

A Data Mining Approach to Predict Student-at-risk

Youyou Zheng, Thanuja Sakruti, Abhishek Uppalapati, Sivakumar Jaganathan

Office of Institutional Research & Effectiveness

University of Connecticut

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, organic chemistry is a required course for undergraduate students in a STEM discipline. It has a very high DFW rate (D=Drop, F=Failure, W=Withdraw). Take Fall 2014 as an example, the average DFW% is 24% at UCONN and there are over 1200 students enrolled in this class. In this study, undergraduate students enrolled during School Year 2010 to 2011 was used to build up the model. SEMMA (Sample, Explore, Modify, Model and Assess) method introduced by SAS Institute Inc. were applied to develop the predictive model. The freshmen SAT scores, campus, semester GPA, financial aid, 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 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,

Table 1: Variables used in the analysis

2. Models

2

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.

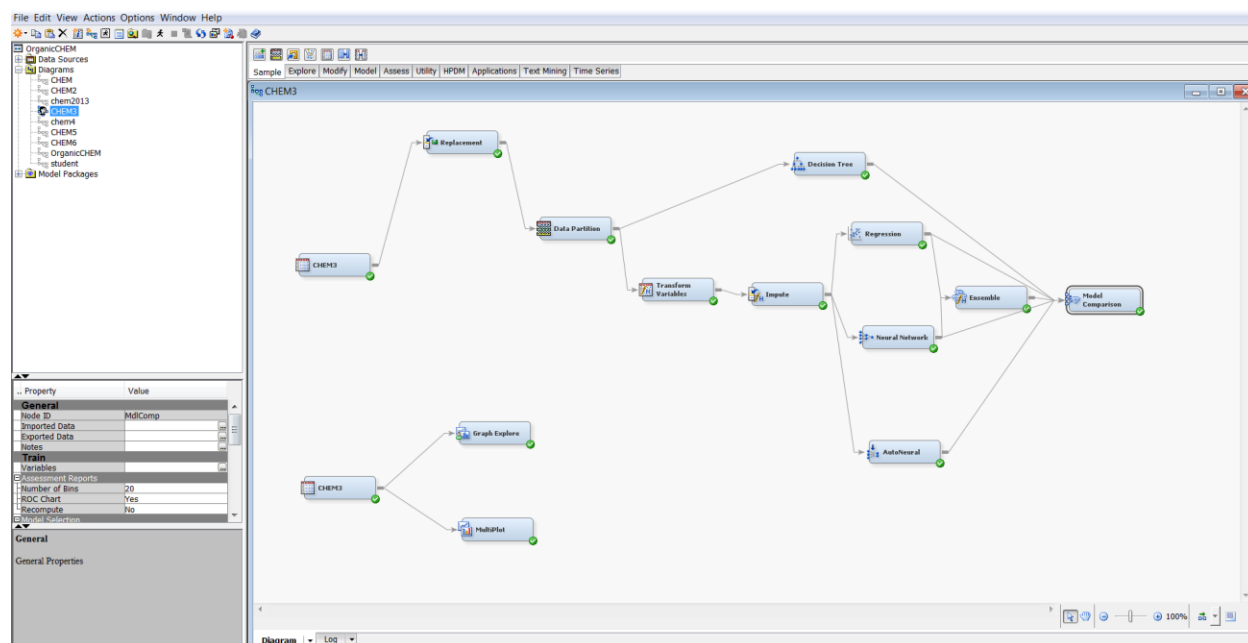


Figure 1: SAS Enterprise Miner process flow chart

Results

1. Decision Tree

Based on the results from the Decision Tree, SAT Math score was of great importance to predict DFW rate of organic chemistry lecture. 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.

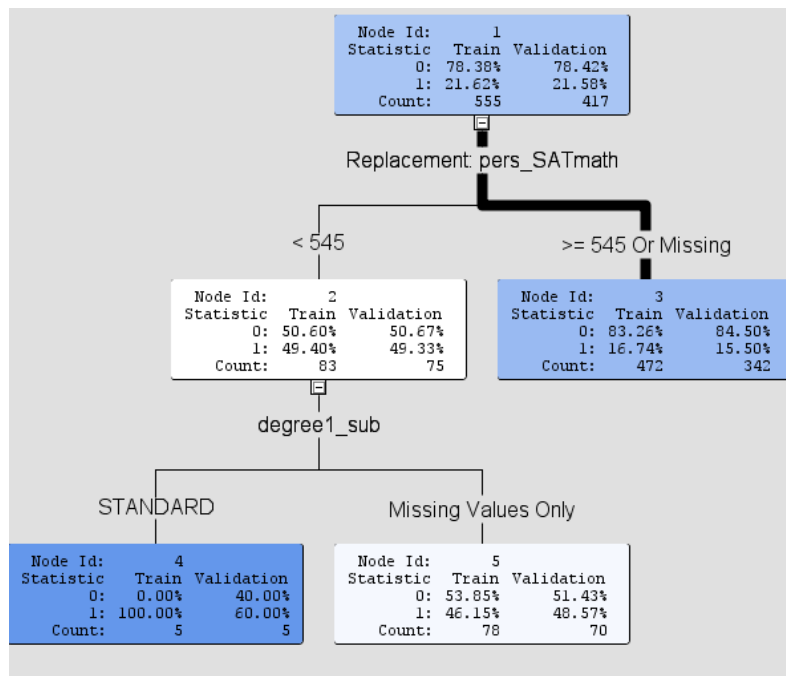


Figure 2: Tree

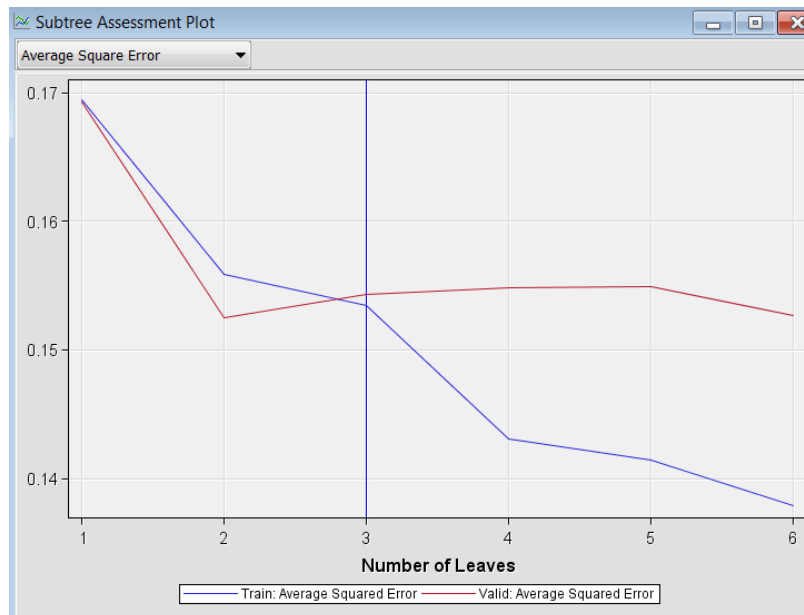


Figure 3: Subtree assessment plot of average square error

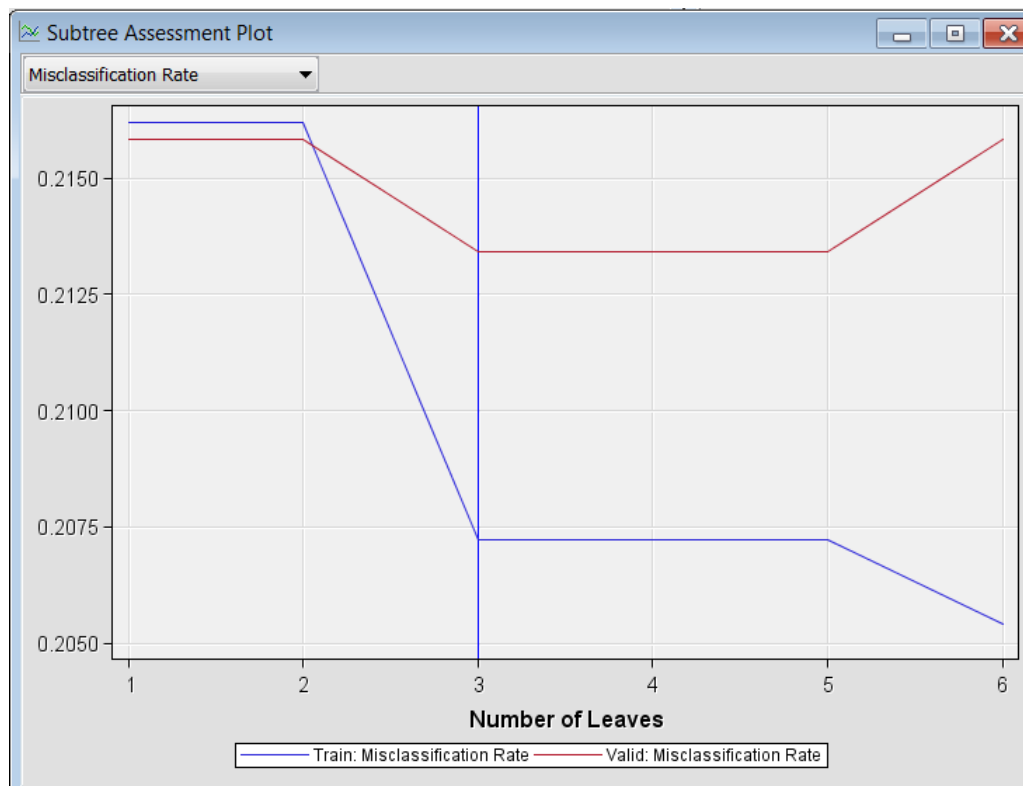


Figure 4: Subtree assessment plot of misclassification rate

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 variables included class campus, students entering term, students load, SAT Math score, residential states, ethnicity, and students major.

Table 2: Type 3 analysis of effects

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Campus	4	810.7654	<.0001
Cterm_Term_Sdesc	3	8.7624	0.0326
FullPart	1	3.4903	0.0617
IMP_REP_Distance_from_School	1	0.1477	0.7007
IMP_REP_EnrlUnitsTaken	0	0.0000	.
IMP_REP_pers_SATmath	1	5.2631	0.0218
IMP_REP_pers_SATverbal	1	0.4763	0.4901
IMP_degreeel_plan	58	471.7988	<.0001
IMP_degreeel_program	1	0.0213	0.8838
IMP_pers_ethnicity	4	8523.3680	<.0001
IMP_pers_state	15	68.0085	<.0001
Plan_Desc	22	2943.9437	<.0001
REP_PELL	1	0.9176	0.3381
TRANSFER	1	0.0021	0.9634
pers_gender	1	0.2202	0.6389

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		AIC	Akaike's Information Criterion	598.6853	.	.
TARGET		ASE	Average Squared Error	0.094386	0.197777	0.168224
TARGET		AVERR	Average Error Function	0.292509	1.352231	1.044702
TARGET		DFE	Degrees of Freedom for Error	418	.	.
TARGET		DFM	Model Degrees of Freedom	137	.	.
TARGET		DFT	Total Degrees of Freedom	555	.	.
TARGET		DIV	Divisor for ASE	1110	834	838
TARGET		ERR	Error Function	324.6853	1127.761	875.4599
TARGET		FPE	Final Prediction Error	0.156256	.	.
TARGET		MAX	Maximum Absolute Error	0.986281	1	1
TARGET		MSE	Mean Square Error	0.125321	0.197777	0.168224
TARGET		NOBS	Sum of Frequencies	555	417	419
TARGET		NW	Number of Estimate Weights	137	.	.
TARGET		RASE	Root Average Sum of Squares	0.307223	0.444722	0.410152
TARGET		RFPE	Root Final Prediction Error	0.395293	.	.
TARGET		RMSE	Root Mean Squared Error	0.354007	0.444722	0.410152
TARGET		SBC	Schwarz's Bayesian Criterion	1190.384	.	.
TARGET		SSE	Sum of Squared Errors	104.7684	164.9463	140.9721
TARGET		SUMW	Sum of Case Weights Times Freq	1110	834	838
TARGET		MISC	Misclassification Rate	0.140541	0.258993	0.221957

Table 3: Fit statistics

3. Neural Network

According to the Iteration Plot, the optimal average square error occurred on the 4th iteration for neural network model.

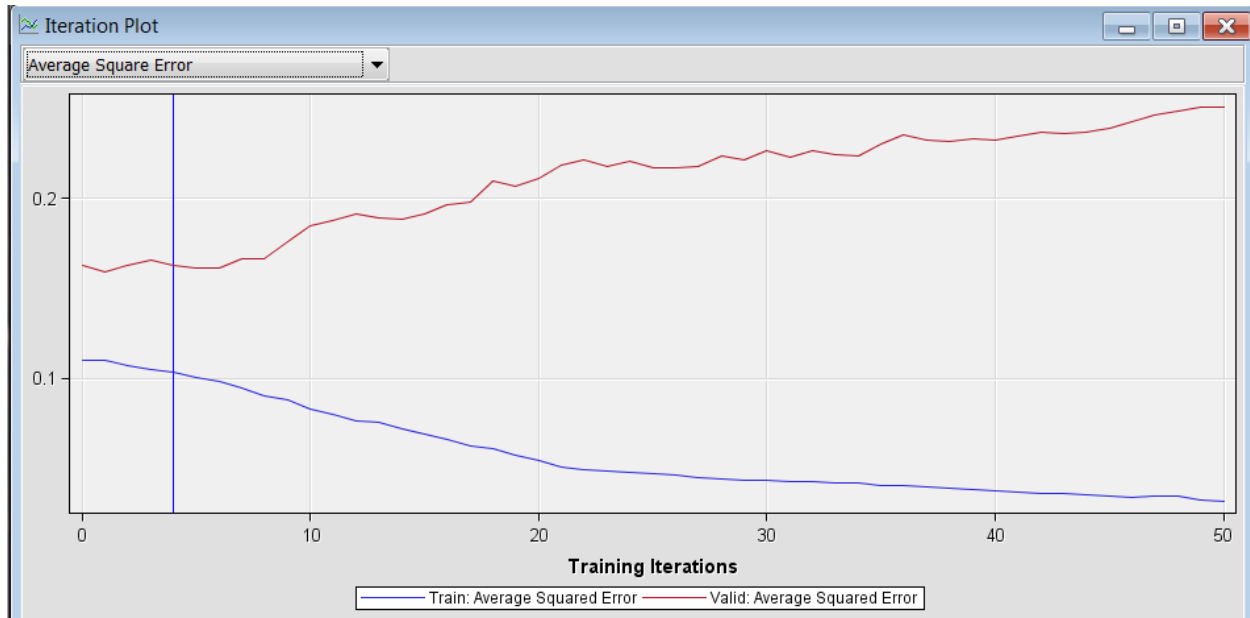


Figure 5: Iteration plot of average square error

Table 4: Fit statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		DFT	Total Degrees of ...	555		
TARGET		DFE	Degrees of Freed...	110		
TARGET		DFM	Model Degrees of ...	445		
TARGET		NW	Number of Estima...	445		
TARGET		AIC	Akaike's Informati...	1253.258		
TARGET		SBC	Schwarz's Bayesi...	3175.198		
TARGET		ASE	Average Squared ...	0.103483	0.162881	0.157476
TARGET		MAX	Maximum Absolut...	0.996176	0.998467	0.998357
TARGET		DIV	Divisor for ASE	1110	834	838
TARGET		NOBS	Sum of Frequenci...	555	417	419
TARGET		RASE	Root Average Squ...	0.321688	0.403585	0.396832
TARGET		SSE	Sum of Squared ...	114.866	135.8426	131.9646
TARGET		SUMW	Sum of Case Wei...	1110	834	838
TARGET		FPE	Final Prediction Er...	0.940753		
TARGET		MSE	Mean Squared Error	0.522118	0.162881	0.157476
TARGET		RFPE	Root Final Predicti...	0.969924		
TARGET		RMSE	Root Mean Squar...	0.722577	0.403585	0.396832
TARGET		AVERR	Average Error Fun...	0.327259	0.5093	0.476724
TARGET		ERR	Error Function	363.2577	424.7559	399.4948
TARGET		MISC	Misclassification ...	0.145946	0.232614	0.231504
TARGET		WRONG	Number of Wrong...	81	97	97

4. Auto Neural Network

The auto Neural Network process gave the optimal average square error on the 5th iteration as shown below.

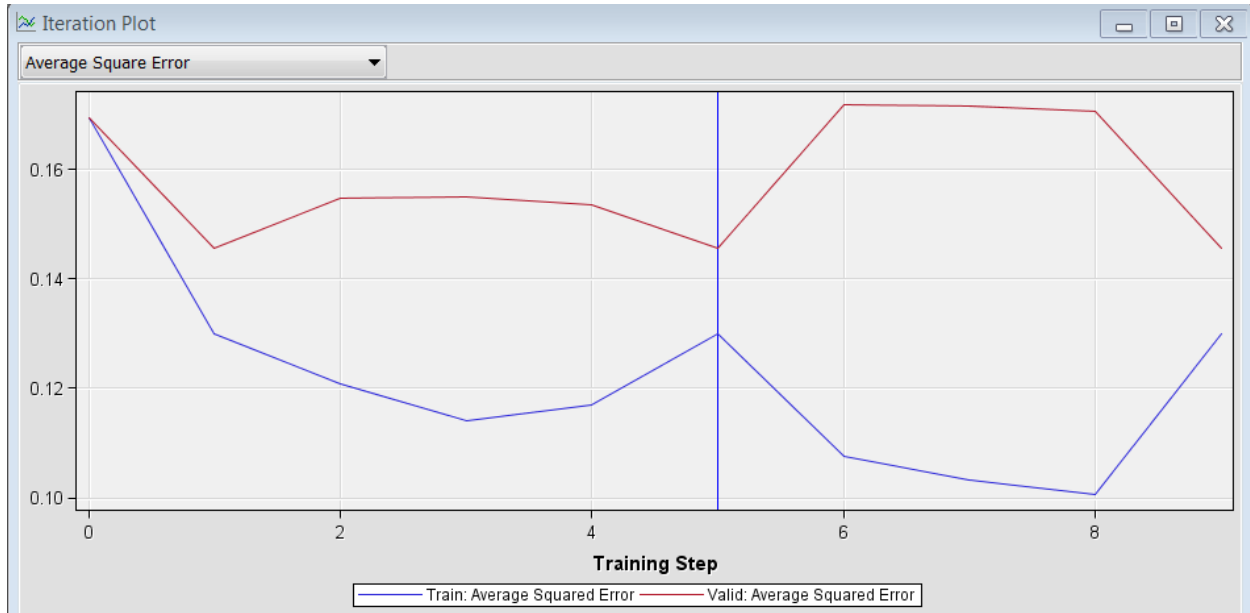


Figure 6: Iteration plot of average square error

Table 5: Fit statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		DFT	Total Degrees of ...	555		
TARGET		DFE	Degrees of Freed...	408		
TARGET		DFM	Model Degrees of ...	147		
TARGET		NW	Number of Estima...	147		
TARGET		AIC	Akaike's Informati...	737.8914		
TARGET		SBC	Schwarz's Bayesi...	1372.78		
TARGET		ASE	Average Squared ...	0.129938	0.145482	0.155006
TARGET		MAX	Maximum Absolut...	0.994707	0.999111	0.984375
TARGET		DIV	Divisor for ASE	1110	834	838
TARGET		NOBS	Sum of Frequenci...	555	417	419
TARGET		RASE	Root Average Squ...	0.360469	0.381421	0.393708
TARGET		SSE	Sum of Squared ...	144.2313	121.3317	129.8953
TARGET		SUMW	Sum of Case Wei...	1110	834	838
TARGET		FPE	Final Prediction Er...	0.22357		
TARGET		MSE	Mean Squared Error	0.176754	0.145482	0.155006
TARGET		RFPE	Root Final Predicti...	0.472832		
TARGET		RMSE	Root Mean Squar...	0.420421	0.381421	0.393708
TARGET		AVERR	Average Error Fun...	0.399902	0.46081	0.462939
TARGET		ERR	Error Function	443.8914	384.3155	387.9427
TARGET		MISC	Misclassification ...	0.196396	0.194245	0.250597
TARGET		WRONG	Number of Wrong...	109	81	105

5. Ensemble

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

Table 6: Fit statistics

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		ASE	Average Squared Error	0.09261	0.168661	0.151376
TARGET		DIV	Divisor for ASE	1110	834	838
TARGET		MAX	Maximum Absolute Error	0.991228	0.995974	0.999178
TARGET		NOBS	Sum of Frequencies	555	417	419
TARGET		RASE	Root Average Squared Error	0.304319	0.410683	0.389071
TARGET		SSE	Sum of Squared Errors	102.7974	140.6629	126.8534
TARGET		DISF	Frequency of Classified Cases	555	417	419
TARGET		MISC	Misclassification Rate	0.115315	0.247002	0.229117
TARGET		WRONG	Number of Wrong Classifications	64	103	96

6. Model Comparison

According to the results from Model Comparison process, Auto Neural Network provided the optimal model. The order from the best to worst performance for the 5 models were as following:

- (1) Auto Neural Network;
- (2) Decision Tree;
- (3) Neural Network;
- (4) Ensemble Model;
- (5) Logistic Regression.

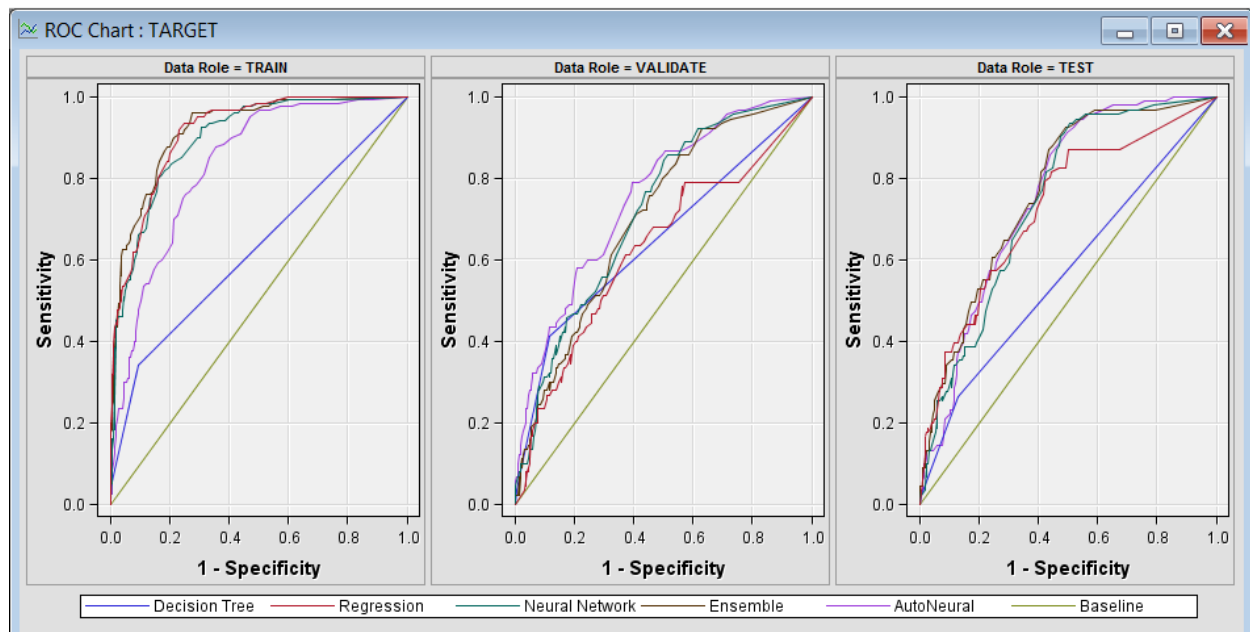


Figure 7: ROC Chart

Table 7: Fit statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	T Train: Misclassification Rate	T Train: Average Squared Error	T T	T T	V Valid: Misclassification Rate	V V Valid: Average Squared Error	V V T	T Test: Misclassification Rate	T T Test: Average Squared Error
Y	AutoNeural	AutoNeural	AutoNeural	TARGET		0.194245	0.196396	0.129938			0.194245	0.145482		0.250597	0.155006
	Tree	Tree	Decision Tree	TARGET		0.213429	0.207207	0.153445			0.213429	0.154354		0.217184	0.170752
	Neural	Neural	Neural Network	TARGET		0.232614	0.145946	0.103483			0.232614	0.162881		0.231504	0.157476
	Ensmbl	Ensmbl	Ensemble	TARGET		0.247002	0.115315	0.09261			0.247002	0.168661		0.229117	0.151376
	Reg	Reg	Regression	TARGET		0.258993	0.140541	0.094386			0.258993	0.197777		0.221957	0.168224

Conclusion and Discussion

SAS Enterprise Miner is a powerful tool for higher education data mining. Neural network is useful for not only enrollment management but also course data analysis. Neural network is normally more efficient with a bigger data set. 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. In the future study, the variables such as first generation, study abroad, and previous course information will be considered. By the development of course data mining, our institution could get a better undergraduate degree awarded predictive model, which could help more students obtain their degree within 4 years.

Acknowledgement

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

Reference

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.oair.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

Obs#	Variable Name	Type	Description
1	Campus	CLASS	Campus Providing the Course
2	ClassCatalogNbr	CLASS	Course Catalog Number
3	ClassComponentCd	CLASS	Course Component Code
4	ClassSubjectCd	CLASS	Course Subject Code
5	Cterm_Term_Sdesc	CLASS	Student Entering Term
6	DEGREE	CLASS	Whether graduated in 4 years?
7	DFW	CLASS	Drop, Fail, and Withdraw
8	EnrlEmplid	CLASS	Student ID
9	EnrlOfficialGrade	CLASS	Course Official Grade
10	FstTimeCat	CLASS	First Time Freshmen, Transfer, Others
11	FullPart	CLASS	Full-time, Part-time
12	Plan_Desc	CLASS	student major description
13	TARGET	CLASS	Whether student gets DFW?
14	TRANSFER	CLASS	Whether student is transfer student
15	degree1_awarded	CLASS	Graduated in 4 years or not
16	degree1_date	CLASS	degree awarded date
17	degree1_plan	CLASS	student major description
18	degree1_program	CLASS	student program description
19	degree1_sub	CLASS	sub-major
20	pers_ethnicity	CLASS	ethnicity
21	pers_gender	CLASS	gender
22	pers_state	CLASS	living state
23	Amount	VAR	How much from the financial aid?
24	ClassTermCd	VAR	Terms providing the course
25	Distance_from_School	VAR	home distance from school
26	EnrlUnitsTaken	VAR	units from the course
27	PELL	VAR	Whether obtained financial aid?
28	gpa_cum	VAR	cumulative GPA
29	gpa_sem	VAR	semester GPA
30	pers_SATmath	VAR	SAT math score
31	pers_SATverbal	VAR	SAT verbal score