5-2010

Predicting Dropout in Online Courses: Comparison of Classification Techniques

Rajeev Bukralia
Dakota State University, bukraliar@pluto.dsu.edu

Predicting Dropout in Online Courses: Comparison of Classification Techniques

Rajeev Bukralia
Dakota State University
bukraliar@pluto.dsu.edu

ABSTRACT

Due to the tremendous growth in e-learning in recent years, there is a need to address the issue of attrition in online courses. Predictive modeling can help identify students who may be "at-risk" to drop out from an online course. This study examines various categorical classification algorithms and evaluates the accuracy of logistic regression (LR), neural networks (Multilayer Perceptron), and support vector machines (SVM) models to predict dropout in online courses. The analyses with LR, MLP, and SVM indicated that current college GPA is the strongest predictor of online course completion.

Keywords

Predictive modeling, academic analytics, classification techniques, online education.

INTRODUCTION

The issues of student retention and dropout are important in traditional face-to-face programs as well as in e-learning programs. It is important to identify at-risk students – or those students who have a greater likelihood of dropping out of a course or program – as it can allow instructors and advisors to proactively implement appropriate retention strategies. Online enrollments have been growing at rates far in excess of the total higher education student population (Allen & Seaman, 2007). A report on e-learning exhibits that 93.5% of 1539 institutions of higher education surveyed believe that the demand for e-learning is growing, and 90.3% of those institutions expect their online enrollment to grow (Allen & Seaman, 2008). Due to tremendous growth in e-learning in recent years, there is a strong need to address the issue of attrition in online courses.

There are studies that have investigated the role of academic and non-academic variables in student retention in face-to-face programs, such as high school GPA and ACT scores – both of which were found to be good predictors of retention rates in face-to-face programs (Lotkowski et al., 2004; Campbell and Oblinger, 2007). Some research specific to e-learning retention rates indicated that females tend to be more successful at online courses than males (Rovai, 2001; Whiteman, 2004). However, another study showed that gender had no correlation with persistence in an e-learning program (Kemp, 2002). Some studies have pointed out the relevance of age as a predictor of dropout rates in e-learning (Whiteman, 2004; Muse, 2003). Diaz, though, found that online students were older than traditional face-to-face students, but there was not a significant correlation between age and retention (Diaz, 2000). External variables such as self-motivation, effective time management, and technical preparation were also found to have an effect on the retention of online students (Diaz, 2000). A study conducted by Park (2007) showed that age, gender, ethnicity, and financial aid eligibility were good predictors of successful course completion. The study conducted by Yu (2007) tied earned hours with student retention in online courses.

Many studies that were conducted to identify high-risk students used statistical models based on logistic regression (Willging and Johnson, 2004; Hopkins, 2008; Pittman, 2008; Araque et al., 2009; Newell, 2007). The rationale for using logistic regression (LR) in the retention problem is that outcome is typically binary (graduated or not graduated) and probability estimates can be calculated for combinations of multiple independent variables (Pittman, 2008). Neural networks have also been used in classification problems of identifying academically at-risk students. Herzog (2006) and Campbell (2008) compared neural networks with regression to estimate student retention and degree-completion time. One study showed both logistic regression and neural networks were similar in performance in predicting graduation (Campbell, 2008); however it was not meant specifically for online students and did not study course level completion/dropout. This study compares the

predictive accuracy of logistic regression, Multilayer Perceptron (MLP), and support vector machines (SVM) models to predict dropout in online courses by analyzing data from Student Information Systems (SIS).

CLASSIFICATION MODEL

The prediction regarding whether a student will complete or drop a course can be categorized as a binary classification model. A literature review can help identify constructs and variables important to course completion or dropout. Since superfluous variables can lead to inaccuracies, it is important to remove them. Stepwise selection is a common approach for removing superfluous variables.

To build a classification model, the following questions should be carefully addressed:

- Which classification algorithms are appropriate in relation to problem characteristics?
- How is the performance of selected classification algorithms evaluated?

The following criteria should be considered to choose appropriate classification algorithms: the nature of the prediction (whether both class labels and probability of class membership are needed), nature of independent variables (continuous, categorical, or both), and the algorithmic approach (black box or white box). Both statistical and machine learning algorithms have been used to build classification models to identify class membership (for example, whether a student completed or dropped the course). Logistic regression, discriminant analysis, and logit analysis are common statistical algorithms used to predict categorical dependent variables. Logistic regression is appropriate if the independent variables are a mix of continuous and categorical variables. Discriminant analysis is appropriate when independent variables (or predictors) are continuous and evenly distributed; while logit analysis is suitable when all of the independent variables are categorical. The following machine learning algorithms are commonly used in classification models: neural networks, decision trees, and support vector machines (SVM). SVM assigns class labels (0 or 1) but it does not provide probability of class membership. In decision trees, continuous variables are implicitly discretized by the splitting process and loose information along the way (Dreiseitl et al., 2003). Since the proposed model uses a mix of continuous and categorical independent variables, LR, MLP, and SVM are better suited to perform binary classification.

The classification models are evaluated using two criteria: discrimination and calibration. Discrimination is used to demonstrate how well two classes (0 or 1) are separated, while calibration provides the accuracy of probability estimates. Common measures of discrimination include: accuracy, precision, specificity, and sensitivity. Sensitivity measures how often we find what we are looking for, while specificity measures how often we find what we are not looking for. Accuracy measures how well a binary classification test correctly identifies or excludes a condition. Precision measures the proportion of the true positives against all the positive results. In a ROC (receiver operating characteristic) analysis, sensitivity (for example, presence of course completion) is plotted against 1-specificity (absence of course completion) for each possible decision threshold. Accuracy can be measured by the area under the ROC curve, so its values for both logistic regression and MLP models are used as a basis of comparison. The performance of each model is assessed by measuring the area under the ROC curve and classification tables. MLP can be prone to overfitting, so a cross-validation strategy should be deployed. The model should be tested on a separate dataset to provide an unbiased estimate of generalization error.

METHODOLOGY

A literature review was conducted to identify constructs and variables from the Student Information System linked to online course dropout. The following constructs were identified for their role in course completion: academic ability, financial support, academic goals, and demographics. Each construct is mapped to their respective variables (see table 1). Stepwide regression was performed to determine the appropriateness of each selected independent variable. The sensitivity of each independent variable was analyzed to predict a binary, dependent variable: Course Completion Status (CCS). A de-identified dataset of 269 students was gathered from undergraduate online courses taught in Spring 2009 at a small Midwest university in the United States. The data were analyzed using LR, MLP, and SVM. The accuracy of the three models was compared on the basis of the classification matrix and the area under the ROC curve.

Dependent Construct	**Dependent Variable**	**Data Type**
Course Completion	• Course Completion Status	Categorical (course completed=1, Dropped=0)
Independent Construct	**Independent Variable(s)**	

Academic Ability	• ACT Comp Score (ACT_Comp) • High school GPA (HS_GPA) • Current College GPA (Current_GPA)	Continuous
Financial Support	• Financial Aid Status (Fin_Aid)	Categorical (Financial Aid=1, No Financial Aid=0)
Academic Goals	• Credit Completed (Cred_Hours_Compl) • Previous Drops (Past_Drops) • Degree Seeking Status (DegSkngStat)	Continuous Continuous Categorical (Degree Seeking=1, Not Degree Seeking=0)
Demographics	• Gender • Age	Categorical (Male=1; Female=0) Continuous

Table 1: Constructs and Variables

DATA ANALYSIS

The study used logistic regression, MLP, and SVM analyses to analyze independent variables to predict a binary, dependent variable: Course Completion Status (CCS). The CCS variable is the indicator of whether a grade was posted at the end of the course (CSS=1), or whether the student withdrew before course completion (CSS=0). All 269 cases were included in Binary LR, MLP, and SVM analyses. Table 2 shows the variable coding for categorical independent variables.

Parameter Coding		Frequency
Gender	Female (0)	208
	Male (1)	61
Degree Seeking Status	Not Degree Seeking (0)	11
	Degree Seeking (1)	258
Financial Aid Status	No Financial Aid (0)	54
	Financial Aid (1)	215

Table 2: Categorical Variables Coding

The logistic regression analysis (table 3) shows that Current College GPA (p-value or sig.=0; Wald=16.81) is the strongest predictor of course completion. Financial Aid Status (p=.024; Wald=5.12) and Degree Seeking Status (p=.084; Wald=2.99) are also strong predictors; while age (p=0.856) and HS_GPA (p=0.731) are found to be weak predictors at 95% CI. The dataset indicates that online courses have a higher percentage of non-traditional or adult learners, thus ACT Score and High School GPA may not be good predictors of course completion for older learners. The Wald estimates provide the importance of each variable in the model. The Wald estimates for current College GPA, Financial Aid Status, and Degree Seeking Status are significantly higher than other independent variables, thus they appear to be stronger predictors of online course completion. The Exp(B) provides odd ratios. Since Current College GPA is a numerical variable, an increase in one point in GPA has a 3.68 times (or 368%) increase in online course completion (95% CI ranging from 1.97 to 6.88). The analysis indicates that a degree seeking student has a 3.18 times (or 318%) greater likelihood of completing an online course. Multicollinearity can create bias in logistic regression, so it is important to check for it. The correlation matrix and standard error (SE) are examined for each variable to spot multicollinearity. SEs (as shown in table 3) are low, implying that multicollinearity does not exist and the model is stable. The analysis shows the prediction accuracy is highest (Hosmer-Lemeshow goodness of fit = 0.734) when all independent variables are included to predict the binary dependent variable. Table 4 is the classification table, which provides information about the overall accuracy to predict students having online course completion using logistic regression model. It provides overall 72.9% accuracy with a predicted probability of 0.5 or greater. The sensitivity is 93.3%, which indicates that the logistic regression model is 93.3% accurate in identifying true positives; however it has low specificity (18.9%), indicating that this model needs improvement in identifying actual negatives (or dropouts). The area under the ROC curve (AUROCC) is another measure of accuracy of a predictive model. It measures discrimination - the ability of the test to correctly classify those with and without course completion. AUROCC

ranges from 0 to 1 and values near 0.5 would mean the model is as good as flipping a coin. The AUROCC is 0.727 (as shown in table 5), which means that this model will assign almost 73% higher probability to students with course completion. The AUROCC shows that predicted probability has at least one tie between the positive actual state group and the negative actual state group, so bias cannot be completely ruled out.

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1	HS_GPA	.128	.373	.118	1	.731	1.137	.547	2.362
	ACT_Comp	-.040	.049	.672	1	.412	.961	.873	1.057
	Current_GPA	1.305	.318	16.819	1	.000	3.688	1.977	6.881
	Cred_Hours_Compl	.003	.004	.625	1	.429	1.003	.996	1.010
	Past_Drops	-.047	.071	.428	1	.513	.954	.830	1.098
	Fin_Aid(1)	.816	.361	5.128	1	.024	2.263	1.116	4.586
	DegSkngStat(1)	1.158	.669	2.993	1	.084	3.184	.857	11.824
	Gender(1)	.504	.372	1.831	1	.176	1.655	.798	3.432
	Age	.008	.045	.033	1	.856	1.008	.924	1.100
	Constant	-4.421	1.927	5.265	1	.022	.012		

Table 3: Regression Table

Observed		Predicted		
		Course Completion Status		Percentage Correct
		Dropped	Completed	
Course Completion Status	Dropped	14	60	18.9
	Completed	13	182	93.3
Overall Percentage				72.9

Table 4: Classification Matrix for Logistic Regression Model[a]

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.727	.034	.000	.661	.793

Table 5: Area Under ROC Curve for Logistic Regression

The Multilayer Perceptron (MLP) is a function of independent variables that minimizes the prediction error of dependent variables. The study employs a cross-validation strategy of 70% training set and 30% testing set. The model uses one hidden layer with hyperbolic tangent function. The case processing summary shows that 193 cases were assigned to the training sample and 76 to the testing sample. The MLP classification table (see table 6) shows that 72.5% of the training cases are classified correctly, corresponding to the 27.5% of incorrect cases. The testing sample is used to validate the model. In the testing set, 75% of cases are correctly classified by the model. In other words, the model is correct about three out of four times. The model works better to identify students with course completion (98.2%) and has a low rate of identifying dropouts (5.3%). The AUROCC for MLP is 0.677, which means this model will assign almost 68% higher probability to subjects with course completion.

Sample	Observed	Predicted		
Training	Dropped	7	48	12.7%
	Completed	5	133	96.4%
	Overall Percent	6.2%	93.8%	72.5%
Testing	Dropped	1	18	5.3%
	Completed	1	56	98.2%
	Overall Percent	2.6%	97.4%	75.0%
Dependent Variable: Course Completion Status				

Table 6: MLP Classification Matrix

	Importance	Normalized Importance
Financial Aid Status	.017	6.0%
Degree Seeking Status	.055	19.1%
Gender	.057	19.7%
High School GPA	.064	22.0%
ACT Comp Score	.175	60.2%
Current College GPA	.290	100.0%
Credit Hours Completed	.045	15.3%
Previous Course Withdrawals	.235	80.9%
Age	.063	21.7%

Table 7: Independent Variable Importance in MLP

SVM are a supervised set of machine learning algorithms that have been used in binary classification, pattern recognition, regression, and time-series forecasting. They can produce a binary classifier through a non-linear mapping of input vectors into the high-dimensional feature space. The poly-kernel SVM algorithm with a 70% training set and 30% testing set provides 75.3% accuracy and the AUROCC of 0.5% (and RMSE=0.49). The relative importance of attribute weight is highest for current college GPA (0.22).

CONCLUSION

The data analysis indicates that the models are somewhat comparable in overall accuracy for predicting online course completion/dropout: MLP and SVM models classified 75% cases accurately, and their accuracy was higher than the logistic regression model (accuracy of 72.9%). The logistic regression performed slightly better in AUROCC (.727 compared to 0.677 in MLP and 0.5 in SVM). Both the LR and MLP models were found to be weak in predicting true negatives (dropout): 18.9% in LR compared to 5.3% in the testing set of MLP. The lower accuracy in predicting true negatives may be due to the small dataset used in the study (n=269). Neural networks require larger datasets for better classification accuracy. The SVM classification accuracy may be improved by feature selection and proper model parameters, so it would be helpful to limit input features. All models gave varied importance to independent variables to classify the binary outcome; however they indicated that the current college GPA is the strongest predictor of online course completion. Age and gender were found to be weak predictors of online course completion, which is contrary to some studies mentioned previously. The models provide a relative importance of each analyzed variable, which can be used to designate a risk score to students based on their probability to drop out. It would be helpful to construct an early alert system using the risk score so that retention personnel could rank and contact the "at risk" students proactively to reduce attrition. There is a need to include additional variables such as course characteristics, technology preparedness, and course usage (or activity in Learning Management Systems), in addition to a larger dataset, to improve the predictive accuracy of the models – which will be addressed in future work.

REFERENCES

1. Allen, I. and Seaman, J. (2007) Online Nation: Five years of growth in online learning. Sloan Consortium.
2. Araque, F., C. Roldán, et al. (2009) Factors influencing university drop out rates. *Computers & Education,* 53, 3, 563-574.
3. Campbell, J (2008) Analysis of institutional data in predicting student retention utilizing knowledge discovery and statistical techniques. Ed.D. dissertation, Northern Arizona University, United States -- Arizona. Retrieved March 12, 2009, from Dissertations & Theses: Full Text database.
4. Campbell P. J. and Oblinger D. (2007) Academic Analytics. p.3. Retrieved from http://net.educause.edu/ir/library/pdf/PUB6101.pdf
5. Dreiseitl, S. and Ohno-Machado, L. (2002) Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics,* 35, 5-6, 352-359.
6. Diaz, D. P. (2002) Online Drop Rates Revisited. *The Technology Source*.
7. Herzog, S. (2006) Estimating student retention and degree-completion time: Decision trees and neural networks vis-a-vis regression. *New Directions for Institutional Research,* 17-33.
8. Hopkins T H. (2008) Early Identification of At-Risk Nursing Students: A Student Support Model. *Journal of Nursing Education*, 47,6, 254-9.
9. Lotkowski, V., Robbins, S., and Noeth, R. (2004) The role of academic and non-academic factors in improving college retention: ACT policy report. ACT. P.vii.
10. Muse, H. (2003) The Web-based community college student: An examination of factors that lead to success and risk. *The Internet and Higher Education*, 6,3, 241−261.
11. Newell, C. (2007) Learner characteristics as predictors of online course completion among nontraditional technical college students. Ed.D. dissertation, University of Georgia.
12. Oblinger, D., and Brian L. (2005) The Myth About E-Learning. *EDUCAUSE Review* 40, 4, 14-15.
13. Park, J.H. (2007) Factors Related to Learner Dropout in Online Learning. *International Research Conference in the Americas of the Academy of Human Resource Development,* Indianapolis, IN.
14. Pittman, Kathleen (2008) Comparison of data mining techniques used to predict student retention. Ph.D. dissertation, Nova Southeastern University, United States -- Florida. Retrieved March 8, 2009, from Dissertations & Theses: Full Text database.
15. Porter, O. F. (1990) Undergraduate completion and persistence at four year colleges and universities: detailed findings. Washington, DC. National Institute of Independent Colleges and Universities.
16. Rovai, A. P. (2003) In search of higher persistence rates in distance education online programs. *The Internet and Higher Education* 6,1, 1-16.
17. Tinto, V. (1987) Leaving College: Rethinking the Causes and Cures of Student Attrition. Chicago: University of Chicago Press.
18. Willging P. and Johnson S. (2004) Factors that influence students' decision to drop out of online courses. *JALN*, 8, 4.
19. Whiteman, J. M. (2004) Factors associated with retention rates in career and technical education teacher preparation web-based courses. Doctoral dissertation, University of Central Florida Orlando, Florida. Retrieved April 4, 2006 from: http://etd.fcla.edu/CF/CFE0000210/Whiteman_JoAnn_M_200412-1EdD.pdf
20. Yu, C. H., S. A. DiGangi, et al. (2007) A Data-Mining Approach to Differentiate Predictors of Retention. *EDUCAUSE Southwest Conference*, Austin, TX.