Paper 2637

Using Analytics to Improve Student Advising System

Youyou Zheng, Thanuja Sakruti

University of Connecticut

ABSTRACT

Student success is one of the most important topics for any institutions. In this paper, the institutional researchers discussed the predictive modeling process that could identify student at risk for a major STEM course at a top public university. SAS® Enterprise Miner and Visual Analytics were applied to predict and visualize the student performance at individual level. This study allows student advisors and administrators to use predictive analytics and provide interventions for at-risk students in a timely manner.

INTRODUCTION

Predictive analytics is the technology that uses historical data, statistical algorithms, and machine learning to predict future outcomes (Mehmed 2003, Tan 2005). Nowadays, predictive analytics has been widely applied to improve organizational operations and reduce risks. With the increasing financial pressure, predictive analytics is used by many state universities to optimize resource allocation, and provide interventions for students at risk (Chang 2009, Bogard 2013). The purpose of this study is to discuss the possibilities to generate a real-time student academic alert system via using SAS® Enterprise Miner and SAS® Visual Analytics. One of a core undergraduate STEM course with relatively higher DFW (Drop, Fail, and Withdraw) rates was chosen for this study. SEMMA (Sample, Explore, Modify, Model and Assess) method introduced by SAS Institute Inc. were applied to develop the predictive models. A wide variety of variables were considered to build the models, such as student academic preparation variables (SAT scores and AP courses), demographic variables (gender, race), and family socioeconomics background (first generation, low income). 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. SAS® Visual Analytics were applied to provide visualization of the predicted grades. The application of using SAS® Enterprise Miner and Visual Analytics to create a student early alert system is going to allow student advisors and administrators to give interventions in a timely manner.

METHODS

1. Selection of Variables

In this study, SAS Enterprise Miner Workstation 14.1 was selected to analyze the data set. As we know, student data includes information in many different areas, for example, student academic performance (AP courses, SAT/ACT, etc.), student finance information (first generation, family annual income, etc.), and student demographic profile (gender, race/ethnicity, etc.). In order to predict students' performance in one of a core STEM courses, variables including class offering campus, SAT scores, gender, ethnicity, and whether students had taken any AP courses were selected (Table 1-1). Student ID number was used as ID. The Target of this analysis was the field demonstrated whether the student got D, F, W or not (1 or 0). The detailed explanation for each variable is shown in the Index. This data set included 1772 observations and 16 fields. The data dictionary is provided in the Index.

Table 1-1: Variables Used in the Analysis

Name	Role	Level	Report	Order	Drop
Career_Level	Input	Nominal	No		No
Class_Campus	Input	Nominal	No		No
LowIncome_Flag	Input	Nominal	No		No
SATverbal	Input	Interval	No		No
Enrollment_Campus	Input	Nominal	No		No
TARGET	Target	Binary	No		No
Age	Input	Interval	No		No
Underrepresented_Flag	Input	Nominal	No		No
SATmath	Input	Interval	No		No
STEM_Flag	Input	Nominal	No		No
ID	ID	Nominal	No		No
Gender	Input	Nominal	No		No
FullPart	Input	Nominal	No		No
Residence	Input	Nominal	No		No
AP_Course	Input	Nominal	No		No
FirstGen_Flag	Input	Nominal	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.

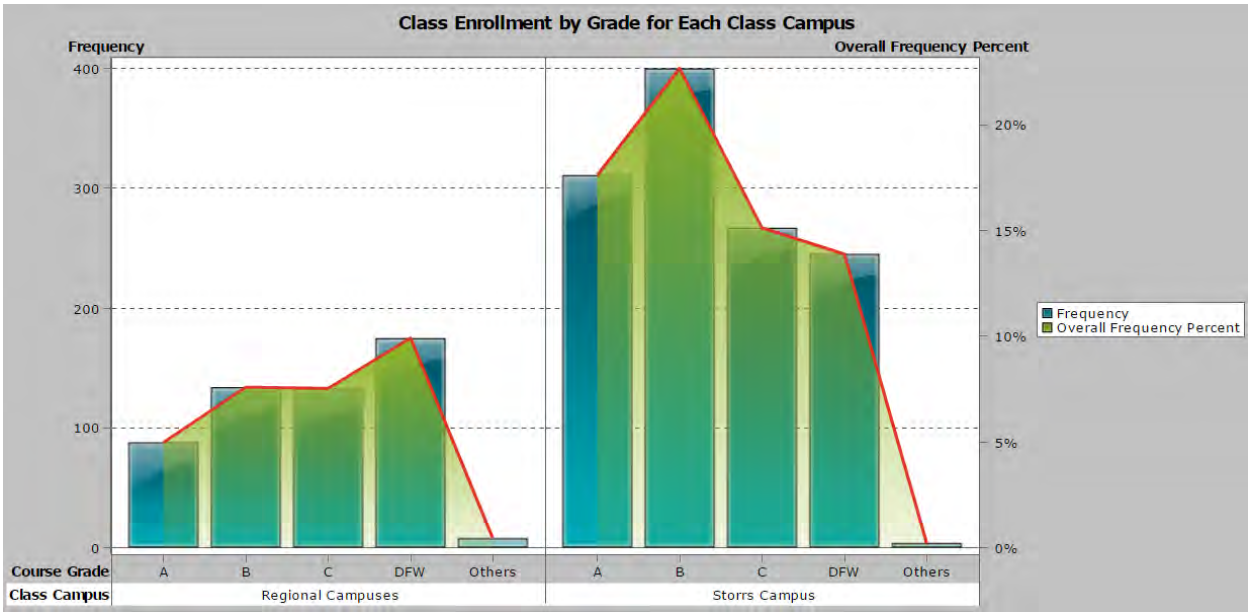

Figure 2-1: Class Enrollment by Grade for Each Class Campus*

*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 their 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.

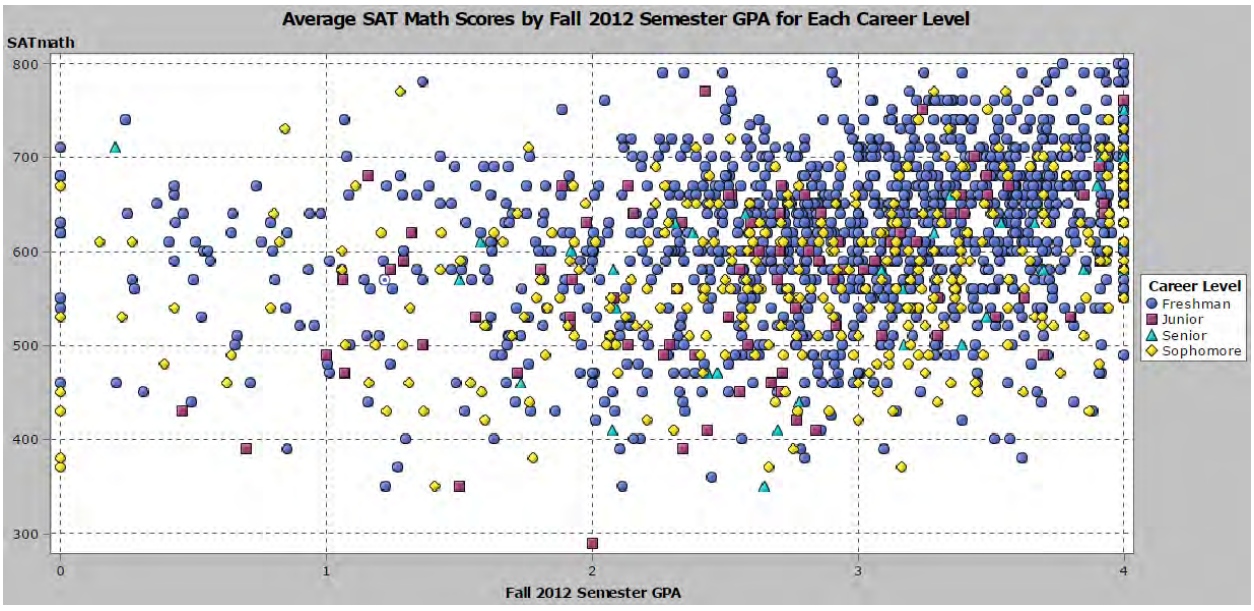

Figure 2-2: SAT Math Scores by Student's Semester GPA for Each Career Level

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 the data set. 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.

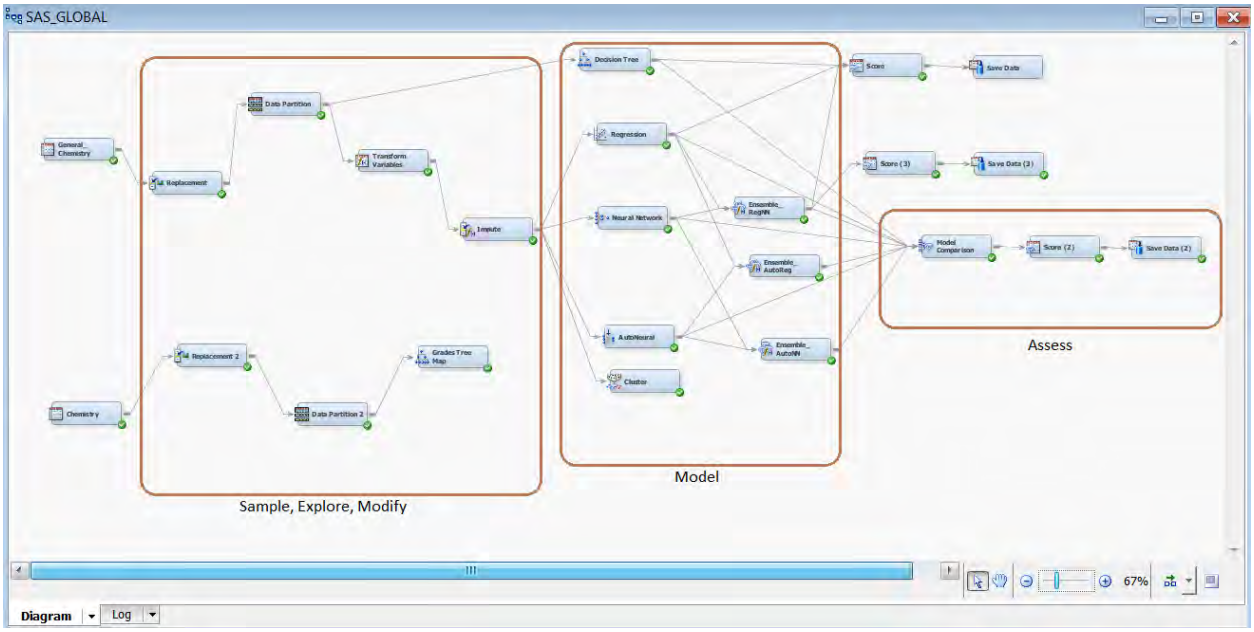

Figure 3-1: SAS® Enterprise Miner Process Flow Chart

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.

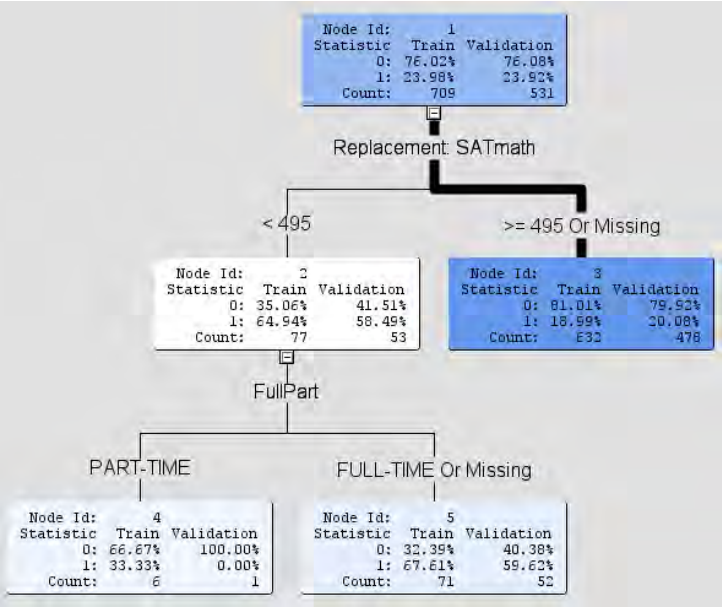

Figure 1-1: Tree

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

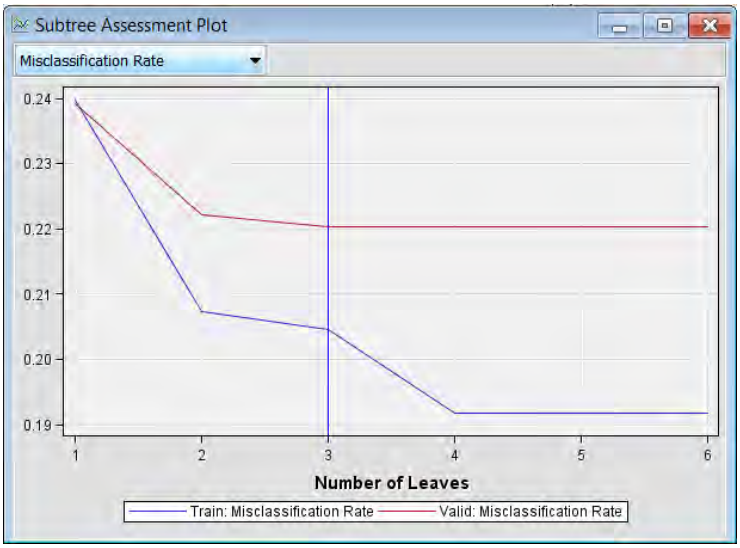

Figure 1-3: Subtree Assessment Plot of Misclassification Rate

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

Effect	DF	Wald Chi-Square	Pr > ChiSq
AP_Course	1	4.9652	0.0259
Career_Level	4	15.2818	0.0042
Class_Campus	1	0.0150	0.9026
Enrollment_Campus	1	0.0146	0.9040
FirstGen_Flag	1	0.3311	0.5650
FullPart	1	0.2510	0.6164
Gender	1	7.4911	0.0062
IMP_REP_SATmath	1	37.2972	<.0001
IMP_REP_SATverbal	1	0.4839	0.4867
LowIncome_Flag	1	0.1404	0.7079
REP_Age	1	0.6450	0.4219
Residence	2	0.4282	0.8073
STEM_Flag	1	0.0012	0.9721
Underrepresented_Flag	1	0.0331	0.8557

Table 2-2: Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		AIC	Akaike's Information Criterion	692.6823	.	.
TARGET		ASE	Average Squared Error	0.150636	0.150937	0.161052
TARGET		AVERR	Average Error Function	0.461694	0.467215	0.504472
TARGET		DFE	Degrees of Freedom for Error	690	.	.
TARGET		DFM	Model Degrees of Freedom	19	.	.
TARGET		DFT	Total Degrees of Freedom	709	.	.
TARGET		DIV	Divisor for ASE	1418	1062	1064
TARGET		ERR	Error Function	654.6823	496.1821	536.7587
TARGET		FPE	Final Prediction Error	0.158932	.	.
TARGET		MAX	Maximum Absolute Error	0.940376	0.954775	0.999786
TARGET		MSE	Mean Square Error	0.154784	0.150937	0.161052
TARGET		NOBS	Sum of Frequencies	709	531	532
TARGET		NW	Number of Estimate Weights	19	.	.
TARGET		RASE	Root Average Sum of Squares	0.388119	0.388506	0.401313
TARGET		RFPE	Root Final Prediction Error	0.398663	.	.
TARGET		RMSE	Root Mean Squared Error	0.393426	0.388506	0.401313
TARGET		SBC	Schwarz's Bayesian Criterion	779.3956	.	.
TARGET		SSE	Sum of Squared Errors	213.6019	160.2954	171.3595
TARGET		SUMW	Sum of Case Weights Times Freq	1418	1062	1064
TARGET		MISC	Misclassification Rate	0.221439	0.214689	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 first iteration for neural network model.

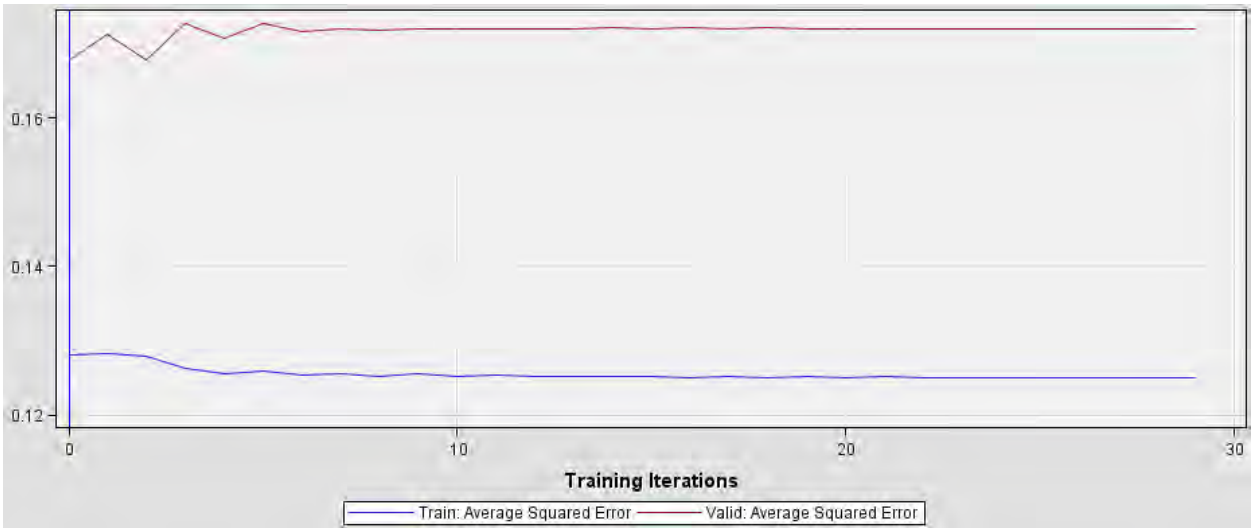

Figure 3-1: Iteration Plot of Average Square Error

Table 3-1: Fit Statistics

Fit Statistics	Statistics Label	Train	Validation	Test
DFT	Total Degrees of Freedom	709	.	.
DFE	Degrees of Freedom for Error	648	.	.
DFM	Model Degrees of Freedom	61	.	.
NW	Number of Estimated Weights	61	.	.
AIC	Akaike's Information Criterion	689.3686	.	.
SBC	Schwarz's Bayesian Criterion	967.7638	.	.
ASE	Average Squared Error	0.128014	0.167824	0.177043
MAX	Maximum Absolute Error	0.959978	0.998928	0.992943
DIV	Divisor for ASE	1418	1062	1064
NOBS	Sum of Frequencies	709	531	532
RASE	Root Average Squared Error	0.35779	0.409664	0.420765
SSE	Sum of Squared Errors	181.5238	178.2296	188.3736
SUMW	Sum of Case Weights Times Freq	1418	1062	1064
FPE	Final Prediction Error	0.152115	.	.
MSE	Mean Squared Error	0.140065	0.167824	0.177043
RFPE	Root Final Prediction Error	0.39002	.	.
RMSE	Root Mean Squared Error	0.374252	0.409664	0.420765
AVERR	Average Error Function	0.400119	0.528737	0.559699
ERR	Error Function	567.3686	561.5185	595.5198
MISC	Misclassification Rate	0.176305	0.248588	0.24812
WRONG	Number of Wrong Classifications	125	132	132

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.

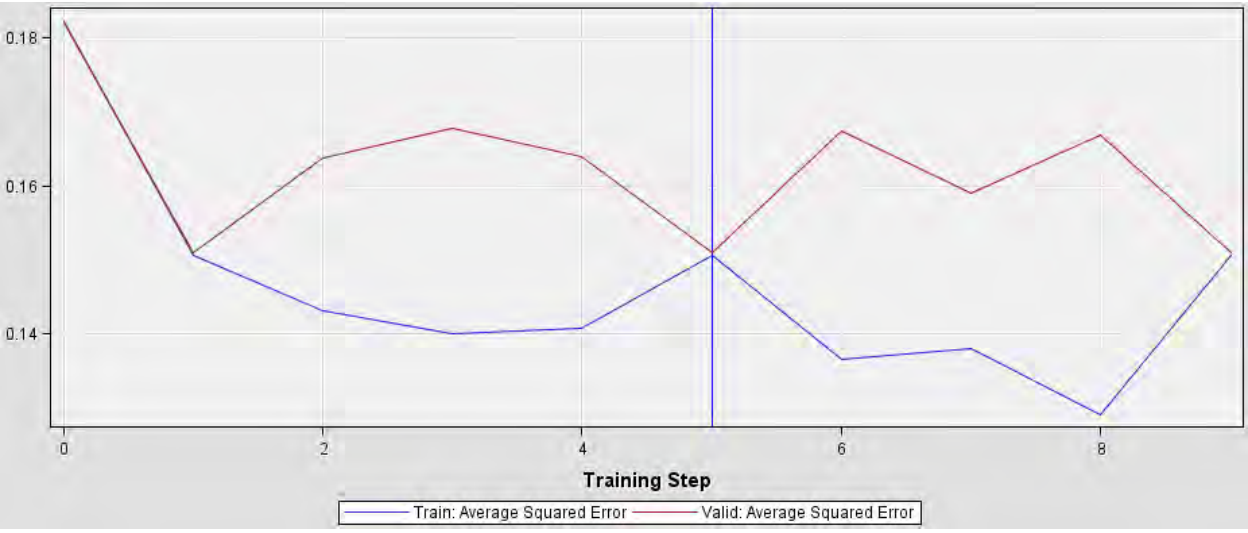

Figure 4-1: Iteration Plot of Average Square Error

Table 4-1: Fit Statistics

Fit Statistics	Statistics Label	Validation	Train	Test
DFT	Total Degrees of Freedom	.	709	.
DFE	Degrees of Freedom for Error	.	690	.
DFM	Model Degrees of Freedom	.	19	.
NW	Number of Estimated Weights	.	19	.
AIC	Akaike's Information Criterion	.	692.6841	.
SBC	Schwarz's Bayesian Criterion	.	779.3974	.
ASE	Average Squared Error	0.150932	0.150619	0.161044
MAX	Maximum Absolute Error	0.954448	0.940206	0.999983
DIV	Divisor for ASE	1062	1418	1064
NOBS	Sum of Frequencies	531	709	532
RASE	Root Average Squared Error	0.3885	0.388097	0.401303
SSE	Sum of Squared Errors	160.2902	213.5783	171.3509
SUMW	Sum of Case Weights Times Freq	1062	1418	1064
FPE	Final Prediction Error	.	0.158914	.
MSE	Mean Squared Error	0.150932	0.154767	0.161044
RFPE	Root Final Prediction Error	.	0.398641	.
RMSE	Root Mean Squared Error	0.3885	0.393404	0.401303
AVERR	Average Error Function	0.46722	0.461695	0.509278
ERR	Error Function	496.1873	654.6841	541.8713
MISC	Misclassification Rate	0.210923	0.221439	0.234962
WRONG	Number of Wrong Classifications	112	157	125

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

Fit Statistics	Statistics Label	Train	Validation	Test
ASE	Average Squared Error	0.1342	0.152903	0.163145
DIV	Divisor for ASE	1418	1062	1064
MAX	Maximum Absolute Error	0.933021	0.963506	0.976785
NOBS	Sum of Frequencies	709	531	532
RASE	Root Average Squared Error	0.366333	0.391028	0.403912
SSE	Sum of Squared Errors	190.2953	162.383	173.5858
DISF	Frequency of Classified Cases	709	531	532
MISC	Misclassification Rate	0.19464	0.229755	0.238722
WRONG	Number of Wrong Classifications	138	122	127

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

Fit Statistics	Statistics Label	Train	Validation	Test
ASE	Average Squared Error	0.134192	0.152897	0.16313
DIV	Divisor for ASE	1418	1062	1064
MAX	Maximum Absolute Error	0.932846	0.963289	0.976753
NOBS	Sum of Frequencies	709	531	532
RASE	Root Average Squared Error	0.366323	0.391021	0.403894
SSE	Sum of Squared Errors	190.2847	162.3768	173.5704
DISF	Frequency of Classified Cases	709	531	532
MISC	Misclassification Rate	0.19464	0.229755	0.238722
WRONG	Number of Wrong Classifications	138	122	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

Fit Statistics	Statistics Label	Train	Validation	Test
ASE	Average Squared Error	0.150628	0.150935	0.161048
DIV	Divisor for ASE	1418	1062	1064
MAX	Maximum Absolute Error	0.940291	0.954612	0.999885
NOBS	Sum of Frequencies	709	531	532
RASE	Root Average Squared Error	0.388108	0.388503	0.401308
SSE	Sum of Squared Errors	213.5899	160.2927	171.3551
DISF	Frequency of Classified Cases	709	531	532
MISC	Misclassification Rate	0.221439	0.212806	0.236842
WRONG	Number of Wrong Classifications	157	113	126

8. Model Comparison and Score

According to the results from Model Comparison process, ensemble model (auto neural network and regression) demonstrates 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 (Auto Neural Network and Regression);
(2) Logistic Regression;
(3) Auto Neural Network;
(4) Decision Tree;
(5) Ensemble (Auto Neural Network and Neural Network);
(6) Ensemble (Auto Neural Network and Regression);
(7) Neural Network.

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.

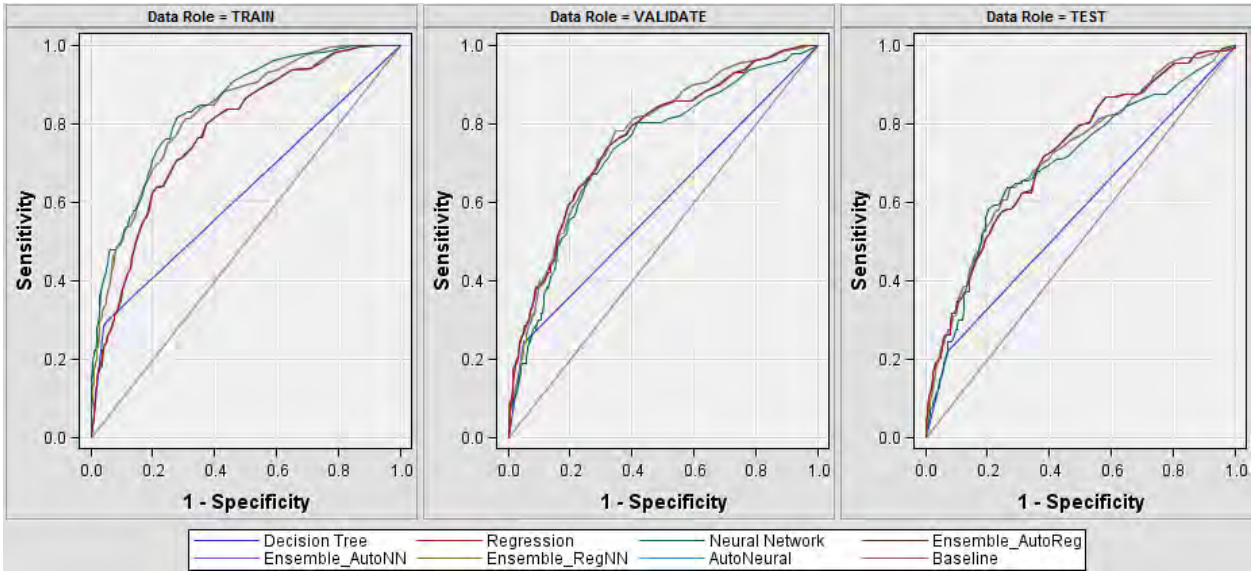

Figure 8-1: ROC Chart

In order to predict the probability of event for each record, the Score Node in SAS® Enterprise Miner was used. Figure 8-2 demonstrates the visualization of grades prediction using training data sets. The scoring code is also available in the node results. The visualization provides an early alert to the advisors of the students that have relatively higher risk to fail in this course. Based on the information from SAS Visual Analytics, students' advisors and faculty members could provide some support to the at-risk students in a timely manner.

Figure 8-2: Model Visualization Using SAS® Visual Analytics

CONCLUSION

SAS® Enterprise Miner and Visual Analytics provide an effective resolution to reduce risk and optimize resources in higher education. Ensemble modeling and logistic regression 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. There are a few things we could do to improve the accuracy of this predictive model. First, clustering analysis could be added into the model. Second, more interaction variables could be added in the data set (For example, first year experience course taken information, early college experience information, etc.). 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.

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.

ACKNOWLEDGMENTS

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.

INDEX

Variable	Description
ID	Student ID
Age	Age
AP_Course	Whether students had taken AP courses before?
Career_Level	Freshman, Sophomore, Junior, Senior
Class_Campus	Class Campuses
Enrollment_Campus	Student Enrolled Campuses
FirstGen_Flag	Whether from first generation family?
FullPart	Full-time, Part-time
Gender	Female, Male
LowIncome_Flag	Whether students came from low income family?
Residence	In-state, Out-of-state
SATmath	SAT Math Score
SATverbal	SAT Verbal Score
STEM_Flag	Whether students enrolled in STEM programs or not?
TARGET	DFW or Not
Underrepresented_Flag	Whether students were Underrepresented Minority

CONTACT INFORMATION

Youyou Zheng
Office of Institutional Research and Effectiveness
University of Connecticut
youyou.zheng@uconn.edu

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