Paper 1337-2017

A Data Mining Approach to Predict Student-at-risk

Youyou Zheng, Thanuja Sakruti, University of Connecticut

ABSTRACT

Student success is one of the most important topics for institutions. In this paper, the institutional researchers discussed the data mining process that could predict student at risk for a major STEM course at a top public university. SAS® Visual Analytics and SAS® Enterprise Miner were used for data visualization and predictive modeling. Several different modeling methods were compared to identify the optimal model.

INTRODUCTION

Data mining is an analysis process to obtain useful information from large data set and unveil its hidden pattern (Mehmed 2003, Tan 2005). It has been successfully applied in the business areas like fraud detection and customer retention for decades. With the increasing amount of educational data, educational data mining has become more and more important to uncover the hidden patterns within the institutional data, so as to support institutional decision making (Luan 2012). However, only very limited studies have been done on educational data mining for institutional decision support. The institutional researchers from Western Kentucky University built up a model to help increasing yield and retention at the University (Bogard 2013). The researcher from the University of California also proposed to apply data mining technique in the college recruitment process to achieve enrollment goals (Chang 2009). Both of the institutions used SAS® Enterprise Miner as their data mining tool. In this study, we are going to use SAS Enterprise Miner to build up the student-at-risk model. At the University of Connecticut (UConn), General Chemistry is a required course for undergraduate students in the STEM disciplines. It has a relatively higher DFW rate (D=Drop, F=Failure, W=Withdraw) compared with other courses. Take Fall 2012 as an example, the average DFW% is 24% at UConn and there are over a thousand students enrolled in this course. In this study, undergraduate students enrolled in Fall 2012 was used to build up the models. SEMMA (Sample, Explore, Modify, Model and Assess) method introduced by SAS Institute Inc. were applied to develop the predictive models. The freshman SAT scores, class and enrollment campus, semester GPA, first generation, low income, and other factors were used to predict students' performance in this course. In the predictive modeling process, several different modeling techniques (decision tree, neural network, ensemble models, and logistic regression) had been compared with each other in order to find an optimal one for our institution. The purpose of this study was to predict student success in the future study so as to improve the education quality in our institution.

METHOD

1. Selection of Variables

In this study, SAS Enterprise Miner Workstation 14.1 was selected to run the analysis. As we know, student data might include information in a variety of areas, for example, student academic performance (GPA, Grades, SAT/ACT, etc.), student finance information (first generation, family annual income, etc.), and student demographic profile (gender, ethnicity, etc.). In order to improve student performance in one of an undergraduate course (General Chemistry), variables including class campus, SAT scores, gender, ethnicity, and students previous semester GPA were selected (Table 1-1). Student ID number was used as ID. The Target of this analysis was the field demonstrated whether the students with D, F, W or not (1 or 0). Students' cumulative and semester GPA from previous semesters were also selected for this study. The detailed explanation for each variable is shown in the index. This data set included 1772 observations and 28 fields. The data dictionary is provided in the Index.

Table 1-1: Variables Used in the Analysis

(none) not Equal to Apply Reset

Columns: Label Mining Basic Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
AP_Course	Input	Nominal	No		No	.	.
Age	Input	Interval	No		No	.	.
CLASS_CAMPUS_CD	Input	Nominal	No		Yes	.	.
CTERM_TERM_CD	Input	Nominal	No		Yes	.	.
CTERM_TERM_SDESC	Input	Nominal	No		Yes	.	.
Career_Level	Input	Nominal	No		No	.	.
Class_Campus	Input	Nominal	No		No	.	.
Enrollment_Campus	Input	Nominal	No		No	.	.
Ethnicity	Input	Nominal	No		No	.	.
FirstGen_Flag	Input	Nominal	No		Yes	.	.
First_Generation	Input	Nominal	No		No	.	.
FullPart	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
ID	ID	Nominal	No		No	.	.
LOAD	Input	Nominal	No		Yes	.	.
LowIncome_Flag	Input	Nominal	No		Yes	.	.
Low_Income	Input	Nominal	No		No	.	.
NSF_STEM_Category	Input	Nominal	No		No	.	.
Residence	Input	Nominal	No		No	.	.
SATmath	Input	Interval	No		No	.	.
SATverbal	Input	Interval	No		No	.	.
STEM_Flag	Input	Nominal	No		No	.	.
Sem_GPA_FS11_CD	Input	Interval	No		No	.	.
Sem_GPA_SP12_CD	Input	Interval	No		No	.	.
TARGET	Target	Binary	No		No	.	.
Underrepresented_Flag	Input	Nominal	No		No	.	.
gpa_sem_FA11	Input	Interval	No		No	.	.
gpa_sem_SP12	Input	Interval	No		No	.	.

2. Data Exploration

The data set was explored via using SAS® Visual Analytics to help understand the relationships among variables and target. The dual-axis bar-line chart [Figure 2-1] presented the class enrollment and grade distribution by the Campuses. The line chart trend represented the overall frequency percentage at both Storrs (Main campus) and Regional campuses. This chart had lattice columns – one for each campus for better visibility of enrollment.

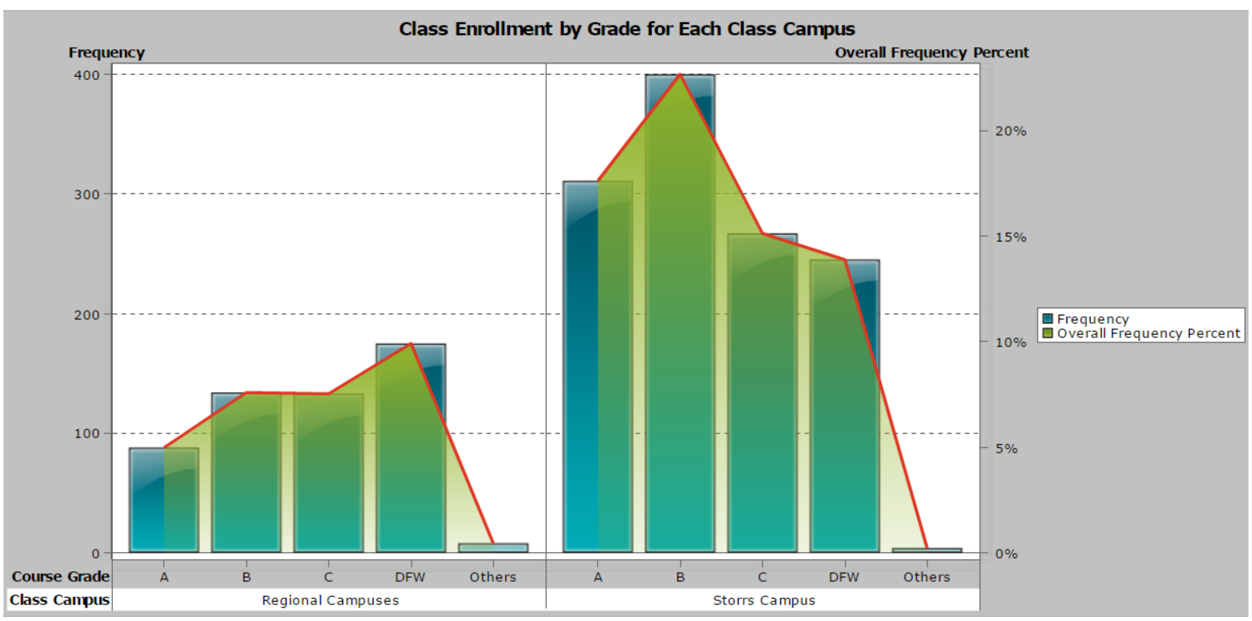

Figure 2-1: Class Enrollment by Grade for Each Class Campus*

*Class Campus: Campus where the class was being held irrespective of where the student enrolled.

The scatter plot [Figure 2-2] presented the relationship between students' entering SAT Math scores and Fall 2012 Semester GPA. The color map indicated the student's career level. Based on the point's location and arrangement, it was noticed that students with better SAT Math scores tended to have better GPAs. Additionally, this course was mostly taken by the freshmen followed by sophomore, junior, and senior respectively.

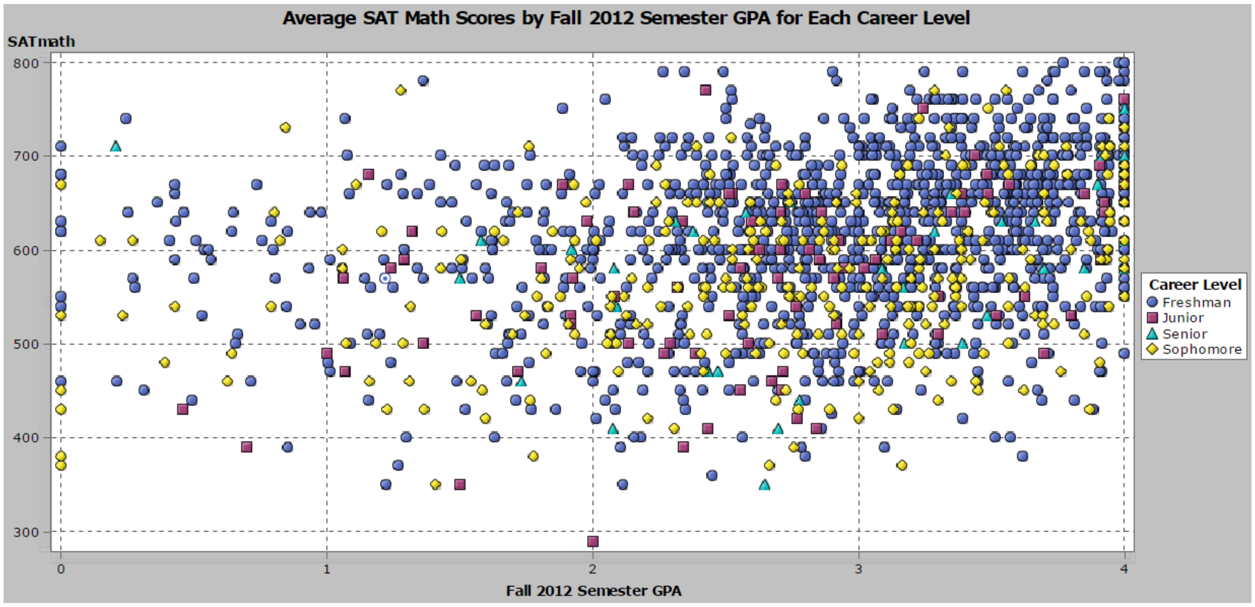

Figure 2-2: SAT Math Scores by Student's Semester GPA for Each Career Level

3. Models

SEMMA (Sample, Explore, Modify, Model and Assess) method introduced by SAS Institute Inc. was applied to develop the predictive models. In this study, the target was a binary variable, and there were many categorical variables. In order to modify the data, replacement was first applied to modify and correct original data. In the Data Partition section, Training, Validation, and Test allocations were automatically set as 40.0%, 30.0%, and 30.0%, respectively.

The Target used in this analysis was a binary variable (0, 1). Misclassification rate was selected to evaluate predictive accuracy of each model. The formula of Misclassification Rate is shown below.

Misclassification Rate = (sum of misclassified records)/(total records) (1)

In the model comparison step, ROC (receiver operating curve) was applied to evaluate model accuracy. ROC presented graphs of Sensitivity by (1-Specificity). Sensitivity gives the probability that a student will have a DFW and the student actually had a DFW. Specificity gives the probability that a student will not have a DFW and the student actually didn't have a DFW. Therefore, one minus specificity gives the probability that a student will have a DFW while the student actually didn't have a DFW. The calculation of sensitivity and specificity was shown as below.

Sensitivity = (True Positive)/(True Positive + False Negative) (2)

Specificity = (True Negative)/(False Positive + True Negative) (3)

Decision Tree methodology was then applied to yield useful information for the following analysis such as neural network and regression. The Impute process was used to take care of missing values in the data set. The Model Comparison node was used to compare the performance of each model.

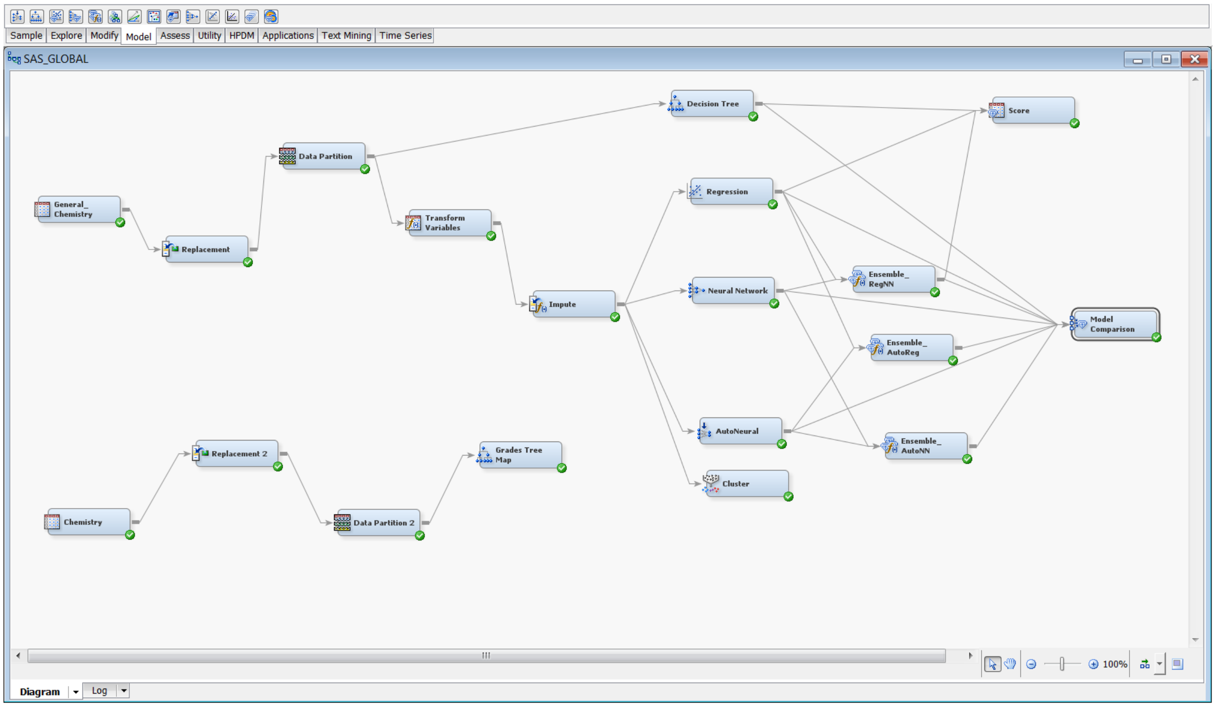

Figure 3-1: SAS® Enterprise Miner Process Flow Chart

RESULTS

1. Decision Tree

Based on the results from the Decision Tree, SAT Math score was of great importance to predict DFW rate of this course. Average square error and misclassification rate were examined to evaluate the decision tree. According to the following results, the optimal tree had about 2 to 3 leaves.

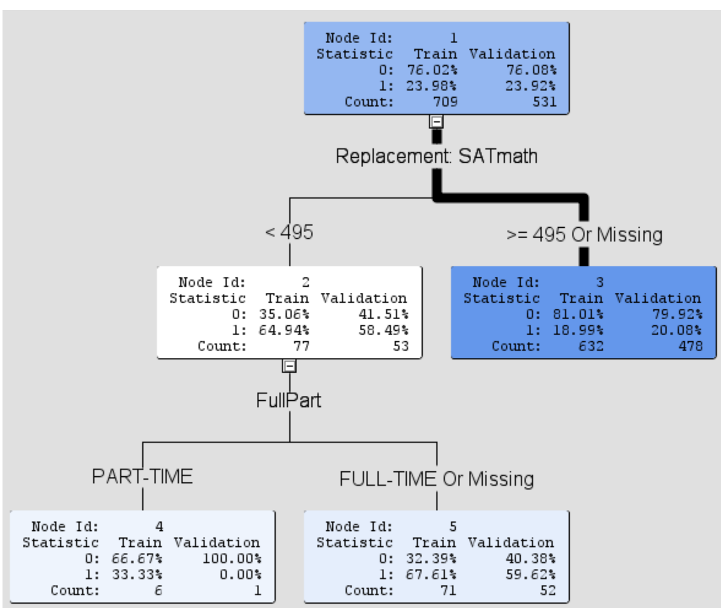

Figure 1-1: Tree

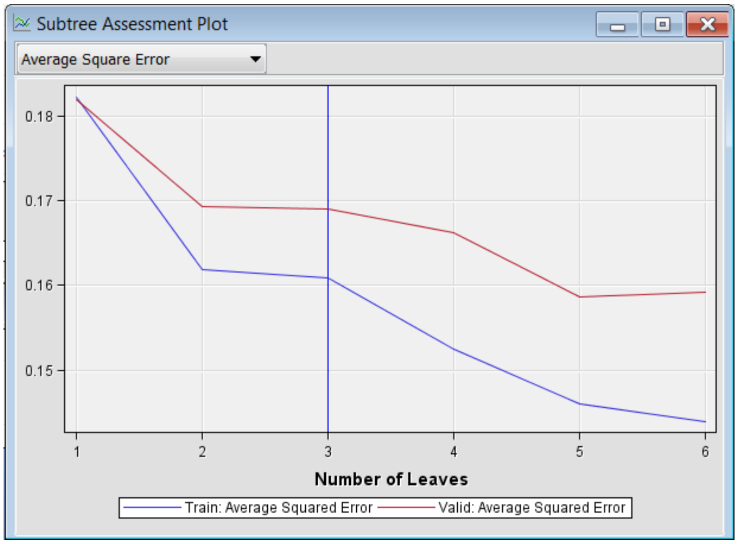

Figure 1-2: Subtree Assessment Plot of Average Square Error

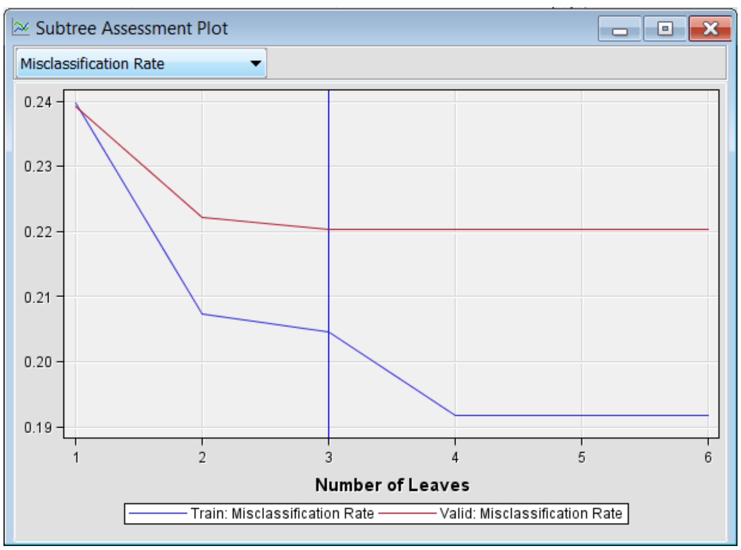

Figure 1-3: Subtree Assessment Plot of Misclassification Rate

2. Regression

The logistic regression model gave the statistical significance of each variable. The variables with p value < 0.1 indicated they were statistically significant. These statistically significant variables included AP Course, Career Level, Gender, and SAT Math Score.

Table 2-1: Type 3 Analysis of Effects

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
AP_Course	1	3.6619	0.0557
Career_Level	4	15.4752	0.0038
Class_Campus	1	0.0132	0.9085
Enrollment_Campus	1	0.0130	0.9093
Ethnicity	7	4.8499	0.6783
FullPart	1	0.4600	0.4976
Gender	1	7.9495	0.0048
IMP_REP_SATmath	1	33.3641	<.0001
IMP_REP_SATverbal	1	0.2262	0.6344
NSF_STEM_Category	10	4.2448	0.9356
REP_Age	1	1.5982	0.2062
REP_Sem_GPA_FS11_CD	1	0.3747	0.5404
REP_Sem_GPA_SP12_CD	1	0.4955	0.4815
Residence	1	0.0348	0.8520
STEM_Flag	0	0.0000	.
Underrepresented_Flag	0	0.0000	.

Table 2-2: Fit Statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		AIC	Akaike's Information Criterion	709.0907	.	.
TARGET		ASE	Average Squared Error	0.147589	0.147675	0.162974
TARGET		AVERR	Average Error Function	0.45352	0.458938	0.511527
TARGET		DFE	Degrees of Freedom for Error	676	.	.
TARGET		DFM	Model Degrees of Freedom	33	.	.
TARGET		DFT	Total Degrees of Freedom	709	.	.
TARGET		DIV	Divisor for ASE	1418	1062	1064
TARGET		ERR	Error Function	643.0907	487.3922	544.2642
TARGET		FPE	Final Prediction Error	0.161999	.	.
TARGET		MAX	Maximum Absolute Error	0.941713	0.961741	0.999837
TARGET		MSE	Mean Square Error	0.154794	0.147675	0.162974
TARGET		NOBS	Sum of Frequencies	709	531	532
TARGET		NW	Number of Estimate Weights	33	.	.
TARGET		RASE	Root Average Sum of Squares	0.384174	0.384285	0.4037
TARGET		RFPE	Root Final Prediction Error	0.402491	.	.
TARGET		RMSE	Root Mean Squared Error	0.393439	0.384285	0.4037
TARGET		SBC	Schwarz's Bayesian Criterion	859.6979	.	.
TARGET		SSE	Sum of Squared Errors	209.2819	156.831	173.4043
TARGET		SUMW	Sum of Case Weights Times Freq	1418	1062	1064
TARGET		MISC	Misclassification Rate	0.22567	0.20904	0.236842

3. Neural Network

In SAS Enterprise Miner, the neural network node provides the possibility to control one hidden layer network. According to the Iteration Plot, the optimal average square error occurred on the 4^{th} iteration for neural network model.

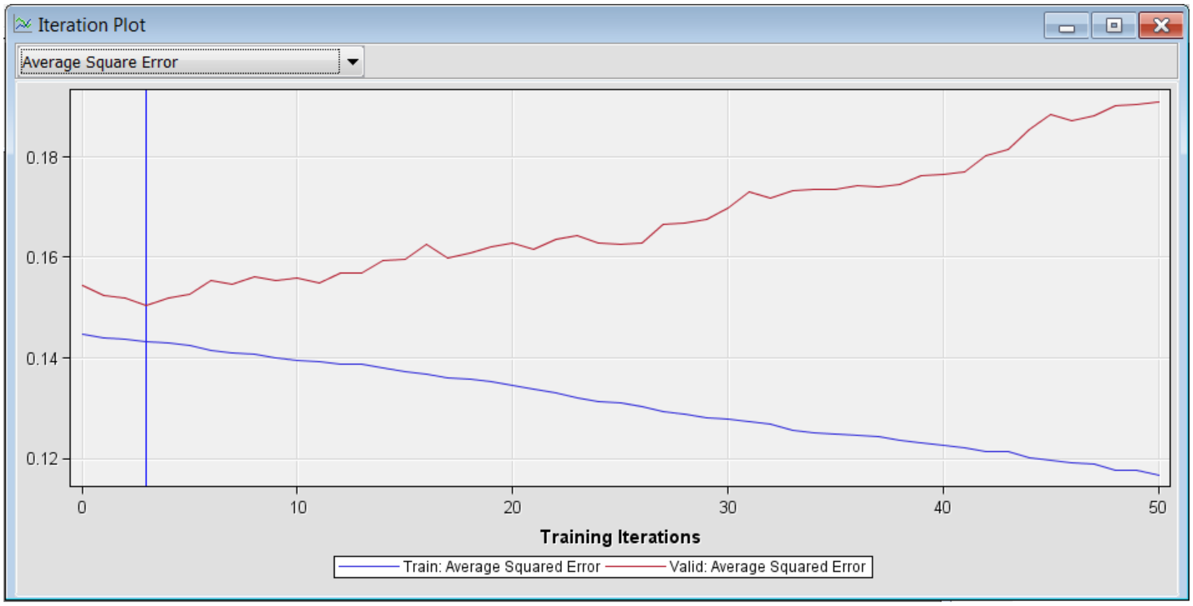

Figure 3-1: Iteration Plot of Average Square Error

Table 3-1: Fit Statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		DFT	Total Degrees of Freedom	709	.	.
TARGET		DFE	Degrees of Freedom for Error	597	.	.
TARGET		DFM	Model Degrees of Freedom	112	.	.
TARGET		NW	Number of Estimated Weights	112	.	.
TARGET		AIC	Akaike's Information Criterion	848.0986	.	.
TARGET		SBC	Schwarz's Bayesian Criterion	1359.25	.	.
TARGET		ASE	Average Squared Error	0.143321	0.150378	0.163717
TARGET		MAX	Maximum Absolute Error	0.96365	0.981702	0.969574
TARGET		DIV	Divisor for ASE	1418	1062	1064
TARGET		NOBS	Sum of Frequencies	709	531	532
TARGET		RASE	Root Average Squared Error	0.378578	0.387786	0.404619
TARGET		SSE	Sum of Squared Errors	203.2293	159.7011	174.1947
TARGET		SUMW	Sum of Case Weights Times Freq	1418	1062	1064
TARGET		FPE	Final Prediction Error	0.197096	.	.
TARGET		MSE	Mean Squared Error	0.170209	0.150378	0.163717
TARGET		RFPE	Root Final Prediction Error	0.443955	.	.
TARGET		RMSE	Root Mean Squared Error	0.412564	0.387786	0.404619
TARGET		AVERR	Average Error Function	0.440126	0.466362	0.498263
TARGET		ERR	Error Function	624.0986	495.276	530.1516
TARGET		MISC	Misclassification Rate	0.215797	0.214689	0.244361
TARGET		WRONG	Number of Wrong Classifications	153	114	130

4. Auto Neural Network

In SAS Enterprise Miner, the Auto Neural node offers the possibility to build a multilayer network. Auto Neural node will automatically test several networks and decide the optimal neural network for the data set. In this study, the Auto Neural Network process gave the optimal average square error on the 5th iteration as shown below.

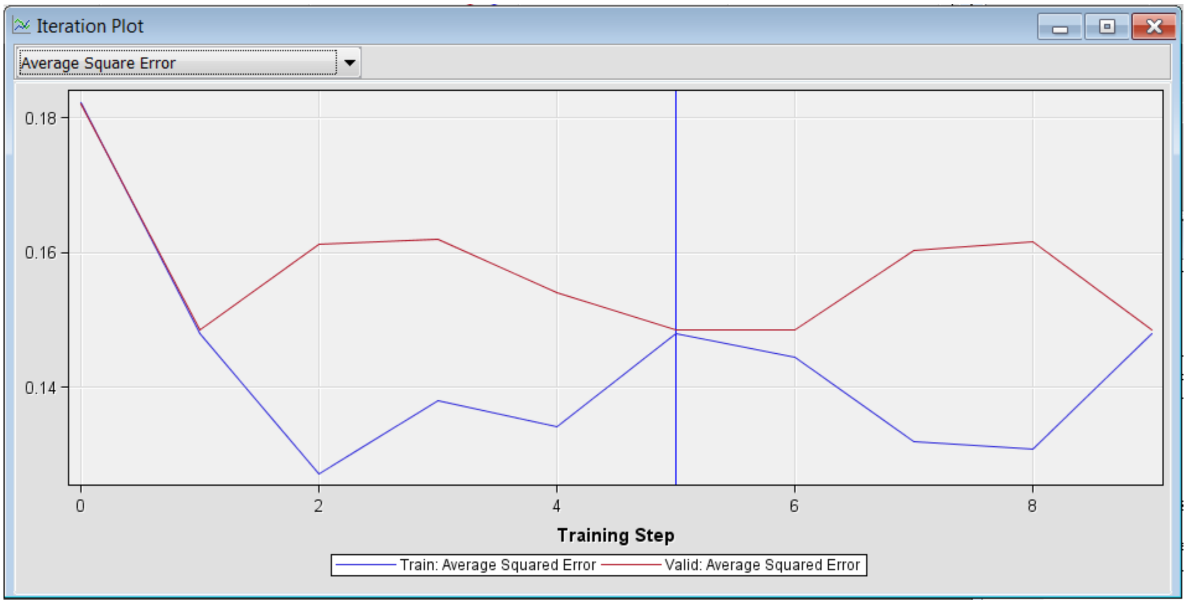

Figure 4-1: Iteration Plot of Average Square Error

Table 4-1: Fit Statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		DFT	Total Degrees of Freedom	709	.	.
TARGET		DFE	Degrees of Freedom for Error	673	.	.
TARGET		DFM	Model Degrees of Freedom	36	.	.
TARGET		NW	Number of Estimated Weights	36	.	.
TARGET		AIC	Akaike's Information Criterion	716.6837	.	.
TARGET		SBC	Schwarz's Bayesian Criterion	880.9825	.	.
TARGET		ASE	Average Squared Error	0.147888	0.148469	0.162015
TARGET		MAX	Maximum Absolute Error	0.938128	0.961511	0.97302
TARGET		DIV	Divisor for ASE	1418	1062	1064
TARGET		NOBS	Sum of Frequencies	709	531	532
TARGET		RASE	Root Average Squared Error	0.384562	0.385317	0.402511
TARGET		SSE	Sum of Squared Errors	209.7053	157.674	172.3839
TARGET		SUMW	Sum of Case Weights Times Freq	1418	1062	1064
TARGET		FPE	Final Prediction Error	0.16371	.	.
TARGET		MSE	Mean Squared Error	0.155799	0.148469	0.162015
TARGET		RFPE	Root Final Prediction Error	0.40461	.	.
TARGET		RMSE	Root Mean Squared Error	0.394714	0.385317	0.402511
TARGET		AVERR	Average Error Function	0.454643	0.461287	0.498672
TARGET		ERR	Error Function	644.6837	489.8865	530.5866
TARGET		MISC	Misclassification Rate	0.221439	0.207156	0.242481
TARGET		WRONG	Number of Wrong Classifications	157	110	129

5. Ensemble (Neural Network and Regression)

Ensemble modeling is capable of synthesizing 2 or more different models, which could improve the accuracy of prediction. In this step, the Ensemble model process combined 2 models including neural network and regression models.

Table 5-1: Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		ASE	Average Squared Error	0.143652	0.147011	0.161487
TARGET		DIV	Divisor for ASE	1418	1062	1064
TARGET		MAX	Maximum Absolute Error	0.952167	0.971722	0.968818
TARGET		NOBS	Sum of Frequencies	709	531	532
TARGET		RASE	Root Average Squared Error	0.379015	0.38342	0.401855
TARGET		SSE	Sum of Squared Errors	203.6988	156.1254	171.8226
TARGET		DISF	Frequency of Classified Cases	709	531	532
TARGET		MISC	Misclassification Rate	0.211566	0.205273	0.240602
TARGET		WRONG	Number of Wrong Classifications	150	109	128

6. Ensemble (Auto Neural Network and Neural Network)

This Ensemble model process combined 2 models including auto neural network and neural network.

Table 6-1: Fit Statistics

Target	Target Label	Fit Statistics ▲	Statistics Label	Train	Validation	Test
TARGET		ASE	Average Squared Error	0.14373	0.147441	0.161022
TARGET		DISF	Frequency of Classified Cases	709	531	532
TARGET		DIV	Divisor for ASE	1418	1062	1064
TARGET		MAX	Maximum Absolute Error	0.950374	0.971436	0.967805
TARGET		MISC	Misclassification Rate	0.211566	0.20904	0.238722
TARGET		NOBS	Sum of Frequencies	709	531	532
TARGET		RASE	Root Average Squared Error	0.379118	0.383981	0.401275
TARGET		SSE	Sum of Squared Errors	203.8095	156.5828	171.3269
TARGET		WRONG	Number of Wrong Classifications	150	111	127

7. Ensemble (Auto Neural Network and Regression)

The Ensemble model process combined 2 models including neural network and regression models.

Table 7-1: Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		ASE	Average Squared Error	0.147638	0.147805	0.162394
TARGET		DIV	Divisor for ASE	1418	1062	1064
TARGET		MAX	Maximum Absolute Error	0.939921	0.961455	0.974032
TARGET		NOBS	Sum of Frequencies	709	531	532
TARGET		RASE	Root Average Squared Error	0.384236	0.384454	0.402981
TARGET		SSE	Sum of Squared Errors	209.3502	156.9684	172.7869
TARGET		DISF	Frequency of Classified Cases	709	531	532
TARGET		MISC	Misclassification Rate	0.222849	0.210923	0.244361
TARGET		WRONG	Number of Wrong Classifications	158	112	130

8. Model Comparison

According to the results from Model Comparison process, Ensemble model (Neural Network and Regression) provided the optimal model. The model selection rule is based on the misclassification rate in

the model validation step. In the validation step, the lower the misclassification rate is, the better the predictive model will be. The order from the best to worst performance for the 7 models were as following:

(1) Ensemble (Neural Network and Regression);
(2) Auto Neural Network;
(3) Ensemble (Auto Neural Network and Neural Network);
(4) Regression;
(5) Ensemble (Auto Neural Network and Regression);
(6) Neural Network;
(7) Decision Tree.

The receiver operating characteristic (ROC) curves indicated the performance of a binary system. As shown in Figure 8-1, the ROC curves provided the optimal models for this analysis.

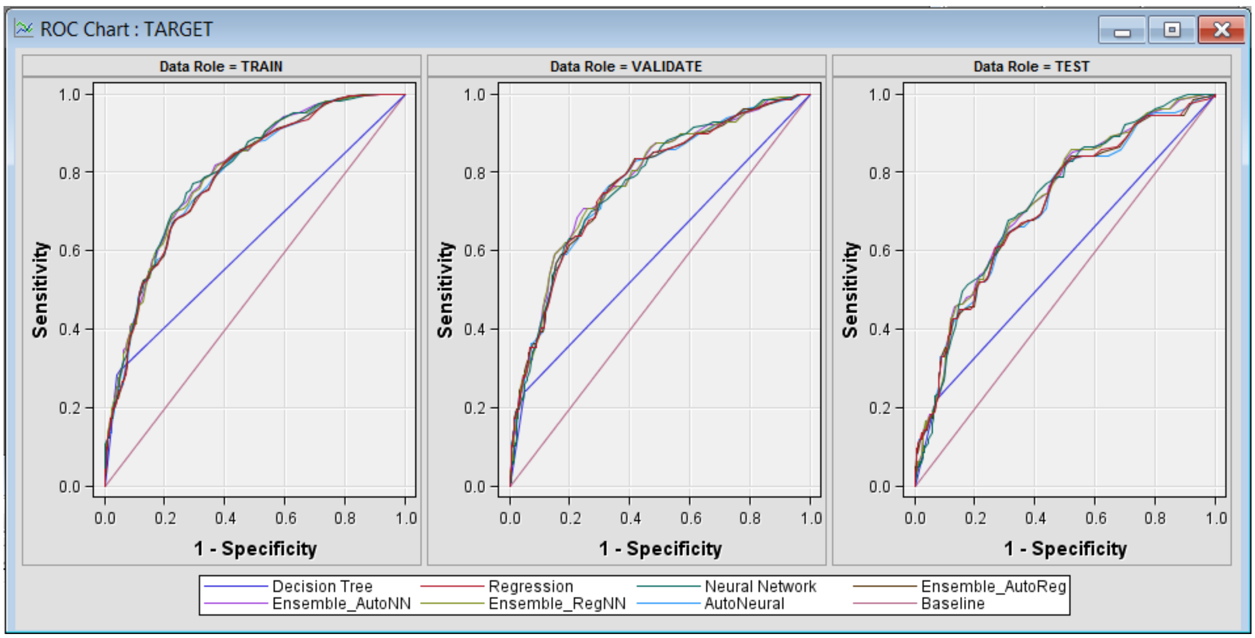

Figure 8-1: ROC Chart

Table 8-1: Fit Statistics

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate ▲	T	Train: Misclassification Rate	T	T	T	T	T	T	V	V	V	V	V	V	T	T	Test: Misclassification Rate	Test: Maximum Absolute Error	Test: Sum of Squared Errors
Y	Ensmbl	Ensmbl	Ensemble RegNN	TARGET		0.205273	...	0.211566	...	...	...	...	...	.	...	...	...	...	...	...	...	.	0.240602	0.968818	171.8
	AutoNeural	AutoNeural	AutoNeural	TARGET		0.207156	...	0.221439	...	...	...	...	...	...	...	...	...	...	...	...	...	...	0.242481	0.97302	172.3
	Ensmbl2	Ensmbl2	Ensemble AutoNN	TARGET		0.20904	...	0.211566	...	...	...	...	...	.	...	...	...	...	...	...	...	.	0.238722	0.967805	171.3
	Reg	Reg	Regression	TARGET		0.20904	...	0.22567	...	...	...	...	...	...	...	...	...	...	...	...	...	...	0.236842	0.999837	173.4
	Ensmbl3	Ensmbl3	Ensemble AutoReg	TARGET		0.210923	...	0.222849	...	...	...	...	...	.	...	...	...	...	...	...	...	.	0.244361	0.974032	172.7
	Neural	Neural	Neural Network	TARGET		0.214689	...	0.215797	...	...	...	...	...	...	...	...	...	...	...	...	...	...	0.244361	0.969574	174.1
	Tree	Tree	Decision Tree	TARGET		0.220339	...	0.204513	...	...	...	...	...	...	...	...	...	...	...	...	...	...	0.242481	0.810127	190.2

CONCLUSION AND DISCUSSION

SAS® Enterprise Miner is a powerful tool for higher education data mining. Ensemble modeling and neural network provide better solutions compared with other modeling methods applied. Neural network is a well-known tool for enrollment management, this study shows it is also powerful in course data analysis. In order to improve the accuracy of the predictive model for university course analysis, more variables and more school years data will be added into the data set for our future work. In the future study, the variables such as early college experience, study abroad and other related information will be considered. The similar predictive modeling methods could be applied in the investigation of graduation, which could potentially help more students obtain their degree within 4 years.

ACKNOWLEDGMENT

This work was done with the great help from Dr. Thulasi Kumar at the OIRA (Office of Institutional Research and Assessment) of George Mason University.

REFERENCES

Bogard, M. (2013). A Data Driven Analytic Strategy for Increasing Yield and Retention at Western Kentucky University Using SAS Enterprise BI and SAS Enterprise Miner. SAS Global Forum 2013**.**

Chang T. (2009). Data Mining: A Magic Technology for College Recruitment. http://www.ocair.org/files/presentations/paper2008_09/tongshan_chang_2009.pdf.

Luan J., Kumar T., Sujitparapitaya S., and Bohannon T. (2012). Exploring and Mining Data. The Handbook of Institutional Research. Howard R.D., McLaughlin G.W., Knight W.E., John Wiley & Sons, Inc.**:** 478-501.

Mehmed, K. (2003). Data Mining: Concepts, Models, Methods, and Algorithms., John Wiley & Sons.

Tan P., Steinbach M., Kumar V. (2005). Introduction to Data Mining, Addison Wesley.

Christie P., Georges J., Thompson J., and Wells C. (2011). Applied Analytics Using SAS® Enterprise Miner™ Course Notes, SAS Institute Inc.

INDEX

Variable	Description
CTERM_TERM_SDESC	Term Description
CTERM_TERM_CD	Term Code
ID	Student ID
LOAD	Full-time, Less than Half, Half-time
Career_Level	Freshman, Sophomore, Junior, Senior
FullPart	Full-time, Part-time
Enrollment_Campus	Student Enrolled Campuses
Gender	Female, Male
Residence	Where is the student from (In-state, Out-of-state)?
Age	Age
Ethnicity	Ethnicity
NSF_STEM_Category	Student STEM Category
Low_Income	Whether from low income family?
First_Generation	Whether from first generation family?
LowIncome_Flag	Low Income Flag
FirstGen_Flag	First Generation Flag
CLASS_CAMPUS_CD	Class Campuses Code
SATmath	SAT Math Score
SATverbal	SAT Verbal Score
gpa_sem_FA11	Fall 2011 Semester GPA
gpa_sem_SP12	Spring 2012 Semester GPA
Underrepresented_Flag	Underrepresented Minority Flag
STEM_Flag	STEM Flag
TARGET	DFW or Not
Class_Campus	Class Campuses
AP_Course	Whether took AP courses before?
Sem_GPA_FS11_CD	Level of Fall 2011 Semester GPA (high, low, not taken)?
Sem_GPA_SP12_CD	Level of Spring 2012 Semester GPA (high, low, not taken)?

CONTACT INFORMATION

Youyou Zheng, Ph.D.
Office of Institutional Research and Effectiveness
University of Connecticut
Email: youyou.zheng@uconn.edu

Thanuja Sakruti
Office of Institutional Research and Effectiveness
University of Connecticut
Email: thanuja.sakruti@uconn.edu