Predicting Student Success. Early for a VTOS Student.

Stephen Colgan

stephen.colgan@inchicore.cdetb.ie

Department of Computing

TU Dublin, Tallaght Campus

Dublin, Ireland

Keith Quille, Jelena Vasic and Seán McHugh

keith.quille, jelena.vasic, sean.mchugh {@TUDublin.ie}

Computing Department

TU Dublin, Tallaght Campus

Dublin, Ireland

Abstract

This study was conducted over two academic cycles, 2017 and 2018, and was based on results from VTOS students (N =70) attending courses provided by a City of Dublin Educational Training Boards (CDETB) centre in south Dublin. The research used machine learning (ML) techniques to develop a prediction model to identify students' success rates and highlighted the students' needs for early intervention of additional services. The sample was sub-divided into 'successful' (n=37) and 'unsuccessful' (n=33) participants. Model selection was developed using current best practice approaches to machine learning, which included data selection and data pre-processing techniques (including attribute selection). Model evaluation and model performance metrics, including bagging and statistical tests, were implemented to investigate the model's application for a larger sample. The results demonstrate that bagging with SMO algorithm produced an accuracy of 64.29 and had the bagging highest sensitivity of ~0.62%, and specificity measure of 0.67%. This model also produced a Root Mean Squared Error of 0.59%. These results had a reliable predictive value for six out of ten students for early intervention and these results indicate increased retention and their individual success probability. The results of this study are significant for predictive data models for a wider student application.

Keywords

Prediction model, data modelling, machine learning, attribute selection, model evaluation, model performance, Support Vector Machines, Sequential minimal optimization (SMO), Support Vector Machines (SVM), predictive value, False Positive, False Negative.

1. Introduction

The study was conducted in a South Dublin based college of Further Education (FE) under the City of Dublin Educational Training Board (CDETB) with students who receive support through the Vocational Training Opportunities Scheme (VTOS). VTOS has two strands of students in attendance, these are Core Student (CS) and

Dispersed Student (DS). A CS is described as a student who attends a specific course whereby all students in the course are also a CS. DS are dispersed throughout the college attending any full-time course offered by the college. A course may have one or more DS or none in attendance. Each centre within the CDETB scheme has been allocated a capitation of VTOS students varying from 20 to 200, of either DS or CS students. The Irish National Framework of Qualifications (NFQ), is a Ten-level framework for the development, recognition and awarding of qualifications in Ireland. Courses provided by the CDETB range from Nine to Eleven modules. Success on a course is based on completion of three mandatory modules and five elective modules, which results in a major award at Quality and Qualifications Ireland (QQI) Level 5. A component award is offered where a student has been successful in individual modules, but not in a mandatory module, or in all the five elective modules. As such this means students do not gain the full Level 5 award.

Student retention on these courses is a concern (Kelly, 2019) among students with lower Leaving Certificate points. Dropout rates are highest among computing students (45%) and in areas such as hospitality, tourism (35%), engineering, manufacturing and construction (33%) (Kelly, 2019). The primary focus of this research paper was to measure the effectiveness of a data prediction model to identify those students who were likely to not complete a course.

2. Literature Review

This literature review will focus on the areas of adult education, specifically on the literature associated with improvement of student attendance and retention, student participation and enhanced learning opportunities. There is little research done on the area of student retention in social welfare funded courses and any Irish research has been conducted in higher education, rather than the further education sector which is the focus of this paper. This review will then discuss the wider research on machine learning and machine learning tools.

2.1 Student Retention Factors

Not every student who enters further education has the intention of completing their course. Students' choices, ambitions, and circumstances may change throughout their

course, a change in direction could possibly be in their own best interests, whereas "some individuals benefits may be gained only by leaving" (Eivers, 2002, p.2). Curzon (1997) maintains that learning may only have meaning and value if it relates to their personal development as a member of society. Yet, some non-completion of courses is unavoidable and should not be viewed as failure by the student, teacher or college (Tinto, 1975). Yorke (1999) noted that many students leave higher education prematurely and at least one-third of students may leave their course within the first twelve months of enrolling.

Non-completion can also be damaging to an institutional and a particular course's reputation. To stem this, in an Irish context, the government has invested heavily in education and are concerned that non-completion represents an inefficient use of limited educational resources and a loss of future skills (HEA, 2019). Although, Finnegan (2010, p. 271) believes "an enhanced and progressive policy and practice of creating, supporting and sustaining communities of learners will be a key intervention and enhance retention".

Often students experience a gap between their expectations and their actual experience of the course (Martinez & Munday, 1998). Reasons for leaving identified by Eivers (2002) include that student's lack an understanding of what the course entails before applying, that college experience is worse than they had expected, and that the workload is heavier than expected. Furthermore, Martinez (1998), Finnegan (2010), and Healy (1999) found students felt that what they wanted, and their capability to do the course, had not been adequately explored in the admissions interview, and that they had not received inadequate support and guidance.

2.2 Machine Learning

Machine learning (ML) and artificial intelligence (AI) are not new ideas. ML is a powerful set of technologies that can help organizations transform their understanding of data (Hurwitz, & Kirsch, 2018). Machine learning is a systematic approach to leveraging advanced algorithms and models to continually train data and test with additional data to begin to apply the most appropriate machine learning algorithms to a problem (Hurwitz & Kirsch, 2018). The difficulty is the selection of the right algorithm, with clean, usable data using the best performing model thus perhaps developing the most generalizable model possible. Machine learning requires a cycle

of data management, modelling, training, and testing (Hurwitz & Kirsch, 2018). The cycle of data management, modelling and testing is the focus of this paper.

3. Methodology

This section will outline the methodological decisions taken to collect data to respond to the research question which sought to predict VTOS students' performance outcomes for early student intervention services. Student records for 2017 and 2018 were the relevant sample for the current research deemed appropriate for the study. Ethical concerns in relation to the anonymity of the student records were removed by ensuring all data was anonymised.

The original data set consisted of one hundred and ten students (N =110). The data was a combination of students starting or finishing a two-year course. In the college, Level 5 and Level 6 courses are combined, or spread over a two-year academic cycle. Students participating in Business and Technology Education Council (BTEC) approved courses have been removed from the data set as they are outside the remit of this study so to have students that are entering a one-year QQI course. Between both years there will be seventy (70) students the following table shows the breakdown.

Class	**Male**	**Female**	**Total**
Successful	13	25	**37**
Unsuccessful	12	20	**33**
Total	**25**	**45**	**70**

Table 1: Breakdown of data instances

As part of the anonymization of the data the following fields were removed Forename, Surname, Address line 1, Address line 2, PPSN, Learner Number, College Number, Centre Number. Later Town was also removed to prevent a student's information being identified. This data anonymization reduced the attributes to 17, which included the target attribute. The target for the purpose of this project were those students who were successful or unsuccessful in receiving a major award, in effect passing eight individual modules for the course that they elected or registered for. A significant time was spent manually addressing missing data. There were

several transcription errors from the source documents (application forms) and omissions, these had to be checked manually from source documents.

A number of baseline model tests were performed using an algorithm like ZeroR and/or NaiveBayes tests. A model's accuracy of ~50% is positive and may mean that it performs better than chance. ZeroR model produced an accuracy ~ 53%. An expected result, as the larger class is 37 / 70, the model should technically not perform less than 50% at this stage. A lower accuracy at this stage could suggest that the data set is biased in some way or further investigation is required. Performing this baseline test gave an indication in the model's accuracy if these attributes removal affected the model's performance. There were no differences in the accuracy of these baseline models with all reaching 52.85% accuracy for either Normalised or Standardised tests. Four of the more commonly used attribute selection tests as outlined by Quille, Bergin & Mooney, 2015; Witten & Frank, 2005; Bergin, 2015 & (Brownlee, 2016) suggest running a number of selections and analysing these for commonality. Using an arbitrary cut off of 0.090 for the CorrelationAttributeEval shows

County	Age	No-Months	Education-Status	>0.090
Nationality	National-Category	Gender	Economic-status	<0.090

Table 2: CorrelationAttributeEval cut of 0.090

Applying arbitrary cut off of 0.05 for InfoGainAttributeEval shows

County	Nationality	Education-status	National-category	Economic-status	>0.05
No-of-Months	Gender	Age			<0.05

Table 3: InfoGainAttributeEval Cut Off 0.05

Further tests were carried out using WrapperSubsetEval with SMO and J48, these results showed

County	Nationality	Age	National-Category	SMO
Age	No-months			J48

Table 4: WrapperSubsetEval with SMO and J48

Comparing the results of the above four tests for commonality attributes that occurred two or more times, resulting in the attribute selection for model selection. Using CorrelationAttributeEval as a baseline the following data sets where produced:

Data Set 1	County	Age	No-Months	Education-Status		
Data Set 2	County	Age	No-Months	Education-Status	Nationality	National-Status

Table 5: Data Sets 1 and 2

The top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM, 2006) were (Steinberg, 2008):

C4.5(J48)	*k*-Means	SVM (SMO)	Apriori	EM	PageRank	AdaBoost ensembled	kNN (IBK)	Naive Bayes	CART

Table 6:Top 10 Data Mining Algorithms identified by IEEEE Internation Conference

However as suggested in previous research (Bergin, 2015), implementation of a blend of algorithms using diverse machine learning techniques, was successful to determine their effectiveness at predicting performance in an introductory programming course. The six different types of algorithms where evaluated are shown in the table below

k-nearest neighbour (iBk)	backpropagation (MultilayerPerceptron)	C4.5 (J48)	logistic regression	naïve Bayes	support vector machines SVM(SMO)

Table 7: Algorithms with Waikato Environment for Knowledge Analysis, WEKA

Quille (2015), suggests that Bootstrapping is a useful tool with small datasets. This is supported by Brownlee, (2016) as ensemble methods combine the predictions from multiple models in order to make more robust predictions. Additionally, the bootstrap procedure may be the best way of estimating error for very small datasets (Witten & Frank, 2005). The researcher assembled the following list of algorithms to test for prediction using data sets 1 and 2.

Rules.ZeroR (rules),	Naïve Bayes (bayes),	SMO (function),	MutliLayer-Preceptron (function),	iBk (lazy),	J48 (trees),	Bagging with SMO (Ensembled)	Bagging with J48).

Table 8: Machine Learning Algorithms to test for prediction on Data views 1 and 2

Data View 1 produced a lower percentage (61.42%) of correctly classified instances by using the algorithm SMO. Data View 2 used the model Bagging with SMO (Ensembled) and produced 64.28% correctly classified instances. This was the highest percentage produced among the models tested.

4. Results

Weka inbuilt feature results suggested that the attribute best suited for predicting if a VTOS student would be successful or not, were the following attributes (data items):

Data Set 2	County	Age	No-Months (in receipt)	Education-Status	Nationality	National-Status

Table 9: Attribute Selection best suited for prediction of success

All selected algorithms were implemented using 10-times 10--fold stratified cross-validation. This is a best practice method to avoid overfitting with machine learning algorithms (Witten I.H & Frank E, 2005). The advantage of this method is that all instances can be used for training and testing thus it reducing bias in partitioning data and increasing overall confidence in the generalizability of the models (Witten I.H & Frank E, 2005). Accuracy, sensitivity and specificity measures for the Algorithms are outlined in **Error! Reference source not found.**. Based upon the accuracy measure, the most successful algorithms in descending order are shown in table 10 and demonstrate that Bagging with SMO produces the highest accuracy.

Model with Data View 2	**% Correctly Identified**	**Sensitivity**	**Specificity**
Bagging with SMO (Ensembled)	0.642857	0.636364	0.648649
SMO (function)	0.585714	0.566666667	0.6
Bagging with J48	0.571429	0.551724138	0.585365854
MutliLayerPreceptron (function)	0.558824	0.517241379	0.58974359
Naïve Bayes (bayes)	0.542857	0.515152	0.567568
J48 (trees)	0.54	0.566666667	0.528571429
Rules.ZeroR (rules)	0.528571	0	0.528571
iBk (lazy)	0.514286	0.482758621	0.536585366

Table 10:Models Performance Confusion Matrix

Although, overall accuracy was important in this study the sensitivity measure was also valuable. Misclassifying a student as Successful, a False Positive (FP) also known as a Type 1 error was far less detrimental than a Type II error or a False Negative (FN) by misclassifying an unsuccessful student as successful. This would omit weak students from receiving additional resources but providing strong students with extra resources unnecessarily could be seen as a waste of teaching resources.

Pezzullo (2019) was used to test the six (6) algorithms using ANOVA and *Tukey ad hoc* testing for validity. The overall result reported $p=0.000$ but a p-value is rarely reported as zero as there is however small the chance that the null hypothesis could be true. The researcher has used *p-value* <0.001 which is statistically highly significant. As the data provided by the VTOS was mostly categorical in nature, it suggested that Bagging with SMO (Ensembled) was the algorithm that produced higher accuracy, specificity and sensitivity.

5. Discussion

The data set provided was from an ETB College of Further Education and the students were registered on VTOS level 5 NFQ courses in 2016 and 2017. This research was focused on creation of a model to predict if students would be successful or unsuccessful using known data mining techniques. The full data set (N = 70), were used as the training instances to validate the Bagging with SMO model using 10-fold cross-validation, confusion matrix analysis reported Bagging with SMO as producing the highest accuracy. Models were further tested using Anova test with Tukey post-hoc testing at a confidence interval of 99%, which demonstrated that Bagging with SMO was statistically highly significant with a *p value* < 0.0001. The model achieved an overall prediction accuracy of 64.28% (25 students were misclassified, 13 False Negative, 12 False Positive). The study shows in this instance that Bagging with SMO produces the highest Accuracy, Specificity and Sensitivity.

There were two unused factors in the final model – gender and economic status. The removal of gender increased the accuracy of the model, and this was similar to research by Quille (2016) in relation to computer game data. During the attribute selection phase, economic status (eg Job Seekers Allowance) gave the lowest ranking and was removed. This study supports that model performance is improved as SMO produced an accuracy of 58.57% and by using Bagging with SMO

(Ensembled) producing an accuracy of 64.29% was achieved showing an improvement of 5.71% in model performance. The researcher acknowledges that 70 students was a small sample and hence the results, while interesting, would need to be tested with a larger data set. This particular study would benefit from the other 13 CDETB centres being involved to ensure the validity and generalisability of the results of this study.

6. Future work

Recruiting is essential for getting students enrolled. But once they are enrolled, what are institutions doing to retain them? According to Tinto (1999), most institutions do not take student retention seriously (Fike, 2008). While big data technologies have demonstrated promise in increasing student retention, ethics are a paramount concern to educators (NMC, et al., 2017). This particular study would benefit from the other 13 CDETB centres being involved so as to help reduce the risks to validity that the small data set produces. The inclusion of Gender especially female should be further studied as no significant increase in accuracy was achieved when added to the PreSS model, but did have value when predicting the performance of only female students (Keith Quille, 2016). As the dataset Gender distribution was 45 Female with 25 Male this is in line with research that females are more likely than males to enrol in community college later in life, and, according to one study, more than four-fifths of females entering college after age 25 are actually re-enrolling (Goldrick-Rab, 2010). With an article in the Irish Times stating the proportion of students completing their third-level course is much higher among females (81 per cent) than males (71 per cent) (O'Brien, 2019). The possible introduction of self-esteem style surveys similar to the pre-processed programming self-esteem data consisted of ten questions (Keith Quille, 2016) or based on the Rosenberg self-esteem questionnaire but modified to reflect a student's perception of their programming ability (Rosenberg 1965). In addition, other elements or combinations from research could be investigated such as other prediction models that include the use of early aptitude tests, examinations or programming tests have only shown value at around eight weeks into the module at the first midterm or have not shown any significant prediction value at all (Porter & Zingaro 2014; Tukiainen & Mönkkönen 2002; Evans & Simkin 1989). However, following a similar structure to the PreSS model to the author's knowledge (including

an extensive review of the literature), is one of the highest prediction systems developed with nearly 80% accuracy, administered at the earliest stage (of four weeks into the module) which has also been validated over a decade (Keith Quille, 2016). The current model dataset is available to use at time of application within the first two to four weeks of the academic term.

7. Conclusion

The study demonstrated the application of machine learning in education with regard to the issue of student retention, that has challenged educators and administrators for many years. This model demonstrated a valid tool to predict student retention and subsequent student success. The study produced a generalizability model predicted that early intervention support would be appropriate for VTOS students, with a 64% accuracy, from attributes (County, Age, No-Months, Education-Status, Nationality and National-Status).

Research suggests that the introduction of data attributes possibly from additional student questionnaire or from early test scores or attendance records may produce additional information that could raise the model's accuracy in predicting student's success rates to avail of early intervention. With each of these factors investigated and accuracies recorded, it must be noted that the results presented must be interpreted with caution due to the size of the data sets used in some experiments and the period in which they were conducted. The current model dataset is available to use at the time of application within the first two to four weeks of the academic term and is able to identify six in ten students correctly and direct them to early intervention services which may increase retention and their individual success.

References

Bergin, Susan, A. M. (2015, December). *Using Machine Learning Techniques to Predict Introductory.* Retrieved February 21, 2012, from http://keithquille.com/PublicationData/MachineLearningTechniques.pdf

Brownlee, J. (2016). *How to Normalize and Standardize Your Machine Learning Data in Weka*. Retrieved April 1, 2018, from https://machinelearningmastery.com/normalize-standardize-machine-learning-data-weka/

Brownlee, J. (2016). *How to Perform Feature Selection With Machine Learning Data in Weka*. Retrieved April 1, 2018, from https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/

Brownlee, J. (2016). *How to Use Ensemble Machine Learning Algorithms in Weka*. Retrieved April 1, 2018, from https://machinelearningmastery.com/use-ensemble-machine-learning-algorithms-weka/

Brownlee, J. (2016). *How To Work Through a Binary Classification Project in Weka Step-By-Step*. Retrieved April 1, 2018, from https://machinelearningmastery.com/binary-classification-tutorial-weka/

Brownlee, J. (2016). *How To Work Through a Multi-Class Classification Project in Weka*. Retrieved March 17, 2018, from https://machinelearningmastery.com/multi-class-classification-tutorial-weka/

Brownlee, J. (June 30, 2016). *How to Better Understand Your Machine Learning Data in Weka*. Retrieved April 1, 2018, from https://machinelearningmastery.com/better-understand-machine-learning-data-weka/

Curzon, L. B. (1997). *Teaching in Further Education. An Outline of Principles and Practice* (5th ed.). London: Cassell.

Eivers, E. R. (2002). *Non-Completion in Institutes of Technology: An Investigation of Preparation, Attitudes and Behaviours among First Year Students.* Retrieved February 20, 2019, from http://www.erc.ie/documents/non-completionit02.pdf

Finnegan, T. F. (2010, August 28). *Retention and Progression in Irish Higher Education.* Retrieved February 20, 2019, from http://mural.maynoothuniversity.ie/2449/1/Access_and_Progression_in_Irish_Higher_Education_Fleming.pdf

HEA. (2019). An Analysis of Completion in Irish Higher Education: 2007/08 Entrants. Dublin: Higher Education Authority.

Healy, M. A. (1999, january 1). *Non-completion in Higher Education: A Study of First Year Students in Three Institutes of Technology.* Retrieved february 20, 2019, from https://www.esri.ie/publications/non-completion-in-higher-education-a-study-of-first-year-students-in-three-institutes

Hurwitz, J., Kirsch, D., (2018). *Machine Learning for Dummies IBM Limited Edition* (First ed.). NJ 07030-5774: John Wiley & Sons, Inc.

Kelly, E. O. (2019, february 21). *HEA: Computing courses have highest level of student drop out.* Retrieved April 18, 2019, from https://www.rte.ie/news/education/2019/0214/1029474-hea-student-study/

Martinez, P. a. (1998). *9,000 Voices: student persistence and dropout in further Education. London, Learning and Skills Development Agency. Further Education Development Agency (FEDA) 1998.* Retrieved february 20, 2019, from https://www.voced.edu.au/content/ngv%3A34027

Pezzullo, J. C. (2019). *Analysis of Variance from Summary Data*. Retrieved 03 25, 2019, from http://statpages.info/anova1sm.html

Quille, K. Bergin S., Mooney A., (2015, July). *PreSS#, A Web-based Educational System to Predict Programming Performance.* Retrieved February 22, 2019, from http://keithquille.com/PublicationData/PreSS.pdf

Quille, K., Bergin S., (2016). *Programming: Further Factors that Influence Success.* Retrieved April 24, 2019, from http://www.keithquille.com/PublicationData/ppig2016.pdf

Rosenberg. (1965). *Society and the adolescent self image.* Princeton Universtiy Press.

Steinberg, e. a. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, Volume 14*(Issue 1), pp 1–37.

Tinto, V. (1975). *Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45, 89-125.* Retrieved February 20, 2019, from https://files.eric.ed.gov/fulltext/ED078802.pdf

Witten I.H & Frank E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann Publishers.

Yorke, M. (1999). *Leaving Early: Undergraduate Non-Completion in Higher Education.* London.: Falmer Press,.