation.5 + Father.s.occupation.6 + Father.s.occupation.7 +
Father.s.occupation.8 + Father.s.occupation.9 + Father.s.occupation.10 +
Admission.grade + Displaced.1 + Educational.special.needs.1 +     Debtor.1 +
Tuition.fees.up.to.date.1 + Gender.1 + Scholarship.holder.1 +
Age.at.enrollment.20.29 + Age.at.enrollment.30. +
Curricular.units.1st.sem..credited..zero +     Curricular.units.1st.sem..enrolled.
+ Curricular.units.1st.sem..evaluations. +     Curricular.units.1st.sem..approved.
+ Curricular.units.1st.sem..grade. +
Curricular.units.1st.sem..without.evaluations. +
Curricular.units.2nd.sem..credited..zero +     Curricular.units.2nd.sem..enrolled.
+ Curricular.units.2nd.sem..evaluations. +     Curricular.units.2nd.sem..approved.
+ Curricular.units.2nd.sem..grade. +
Curricular.units.2nd.sem..without.evaluations.zero + Unemployment.rate +
Inflation.rate + GDP
BP = 359.88, df = 81, p-value < 0.00000000000000022
```

Linktest for the first model with all significant variables

```
Call:
```

```
glm(formula = y ~ yhat + yhat2, family = binomial(link = model$family$link))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7667  -0.0143   0.2320   0.3830   3.9514

Coefficients:
            Estimate Std. Error z value             Pr(>|z|)
(Intercept)  0.10129    0.08399   1.206              0.22779
yhat         1.07647    0.05407  19.911 < 0.0000000000000002 ***
yhat2       -0.04797    0.01792  -2.676              0.00745 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4633.0  on 3489  degrees of freedom
Residual deviance: 1615.6  on 3487  degrees of freedom
AIC: 1621.6

Number of Fisher Scoring iterations: 9
```

Linktest for model with corrected form with interactions (final model)

```
Call:
glm(formula = y ~ yhat + yhat2, family = binomial(link = model$family$link))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7432  -0.0180   0.2265   0.3899   4.0428

Coefficients:
            Estimate Std. Error z value            Pr(>|z|)
(Intercept)  0.07364    0.08506   0.866              0.3867
yhat         1.04753    0.05076  20.638 <0.0000000000000002 ***
yhat2       -0.03285    0.01758  -1.869              0.0616 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4633.0  on 3489  degrees of freedom
Residual deviance: 1627.3  on 3487  degrees of freedom
AIC: 1633.3

Number of Fisher Scoring iterations: 9
```

Final logit Hosmer-Lemeshow and Osius-Rojek tests output

	test	stat	val	df	pVal
1:	HL	chiSq	37.23442804	8	0.00001042063076546915494796120 8
2:	mHL	F	15.19837849	9	0.00000000000000000000001874763
3:	OsRo	Z	0.06533282	NA	0.94790901300743013457861252391 0
4:	SstPgeq0.5	Z	2.72178299	NA	0.00649307644768201859541978393 5
5:	SstPl0.5	Z	0.87741083	NA	0.38026353110772487253754547964 5
6:	SstBoth	chiSq	8.17795244	2	0.01675637964179456373958743142 8
7:	SllPgeq0.5	chiSq	7.16565079	1	0.00743127257876197676311758044 7
8:	SllPl0.5	chiSq	0.92360832	1	0.33652932664737272849819760267 6
9:	SllBoth	chiSq	7.58219202	2	0.02257085036392890778356168368 6

Final logit Pseudo R2 statistics

McFadden	Adj.McFadden	Cox.Snell	Nagelkerke	McKelvey.Zavoina	Effron	Count	Adj.Count	AIC
0.6479259	0.6375655	0.5768920	0.7850332	0.8619838	0.7119910	0.9131805	0.7711480	1677.1617942

Corrected.AIC
1677.4803170