

**PREDICTING AN INDIVIDUAL RETENTION RATE  
USING A STATISTICAL ANALYSIS**

**HYUN-JOO KIM**

**Table of Content**

1. Introduction .....	3
2. Data and Analysis .....	4
2.1. Data .....	4
2.2. Analysis .....	5
2.3. Prediction and application .....	9
2.4. Model Validation .....	11
3. Conclusion .....	13
4. Future Study .....	14
Appendix .....	15
Reference .....	16

## 1. Introduction

Retention rate is a measure of student success in many universities. Retention is important because it implies how well a university and its community serve students, and it has an important role in the quality of the education. For the last few decades, there have been many studies related to the retention issue with many factors including gender (Spady (1971), Pascarella, Duby, and Iverson (1983), Stage and Hossler (1989)), race (Braxton, Duster, and Pascarella (1988), Stage and Hossler (1989)), the parents' educational background (Pascarella and Terenzini (1978), Stage (1988)), a student's educational aspiration (Bean (1982), Metzner and Bean (1987), college GPA (Bean (1982,1983), Cabrera et al., (1992, 1993)), the effect of financial aid (DesJardins et al. (1999), Hochstein and Butler (1983)), and many others. Truman State University is not an exception. For the last several years, there have been many studies about student retention, and we are trying to improve the current retention in Truman. (Ishiyama, 2000)

In this project, the relationship between an individual's retention rate and the possible causes discussed in some of these internal and external studies is statistically analyzed hoping that this will provide better understanding of incoming students and their needs, and ultimately increases the overall retention rate by providing appropriate help. This is based on the idea that if we can identify the students who have higher potential of leaving Truman earlier and also know the reasons of their struggle, we will be able to give more appropriate, more efficient, and more timely help to the students. The early identification is important especially because many students decide to leave Truman in the very first couple weeks of their freshman year. If we can act sooner, we will be able to prevent students from leaving early. Statistical analysis on freshmen data and their retention can provide such early identification.

Many previous studies have dealt with the relationship between retention and individual factors. This project is to find the relationship of retention and the potential factors collectively, so that the interactions and correlation between factors can also be taken into account. Pearson correlation is studied between the retention variable and each individual factor. An optimal set of factors that determine an individual's retention rate is found by a Logistic regression model. Whether an individual student is highly likely to leave Truman or not is predicted using the model developed and compared to the actual retention. The final model is analyzed using the model validation procedures including cross-tabulation and ROC curve.

## 2. Data and Analysis

### 2.1. Data

There are many factors that might affect students' decision about their staying or leaving Truman. Some students who are far away from their hometown might have a hard time adjusting themselves at the university. Some might find the courses to be too difficult. Some might lose their scholarship and become financially jeopardized. Some students might have a hard time finding the personal connection during the beginning of their college life and decide to try out other places. In this project, many possible factors are considered and analyzed with the retention, including financial aspects (family income, financial aid, how many hours working), academic aspects (expect to change a major, high school GPA, ACT composition scores, how many hours studying, and pursuing advanced degree), and social aspects (distance from home, ethnicity, gender, how many hours socializing, how many hours exercising, and whether or not a student is the first generation in a college). Table 1 summarizes the possible covariates (factors).

**Table 1. Possible factors on the retention**

<b>Financial factor</b>	<b>Academic factor</b>	<b>Social factor</b>
Family income	Expect to change major	Hometown (distance from Kirksville)
Financial aid	High school GPA	Ethnicity
How many hours working	ACT (composition)	Sex
	How many hours studying	How many hours socializing
	Pursue advanced degree	How many hours exercising
		First generation

Note that the number of financial aid sources is used for the financial aid variable rather than the amount that he/she received. The expected to change major variable has a value of 1 for no chance, 2 for little chance, 3 for some chance, and 4 for good chance to change major. Pursuing advanced degree is a categorical variable with 0 if a student is pursuing up to bachelor degree or 1 if pursue beyond bachelor degree. Race variable is characterized by White, African American, American Indian, Asian, and others. Gender is 0 for male and 1 for female. First generation is 1

for the first generation college student or 0 otherwise. Note that covariates are mixture of continuous, ordinal, or categorical variables.

2 year retention is the response variable. If a student is still in Truman in the beginning of her/his junior year, the response variable (retention) is 1, or 0 is recorded otherwise. 2 year retention is important because students often make their decision in 2 years rather than later in their junior or senior year.

CIRP(Cooperative Institutional Research Program, freshman survey) and CSEQ (College Student Experiences Questionnaire) data is organized by Dr. John Ishiyama (TSU, Political Science) and his students (Ishiyama, 2000). The analysis is based on this data. The freshman data from 1996 to 1997 (with 2 year retention in the beginning of the year 1998, and 1999) is used to develop a statistical model.

## **2.2. Analysis**

Initially, the correlations between the response variable and each factor are studied. Pearson correlation shows that family income (.071), hours working for pay (-.062), change of major (.068), high school GPA (.131), ACT (.101), study time (.079), first generation (-.087) are highly correlated to the retention. Pursuing highest degree, race, and socializing hours are moderately, and financial aid, gender, and exercising hours are weakly correlated with the retention.

A logistic regression model is developed to study the relations between a binary response variable and possible covariates. This is an appropriate statistical method for the current data, since the two year retention is a binary variable (0 or 1), and we are looking for the explanation of the relationship between retention and many covariates. Various model selection procedures are run including backward stepwise and forward stepwise regression. These are very common model selection procedures that either eliminate unnecessary variables from the model or include necessary variables in the model. Various significance levels are also used. The following table summarizes the variables chosen by some of the model selection procedure and significance levels.

**Table 2. Model selection by stepwise selection procedure**

<b>Model selection procedure</b>	<b>Chosen variables</b>
Backward stepwise ( $\alpha = .05$ )	Income, change of major, high school GPA, ACT composition scores, hour spent studying, pursuing highest degree, sex, first generation
Backward stepwise ( $\alpha = .1$ )	Income, change of major, high school GPA, ACT composition scores, hour spent studying
Forward stepwise ( $\alpha = .05$ )	Income, change of major, high school GPA, ACT composition scores, hour spent studying
Forward stepwise ( $\alpha = .1$ )	Income, change of major, high school GPA, ACT composition scores, hour spent studying, pursuing highest degree, sex, first generation

An optimal set of factors is not clear in this case since different procedures give a different set of choices for possible optimal set. In particular, pursuit of highest degree, sex, and first generation variables could be or could not be in the model depending on which model selection procedure is used. In fact, these model selection procedures are known to be less reliable than other more sophisticated statistical model selection tools. For more reliable model selection, many other model selection criteria have been recently studied and proposed.

AIC (Akaike Information Criteria, Akaike, 1973) and AICc (corrected AIC, Hurvich & Tsai, 1989) are two of the most well known model selection criteria that can be applied in logistic regression model. Both measures estimate the difference between the unknown true model and the current candidate model. Thus, the smaller values of these criteria imply that candidate model is closer to the true model; thus, the better model. AIC and AICc are computed for the possible candidate models and the results are provided in Table 3.

Note that between five and ten variables are considered for the candidate models. It is pretty certain that 5 variables: income, expect to change major, high school GPA, ACT composition score, hours spent studying are important in the model (all the models include these 5 variables.) However, it is not so obvious whether the other 5 variables (the five variables with the next smallest p value), pursuit of highest degree, sex, first generation, financial aid, Asian or some of them, should be included in the model or not. Table 3 summarizes the results of model selection

of these candidate models. Model 1 includes five essential variables in the model. Model 2 and 3 are the best two models among the models with 6 covariate variables (5 essential variables and one of the unobvious five variables). Model 4 and 5 are the best two models among with the models 7 covariate variables (5 essential variables and two of the possible 5 variables), and so on.

**Table 3. Model selection from AIC, AICc method**

	Model	AIC	AICc
1	Income, change of major, high school GPA, ACT composition scores, hour spent studying	2273.135	2273.172
2	Income, change of major, high school GPA, ACT composition scores, hour spent studying, pursuing highest degree	1991.843	1991.893
3	Income, change of major, high school GPA, ACT composition scores, hour spent studying, financial aid	1703.089	1703.139
<b>4</b>	<b>Income, change of major, high school GPA, ACT composition scores, hour spent studying, pursuing highest degree, financial aid</b>	<b>1491.680</b>	<b>1491.744</b>
5	Income, change of major, high school GPA, ACT composition scores, hour spent studying, first generation, financial aid	1699.966	1700.030
6	<b>Income, change of major, high school GPA, ACT composition scores, hour spent studying, pursuing highest degree, sex, financial aid</b>	<b>1490.593</b>	<b>1490.673</b>
7	Income, change of major, high school GPA, ACT composition scores, hour spent studying, sex, first generation, financial aid	1700.342	1700.422
<b>8</b>	<b>Income, change of major, high school GPA, ACT composition scores, hour spent studying, pursuing highest degree, sex, first generation, financial aid</b>	<b><u>1488.883</u></b>	<b><u>1488.981</u></b>
9	Income, change of major, high school GPA, ACT composition scores, hour spent studying, pursuing highest degree, sex, first generation, financial aid, Asian	1490.036	1490.154

Note that the model with variables of income, change of major, high school GPA, ACT composition score, hours spent studying, pursuit highest degree, sex, first generation, and financial aid has the smallest AIC and AICc. This means that the model 8 is the closest model to the unknown true model. Note that model selection procedure gives slightly different results from AIC and AICc result. This is also different from Pearson correlation study. Statistically, AIC and AICc result is considered to be more reliable than others. Thus, we will use the model 8 in Table 3 for the final model. (Note that since model 4 or 6 in Table 3 have similar AIC and AICc values as the model 8 and have less number of variables than the model 8, one might have similar prediction results using model 4 and 6 in Table 3.) The final model includes the following factors.

**Table 4. Variable included in the final model**

<b>Financial factor</b>	<b>Academic factor</b>	<b>Social factor</b>
Family income Financial Aid	Expect to change major High school GPA ACT (sub scores) How many hours studying Pursue advanced degree	Sex First generation

The general formula for logistic regression is

$$\ln \left( \frac{p(x)}{1-p(x)} \right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k + e,$$

where  $p(x)$  is the probability that a student retains at Truman after 2 years. From Table 5 (model coefficient), the final model equation can be written as

$$\begin{aligned} \ln \left( \frac{p(x)}{1-p(x)} \right) = & -3.295 + .054 \text{ Income} + .252 \text{ Change of major} + .234 \text{ High school GPA} \\ & + .035 \text{ ACT composition} + .138 \text{ Hour spent studying} + .023 \text{ Financial aid} \\ & + .264 \text{ Pursing highest degree} - .227 \text{ Sex} - .275 \text{ Fist generation.} \end{aligned}$$

Note that there are a few variables with larger  $p$  values than an usual significance level 0.05. However, again, AIC and AICc chose this model as the closest model to the true model, thus these variables are kept in the final model.

**Table 5. Model Coefficient**

	B	S.E.	Wald	Df	Sig.	Exp(B)
INCOME	.054	.028	3.608	1	.057	1.055
CHANGEM	.252	.071	12.476	1	.000	1.286
HSGPA	.234	.063	13.754	1	.000	1.264
ACTCOMP	.035	.023	2.226	1	.136	1.035
STUDING	.138	.047	8.733	1	.003	1.148
AIDGRANT	.023	.025	.849	1	.357	1.023
HIGHEST	.264	.168	2.471	1	.116	1.302
SEX	-.227	.140	2.616	1	.106	.797
FIRSTGEN	-.275	.142	3.747	1	.053	.760
Constant	-3.295	.732	20.266	1	.000	.037

From the previous logistic regression equation, several conclusions can be drawn.

- Sex and first generation variables have negative coefficients. This means that female students have smaller probabilities of staying at Truman than male students, and first generation students have smaller probabilities staying at Truman than non first generation students.
- The rest of the variables have positive coefficients. Students are more likely to stay at Truman if family income is higher, more likely to change their major, higher high school GPA and ACT composition score, more studying time, larger number of financial aid sources, and more likely to pursue a higher degree.
- Generally, academic factors have an important effect on retention. Expect to change major, high school GPA, and study hour especially have a great effect on the retention. Indeed, all the academic factors have statistically significant effect on retention.
- Family income and financial aid have significant effects on retention among financial factor. Sex and first generation have significant effects on retention among social factor.

### **2.3. Prediction and application**

Using the logistic regression model developed in the previous section, we can predict an individual's retention rate with the chosen covariates (factors). Once we recognize a student with a low potential retention rate, both academic and RCP advisors will be able to give appropriate advice for the student. This may influence the student positively and encourage them to stay at Truman. Some of the prediction examples are given in Table 6.

**Table 6. Individual actual and predicted retention rate**

	Actual Retention	Retention probability	Income	Change major	HGPA	ACT comp	Study hour	Financial Aid	Degree pursue	Sex	First generation
1	1	.702	6	3	7	24	3	16	1	2	0
2	1	.832	11	4	7	31	3	15	1	2	0
3	0	.699	8	2	6	34	1	19	1	1	0
4	0	.759	11	2	8	31	4	13	1	2	1
5	1	.704	7	3	8	23	4	13	0	1	1
<b>6</b>	<b>0</b>	<b>.443</b>	<b>7</b>	<b>4</b>	<b>5</b>	<b>22</b>	<b>2</b>	<b>9</b>	<b>0</b>	<b>2</b>	<b>1</b>
<b>7</b>	<b>0</b>	<b>.498</b>	<b>10</b>	<b>1</b>	<b>6</b>	<b>30</b>	<b>1</b>	<b>17</b>	<b>1</b>	<b>2</b>	<b>1</b>
8	1	.676	7	3	5	27	5	13	1	2	0
9	1	.670	9	3	6	22	4	12	0	1	0
10	0	.718	8	3	7	26	5	12	0	2	0

Actual retention has a value of 1 if a student retains after 2 years, or 0 otherwise. Retention probability is the predicted probability that the student remains at Truman. Generally, if the predicted retention probability is less than .5, the student is considered to have a strong potential to leave Truman. In Table 6, individual number 1 and 2 have pretty high retention probability, and they actually stay at Truman after 2 years (correctly specified). On the other hand, individual number 3, 4, and 10 left Truman even though they have higher than .5 retention probability (incorrectly specified). Also, individual number 6 and 7 has predicted retention probability of less than .5, and these students actually left Truman before their junior year. .5 is a common cutoff point, and one can say that the student with a predicted retention rate lower than .5 is highly likely to leave Truman in 2 years.

Once a student is recognized as the one who is highly likely to leave Truman, each factor of the student can be checked and some individual help can be provided. For example, the number 6 student has less than .5 predicted retention probability, and this student has a lower high school GPA, ACT composition score, smaller amount of study hour, and smaller number of financial aid. This student can be encouraged to study more hours and can be given information about many financial aids available. Student number 7 also has a predicted retention probability smaller

than .5. This student has smaller amount of studying hour and small chance to change her major. Increasing study hour and emphasizing the benefits and philosophy of liberal arts education might help this particular student.

#### 2.4. Model Validation

The model validation procedure is based on 1996-1997 freshmen data. Originally, 1998 data was going to be implemented as a cross-fold validation (based on a separate data set for the validation of the data used for the model building). Unfortunately, the 1998 data are not available yet. As an alternative, 1996-1997 data set is implemented. This analysis provides similar insight to the model prediction performance. Using the current logistic regression model the prediction of the retention is analyzed. In the validation procedure, a few other final candidate models are also discussed including: the models suggested by Pearson correlation, model selection procedure, and model selection criteria. 0.5 is a common for the cutoff point in practice. If the predicted probability is larger than 0.5, the individual is predicted to remain at the University, and if it is smaller than 0.5, predicted to leave the university. Note that one may use different cutoff points depending on which error is considered to be more crucial. The 2 by 2 cross-tabulation from the final model is shown in Table 7.

**Table 7. Cross-tabulation**

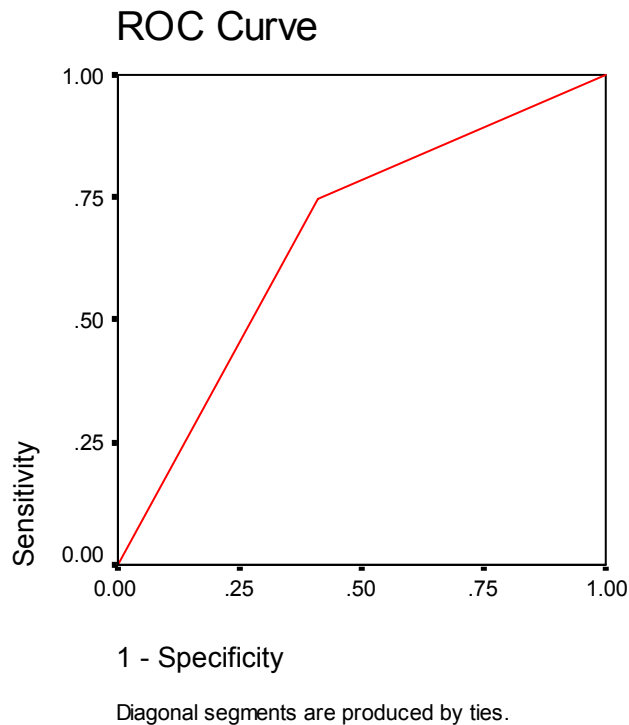
		Predicted retention		Total
		Leave	Stay	
2 year retention	Leave	20	331	351
	Stay	14	979	993
Total		34	1310	1344

Here, 20 and 979 represent the frequencies of correctly classified cases, and 331 and 14 represent the frequencies of misclassified cases. The matching coefficient,  $(20+979)/1344 = .743$ , represents the proportion of correctly classified cases. Different predicted models and cutoff points may cause different correctly classified or misclassified results, thus different matching coefficients. Higher matching coefficients implies a better model prediction. The cross-tabulation of the other candidate models are given in Appendix. Model 1, 2, and 3 in Appendix have the matching coefficient 72.9%, 73.4%, and 73% respectively. The current model provides the best correct prediction rate with the matching coefficient 74.3%. Note that the total numbers in the

cross-tabulation are different. For example, the current model has total 1344 data values, and model 1, 2, and 3 in Appendix have 2007, 1764, and 2013 data values respectively. This is because the missing values of independent variables causes missing predicted values, thus lose data points.

ROC (Receiver Operating Characteristic, well known data validation method for categorical data) curve is a summary statistic derived from several 2 by 2 contingency tables, where each contingency table corresponds to a different simulated scenario of the retention (Pontius And Schneider, 2001). Here, the “Sensitivity” is the probability that the model correctly classified the individual who is staying at the university, and the “Specificity” is the probability that the model correctly classified the individual who is leaving the university. In practice, the larger area under the ROC curve implies better prediction. The ROC curve for the current model is given in Figure 1.

**Figure 1.**



ROC curves for other candidate models are also available in Appendix. SPSS provide that the current model has the largest area under the curve, .668 comparing to the area .581, .643, and .571 for model 1, 2, and 3 in Appendix, respectively.

In summary, we observed that the current predicted model has the highest matching percentage from the cross-tabulation and the largest area under the ROC curve among other candidate models. Both cross-tabulation and ROC results support that the current model predicts the actual retention better than the other candidate models, and about 74.3% of time, it predicts the retention correctly.

### **3. Conclusion**

An early identification of a student with a high probability of leaving Truman will benefit both the university and the student. The university community will be able to recognize such students and provide more efficient and immediate help for them, so that we have a better chance to keep valuable students. Using logistic regression analysis, we found out the family income, change of major, high school GPA, ACT composition score, hours spent studying, pursuit highest degree, sex, first generation, and financial aid affect whether students stay at Truman or not. Among those variables, academic factors seem to have a major effect on retention. As family income, high school GPA, ACT composition, and studying hour increase and as they plan to pursue higher degree than undergraduate degree, students are less likely to leave Truman. Female and first generation students are more likely to leave Truman. Also, students who are more certain about their major in the very beginning of their freshman year are more likely to leave Truman.

The logistic regression model developed in previously chapter will provide the predicted retention probability for each student, and we can identify students who have potential to leave. Once a student with a low potential retention rate is recognized, the student will be able to receive the right kind of advice and assistance from the university. This may positively influence the student and the university.

#### **4. Future Study**

Validation based on 1998 data will provide the information of how the current model performs in predicting retention for the following year. Note that though there are possible changes in retention phenomenon between 1996-1997 and 1998 freshman students. In other words, the best predicted model for 1996-1997 retention is not necessarily the best predicted model for 1998 retention. In fact, it will be interesting to see how well the current model fits and predicts more recent year data by developing a logistic regression model from the more current data and finding if there is any change in time.

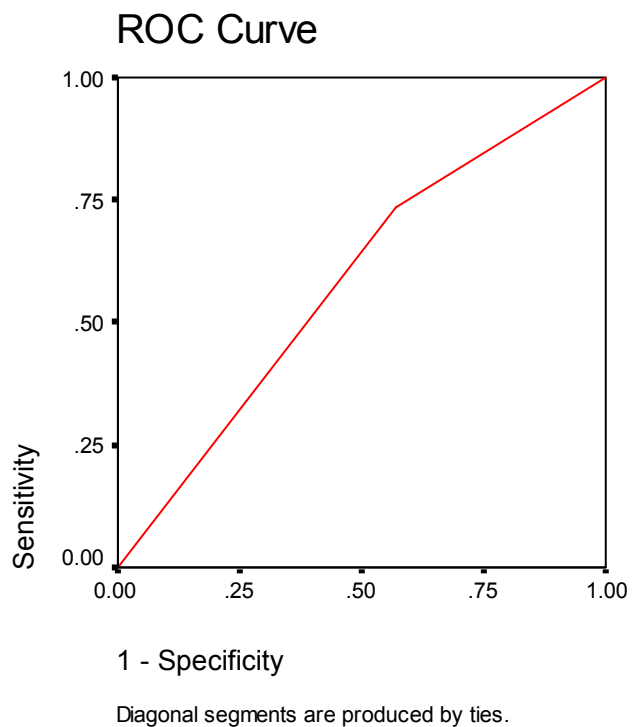
Using 1996 and 1997 data a Jack-knife procedure can be done as an alternative model validation. According to the procedure single observations are withheld from the data set, the model built, and the observation then tested in the model. This process occurs until all observations have been tested. This is computationally extensive and commonly used especially when the data set is small.

## Appendix.

1. Model includes independent variables: Income, Working hour, Change major, High school GPA, ACT, Study hour, First generation (Suggested by Pearson correlation)

### Cross-tabulation

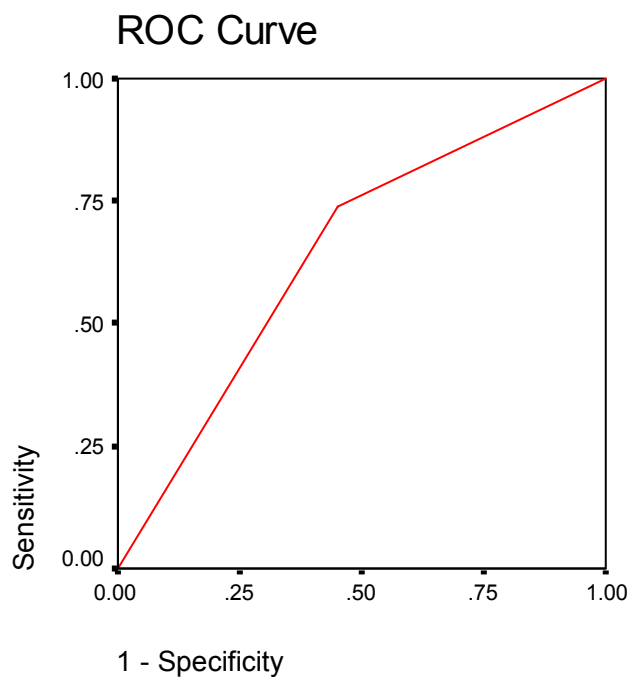
		Predicted retention		Total
		Leave	Stay	
2 year retention	Leave	12	527	539
	Stay	16	1452	1468
Total		28	1979	2007



2. Model includes independent variables: Income, Change major, High school GPA, ACT, Study hour, Pursue highest degree, First generation (Suggested by Backward stepwise ( $= .05$ ), Forward stepwise ( $= .1$ ))

### Cross-tabulation

		Predicted retention		Total
		Leave	Stay	
2 year retention	Leave	17	456	473
	Stay	14	1277	1291
Total		31	1733	1764

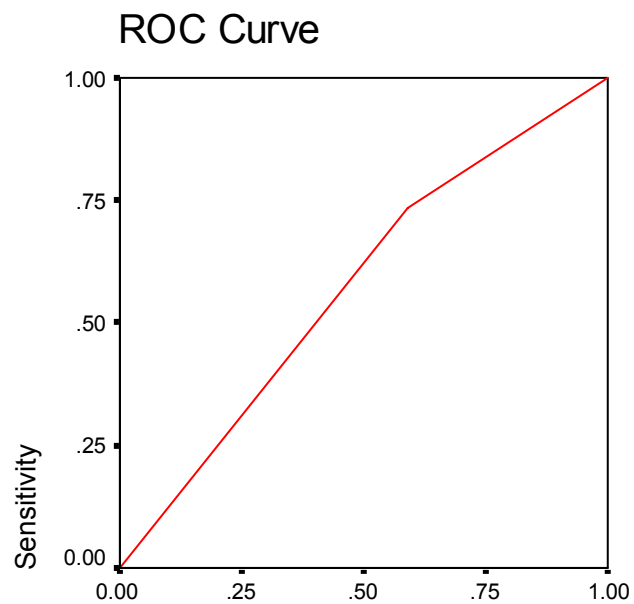


Diagonal segments are produced by ties.

3. Model includes independent variables: Income, Change major, High school GPA, ACT, Study hour (Suggested by Backward stepwise ( $= .1$ ), Forward stepwise ( $= .05$ ))

### Cross-tabulation

		Predicted retention		Total
		Leave	Stay	
2 year retention	Leave	9	531	540
	Stay	13	1460	1473
Total		22	1991	2013



1 - Specificity

Diagonal segments are produced by ties.

## Reference

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In : B. N. Petrov and F. Csaki, eds. *2<sup>nd</sup> International Symposium on Information Theory*, 267-281. Akademia Kiado, Budapest.
- Bean, J. (1982). Student attrition, intentions, and confidence: Interaction effects in a path model. *Research in Higher Education*, 17(4), 291-320.
- Bean, J. (1983). The application of model of turnover in work organizations to the student attrition process. *Review of Higher Education*, 6(2), 129-148.
- Braxton, J. M., Brier, E. M., & Hossler, D. (1988). The influence of student problems on student withdrawal decisions: An autopsy on autopsy studies. *Research in Education*, 28(2), 241-253.
- Cabrera, A. F., Castaneda, M. B., Nora, A., & Hengstler, D. (1992). The convergence between two theories of college persistence. *Journal of Higher Education*, 63(2), 143-164.
- Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *Journal of Higher Education*, 64(2), 123-139.
- DesJardins, S. L., Ahlburg, D. A., & McCall, B. P. (1999). An event history model of student departure. *Economics of Education Review*, 18(3), 375-390.
- Hochstein, S. K., & Butler, R. R. (1983). The effects of the composition of a financial aids package on student retention. *Journal of Student Financial Aid*, 13(1), 21-27.
- Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples, *Biometrika*, 76, 297-307.
- Ishiyama, J. (2000). The factors affecting retention, graduation and satisfaction rates at Truman State University: an initial empirical inquiry, *Truman State University*.
- Metner, B. S., & Bean, J. (1987). The estimation of a conceptual model of nontraditional undergraduate student attrition. *Research in Higher Education*, 27(1), 15-38.
- Pascarella, E. T., Duby, P., & Iverson, B. (1983). A test and reconceptualization of a theoretical model of college withdrawal in a commuter institution setting. *Sociology of Education*, 56(2), 88-100.
- Pascarella, E. T., & Terenzini, P. (1978). The relation of students' precollege characteristics and freshman year experience to voluntary attrition. *Research in Higher Education*, 9(4), 347-366.
- Pontius, R. & Schneider, L. (2001). Land-cover change model validation by an ROC method for the Ipswich watershed, *Agriculture, Ecosystems & Environment*, 85, 239-248.
- Spady, W. G. (1971). Dropout from higher education: Toward an empirical model. *Interchange*, 2(3), 38-63.
- Stage, F. K. (1988). University attrition: LISREL with logistic regression for the persistence criterion. *Research in Higher Education*, 29(4), 343-357.
- Stage, F. K., & Hossler, . (1989). Differences in family influences on college attendance plans for male and female ninth graders. *Research in Higher Education*, 30(3), 301-315.