SESUG Paper 243-2018

Predicting the Risk of Attrition for Undergraduate Students using SAS® Enterprise Miner™

Author: Sivateja Reddy Kandula,

Oklahoma State University, Stillwater

ABSTRACT

A more substantial number of undergraduate students drop out of college before their graduation despite the efforts of college management to improve the retention rates. In order to strengthen these retention rates, it is essential to identify the significant factors that contribute towards students' attrition. In this paper, I will concentrate on various aspects that play an important role in a student's decision to drop out of a college. According to surveys on student attrition, it is evident that students' incoming credentials such as high school grades, demographic factors such as gender, race, and distance from the college, financial factors, socio-economic factors, academic performance, and campus involvement of students play a pivotal role in students' decision to attrite from college. In this paper, I will mainly concentrate on students' attrition in the third semester meaning how likely a student drops out or transfer from a college before reaching the third semester. For this project the data is acquired from a large mid-west university for Fall 2016 - Fall 2017 and the model is used to score data from Spring 2017 - Spring 2018. Variables mentioned above will be used along with some calculated fields to predict college attrition/retention.

This paper focuses on analyzing student applicant data and their campus involvement within the first two semesters using SAS Enterprise Guide 7.1, SAS Enterprise Miner 14.1. This project will determine the probability of attrition of each student. Results from this study will help university officials provide services to those students who may be at risk for drop out.

Keywords: Data Mining, Machine Learning, Predictive Modeling, Student Attrition, Graduation

INTRODUCTION

The current schema of higher education demands that universities strive to compete across several different offerings. Catering to the needs of changing the population of students with diverse priorities and backgrounds is a progressively complicated process that goes well beyond initial enrollment. Today's undergraduate students have more options than ever when it comes to choosing a university leading to an increase in dropping out or transferring from one university to another. This phenomenon is called student attrition, that is students leaving the program of study before they graduate.

Student retention rate is one of the key indicators of student success used by many universities. Measuring this rate is essential because it can also evaluate university success in retaining students based on the quality of education, curriculum, affordability, brand, welcoming international students and many other. Student retention is challenging for many universities as there an ample number of reasons why a student attrite from college before graduation. On the other hand, students may think transferring schools is not a big deal, but it can be very detrimental to a student advancing. Not only does a transfer cost, money and time, it also makes it more challenging for students to finish the academic program on time or at all.

This paper mainly focusses on building different predictive models like Logistic Regression, Decision Trees, Neural Network, Random Forest, Support Vector Machine (SVM), Bayesian Network Classifier and Gradient Boosting and comparing the results between models.

DATA MODELING

Gathered data from Information Research and Information Management (IRIM) Oklahoma State University (OSU), Stillwater for the semesters Fall 2016, Spring 2017, Fall 2017, and Spring 2017. This data contains demographic information like race, gender, birth date, resident status, and identification as Hispanic, pre-college information like High School GPA and other courses GPA if available, family information related to OSU legacy, courses student enrolled in like college, degree, and major, amount they owe to college, academic information like GPA for enrolled semesters and credits earned, other campus involvement like events they participated during college, athletes played, interaction between students and college before their admission. The training data includes the fall of 2016 and spring of 2017. The data was put together by series of DATA steps, and PROC SQL joins and based on Student_ID. The below table provides the list of input variables that were passed into the model.

Variable Name	Variable Description
Gender	Gender of students
Age	Age of students
Race	Race of students
Region	Region students come from
Hispanic	Y/N if student is Hispanic
HS_GPA	High School GPA
HS_Units	High School total credits
Application_Student_Type	Transfer, Freshmen etc.
First_Generation	First Generation Students
OSU_Legacy	Yes/No if parents are legacy of OSU
College	College they enrolled
Major	Major concentration
Degree	Degree they were interested
App_Entry	When they applied to OSU
Interactions	Interactions between university and students
Student_initialised	How many students initialized interactions
events	Number of events they participated in college
Athletes_Played	Number of Athletes played in college
Emails_Sent	Emails college sent to students before admission
Responded_Mails	Emails responded by student before admission
Total_Hours_Enrolled	Number of hours enrolled for two semesters
Fall_credits	Credits earned in Fall Semester
Spring_credits	Credits earned in Spring Semester
GPA	Total GPA for two semesters
Account_Balance	Balance they owe to college
Attrition	Target variable (1 – Attrited; 0 – Not Attrited)

Table 1: Input Variables

For several student records, there are missing values in individual credits and GPA information of their high-school education, to combat this problem I have used HS_GPA and total credits they earned. Campus involvement data like events they attended to these stored as per each entry so have to aggregate data against each student and join its training data. As far it goes for GPA data I have combined the GPA for both semesters and generated a single column of GPA that is aggregate for both semesters. As for the target variable, created an Attrition column by looking up whether a student is enrolled in third semester (set to 0) if not (set to 1). Attrition column acts as the target variable for the model to classify. In target variable there are 1038 dropout students (Attrition = 1) and 5550 retained students (Attrition = 0). We can see that dropout students account for about 16% of total students enrolled and hence we must oversample the data to eliminate the bias prediction. We can adjust the prior probabilities in SAS Miner using the decisions tab under properties panel of the data set.

DATA ROLES IN SAS MINER

Based on the values of independent variables I have assigned appropriate roles in SAS Enterprise Miner. Assigned Student variable to ID as a role, and Attrition variable to target as role and Level to Binary. The below figure shows the roles and their measurement levels assigned to each of the independent variables.

Name	Role	Level
STUDENT_ID	ID	Nominal
interactions	Input	Interval
HS_Units	Input	Interval
MAJOR	Input	Nominal
OSU_Legacy	Input	Nominal
Hispanic	Input	Nominal
GPA	Input	Interval
HS_GPA	Input	Interval
spring_credits	Input	Interval
responded_mails	Input	Interval
Total_Hours_Enrolled	Input	Interval
student_initiated	Input	Interval
Region	Input	Nominal
Race	Input	Nominal
Residency	Input	Nominal
Banner_Admitted_Residency	Input	Nominal
Athletes_played	Input	Interval
Date_Submitted	Input	Nominal
COLLEGE	Input	Nominal
App_Entry	Input	Nominal
ACCOUNT_BALANCE	Input	Interval
Application_Student_Type	Input	Nominal
First_Generation	Input	Nominal
fall_credits	Input	Interval
Gender	Input	Nominal
events	Input	Interval
emails_sent	Input	Interval
DEGREE	Input	Nominal
fall_gpa	Rejected	Interval
spring_gpa	Rejected	Interval
Attrition	Target	Binary

Figure 1: Roles and Levels of input variables in SAS Miner

DESCRIPTIVE STATISTICS

In SAS® Enterprise Miner™ Stat Explorer node is used to generate descriptive statistics of input variables. The following figure shows the basic summary statistics of all class and interval variables.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	App_Entry	INPUT	7	1	Fall 2016	82.07	Spring 2016	10.78
TRAIN	Application_Student_Type	INPUT	7	0	Freshman	59.41	Transfer 2	27.55
TRAIN	Banner_Admitted_Residency	INPUT	4	1004	In-state Resident	58.88	Out of State	22.81
TRAIN	COLLEGE	INPUT	8	0	AS	21.21	EN	19.26
TRAIN	DEGREE	INPUT	24	0	BS	19.47	B	17.40
TRAIN	Date_Submitted	INPUT	485	0	01Jul2015	1.23	01Feb2016	0.97
TRAIN	First_Generation	INPUT	3	3	No	82.26	Yes	17.70
TRAIN	Gender	INPUT	2	0	M	50.05	F	49.95
TRAIN	Hispanic	INPUT	3	62	No	90.50	Yes	8.56
TRAIN	MAJOR	INPUT	90	0	UND	15.22	ANSI	5.25
TRAIN	OSU_Legacy	INPUT	3	76	No	73.77	Yes	25.08
TRAIN	Race	INPUT	6	167	White	75.06	American Indian or Alaska Native	10.58
TRAIN	Region	INPUT	94	46	OK	73.19	TX	14.89
TRAIN	Residency	INPUT	3	4831		73.33	In-State	19.81
TRAIN	Attrition	TARGET	2	0	0	84.24	1	15.76

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
ACCOUNT_BALANCE	INPUT	1116.709	3525.135	6588	0	0	5.08	66233.22	7.732107	87.67849
Athletes_played	INPUT	0.029447	0.16907	6588	0	0	0	1	5.568054	29.01203
GPA	INPUT	2.953569	0.860215	6582	6	0	3.125	4	-1.09011	1.01694
HS_GPA	INPUT	3.596401	0.484181	4162	2426	1.75	3.67	4.73	-0.50353	-0.16081
HS_Units	INPUT	14.88614	0.505741	4163	2425	0	15	18	-10.2468	213.5516
Total_Hours_Enrolled	INPUT	27.08075	5.763296	6588	0	2	28	44	-1.66049	3.302681
emails_sent	INPUT	65.79432	38.25297	6588	0	0	70	195	-0.10663	-1.06126
events	INPUT	0.74742	0.848802	6588	0	0	1	6	1.258566	2.129513
fall_credits	INPUT	12.40437	3.416771	6588	0	0	13	21	-1.32957	1.923207
interactions	INPUT	7.156193	6.088716	6588	0	0	6	70	2.039878	8.82162
responded_mails	INPUT	35.64056	25.42817	6588	0	0	32	124	0.574558	-0.46883
spring_credits	INPUT	12.64648	4.028622	6588	0	0	13	23	-1.19331	1.343469
student_initiated	INPUT	0.985124	2.053133	6588	0	0	0	33	4.452084	31.12805

Figure 2: Summary Statistics of input variables

The above figure showing the summary statistics of class variables says that data is cleaned and handled as required. Except for few variables like Date_Submitted, Major, and Region all the class variables data is complete.

For interval variables especially for HS_GPA and HS_Units there are a lot of missing values in the dataset, so we need to find proper imputation techniques to address this issue when we are applying logistic regression techniques on these variables whereas other algorithms like Random Forest, Decision Trees, and Gradient Boosting are not affected by missing data. Also, the skewness of few variables like HS_Units, Account_Balance, and student_initialized variables is a bit high, and so appropriate transformation is necessary, and those are decsribed in the later section in this paper.

Stat Explorer node also provides chi-square statistics for top 20 variables using Cramer's V statistic along with variable worth for top 20 variables using GINI split statistics that is generated by building a decision tree of depth 1. The following figure provides the information generated by the Stat Explorer node.

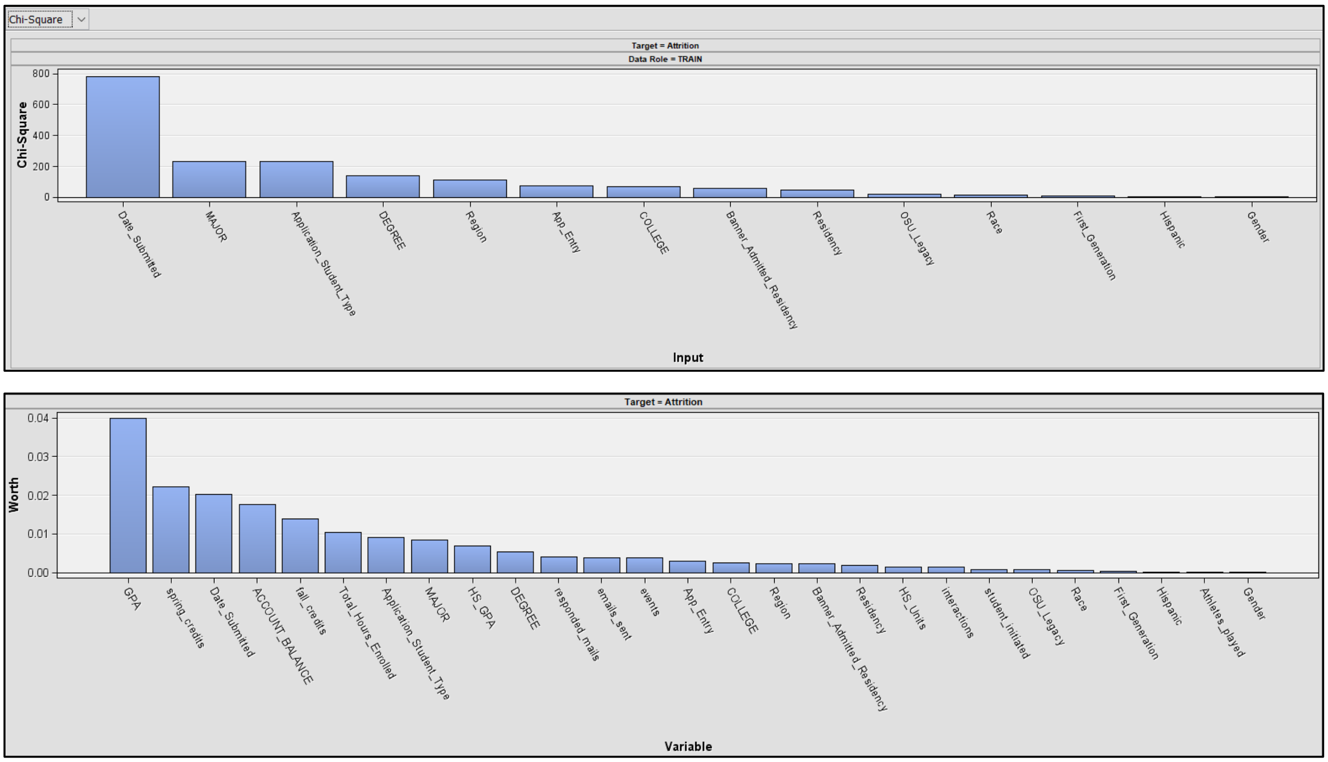

Figure 3: Chi-square and variable worth for input variables

From the above figure, one can observe the log worth of variables in determining the probability of attrition. Higher the worth of variable higher are the chances of deciding the target variable using this variable information.

From the figure 3, we can observe the variables GPA (cumulative GPA for both the semesters) is having the highest log worth value as compared to other variables. Other relevant variables are the number of credits student earned in the second semester, the date that he applied for college, the amount that he owes to college and credits he gained in fall semester can be used to determine the target variable (probability of attrition).

Interesting enough the campus involvement of students like responded emails, total emails sent to a student, events that he participated in college, and application entry date also seems to have some impact on attrition prediction, while variables such as OSU_legacy, race, first-generation students and so on have minimal impacts.

BUILDING PREDICTIVE MODELS USING SAS® ENTERPRISE MINER™

For this paper, SAS® Enterprise Miner™ read in a training dataset, and several machine learning algorithms are applied on it to classify the attrition of students. Models used in this paper include Logistic Regression, Decision Tree, Neural Networks, Random Forest, Support Vector Machine (SVM), Bayesian Network Classifier and Gradient Boosting. The following Figure shows the model diagram.

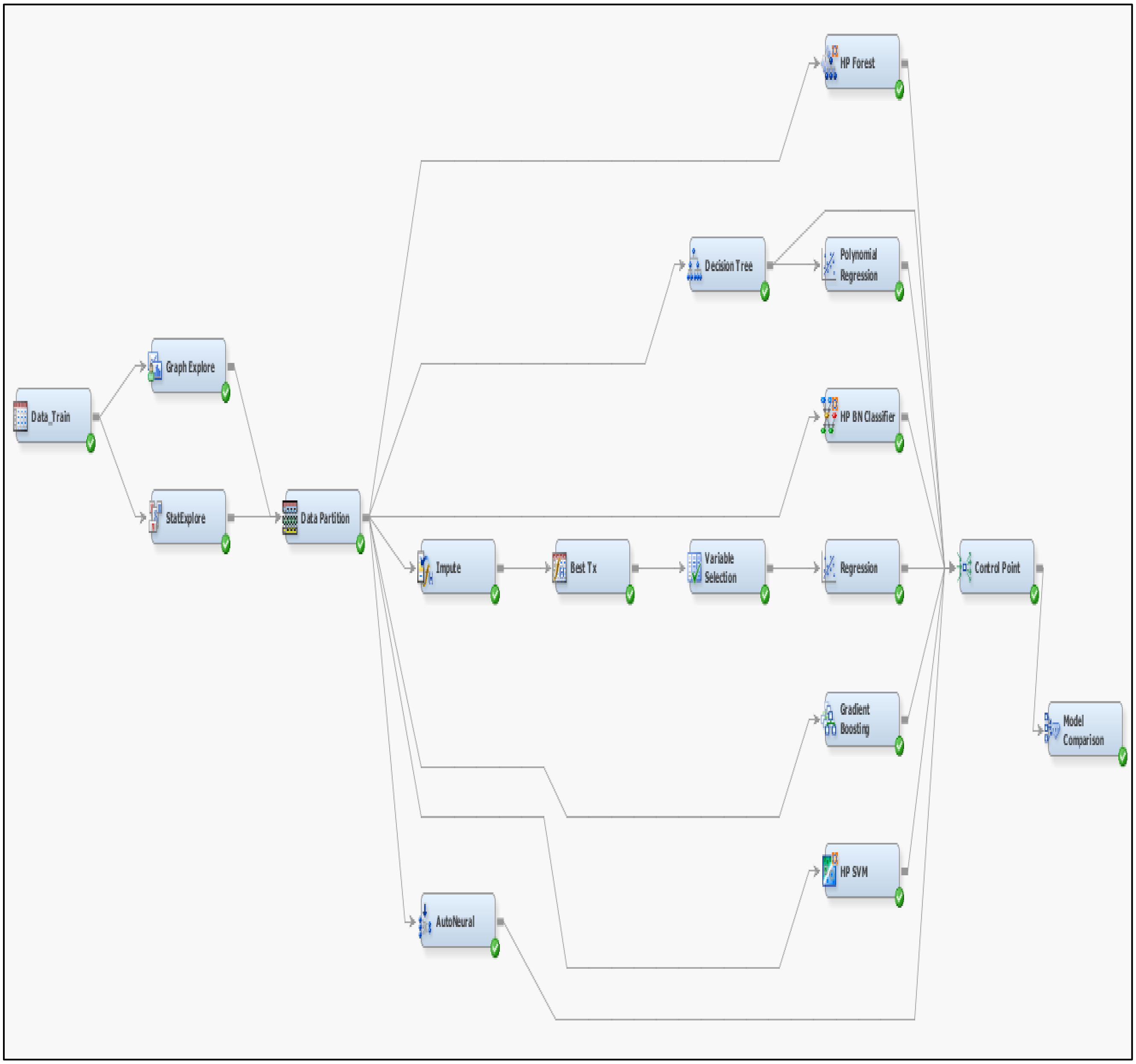

Figure 4: Model Diagram

Graph Explorer node is used to generate a variety of graphs based on different variables. Training data has been split by 70% to train the model and 30% to validate the model. Transformation Variables node is used to transform variables if there are any distortions in the distribution of variables that might cause noise in model results. To nullify this effect, I have considered using 'Best' Transformation meaning it performs several transformations and applies that has the best Chi-Squared test for the target. Initially with the default setting the modeling nodes provided the basis for an easy-to-implement predictive model; however, some adjustments were made to improve model performance. In this binary attrition prediction model total of six machine learning algorithms are used.

The stepwise regression model used a validation error as the selection criterion, which minimized error and overfitting. The Random Forest model used variable importance as Loss Reduction which helps to bring the significant variable as the first split. For Gradient Boosting the maximum depth has been set to 10 to improve classification sensitivity and Number of surrogate rules to two which can handle the missing values in primary predictor variable. To access the interaction between input variables polynomial regression is also used to look for the second order interaction (two variable interactions) and include effects of any such significant combinations. Decision Tree is used for variable selection before polynomial regression. The attributes like spring and fall GPA and credits enrolled, amount a student owe to college, application type (transfer, freshman etc.), students incoming credentials and campus involvement came out as important variables and it is quite understandable. As a student with good academic background and performing well in his program, low owed balance and good campus involvement means the student is most likely to stay with the university. On the other hand, a struggling student with low GPA, low campus involvement is a trigger that student is likely to dropout from college.

MODEL COMPARISONS

After a series of changes to default models, all the models are assessed by model comparison node in SAS Miner based on ROC on validation data. The following figure shows the model comparison results when Validation ROC is used as the selection criterion.

Selected Model ▼	Predecessor Node	Model Node	Model Descriptio n	Selection Criterion: Valid: Roc Index
Y	HPDMForest	HPDMFo...	HP Forest	0.819
	Tree	Tree	Decision ...	0.773
	Reg	Reg	Regressi...	0.773
	Boost	Boost	Gradient ...	0.761
	HPBNC	HPBNC	HP BN Cl...	0.697
	HPSVM	HPSVM	HP SVM	0.632
	Reg2	Reg2	Polynomi...	0.596
	AutoNeural	AutoNeural	AutoNeural	0.513

Figure 5: Model Comparison result; Selection Statistic: Validation ROC

Random Forest model has proven to be the champion model based on the above comparison results because it has the highest validation ROC index of 0.819. The second-best model is a tie between Decision Tree and Logistic Regression. While ROC index is a good indicator or model's overall prediction power, for more insights, we look at sensitivity/specificity numbers. Sensitivity is how well the model identifies the target = 1 students, that is the ability to determine the dropout students correctly. On the other hand, specificity is the ability to identify retained students by the university. The following figure is showing the assessment of models based on sensitivity and specificity.

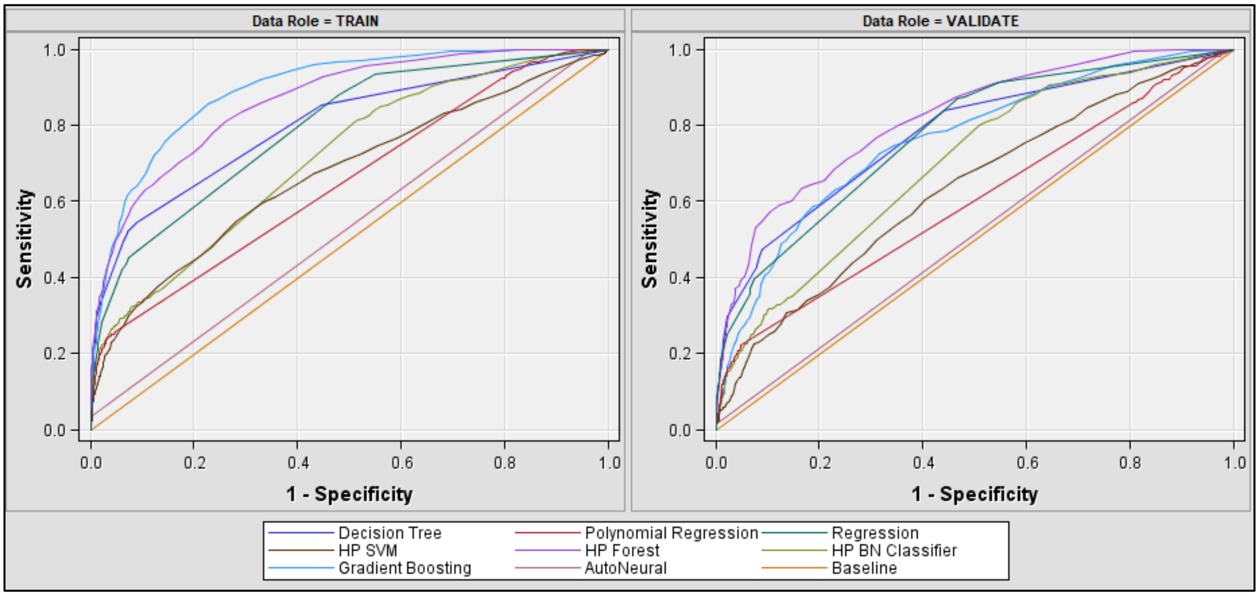

Figure 6: Comparisons between models

Based on above figure we can see that Decision Tree has better ROC index compared to Random Forest but with the validation, ROC is well below Random Forest meaning some amount of overfitting can be observed with Decision Tree model. On the contrary, Random Forest is showing similar results for both training and validation data proving to be a stable model.

OVERVIEW OF CUTOFF NODE

Now that I have identified the champion model, my next task is to find the optimal cutoff value. The cut-off node in SAS Miner can be used for this case (connect this cutoff node to model comparison node). This node provides tabular and graphical information to determine the appropriate cutoff probabilities for decision making with binary target models. This optimal cutoff can help minimize the risk of generating high false positives and high false negatives. For this, I have used Cutoff Method as Event Precision Equal Recall which can fetch the option cutoff which can balance both overall precision rate and balance between True Positives and False Positives. The following figure shows the setting used in this model.

General	
Node ID	CUT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Depth Scale %	1
Score	
Cutoff Method	Event Precision Equal Recall
Cutoff User Input	0.5
Status	
Create Time	7/28/18 7:40 PM
Run ID	77569588-e173-4dd9-a921-8
Last Error	
Last Status	Complete
Last Run Time	7/28/18 8:03 PM
Run Duration	0 Hr. 0 Min. 3.70 Sec.
Grid Host	
User-Added Node	No

Figure 7: Cutoff Node Properties

After running the cutoff node, we can see the optimal cutoff value suggested by the node in the Overall Rates chart. The following figure provides the Overall Rates chart.

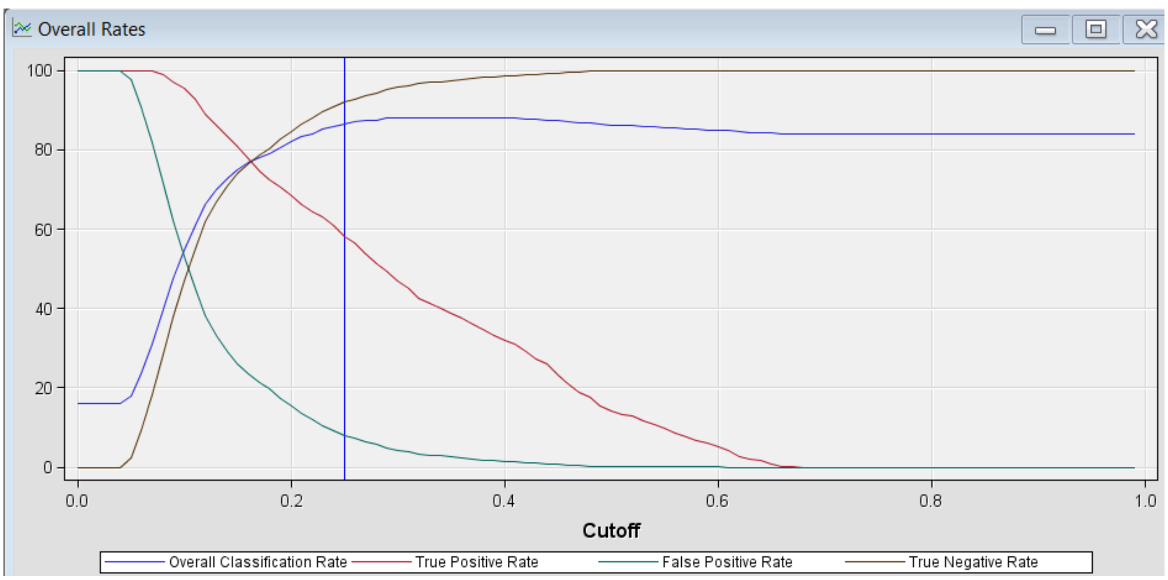

Figure 8: Overall Rates Chart

Based on the above figure, I have identified the optimal cutoff to use for scoring new data which is 0.25.

MODEL RESULTS

Champion Model	Random Forest
Validation ROC Index	0.819
Misclassification Rate	14.3%

Table 2: Model Summary

Variable Importance:

Random Forest provides variables that are important in determining the probability of attrition based on number splitting rules that variables is involved with. The below figure provides the ranking of variable importance based on splitting rules.

Variable Name	Number of Splitting Rules
spring gpa	190
fall gpa	151
Application Student Type	133
spring credits	122
ACCOUNT BALANCE	89
fall credits	87
App Entry	74
HS GPA	72
DEGREE	67
Total Hours Enrolled	61
COLLEGE	56
emails sent	51
HS Units	47
MAJOR	44
Residency	42
events	41
responded mails	37
Banner Admitted Residency	36
Race	31
student initiated	30
OSU Legacy	26
Date Submitted	25
interactions	19
First Generation	18
Region	18
Hispanic	15
Athletes played	14
Gender	10

Figure 9: Variable Importance

This model can be used to provide the attrition probability of students that are likely to attrite after the first two semesters of college. From figure 9, the GPA of two semesters, Application Student Type, credits they earned in the last two semesters seems to be providing more information than one can perceive in determining the probability of attrition. With the help of this model, the university can divide the students into specific tiers based on the above factors and counsel those students who are having a high probability of dropping out and try to retain them.

APPLYING THE PREDICTIVE MODELS TO SCORE DATA

Now that I identified optimal cutoff, I can use score node to score the data and determine the probability of student attrition for Spring 2018. But I observed the cutoff value is not applied to score node directly, and I came up with two methods using a SAS code to implement the newly determined cutoff value or change the cutoff method to user input and change the cutoff value manually to 0.25. Both the approaches seem to work. The following figure shows how to apply the predicted models to score new data.

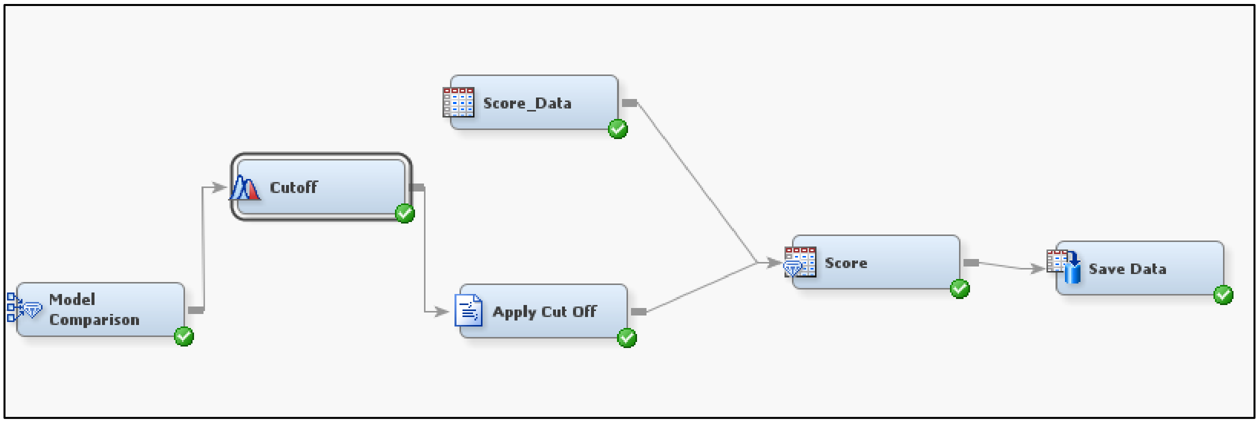

Figure 10: Model Diagram of Scoring new data

CONCLUSION

SAS® Enterprise Miner ™ and SAS® Enterprise Guide™ are powerful tools for analyzing higher education data. Many types of research have shown that combining data mining and machine-learning techniques can be quite handy in many problems in the real world and the attrition problem for the university is no different. This study developed a student attrition model that predicts a student's risk of attrition in the third semester. By using principles of machine learning, this paper has identified the influential variables for student attrition, and Information Research and Information Management can use the probability of student attrition and this information at OSU, Stillwater. By observing the significant variables, we can say that student academic performance, campus involvement, amount they owe to the university, and pre-academic credentials play an influential role as compared to their demographic attributes such as Gender, Age, Race, and Region. The university can also plan to have some counseling sessions or mentor program that can help struggling students regarding their academic performance, participation in college events and make their life comfortable so that they can complete their graduation without dropping out from college.

REFERENCES

Bogard, M., James, C., Helbig, & Huff, G. (2012). *Using SAS® Enterprise BI and SAS ® Enterprise Miner™ to reduce student attrition*. Paper presented at the 2012 SAS Global Forum, Orlando, FL.

Berry, M., & Linoff, G. (2008). Using validation data in Enterprise Miner. *Data Miners Blog.* Retrieved from http://blog.data-miners.com/2008/04/using-validation-data-in-enterprise.html.

Givens, D. (2015). *Kentucky Senate Joint Resolution 106*. Accessed 2017. Retrieved from http://www.lrc.ky.gov/record/15RS/SJ106/bill.pdf.

Payne, B., & Boelscher, S. (2016). *2016-2018 Postsecondary education budget recommendation institutional operating funds.* Accessed January 2017. Retrieved from https://v3.boardbook.org/Public/PublicItemDownload.aspx?ik=37872824.

Shah, Y. (2012). *Use of cutoff and SAS® code nodes in SAS® Enterprise Miner™ to determine appropriate probability cutoff point for decision making with binary target models.* Paper presented at the 2012 SAS Global Forum, Orlando, FL.

Dick, J. (2016, January 28). WKU faces cuts in governor's budget proposal. *The College Heights Herald.* Retrieved from http://wkuherald.com/news/wku-faces-cuts-in-governor-s-budget-proposal/article_14bb48ba-c55d-11e5-9527-dfb8d4276b70.html.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, *53*(3), 950-965.

Maldonado, M., Dean, J., Czika, W., & Haller, S. (2014). *Leveraging Ensemble Models in SAS® Enterprise Miner™.* Paper presented at the 2014 SAS Global Forum, Washington, D.C.

Gönen, M. (2006). Receiver operating characteristic (ROC) curves. *SAS Users Group International (SUGI)*, *31*, 210-231.

Luan, J. (2002). Data mining and its applications in higher education. *New directions for institutional research*, *2002*(113), 17-36.

SAS® Institute Inc. (2011). *Create a gradient boosting model of the data.* Getting started with SAS® Enterprise Miner™ 7.1. Accessed February 2017. Retrieved from https://support.sas.com/documentation/cdl/en/emgsj/64144/HTML/default/viewer.htm#p03iy98sk0c9bvn1r6x7ppx8uj08.htm

Zhang, J., and Mani, I. (2003). "kNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction." In *Proceedings of the Twentieth International Conference on Machine Learning* (ICML-2003), Workshop on Learning from Imbalanced Data Sets II. Palo Alto, CA: AAAI Press.

Zou, H., and Hastie, T. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society B* 67:301–320.

ACKNOWLEDGMENTS

Thanks to Information Research and Information Management of Oklahoma State University, Stillwater for allowing me to work on this data and providing timely inputs. Thanks to my Professors Dr. Goutam Chakraborty and Dr. Miriam McGaugh for tutoring the concepts of Data Mining and Machine Learning.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sivateja Reddy kandula
Graduate Student of Business Analytics (Class of 2019)
Oklahoma State University, Stillwater
+1 (405)334-6904
Sivateja.kandula@okstate.edu
https://www.linkedin.com/in/sivatejak/