DETERMINING FACTORS AFFECTING STUDENT RETENTION IN A HIGHER EDUCATION INSTITUTE IN TAIWAN AND BUILDING A PREDICTION MODEL USING LOGISTIC REGRESSION AND SUPPORT VECTOR MACHINE

FUMEI WENG [1, 2], FRANCE CHEONG [1], CHRISTOPHER CHEONG [1]

[1] *School of Business Information Technology, RMIT University, 239-251 Bourke Street, Melbourne, VIC, 3000, AU*

[2] *Department of Information Management, WuFeng Institute of Technology, No.117, Sec. 2, Jianguo Rd., Minsyong Township, Chiayi County 621, Taiwan*

The purpose of this research is to select the attributes that best influence student retention and build a model that can predict student retention. A longitudinal dataset for 2353 college students from a private higher education institute in Taiwan for the 2003 to 2005 academic years was used for the study. Cross-tabulation and Pearson tests were used for attribute selection. Two prediction models using the same attributes were built: one using logistic regression and another one using SVM. Logistic regression found five attributes to significantly affect to dropout: major, residence, GPA, loan and absenteeism. The SVM prediction model was found to have better prediction accuracy than logistic regression.

1. Introduction

Student retention in higher education institutes and universities has long been a concern over the last two decades. In the U.S., five-year graduation rates dropped about 6% from 58% to 52% in the 1980s and 1990s (Mortenson 1998). In the UK, the rate of non-completion of degrees was 13% and 17% in 1982 ~ 1983 and 1997 ~ 1998, respectively (Select committee on Education and Employment 2001). Institutes need to know which students will enroll, need assistance to complete their degree requirements and drop out.

In the last couple years, there was a revolution in education institutes in Taiwan. The number of institutes of higher education has increased and birth rate has reduced. Student retention is more efficient than recruiting students for maintaining student population in Taiwan.

Published studies on student retention have focused on examining the relationship among demographic variables, causes of student attrition, specific campus programs, and teaching techniques. Logistic regression is the most commonly used technique on student retention, but often criticized due to its strong model assumptions like variation homogeneity and linear relationships (Lee et al. 2006). In contrast to traditional statistical techniques, the machine

learning technique, Support Vector Machine (SVM) derived from neural networks technique, does not require knowledge of the relationships between input and output variables (Huang et al. 2004). In order to provide accurate information using students' attributes for school administrators to design intervention and make decision, there is a need to build a better model on student retention.

2. Literature review

There has been growing interests in the construction of models and theories of student retention to explain the complex interactions of factors that affect student persistence or dropout (Mannan 2007). Studies focused on specific factors on retention have been investigated, for example, student race and gender (Leppel 2002), major choice, financial status (St. John et al. 2004), college grade point average (GPA) (Mannan 2007), and admission status. Individual student's demographic and academic performance variables are generally available from the student information system or enrollment system at each institute.

Logistic regression has often been used in student retention studies in higher education to deal with dichotomous dependent variables (Peng et al. 2002). Factors such as: gender (Leppel 2002), age, ethnicity, marital status, number of children, hours working, high school GPA, and first-quarter college GPA (Murtaugh, Burns & Schuster 1999) have been used with logistic regression models to examine their relationships with dropout.

An appropriate technique for classification problem, Support Vector Machine (SVM), was developed to solve classification problems (Vapnik & Smola 1996). SVM has been applied to financial time series forecasting and was found to have better predicting accuracy than others (Tay & Cao 2001). In a bankruptcy prediction model, SVM was found to be superior than neural networks (Shin, Lee & Kim 2004). Although SVM has yielded excellent generalization performance on a wide range of problems, it has not been used for predicting student retention in higher education (Huang et al. 2004).

The object of this study is to discover the determinants of student retention and use these factors to predict student retention in logistic regression and SVM models.

3. Methodology

The dataset used in this study was obtained from a private higher education institute in Taiwan, covering all 4-year college students and the time period was

from 2003 till 2005. It included 14 independent variables and one dependent variable. Independent variables are: demographic variables (major, sex, age, pre-college school type, entrance test score, admission basis and residence), academic performance (first semester credits, first semester GPA, second semester credits, second semester GPA and absence of class), financial variables (tuition deduction and financial loan). The dependent variable is dropout from college.

Table 1. Variables selected

Variables	Values and description
Major	1 = engineering, 2 = business, 3 = social science, 4 =security
Sex	1 = male, 2 = female
Age	1 = < 18-yrs, 2 = [19-21], 3 = > 22-yrs.
Pre-college school type	1 = high school, 2 = vocational high school
Entrance test score	0.0 - 100.0
Special admission status	1 = general admission, 2 = special admission
Zip	1 = north, 2 = middle, 3 = local, 4 = south, 5 = east
First semester credits (FSCRD)	1 = < 18 credits, 2 = [19 - 22], 3 = > 23 credits
First semester GPA (FSGPA)	1 = < 59.9, 2 = [60 - 74.9], 3 = [75 - 84.9], 4 = > 85
Second semester credits (SSCRD)	1 = < 18 credits, 2 = [19 - 22], 3 = > 23 credits
Second semester GPA (SSGPA)	1 = < 59.9, 2 = [60 - 74.9], 3 = [75 - 84.9], 4 = > 85
Tuition deduction	0 = without tuition deduction, 1 = with tuition deduction
Loan	0 = without loan, 1 = with loan
Absence	0 = < 10 classes, 2 = [11 – 20], 3 = > 21
Dropout	0 = dropout, 1 = persistence

The 14 variables shown in Table 1 were either cross-tabulated or correlation tested with the dropout categories to ensure a relationship between the independent variable and the dependent variable. After having obtained the predictor variables, they were used to create the logistic regression and SVM models. Regarding the validation of the SVM model, ten-cross fold validation was adopted as it is the standard way of measuring the error rate of a learning scheme on a particular dataset (Witten & Frank 2005). The SVM parameters used in this study were *c* (regularization parameter) =10 and *gamma* (kernel parameter for radial basis function) = 0.00043 (1/2353).

In order to validate the proposed models, the dataset was partitioned into three cohorts based on academic year, 2003, 2004, and 2005, as academic year

had no significant relationship with the dependent variable in the preliminary test. Each cohort was used as the training set for creating the prediction models, and the remaining two cohorts were used as the testing set for validating the models. After the two models were created, prediction accuracies of logistic regression and SVM were compared.

4. Results and discussion

4.1. *Descriptive statistics*

From the three cohorts partitioned by academic year, the number of students and the dropout rates were computed and are shown in Table 2. The average dropout rate was found to be 20.7%.

Table 2. Number of students and dropout rates in three cohorts

Cohort	Number of Dropouts	Number of Persistence	Total (Number)	Dropout rate (%)
Year 2003	139	620	759	18.3
Year 2004	170	599	769	22.1
Year 2005	179	646	825	21.7
Total	488	1865	2353	---
Average	162	622	784	20.7

Fourteen variables were either cross-tabulated or Pearson tested and eight predictors: *Major*, *Sex*, *Age*, *Zip*, *SSCRD*, *SSGPA*, *Loan* and *Absence*, were obtained in the proposed models based on statistical significance.

The descriptive statistics for student retention are shown in Table 3. There was a lower dropout rate in business and security majors (19.1% for business and 11.2% for security). Compared to sex, age, and loan, males older than 22 years and students without loan had a higher dropout rate than those younger than 22-years-old females, and students with loan, respectively. Locally resident students had a lower dropout rate than non-local ones. The *SSGPA* had a significant influence on student dropout. The dropout rate decreased with increasing *SSGPA* ranging from 66.6% to 7.9%. Similarly, the dropout rate also decreased with the decreasing absenteeism ranging from 75% to 18.1%.

4.2. *Logistic regression*

The coefficients of logistic regression model created using the eight variables are shown in Table 4. The coefficient (*B*) shows the relationship between the independent variables and the dependent variable. *Exp*(*B*) represents the ratio of change in the odds of the event of interest for a one-unit change in the predictor.

For example, the *Exp*(*B*) of *SSGPA* strongly affects dropout in three cohorts as *Exp*(*B*) for *SSGPA* is equal to 2.82, meaning that the odds of *SSGPA* for a dropout student is 2.82 times the persistent student.

Table 3. Descriptive statistics of student retention

Variable		Number of Dropouts	Dropout rate (%)	Number of Persistence	Persistence rate (%)	Total (Number)
Major	Engineering	234	23.1	778	76.9	1012
	Business	126	19.1	535	80.9	661
	Social science	89	26.8	243	73.2	332
	Security	39	11.2	309	88.8	348
Sex	Male	375	22.2	1318	77.8	1693
	Female	113	17.1	547	82.9	660
Age	< 18	251	19.5	1038	80.5	1289
	19 - 21	224	21.9	801	78.1	1025
	> 22	13	33.3	26	66.7	39
Zip	North	140	25.9	401	74.1	541
	Middle	124	23.8	397	76.2	521
	Local	129	13.3	844	86.7	973
	South	77	30.6	175	69.4	252
	East	18	27.3	48	72.7	66
SSCRD	< 17	360	22.7	1227	77.3	1587
	> 18	128	16.7	638	83.3	766
SSGPA	< 59.9	232	66.7	116	33.3	348
	60 - 74.99	104	16.5	528	83.5	632
	75 - 84.99	124	12.2	893	87.8	1017
	> 85	28	7.9	328	92.1	356
Loan	Without loan	372	25.6	1081	74.4	1453
	With loan	116	12.9	784	87.1	900
Absence	< 10	381	18.1	1727	81.9	2108
	11 - 20	71	36.0	126	64.0	197
	> 21	36	75.0	12	25.0	48
Total		488	20.7	1865	79.3	2353

Five variables, namely: *Major*, *Zip*, *SSGPA*, *Loan*, and *Absence* were found to have a significant influence on *Dropout* in three cohorts. However, two variables, *Sex* and *Zip* (local), showed a significant influence on *Dropout* in year 2005 only. For each cohort, major in engineering, business and social science were negatively associated with the probability of retention. This indicates that students' major has a significant influence on student retention. This result is consisted with St. John et al's study (2004).

SSGPA also had a significant influence on student retention. Students whose second semester GPA was less than 60 (i.e. failed) were more likely to dropout. Absenteeism also had a negative relationship with dropout. The higher the absenteeism is, the less likely the student is to persist. Local residence had lower dropout rate than non-local residence.

Table 4. Coefficients of logistic regression for student retention

Variables	2003			2004			2005		
	B	Sig.	Exp(B)	B	Sig.	Exp(B)	B	Sig.	Exp(B)
Major		***			***			***	
Major (engineering)	-.90	***	.40	-.86	**	.42	-.54	***	.58
Major (business)	-.81	***	.44	-1.16	**	.31	-.43	**	.64
Major(social science)	-1.29	***	.27	-1.54	***	.21	-.97	***	.37
Sex (male)	.27		1.31	.05		1.05	.30	**	1.35
Age			.40						
Age (> 22)	-.90		.71	.08		1.09	-.55		.57
Age (< 18)	.09		1.10	-.14		.86	.06		1.07
Zip		***			***			***	
Zip (north)	-.49		.61	-.69		.49	-.53		.58
Zip (middle)	-.09		.90	-.49		.61	-.41		.66
Zip (local)	-.90		1.90	.47	*	1.60	.28	*	1.33
Zip (south)	-.55		.57	-.34		.70	-.68		.50
SSCRD	.17		1.18	.03		1.03	.20		1.22
SSGPA	1.03	***	2.82	.96	***	2.62	1.09	***	2.97
Loan (without)	-.45	***	.63	-.09		.90	-.75	***	.46
Absence	-.32	***	.72				-.44	***	.64

*Coefficient significant at .05; **coefficient significant at .01, and ***coefficient significant at .001

4.3. *Prediction performance of logistic regression and SVM*

The Hosmer-Lemeshow statistics used in logistic regression indicated a good fit as the significance value was greater than 0.05. Pseudo R^2 is a statistic with a scale ranging from 0 to 1 (Stratton, O'Toole & Wetzel 2008). The goodness-of-fit of logistic regression was found to be statistically significant at α = 0.10 based on the chi-square test of the overall model adequacy as shown in Table 5. The values of R^2 were 0.37, 0.33, and 0.29 for the three cohorts, respectively.

While training, logistic regression and SVM achieved student retention prediction accuracies with an average of 72.5% and 80.1%, respectively. This indicates that SVM has better prediction accuracy for student retention than logistic regression.

Table 5. Goodness-of-fit of logistic regression and performance of logistic regression and SVM

Cohort	Hosmer& Lemeshow statistics	R^2	Accuracy (%)	
			Logistic regression	SVM
2003	0.63	0.37	71.4	78.1
2004	0.30	0.33	72.7	80.7
2005	0.65	0.29	73.3	81.4
Average	0.53	0.33	72.5	80.1

5. Conclusion

A dataset containing the records of 2353 students from 2003 to 2005 was obtained from a higher education institute in Taiwan. The dataset contained fourteen student attributes and the dropout rate was analyzed and found to be 20.7%. Cross-tabulation and Pearson tests were used to perform attributes selection and eight attributes were found to significantly determine students' dropout. They were: major, sex, age, zip, second semester GPA, second semester credits, loan and absence. The eight variables were used by both logistic regression and SVM to predict students' dropout. The main predictors of student retention in higher education institute in Taiwan obtained from logistic regression were five variables which were: major, residence, second semester GPA, loan, and absence of class. SVM was found to outperform logistic regression (80.1% vs. 72.5%).

References

1. Z. Huang, H. Chen, C. Hsu, W. Chen and S. Wu, *Decision Support Systems.* 37. 4, 543 (2004).
2. T. Lee, C. Chiu, Y. Chou and C. Lu, *Computational statistics & data analysis.* 50, 18 (2006).
3. K. Leppel, *Review of higher education.* 25.4, 18 (2002).
4. T. Mortenson, *Postsecondary education opportunity.* 73, 10 (1998).
5. P. Murtaugh, LD. Burns and J. Schuster, *Research in higher education.* 40.3, 17 (1999).
6. C. Peng, T. So, F. Stage and E. St.John, *Research in higher education.* 43.3, 35 (2002).
7. Select committee on Education and Employment 2001, 'the Select committee on Education and Employment', 6, 1.11(2001).
8. K. Shin,T. Lee and H. Kim, *Expert systems with applications*. 28.1, 9 (2004).
9. E. St.John, S. Hu, A. Simmons, D. Carter and J. Weber, *Research in higher education.* 45.3, 44 (2004).
10. L. Stratton, D. O'Toole and J. Wetzel, *Economics of Education Review.* 27.3, 319 (2008).
11. F. Tay and L. Cao, *The international journal of management science*, 29, 19 (2001).
12. V. Vapnik and A. Smola, *Advances in neural information processings systems*, MIT Press, Cambridge, MA (1996).
13. I. Witten, and E. Frank,*Data mining: practical machine learning tools and techniques*, Elsevier Inc., San Francisco (2005).