Analysis of dropout in engineering BSc using logistic mixed-effect models

Analisi dell'abbandono universitario nelle lauree di primo livello di ingegneria utilizzando modelli logistici a effetti misti

Luca Fontana and Anna Maria Paganoni

Abstract The main goal of this report is to apply a logistic mixed-effect model to analyse the relationship between the success probability in getting the BSc engineering degree in Politecnico di Milano and different sets of covariates, using as grouping factor the engineering programme attended. The dataset of interest contains detailed information about more than 18,000 students enrolled in BSc from 2010 to 2013. This analysis is performed within the Student Profile for Enhancing Tutoring Engineering (SPEET) ERASMUS+ project that involves Politecnico di Milano and five other european engineering universities, aimed at opening a new perspective to university tutoring systems.

Abstract *L'obiettivo di questo report è quello di applicare un modello logistico a effetti misti per analizzare la possibile dipendenza tra la probabilità di concludere con successo la Laurea di Primo Livello in Ingegneria al Politecnico di Milano e diversi insiemi di covariate, raggruppando gli studenti per corso di studi frequentato. Il dataset utilizzato contiene informazioni dettagliate riguardo più di 18,000 studenti immatricolati tra il 2010 e il 2013. Questa analisi è parte del progetto Student Profile for Enhancing Tutoring Engineering (SPEET), una collaborazione ERASMUS+ tra il Politecnico di Milano e cinque altri atenei europei di ingegneria.*

Key words: academic data; engineering programmes; university tutoring systems; generalized linear mixed-effects model; dropout prediction.

Luca Fontana
MOX, Dipartimento di Matematica, Politecnico di Milano,
P.za Leonardo da Vinci 32, 20133 Milano (Italy)
e-mail: luca11.fontana@mail.polimi.it

Anna Maria Paganoni
MOX, Dipartimento di Matematica, Politecnico di Milano,
P.za Leonardo da Vinci 32, 20133 Milano (Italy)
e-mail: anna.paganoni@polimi.it

1 Introduction

The present work is a first step of statistical analysis of academic data related to engineering students attending Bachelor of Science Degree in Politecnico di Milano. This analysis is performed within the Student Profile for Enhancing Tutoring Engineering (SPEET) ERASMUS+ project that involves Politecnico di Milano, University of Galati (Romania), Escola d'Enginyeria UAB (Spain), Instituto Politecnico de Braganca (Portugal), Opole University of Technology (Poland) and Universidad de Leon (Spain). The project novelty emerges from the potential synergy among the huge amount of academic data actually existing at the academic departments and the maturity of data science algorithms and tools to analyse and extract information from those data. SPEET project aims to process the academic data to identify different student profiles to provide them with a personal tutoring service [6]. Despite many possibilities we have chosen to start the project by analyzing the distinction between students who complete their programme and those who instead decide to abandon studies [2]. This choice is based on the fact that across all SPEET partners almost a student out of two resigns his engineering studies before obtaining the BSc degree, and this phenomenon marks Politecnico di Milano in a remarkable way. The student profiles we are referring to within the SPEET project scope are:

- *dropout*: careers permanently finished for any reason other than the achievement of the BSc degree;
- *graduate*: careers definitely closed with the achievement of academic qualification, sooner or later.

Mixed-effects models are commonly employed in the analysis of grouped or clustered data, where observations in a cluster cannot reasonably be assumed to be independent of one-another: we use a mixed model in which engineering students are nested within the programmes they are attending.

2 Dataset and Model

The dataset consists of $18,612$ careers that began from A.Y. 2010/2011 to A.Y. 2013/2014, from 19 different engineering programmes at Politecnico di Milano. Collected data include degree information, collateral information regarding the students' background (nationality, previous studies, age, ...) as well as student performance on every subject of his study plan (subject score, subject year, subject semester, ...). The variables we are including in the model are described in table 1.

We now analyze the relationship between the success probability in getting the degree and a set of explainatory variables. Our response variable is the career `status`, a two-level factor we coded as a binary variable:

- `status` $= 1$ for careers definitely closed with the achievement of academic qualification (factor level = *graduate*)

- `status` $= 0$ if the career is permanently finished for any reason other than the achievement of the BSc degree (factor level = *dropout*).

Since those careers belong to 19 different engineering programmes, the choice of a Logit Mixed-Effects Model in which students are nested within programmes is justified [1]. The influence of the programme on the linear predictor is modeled through a single random effect on the intercept. Let y_{ij} denote observation j in group i ($i = 1:19, j = 1:n_i$). The model formulation in extended form is:

$$\begin{aligned} Y_{ij} &\sim Bernoulli(p_{ij}) \\ p_{ij} &= E[Y_{ij}|b_i] = P(y_{ij} = 1|b_i) \\ logit(p_{ij}) &= \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \sum_{h=0}^{p} x_{ijh}\beta_k + b_i \\ b_i &\sim N(0,\sigma^2) \\ b_i, b_{i'} &\text{ are independent for } i \neq i' \end{aligned} \tag{1}$$

where x_{ijh} represents the values of explanatory variables for fixed effects model parameters; $\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \ldots \quad \beta_p]^T$ is the (p+1)-dimensional vector of fixed effects and b_i is the scalar random effect on the intercept of the linear predictor for observations in group i. We can fit and analyze models of this type using the `lme4` library [3] in the open-source statistical software `R` [5].

As SPEET main goal is to analyze the data trail of the student in real time in order to know as in advance as possible which profile the student belongs to, we decided to keep the set of covariates that are available at the time of the enrollment, and three more variables that could be recorded after the first semester of the first year of study. Thus, in the fixed-effect part of the model we consider the covariates that are described in table 1.

3 Results

We decide to randomly split the dataset into training and evaluation subsets, with a ratio of 80% for training and 20% for evaluation. Not all covariates turn out to be significant. We use stepwise backward elimination to obtain a reduced model: regressors `PreviousStudies` and `Nationality` are removed from the full model through this procedure. We can make the following considerations about fixed-effect parameters estimation:

- female students outperform their male counterpart: being a men penalize the log odds by 0.311. This information suggest that female students who decide to enroll in engineering degree are more resolute in their choice.
- The relationship between the linear predictor and the age at the time of the admission is negative. Generally, the aged students may have less time to devote to studies and this may affects their performance ($\hat{\beta} = -0.201$).

Variable	Description	Type of variable
`Sex`	sex	factor (2 levels: M, F)
`Nationality`	nationality	factor (2 levels: Italian, Foreigner)
`PreviousStudies`	high school studies	factor (3 levels: Liceo Scientifico, Istituto Tecnico, Other)
`AdmissionScore`	PoliMi admission test result	real number [0,100]
`AccessToStudiesAge`	age at the beginning of the BSc studies at PoliMi	natural number
`WeightedAvgEval1.1`	weighted average (by ECTS) of the evaluations during the first semester of the first year	real number [18,30]
`AvgAttempts1.1`	average number of attempts to be evaluated for both passed and not passed exams, during the first semester of the first year	real number [0,3]
`TotalCredits1.1`	number of ECTS credits obtained by the student during the first semester of the first year	natural number

Table 1: *Set of used covariates for the GLM model* (1)

- PoliMi admission test has a positive influence on the transformed response: a unit increase in the score improves the log odds by 0.008.
- The weighted average of the evaluations in the first semester shows strong positive effect on the transformed response: this result is reasonable and realistic. A unit increase improves the log odds by 0.060.
- `AverageExamAttempts1.1` has positive effect on the response: this information may suggest that if a student do not give up and tries to complete all his exams during the first semester of the first year, even after failing more than once, he is more likely to end successfully his programme ($\hat{\beta} = 0.243$).
- The effect of regressor `TotalCreditsObtained1.1` is positive as expected: a better performance during the first semester of the first year greatly improves the success probability in getting the degree ($\hat{\beta} = 0.196$).

Parameter	Estimate	P-value
`(Intercept)`	0.664	0.2354
`Sex(male)`	-0.311	0.0002
`AdmissionScore`	0.008	0.0124
`AccessToStudiesAge`	-0.201	4.87×10^{-11}
`WeightedAverageEvaluations1.1`	0.060	$<2 \times 10^{-16}$
`AverageExamAttempts1.1`	0.243	3.49×10^{-7}
`TotalCreditsObtained1.1`	0.196	$<2 \times 10^{-16}$

Table 2: *Fixed-effect coefficient estimates and P-values of model* (1)

As next step we underline the differences among the study programmes. Figure 1 shows the estimated random effects for all 19 groups in the dataset. Many of the 95% confidence intervals for $\hat{b}_i$ do not overlap the vertical line at zero, underlining substantial differences between the programmes. For example, level *Environmental and Land Planning Engineering* has the highest positive effect on the intercept: being a student from this programme improves the log odds by 1.486. On the contrary, studying *Civil Engineering* penalizes the log odds by 1.008.

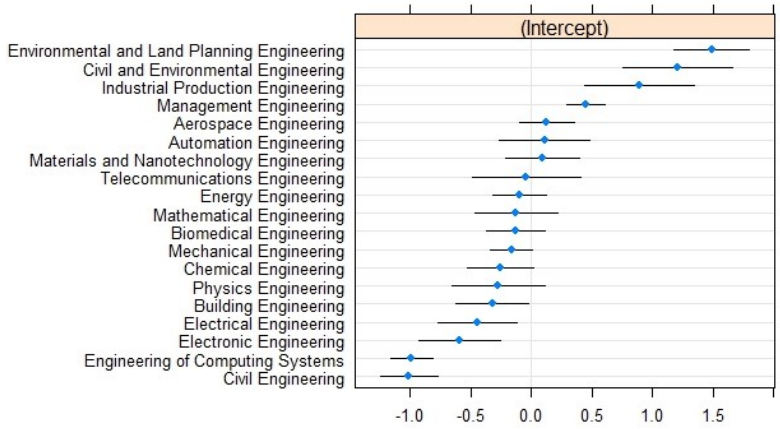

Fig. 1: *Random effect of the degree programme on the intercept in model* (1)

By using a multilevel model we can account for the interdependence of observations by partitioning the total variance into different components of variation due to the various cluster levels in the data. The intraclass correlation is equal to the percentage of variation that is found at the higher level and it is generally called the Variance Partition Coefficient [4]. Using the *latent variable approach* method the VPC is constant across all individuals and it is defined as $VPC = \frac{\sigma_b^2}{\sigma_b^2+\sigma_{lat}^2}$. Since the variance of the standard logistic distribution is $\sigma_{lat}^2 = \pi^2/3$ and the estimated variance is $\hat{\sigma}_b^2 = 0.4245$, the estimated VPC is equal to 0.1143. This means that 11.4% of variation in the response is attributed to the classification by degree type.

As last step, we use this model for classification and prediction of success probability. After fixing the optimal threshold value of $p_0 = 0.6$ using ROC curve analysis, we evaluate the predictive performance of the model by computing average classification indexes. We repeat 20 times the following procedure:

- randomly split the observations in training set (80% of the full dataset) and test set (20% of the full dataset)
- using the training set, fit the logit mixed-effect model and estimate its parameters
- estimate the success probability of observations in the test set and classify them using the optimal threshold value
- build the classification table and compute accuracy, sensitivity and specificity.

At the end of the 20 iterations we compute the average accuracy, sensitivity and specificity and their standard deviation, reported in table 3. High values of accuracy, sensitivity and specificity point to a good performance of the model. In addition, the model performance is notably robust, as highlighted by the low standard deviation of all performance indexes.

Index	**Mean**	**Std deviation**
Accuracy	0.899	0.0045
Sensitivity	0.925	0.0050
Specificity	0.850	0.0089

Table 3: *Performance indexes of a classifier based on the GLM* (1)

4 Conclusions

As far as the SPEET consortium knowledge, this is one of the first experiences of Learning Analytics at university level in Italy. Using predictive analytics we can give our educational institutes insights in future students outcomes: this predictions can be used to change particular programmes and deliver an optimal learning environment for the students.

Other further studies are being conducted within this project, proposing alternative nonparametric modeling in order to test the validity of the mixed-effect model and to analyse advantages and drawbacks of both methods. An immediate step further is extending the student profiling to other SPEET partners and compare the differences (if any) that arise at country level.

References

1. Agresti A., *Categorical Data Analysis*. 2nd ed. Wiley Series in Probability and Statistics. Wiley, 2007.
2. Barbu M., Vilanova R., Lopez Vicario J., Varanda M.J., Alves P., Podpora M., Prada M.A., Moran A., Torrebruno A., Marin S. and R. Tocu R., *SPEET Intellectual Output # 1, Data Mining Tool for Academic Data Exploitation, Literature review and first architecture proposal*, ERASMUS+ KA2 / KA203 (2017).
3. Bates D. *Lme4: Mixed-Effects Modeling With R*. 2010. `http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf`.
4. Goldstein H., Browne W., and Rasbash J.. *Partitioning Variation in Multilevel Models*. In: Understanding Statistics 1.4 (2002), pp. 223-231.
5. R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. `http://www.R-project.org/`
6. *SPEET, proposal for strategic partnerships (proposal narrative)* (2016) `https://www.speet-project.com/the-project`