USE OF ARTIFICIAL INTELLIGENCE TO PREDICT UNIVERSITY DROPOUT: A QUANTITATIVE RESEARCH

Francesco Agrusti, Gianmarco Bonavolontà, Mauro Mezzini, Roma Tre University, Italy

Abstract

The main aim of the research is to predict, as early as possible, which student will drop out in the Higher Education (HE) context. Artificial Intelligence (AI) is used for replacing repetitive human activities, e.g. in the field of autonomous driving or for the task of classification pictures. In these areas IA competes with the man with fairly satisfactory results and, in the case of college dropout, it is extremely unlikely that an experienced teacher can “predict” the student’s academic success based on only data provided by administrative offices. In this study used administrative data of about 6,000 students enrolled in the Department of Education of the University of Roma Tre to train convolutive neural nets (RNC). The trained network provides a probabilistic indicating, for each student, the probability of abandonment. Then, the trained network provides a predictive model that predicts whether the student will dropout. The accuracy of the obtained deep learning models ranged from 67.1% for the first-year students up to 94.3% for the third-year students.

Introduction

In the comparative study on dropping out of higher education in Europe conducted by Vossensteyn and other researchers (2015), it was found that successful studies are seen as a crucial factor for personal success in 28 of the 35 participating countries. Early recognition of dropout is a key prerequisite for reducing dropout rates: several studies highlight the importance of monitoring individual and social characteristics of students as they have a strong impact on the probability of success of students in higher education. A key objective of the Europe 2020 strategy is in fact to reduce drop-out rates by seeking to achieve at least 40% of 30-34 year olds completing higher education (Vossensteyn et al., 2015). As reported in the literature, students generally leave during their first year of university (Larsen et al., 2013), immediately after upper secondary school: in this period, they must develop their sense of responsibility and self-regulation (Pintrich & Zusho, 2002). Individual skills and dispositions are investigated in different psychological and pedagogical models in relation to the phenomenon of early abandonment in terms of
personality characteristics (Pincus, 1980). Numerous studies have explored the impact of the economic and social status of students (e.g. race or income) and the organisational services provided to students by the university (e.g. faculty-student relationship) on the drop-out rate (Pincus, 1980). For decades, one of the most used and discussed models has been Tinto’s “student integration” model, which underlines the importance of the academic and social integration of students in predicting the phenomenon of early school leaving (Tinto, 2010). One of the other main models is the one proposed by Bean (1988), the “student attrition” model, based on the attitude-behaviour of the student, which measures individual and institutional factors and evaluates their interactions in order to predict university dropout. Another interesting model of student/institution integration is the Pascarella model (Pascarella & Terenzini, 2005), which emphasizes the cruciality for student success of having informal contacts with teachers. In other words, in this model, background characteristics interact with institutional factors influencing student satisfaction with the university. Numerous studies have demonstrated the positive effects of student-university interaction on persistence (Pascarella & Terenzini, 2005). Event history modelling is another model much discussed in literature: proposed by Des Jardins, Albourg, and Mccallan (1999), this model takes into account the role of the succession of different events in the different stages of the student’s educational career, changing the importance of factors from year to year, depending on the time period. In all these models, the relationship between students and institutions is crucial to reduce drop-out rates and several variables have been identified to improve student retention (Siri, 2015). In Italy, due to the very high drop-out rates of university students (ANVUR, 2018), several specific studies were conducted (Burgalassi et al., 2016) which confirmed the value of the baccalaureate vote (and of the entry skills of students more generally) together with the socio-demographic traits of the students (mostly the socio-economic context) as valid indicators of university drop-out compared to the outcome of the first year of study. Many of the models and studies conducted, both national and international, have presented different analyses from the psychological point of view, building psychological-motivational models focused on expectation, reasons for involvement, personal value and motivation in general. These models and surveys all involve the collection of data by interviewing students directly, through the use of tools (usually questionnaires) specially administered. The study presented in this article, however, aims to use only the data available in any university statistical office, without therefore, at least at this stage of research, interviewing students directly. In this regard, it was decided to proceed to the analysis of these data through the use of Artificial Intelligence (AI). Today, AI is used to replace human activities that are repetitive, for example, in the field of autonomous driving or for the task of classifying images. In these areas, IA competes with man with quite satisfactory results and, in the case of abandonment of the educational system, it is extremely unlikely that an experienced teacher will be able to “predict” the educational
success of the student on the basis of data provided by the administrative offices. These recent advances on neural networks have shown that AI may be able to compete (or even exceed) with human capabilities in the tasks of classification and recognition. Here below are then first shown some of the most important studies obtained thanks to the IA, on the prediction of university dropout. Then the metrics for the evaluation of these models and then the methodology used and the results obtained in our research are presented. Preliminary conclusions on the study are therefore briefly drawn.

**State of the Art**

Many research projects have used data mining techniques to study the Dropout phenomenon. Specifically, in this section we will discuss work that has investigated university dropout by developing predictive models through EDM (Educational Data Mining), or the use of data mining in education, applying computer methods to analyse large data collections. From the analysis of the literature it emerged that the decision tree algorithm (DT) is the one most used for the development of predictive models aimed at identifying university dropout. A research project funded by the Colombian Ministry of Education tried to identify predictive models of early school leaving by analysing 62 attributes belonging to socio-economic, academic and institutional data. Also in this case a decision tree has been implemented (algorithm J48) and for the validation of the model the cross-validation folder has been used with an accuracy of more than 80% (Pereira et al., 2013). Similarly, research was conducted in India to develop a DT based on the ID3 algorithm that could predict students dropping out of university. The study is based on the analysis of 32 variables on a sample of 240 students selected through a survey. Model performance was evaluated using the accuracy index, accuracy, recall and F-measure (Sivakumar et al., 2016). In 2018, research presented a classification based on the DT algorithm. The study analyses 5288 cases of students belonging to the Chilean public university (cohorts of students belonging to 44 university courses in the fields of humanities, arts, education, engineering and health). The attributes selected for the analysis are related to the student’s demographic variables, economic situation, and data on previous academic performance prior to his or her admission to university. The accuracy index of the best model developed was 87.2% (Ramírez & Grandón, 2018). In addition to the DT, other classification methods have been used in order to implement models for the prediction of university dropout. Some researchers have used specific methodologies such as CRISP-DM (Cross Industry Standard Process for Data Mining), to predict at the end of the first semester students at risk of dropping out. The dataset consists of over 25 thousand students and 39 variables for each student and the algorithms used are: DT, artificial neural networks (ANN) and logit model (LR). The results show an accuracy of 81.2% for the model developed with ANN (Delen, 2011). Similarly, a research conducted at the University of Genoa, employed the ANNs to detect students at risk of
A quantitative research at Roma Tre University

One of the most important problems in the field of IA is the problem of classification (LeCun et al., 2015). In this problem you have an object, which can be an image, a sound or a sentence and you want to associate to this object a class taken within a finite set K of classes. A neural network (RN) can be seen as a function \( \varphi \) that takes an input from a vector n-dimensional \( x \) and produces a value, called the prediction of \( x \). The prediction is correct when \( \varphi(x) = f(x) \) and otherwise incorrect. Contrary to the classic programming paradigm, where the programmer to design an algorithm must have a deep and complete knowledge...

Agrusti, F., Bonavolontà, G., & Mezzini, M. Use of Artificial Intelligence to Predict University Dropout: A Quantitative Research

Agrusti, F., Bonavolontà, G., & Mezzini, M. Use of Artificial Intelligence to Predict University Dropout: A Quantitative Research

A quantitative research at Roma Tre University

One of the most important problems in the field of IA is the problem of classification (LeCun et al., 2015). In this problem you have an object, which can be an image, a sound or a sentence and you want to associate to this object a class taken within a finite set K of classes. A neural network (RN) can be seen as a function \( \varphi \) that takes an input from a vector n-dimensional \( x \) and produces a value, called the prediction of \( x \). The prediction is correct when \( \varphi(x) = f(x) \) and otherwise incorrect. Contrary to the classic programming paradigm, where the programmer to design an algorithm must have a deep and complete knowledge...
of the problem of interest such as in (Malvestuto, Mezzini, & Moscarini, 2011; Mezzini, 2010; 2011; 2012; 2016; 2018; Mezzini & Moscarini, 2015; 2016), to implement an RN the programmer may also be completely unaware of the mechanism or semantics of classification.

We collected, from the administration office of Roma Tre University, a dataset of students enrolled in the Department of Education (DE). The years of enrolment ranges from 2009 up to 2014 comprising a total of 6078 students. We found that 649 of all students were still active at the time when we acquired the dataset (August 2018), while the remaining 5429 closed the course of their studies either because they graduated or because they dropped out or by other reasons, explained later. We refer to this set of students as the no active students. Note that in the following when we will refer to the enrolment year (or simply the year) of a student we mean the number of years passed since her/his first enrolment to university, that is, we refer to an integer value between 0 and 9 since no student is enrolled for more than 9 years. In general, each of the no active student is classified in two different classes: Graduated and Dropout. We excluded later all students which do not classified in these two classes, like for example students who changed faculty within the R3U or went to another university. The number of such students is 118. The number of graduated students is 2833 while the number of who dropped out is 2478. We obtained, from the R3U’s administrative office, most of the (out of what were available) administrative fields of all students. The attributes relative to the student’s academic career are the following: Exam name, Score of the exam, Maximum score of the exam, ECTS of the exam, Exam date (month/day), Academic year, Type of validation. They represent the attributes relative to each test or exam given by the student. Note that the field “ECTS of the exam” refer to the European Credit Transfer and Accumulation System. In order to construct the training set all the domains of the dataset are converted, using an arbitrary bijective function, to a non-negative integer domain. For example, the domain of the attribute GENDER, was converted to the domain \{0,1\} where 0 correspond to “male” and 1 to “female”. We created a table STUDENT, whose schema S contains all the attributes provided by administrative offices. We limited our tests only to the students that are still active at the year 3 because after that year the number of those students dropping out to university is very small and not significant from statistical and/or practical purposes. If a student ends his/her career in the year \( z \), \( 0 \leq z < y \), then \( f_y \) will take the value \( \delta \) for every year \( z < y \leq 3 \). The value of \( \delta \), which was arbitrarily chosen to be equal to \(-1\), can be considered as a NULL value and it does not appear in the original domain of any field on the scheme S. Furthermore, for each year of enrolment \( y \in \{1, 2, 3\} \) an integer \( m \) is set to represent the maximum number of exams sustained by any student on the year of enrolment \( y \). We found that \( m_1 = 24, m_2 = 19 \) and \( m_3 = 23 \). Thus, for any field in List 3, for each year \( y \) and for each \( z, 0 \leq z \leq m_y \), we added a field denoted as \( g_{y,z} \). If a student in the year \( y > 0 \) of her/his academic career
We build a table called Y_LABEL containing two attributes: STUDENTID and Dropout, where the last represents the label of each student. It has a numerical domain with the following meanings: 0, if the student graduated, 1, if the student dropped out. From the table student described above, we derived three type of tables denoted as STUDENT_A_x, STUDENT_B_y and STUDENT_C_z for $0 \leq x \leq 3$ where $x$ is the number of years from the first enrolment.
In the schema of tables STUDENT_Ax we added all the attributes in List 1 and all the attributes in List 2 (of the type $f_y$), and all the attributes of List 3 (of type $g_{y,z}$) for all $y = 0, \ldots, x$.

The tables denoted as STUDENT_Bx, $x = 0, \ldots, 3$, contain only the attributes of List 1 and List 2. That is, we considered in these tables only administrative fields and we excluded the fields related to the academic careers of the students (the ones of type $g_{y,z}$).

The tables STUDENT_Cx, $x = 0, \ldots, 3$, have been constructed in the following way. We computed, for each student, the following aggregate statistics: \textsc{diffYear} and for each year $x > 0$, \textsc{NumberExams$_x$, avgScore$_x$} and \textsc{sumETCS$_x$}. The first statistic contains the value \textsc{Year of Birth – Year of Beginning of Studies} – 19 that is, the difference in years between the age of the student (at the date of the enrolment) and 19. The other statistics contains, for each student and for each year $x = 1, 2, 3$ respectively, the number of exams successfully passed, the average score of the exams successfully passed and the sum of the ECTS gained. We thus obtained the schema of STUDENT_Cx by adding to the schema of each table STUDENT_Bx, all the above four fields. The idea we want to test here is whether it is better and effective to use only some significant aggregate statistics or, instead, it is better and effective to use all the attributes relative to the academic career (like in tables STUDENT_Ax).

For the tests of both CNN and BN we choose a random permutation of all no active students. Next, we partitioned all students in twelve different mutually disjoint groups containing approximately 450 students each thus obtaining a partition $\mathcal{P} = \{P_0, P_1, \ldots, P_{11}\}$. For all $0 \leq i \leq 11$ the group $P_i$ is used as a validation set $V_i$ and the group $P_{i+1 \mod 12}$ as a test set $T_i$ and the students in the remaining groups, as the training set $A_i$. In the validation set for the year $x$ we put only the students who, at that year of enrolment, were still active. We trained three models based on the CNN architectures mentioned above by taking from each of the table above (A or B or C) the training, validation and test sets from the partition $\mathcal{P}$. We got data from a total of 43200 epochs. For each epoch the confusion matrix of both the validation and the test sets were produced. We found that the $F_1$ measure was the better indicator for the selection of the best model. We calculated the accuracy for the validation set, for the training set and the $F_1$ measure. Training and validation data were taken from the table STUDENT_Bx. Furthermore, we computed the value of the $F_1$ measure for the year 1 and for the years 2 and 3 for the three different tables STUDENT_Ax, STUDENT_Bx and STUDENT_Cx. We observe that in all three cases the value of the $F_1$ measure relative at the table STUDENT_Bx is always worse in every year. This clearly shows that using only administrative data gives very poor performance in predicting the dropout of a student. In Table 1 we report the data of the confusion matrix, for both validation and test sets, in which the validation set, among the twelve possible different sets of the partition $\mathcal{P}$, achieved the best score on the $F_1$ measure.
Conclusions

We explored the effectiveness of predicting the dropout from university using three different sets of features. The first one, containing all the academic and administrative features (tables STUDENT_Ax). The second one, containing only administrative features (tables STUDENT_Bx) and the third (tables STUDENT_Cx) containing the administrative features and 3 aggregate statistics about the academic career of the students. The experiment showed that using only administrative features does not give good results and the models using only them are always outperformed by models using also the academic career features or aggregate statistics. Furthermore, the models using, besides administrative features, also aggregate statistics perform slightly worse than the models using only and all the academic careers features. From all the above discussion we clearly conclude that the more accurate data we have the more precise and effective the model’s predictions could be. Since it is not required that the prediction process is made in real time, we can train hundreds of models and make multiple prediction in order to reduce the random variation found in the early phase of training. Clearly the system can be made finer by introducing a prediction model every semester or even every trimester or it can be extended to other faculty or other types of students.

References


Agrusti, F., Bonavolontà, G., & Mezzini, M.

Use of Artificial Intelligence to Predict University Dropout: A Quantitative Research


Agrusti, F., Bonavolontà, G., & Mezzini, M.  
**Use of Artificial Intelligence to Predict University Dropout: A Quantitative Research**


