4th International Conference on Higher Education Advances (HEAd'18)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/HEAd18.2018.8018

Factors that Influence Student Retention

Devereux, Aisling[a]; Hofmann, Markus[b]
[a]Athlone Institute of Technology, Ireland, [b] Department of Informatics, Institute of Technology Blanchardstown, Ireland

Abstract

With the increase in enrolment figures from second level education to third level education over the last number of decades, non-progression rates continue to give cause for concern in certain levels and disciplines. It has been widely argued that in addition to increasing enrolment numbers, higher education must also be concerned with the success of these students. In both the Irish and the international sector, the negative consequences of non-progression has been highlighted, not just on a societal level, but also for the students themselves. It is crucial for first-year student experience to have a positive experience and be fully supported in achieving the goals of higher education. From researching several reports in the area of retention and in particular the reports published by the Irish Higher Education Authority and the National Forum for the Enhancement of Teaching and Learning in Higher Education in this area, it is clear that there is a need to analyse the data available and present the findings in a clear way to the key decision makers to allow for early intervention. This paper uses the different phases of the CRISP-DM methodology and applies data mining techniques and models to a real student dataset with the aim to predict the students that will progress.

Keywords: *Learning analytics; Data Mining; Higher Education; Retention.*

1. Introduction

Student retention is a big issue in the Higher Education sector both at a national and international level. A big challenge for the Higher Educational sector is finding students that are at risk of not completing their studies and dropping out before they become a statistic. Learning Analytics (LA) and Data Mining can help to identify such students months before they drop out according to O Farrell (2016). This paper investigates the issue of the ever increase in enrolment figures from 2nd level to 3rd level education, high non-progression rates (see Table 1) on level 6 programmes in Institutes of Technologies Ireland (IOTI) and the increase in the number of students withdrawing from third level education throughout the Institute of Technology and the University sector in Ireland and Universities abroad (Frawley, Pigott, & Carroll, 2017). Retention is an issue that is being focused on currently at the Athlone Institute of Technology (AIT). PricewaterhouseCooper (PwC) auditors were commissioned to complete a review on retention (March 2017) in AIT. One of its objectives is to gain an understanding of progression rates for each faculty and the strategies employed. Using data mining techniques and models, the information captured on various systems employed in AIT is explored to find the most influential attributes to predict students that will progress.

Table 1. Non-progression rates by level.

Sector	Level	Most Common Points Attained	% Non Progression
Institutes of Technology	Level 6	250-300	25%
	Level 7	250-300	26%
	Level 8	300-350	16%
	L8 3 yr duration	300-350	16%
	L8 4 yr duration	300-350	16%

Source: Frawley, Pigott, & Carroll, (2017).

2. Background

2.1 Overview of Data

Two datasets are used in the project (2014/15 and 2015/16). Both datasets contain the same variables except for the class label (Progress) included in the 2014/15 dataset. This class label is the outcome that will be predicted in the 2015/16 dataset and describes whether the student will progress or not progress into the second year of the course.

The 2014/15 dataset contains 1,118 examples with 2 special attributes and 62 regular attributes. The special attributes are the Spriden PIDM which is the unique identifier for each student and the class label (Progress). The 2015/16 dataset contains 1,041 examples with 1 special attributes (Spriden PIDM) and 62 regular attributes.

The data is sourced from many of the internal systems in the AIT. This includes the student record system which store the students' personal details, admission records, registration information, grant records, bio/demographic information, examination results and student account information; Moodle data; data from the library system; data from the disability office; the Student Resource Centre; and the Central Admissions Office (CAO) providing Maths Points, English Points, Leaving Cert Score, Acceptance Round, Acceptance Date and Course Preference Number. All student data was anonymised during this project.

3. Related Work

Learning Analytics and Educational Data Mining are emerging disciplines (Agudo-Peregrina, Iglesias-Pradas, Conde-González, & Hernández-García, 2014), concerned with developing techniques for exploring the different types of unique data that come from the educational context. O Farrell (2018) mentions that the most widely-used source of data is student interactions within the virtual learning environment (VLE). VLE systems are online platforms that accumulate a vast amount of information (Thakur, Olama, McNair, Sukumar, & Studham, 2014) which is extremely useful for analysing students' behaviour and trends. This type of analysis could be very beneficial to the Higher Educational sector in Ireland. It is evident there is a link between academic performance and Moodle activity usage but according to (Casey, Gibson, & Paris, 2010) in their research, this is at a basic level. Activity log data can provide an opportunity to address some of the critical challenges within the Higher Education sector such as high drop-out rates (Siemens & Long, 2011), (Thakur, Olama, McNair, Sukumar, & Studham, 2014) and (Azcona, Corrigan, Scanlon, & Smeaton, 2017). O Farrell (2018) discuss in a report that for all the benefits that learning analytics can provide within the educational domain, it is just a resource for providing insights, uncovering hidden patterns in data and providing answers. In order to enhance teaching and learning, learning analytics must be used effectively and when this is the case, it can become an essential and invaluable tool for supporting and informing successful policies such as a retention strategy. Thoroughly tracking and assess all students' activities while evaluating the structure and contents of courses and its effectiveness for the learning process (Zorrilla, Menasalvas, Marin, Mora, & Segovia, 2005) can pose both, an opportunity and a challenge. A very promising area for attaining this objective is the use of data mining (Zaiane, 2001).

4. Methodology

The Cross-Industry Process for Data Mining (CRISP-DM) methodology provides a structured approach to planning a data mining project. The following stages will be reviewed during the lifecycle of this project: Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment. The following listing outlines the data mining objectives for this project:

- Collect and clean data for the 2014/15 and 2015/16 academic years.
- Explore / Visualise the data to identify factors predictive of students' success at AIT.
- Predictive statistical models and data mining techniques to model students progress:
    - ROC curves are applied and compared to the unmodified dataset to check which algorithm best suited the data and again after the data preparation stage.
    - Train a model using 2014/15 dataset.
    - Test and evaluate performance of the models on the unlabelled 2015/16 dataset.

5. Results

During the data exploration phase of this project all of the attributes are further investigated. The following attribute *Moodle Usage* is an example of this and turns out to be one of the most useful attributes for predicting the class label.

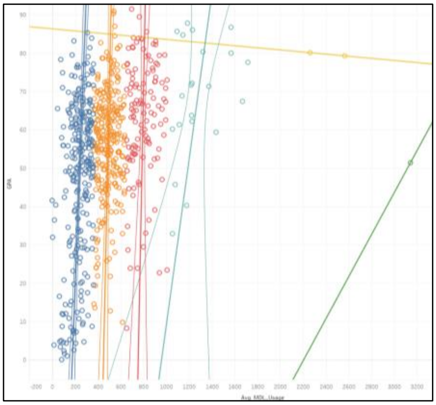


Figure 1. Average Moodle Usage with GPA.

5.1. Moodle Usage

Taking a look at the attribute *Moodle Usage*, gives us an insight into how many times a student has logged onto their Moodle account in the 2014/15 academic year. Within the data it is evident that students with the higher moodle usage have a higher GPA. Clustering has been added to this graph which indicates the number of different groupings (see Figure 1). Cluster 5 and Cluster 6 have a few unusual points which could be potential outliers. Outliers are not always errors but they may skew the mean and standard deviation if there are many of them. If the values are more then +/-3 times the standard deviation from the mean then outlier detection methods need to implemented. The standard deviation is 343.7 for this attribute and the values for some of the points are greater than 1,000. Knowing the data, these outliers are not errors. There are a number of students that have high usage on Moodle.

5.2 Model Evaluation

Most of the models using both Method 1 and Method 2 have resulted in high accuracy, precision, R^2 values and fairly good predictions for the 2014/15 labelled dataset which is evident from the results in Table 2.

Method 1: AcadYr and CourseYr are excluded from the dataset. These attributes exhibit very low variance so are not useful in the dataset when trying to predict the class label.

Method 2: (Same attributes as Method 1) with Remove Correlated and Remove Useless Attributes algorithms applied.

Table 2. Summary of results from models tested on 2014/15 dataset using RapidMiner

Method 1	**14/15 Model**	**R^2**	**Accuracy**	**Percision**	**RMSE**	**AUC**	**15/16 Misclassified**
							Progress / NotProgress
Select Attribute	Decision Tree	0.986	99.82+/- 0.36	99.41+/- 1.76	0.019+/- 0.038	0.5	0 / 15
	Naïve Bayes(Kernel)	0.888	98.48+/- 1.39	94.28+/- 6.74	0.101+/- 0.052	0.997	6 / 48
	Logistic Regression	0.937	99.2 +/- 0.84	99.33+/- 2.00	0.054+/- 0.058	1	1 / 33
	k-NN	0.159	88.01+/- 1.39	89.67+/- 15.45	0.299+/- 0.013	0.83	16 / 397
	Linear Regression	0.772	97.07+/- 0.27	98.43+/- 0.70	0.400+/- 0.000	0.998	6 / 265
Method 2	**14/15 Model**	**R^2**	**Accuracy**	**Percision**	**RMSE**	**AUC**	**15/16 Misclassified**
Remove Correlated Attributes	Decision Tree	0.993	99.82+/- 0.36	99.41+/- 1.76	0.019+/- 0.038	0.5	0 / 15
	Naïve Bayes(Kernel)	0.825	97.67+/- 0.72	93.09+/- 4.92	0.147+/- 0.022	0.976	16 / 61
	Logistic Regression	0.476	92.22+/- 2.26	74.36+/- 8.46	0.330+/- 0.020	0.927	7 / 282
	k-NN	0	85.42+/- 0.42	0	0.229+/- 0.013	0.656	11 / 830
	Linear Regression	0.765	96.98+/- 0.13	98.83+/- 0.84	0.399+/- 0.000	0.999	6 / 91

k-fold cross validation is used in all models. This divides the training dataset into k=10 separate folds. Each time the algorithm is run, it will be trained on 90% of the data and tested on 10%, and each run of the algorithm will change which 10% of the data the algorithm is tested on. Using this method of cross-validation the entire 2014/15 dataset is used. The full 2014/15 dataset is used to train the model and then the 2015/16 unseen dataset will be used to test the model. When the trained model is applied to the unseen dataset the results will be evaluated using the unlabelled 2015/16 dataset for validation.

The decision tree is the first model that is tested on the 2014/15 dataset. Different parameters were applied to the decision tree. The confidence level was changed, pruning and pre-pruning set to on and off, information gain and gini index were tested, the minimal leaf size was changed from 4 to 8 and the minimal leaf size was changed from 2 to 4 but the output remained unchanged. The Remove Correlated Attributes operator (Method 2) was applied to the dataset but this did not change the result either. It is evident from the results in Table 2 that the Decision Tree is the most accurate and robust model, displaying high values for R^2 but this is taking the attribute GPA and always splitting at the greater than 40 and less than 40 in all cases. If the GPA is omitted the accuracy of the model will reduce to 90% and the AUC to 0.720 which is considerably lower. k-NN produced the model with the lowest accuracy using both methods. This model has a very low R^2 using both methods and has misclassified a large proportion of the 2015/16 dataset. Naive Bayes (Kernel) with the estimation mode set to greedy produces good prediction and has a high R^2. Linear regression has high precision values and high values for AUC in both models with the R^2 of 77%.

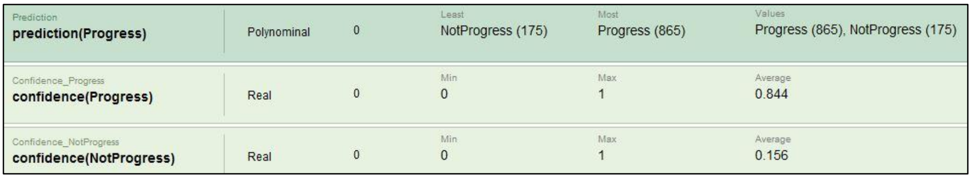
5.3 Logistic Regression - Results using Method 1

In relation to the accuracy of the different models and in order to evaluate whether the business objective was met, the logistical regression model displayed the most accurate results with both accuracy and precision at 99% and a high R^2 of 94% when applied to the 2014/15 training dataset (see Table 3).

Table 3. Logistic Regression performance on labelled 2014/15 dataset using RapidMiner

accuracy: 99.19% +/- 0.63%	**true Progress**	**true NotProgress**	**class precision**
pred. Progress	954	8	99.17%
pred. NotProgress	1	155	99.36%
class recall	99.90%	95.09%	

After applying the trained model to the unseen 2015/16 dataset (see Figure 2) the number of students predicted to Progress is 865 and NotProgress is 175.



Prediction **prediction(Progress)**	Polynominal	0	Least NotProgress (175)	Most Progress (865)	Values Progress (865), NotProgress (175)
Confidence_Progress **confidence(Progress)**	Real	0	Min 0	Max 1	Average 0.844
Confidence_NotProgress **confidence(NotProgress)**	Real	0	Min 0	Max 1	Average 0.156

Figure 2. Logistic Regression Model on unlabelled 2015/16 dataset – Predicted outcome

Table 4 displays a summary of results. The first column shows the class label of which there are 1,040 first year students. The next column 'Predicted' displays the predicted probability of the occurrence of the class label from the unlabelled 2015/16 data model (see Figure 2). The next column 'Actual' contains the exact number of student that progressed from first year using the 2015/16 dataset by checking their GPA and finally the last column gives the number of predicted outcomes that were misclassified. Looking at the misclassifications, one student was misclassified as Progress when they actually had a GPA under 40 and 33 students were misclassified as Not Progress when their actual GPA was greater than 40. The results for this model is showing good levels of accuracy and the probable classification using logistic regression of the labels is high.

Table 4. Summary of Predicted class label 2015/16 validated against actual 2015/16 dataset

Class (1040)	Predicted (unlabelled dataset 2015/16)	Actual (2015/16 Student Results)	Misclassified
Progress	865	897	1
Not Progress	175	143	33

6. Conclusion

The aspiration of this project was to provide a stepping stone using the outcome from the results in the data to start a bigger discussion in the third level institute around retention policies and cross-departmental initiatives. Working with real data had its advantages. It was difficult to spot some useful patterns in the dataset during the initial data exploration stage due to its complexity. There were problems with data files at the beginning of the project. The main issue with the dataset was the inconsistency of data that is recorded across the various systems in AIT. Going forward more emphasis should be placed on the data collection ensuring data quality and integrity.

It is evident from the results section that implementing data mining techniques on educational datasets can result in good models for predicting student progression rates based on the most influential attributes captured from the relevant data systems. Using the data mining modelling technique of logistic regression and applying to the unlabelled 2015/16 dataset and validating against the actual values from the 2015/16 dataset it is clear that the number of misclassifications are minimal meaning that the predictive modelling has evident abilities to be applied to new and unseen data and therefore can be used to identify potential drop outs earlier than without using the advanced modelling techniques.

References

Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á., & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in human behavior, 31*, 542-550.

Azcona, D., Corrigan, O., Scanlon, P., & Smeaton, A. F. (2017). Innovative learning analytics research at a data-driven HEI. *Editorial Universitat Politecnica de Valencia*.

Casey, K., Gibson, P., & Paris, I. S. (2010). Mining moodle to understand student behaviour. *International Conference on Engaging Pedagogy. (ICEP10)*, National University of Ireland Maynooth. Retrieved from http://www-public. tem-tsp. eu/~ gibson/Research/Publications/E-Copies/ICEP10. pdf.

Frawley, D., Pigott, V., & Carroll, D. (2017, March). A Study of Progression in Irish Higher Education 2012/13 to 2013/14. Dublin: Higher Education Authority.

O Farrell, L. (2016). Learning Analytics and Educational Data Mining for Learning Impact.

O Farrell, L. (2018). Using Learning Analytics to support the enhancement of Teaching and Learning in Higher Education. National Forum for the Enhancement of Teaching and Learning in Higher Education.

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and education. *EDUCAUSE review, 46(5)*, 30.

Thakur, G., Olama, M. M., McNair, W., Sukumar, S. R., & Studham, S. (2014, January). Towards adaptive educational assessments: predicting student performance using temporal stability and data analytics in learning management systems. In Proceedings 20th ACM SIGKDD conference on knowledge discovery and data mining.

Zaiane, O. (2001). Web usage mining for a better web-based learning environment.

Zorrilla, M. E., Menasalvas, E., Marin, D., Mora, E., & Segovia, J. (2005, February). Web usage mining project for improving web-based learning sites. In International Conference on Computer Aided Systems Theory (pp. 205-210). Springer, Berlin, Heidelberg.