Trabalho apresentado no XLI CNMAC, Unicamp - Campinas - SP, 2022.

Modelling Academic Performance: A Case Study on Engineering Courses

Hans Rolan E. Mersch Fernandez[1], Carlos Sauer, Jose Rivas , Diego H. Stalder
Facultad de Ingeniería, Universidad Nacional de Asunción, Paraguay

Engineering education is of great importance to train future leaders and innovators in finding solutions for problems arising in our fast changing world. Due to the challenging nature of engineering education, an alarming number of undergraduate engineering students do not move to degree completion in curricular planned timeframes [1]. It is, therefore, necessary to early identify students at risk of failing and design strategies to support them until completion.

Mathematical models can be used to extract information from students grade records (data) to perform predictions of academic performance (see [2, 3]). Classical Machine learning techniques, such as logistic regression (LR) models and multilayer perceptron (MLP) neural networks, can be used to fit the data and identify students at risk of failing. This work applies LR models and MLP neural networks to the academic records of this School.

Engineering students anonymized academic records (190108, in total) from 2012 to 2019 were provided by the Engineering School of the National University of Asunción.

The variables considered in this work were: the course name (461 classes), the engineering career (7 classes), the year when the course was taken, the midterm exams scores (1P$\in \mathbb{R}$ and 2P$\in \mathbb{R}$), a binary variable(2nd Try, true/false) to identify students who already took and failed the course in the past, the Lab/Workshop(L/W $\in \mathbb{R}$) scores, and the final exam results (pass/fail).

The LR is a model that fits the probability $p(x)$ of a discrete outcome given a set of input variables x. The LR models a binary outcome. Let y be a binary variable that can take two values: the student will approve/fail the course. The LR is a non-linear transformation and is expressed as follows:

$$z = \sum_{i=0}^{p} \theta_i x_i = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_p x_p = \theta_0 + X^T \Theta, \quad p(x) = \sigma(z) = \frac{1}{1+e^{-z}}, \tag{1}$$

where $\Theta \in \mathbb{R}^p$ and θ_0 are the model free parameters. Finally a certain threshold should be defined (e.g. 0.5) to predict $y = 1$ when $p(x) \geq 0.5$ and $y = 0$ when $p(x) < 0.5$.

The MLP is a class of feedforward artificial neural network (ANN) which can be interpreted as a generalization of several LR. The output of one single layer is computed as: $y = f(\theta_0 + X^T \Theta)$, where $\Theta \in \mathbb{R}^p$ are the parameters (weights), $\theta_0 \in \mathbb{R}$ the a constant, f is the activation function, eg. sigmoid, relu, etc. Let c_h be the number of neurons of one given layer h, $z_j = f(\theta_0^j + X_j^T \Theta^j)$, where $j \in c_1, c_2, ..., c_h, ..., c_{n+1}$, where $c_h \in [1, n+1]$ and $c_0 = p, c_{n+1} = q$ are respectively the number of input and output variables. Therefore the output of the MLP neural network can be computed as a multivariate function $F(X)$ which maps $X \in \mathbb{R}^p$ into $Y \in \mathbb{R}^q$ (see for more details [4]):

$$y_j = F(x_i) = f\Bigg(\theta_0^j + f\Big(... + f\Big(\theta_0^i + X^T \Theta^i\Big)\Theta^i + ... + \Big)\Theta^j\Bigg) \tag{2}$$

where $j \in 1, ..., q$ and $i \in 1, ..., p$.

[1] hmersch@fiuna.edu.py

Once the models are defined, the next step consists in estimating the parameters to fit the data. It is not known which variables have a stronger influence on the chances of a student passing the course. A set of LR models were defined to analyze their performance using a hold-out method i.e. 80% of records to estimate parameters (the training sample) and 20% to validate the predictions (the test sample). The cost function considered to fit the model parameters was the binary cross-entropy. Then the best LR model was selected and compared to the results obtained by a nonlinear classifier i.e. a MLP neural network. We use the Accuracy Score (ACC) and the Matthews Correlation Coefficient (MCC) to evaluate and compare the models. The ACC is the number of correct predictions made divided by the total number of records in the test sample. The MCC takes into account true and false positives, as well as true and false negatives [5].

Table 1 presents the estimated coefficients for each parameter and their corresponding accuracy. All models have an accuracy higher than 79%. Model 1 has the lowest performance and considers information obtained at the end of the semester e.g. L/W and 2P. Overall, the best performing one was Model 5, where the most important variables were the 2nd Try and 1P scores. Note that these variables are available for early prediction i.e. after the midterm exam.

Table 1: Parameters and accuracy evaluation for each LR model

Model	θ_0	Course	Career	2nd Try	Year	1P	2P	L/W	ACC	MCC
1	-2.152	0.879	-0.160			1.703	3.387	0.193	79.4%	0.577
2	-1.207	0.348	-0.124		-2.221	1.759	3.487	0.289	79.7%	0.584
3	-6.153	0.387	0.020	6.119		0.902	1.721	1.721	85.1%	0.717
4	-5.322	0.437	0.093	6.520	-2.514	1.771		0.367	85.7%	0.723
5	-5.276	0.476	0.097	6.592	-2.494	1.976			85.8%	0.726

The second experiment was performed considering the five parameters from the best performing model i.e. Model 5. An MLP was trained using tensor-flow and a stochastic gradient descent method (called Adam) [6]. The MLP network has five layers with 10, 20, 10, 10, and 2 neurons respectively. The total number of free parameters was 622. The MLP model has improved the prediction accuracy and MCC from 85.83 % to 88.33%, and 0.726 to 0.771, respectively.

The results we obtained indicate that it is possible to extract relevant information from academic records history, to support decision-making. Since one of the best models does not consider the information available after the first evaluation (1P), we can use that model for early detection of students who may fail to pass the course, and to take preventive actions. Hence, the results motivate the authors to continue improving this work by including new variables and complexity in the models.

References

[1] Alize J Trinquet et al. “Student Grade Prediction Based Upon Prerequisite Lab or Topic Course Performance”. In: (2018).

[2] Zafar Iqbal et al. “Machine learning based student grade prediction: A case study”. In: **arXiv preprint arXiv:1708.08744** (2017).

[3] Peggy C Boylan-Ashraf and John R Haughery. “Failure Rates in Engineering: Does It Have to Do with Class Size?” In: **2018 ASEE Annual Conference & Exposition**. 2018.

[4] Castro Gbememali Hounmenou, Kossi Essona Gneyou, and Romain Lucas GLELE KAKAÏ. “A Formalism of the General Mathematical Expression of Multilayer Perceptron Neural Networks”. In: (2021).

[5] Davide Chicco. “Ten quick tips for machine learning in computational biology”. In: **BioData mining** 10.1 (2017), pp. 1–17.

[6] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: **CoRR** abs/1412.6980 (2015).