Paper 788-2017

Examining Higher Education Performance Metrics with SAS® Enterprise Miner™ and SAS® Visual Analytics™

Taylor Blaetz, M.S., Western Kentucky University; Bowling Green, KY
Tuesdi Helbig, Ph.D., Western Kentucky University; Bowling Green, KY
Gina Huff, Western Kentucky University; Bowling Green, KY
Matt Bogard, Western Kentucky University; Bowling Green, KY

ABSTRACT

Given the proposed budget cuts to higher education in the state of Kentucky, public universities will likely be awarded financial appropriations based on several performance metrics. The purpose of this project was to conceptualize, design, and implement predictive models that addressed two of the state's metrics: six-year graduation rate and fall-to-fall retention for freshmen. The Western Kentucky University (WKU) Office of Institutional Research analyzed five years' worth of data on first time, full time bachelor's degree seeking students. Two predictive models evaluated and scored current students on their likelihood to stay enrolled and their chances of graduating on time. Following an ensemble of machine-learning assessments, the scored data were imported into SAS® Visual Analytics, where interactive reports allowed users to easily identify which students were at a high risk for attrition or at risk of not graduating on time.

Keywords: predictive modeling, higher education, institutional research, retention, graduation, data visualization

INTRODUCTION

During the 2015-2016 General Assembly of the Kentucky State Legislature, Senate Joint Resolution 106 (SJR106) was introduced. If enacted, this joint resolution would require the Kentucky Council on Postsecondary Education (CPE) “to develop a performance-based and outcomes-based funding model for the public postsecondary education institutions as part of its biennial budget request to the Governor and General Assembly” (Givens, 2015). Appropriations for public higher-education institutions would be contingent on performance and outcome-based metrics.

The 2016-2018 CPE operating funds report suggested that approximately $43,388,500.00 be allocated for 2016-2017 performance funding, and approximately $86,737,000.00 for 2017-2018. These funds would “…support a new performance funding approach that provides financial incentives for the postsecondary institutions to accelerate improvement on key student success measures” (Payne & Boelscher, 2016). Per the CPE report, the list of desired performance metrics included:

- Number of degrees and credentials awarded
- First to second year retention rates
- Percentages of students earning at least 30 credit hours per academic year
- Number of students passing the 60-credit hour threshold
- Number of students passing the 90-credit hour threshold
- Number of underprepared students completing credit bearing math courses
- Number of underprepared students completing credit bearing English courses
- Six-year graduation rates at four-year institutions; three-year rate for community colleges
- Number of underrepresented minorities and low-income students

The generation of the performance funding pool required significant budget cuts to public higher-education institutions. At Western Kentucky University (WKU), the operating budget suffered an immediate four-and-a-half percent reduction in state funds as the result of a gubernatorial executive order (Dick, 2016). Although the state judiciary has since overruled the order, a statewide nine percent reduction continues to affect the university budget. The WKU Office of Institutional Research decided to address these budget reductions by analyzing the historical university data with hopes that novel enrollment and graduation trends would reveal themselves. If discovered, new trends would facilitate actionable insight that would increase the likelihood of improving performance on the state metrics, and in turn, increase the likelihood of receiving performance funding in addition to ongoing tuition revenue.

This paper provides a methodological update to Bogard, James, Helbig, and Huff (2012) and addresses two of the aforementioned performance metrics: retention rates and six-year graduation rates. In their paper, Bogard and colleagues developed several predictive models for addressing student attrition and provided descriptions of the various machine-learning techniques used in their models. The authors further discussed the implementation of the predicted results into SAS® Enterprise BI dashboards. Because the WKU Office of Institutional Research has since replaced SAS® Enterprise BI with SAS® Visual Analytics™, the current paper provides updated models based on more recent student data and examples of the integration into SAS® Visual Analytics™.

DATA AGGREGATION

The models discussed in this paper focused on first time, full time bachelor's degree seeking students from a block of fall semesters. The retention-training sample included students from the falls of 2010 through 2015, with a scoring dataset from fall of 2016; the graduation-training sample included students from the falls of 2004 through 2009, with a scoring sample from the fall of 2016. In order to capture a full six-year graduation window, older graduation data were used (i.e., students who started in the fall of 2012 have not yet had a full six-year window). Data sets were compiled using a series of DATA steps and PROC SQL joins and were based on student records. Numerous demographic variables and several academic success variables trained the predictive models. Table 1 provides several examples of the variables used in this project.

Variable Name	Variable Description
DW_SEX	Gender
DW_AGE	Age
RACE	Race
FIRST_GEN	First generation student
ONCAMP	Lives on campus
DW_MAJR_CODE_1	First major code
HI_ACT_COMP	ACT composite score
DW_INST_TGPA_HOURS_EARNED	Credit hours earned first term
DW_INST_TGPA_HOURS_ATTEMPTED	Credit hours attempted first term
DW_INST_TGPA_GPA	First term GPA
HIT_15HOURS_FIRSTTERM	Earned 15 credit hours first term
COUNT_A	Number of courses with attendance problems at fifth week assessment
COUNT_DF	Number of courses with D/F grade at fifth week assessment
ATTEMPTED_EARNED_RATIO	Ratio of earned to attempted first term credit hours

Table 1. Example variables

On several student files, composite ACT and high school GPA variables lacked data. For ACT scores, an SAT to ACT conversion populated the missing values. For students lacking both an ACT and SAT score, a SQL macro variable calculated the average of all non-missing values and populated the missing data using an IF-THEN statement. By replacing the ACT variable with high school GPA, the code would also populate missing high school GPAs:

```
/*CAN SUBSTITUTE HS_GPA FOR ALL ACT VARIABLES*/

PROC SQL NOPRINT;
        SELECT AVG(HI_ACT_COMP)
        INTO: AVERAGE_ACT
        FROM S22_ACT;
QUIT;

%PUT &AVERAGE_ACT.;

DATA S23_FULL_ACT;
        SET S22_ACT;
        IF HI_ACT_COMP = . THEN HI_ACT_COMP = &AVERAGE_ACT.;
RUN;
```

Although PROC HPIMPUTE in Base SAS® or the imputation node in SAS® Enterprise Miner™ could be used to generate missing values, the above code provided an additional programmatic approach based on the average value.

BUILDING PREDICTIVE MODELS USING SAS® ENTERPRISE MINER™

For this project, SAS® Enterprise Miner™ read in a training set and a scoring set, and included two predictive models: a fall-to-fall retention model and a six-year graduation rate model. Both models implemented the S.E.M.M.A. approach (i.e., sample, explore, modify, model, assess), and predicted a binary target variable. The models in this paper successfully predicted their respective targets.

FALL-TO-FALL RETENTION MODEL

The retention target in the sample data was populated with either a “Y” or an “N” based on whether the student retained to the second fall, however the retention target variable was absent from the scoring dataset. Exploration of the data included the use of three nodes: Stat Explore, Graph Explore, and Multi Plot.

The Stat Explore node provided several helpful tables. Of most interest was the output window, which provided descriptive statistics for the training data. Provided were a frequency count of the binary target variable (see Figure 2) and central tendency, dispersion, and distribution statistics for an assortment of interval variables (see Figure 3). These statistics provided a basic summary of the data prior to constructing the model diagram.

The Graph Explore node generated a variety of graphs based on different variables and the Multi Plot node provided visualizations of the interactions between input variables and the binary target (see Figure 4).

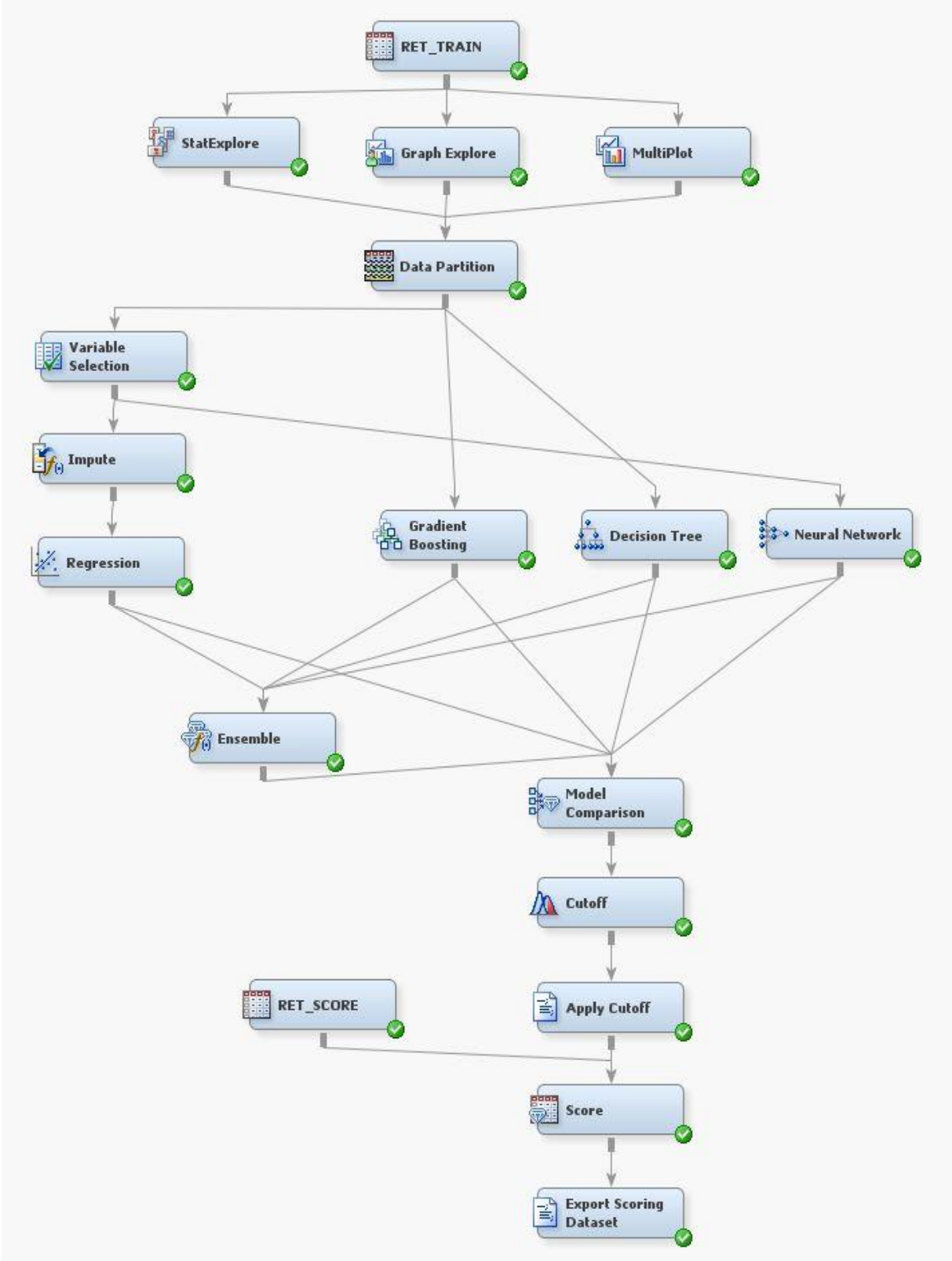

Figure 1. Screen capture of the retention model

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	RETAINED_1YR	TARGET	Y	11513	71.8933
TRAIN	RETAINED_1YR	TARGET	N	4501	28.1067

Figure 2. Frequency count for the binary target variable

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
AGE	INPUT	18.44817	1.631867	16014	0	15	18	66	10.29076	160.0141
ATTEMPTED_HOURS_FIRSTTERM	INPUT	14.58271	1.714747	15995	19	1	15	26	-1.44488	7.042781
EARNEDHOURS_AFTER_FIRSTTERM	INPUT	12.104	4.420623	15995	19	0	13	21	-1.43015	1.275549
GPA_AFTER_FIRSTTERM	INPUT	2.774603	1.087372	15790	224	0	3	4	-1.05252	0.366124
HI_ACT_COMP	INPUT	23.02047	4.052401	16014	0	11	23	36	0.375863	-0.36664
HS_GPA	INPUT	3.313705	0.533016	15788	226	1.22	3.39	4	-0.62595	-0.30788
OTHER_CREDIT_HRS	INPUT	3.724128	8.41447	16014	0	0	0	152	4.214887	29.26248
WKU_DUAL_CREDIT_HRS	INPUT	1.024728	3.140561	16014	0	0	0	31	3.948004	17.96427

Figure 3. Descriptive statistics for variable defined as interval

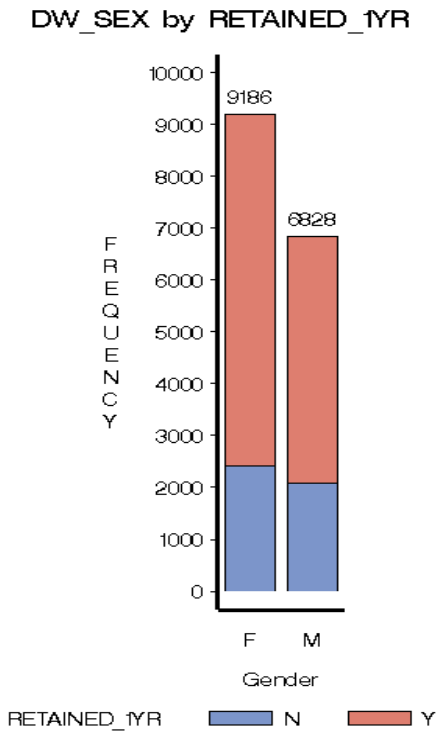

Figure 4. Multi Plot node output

The modification component of S.E.M.M.A started after exploring the data. The data were partitioned using the data partition node, which split the sample into a 60% training partition and a 40% validation partition. Using the variable selection node, an R^2 assessment determined variable inputs for two of the modeling nodes (see Figure 6). The imputation node populated values when more than an adjustable percentage of the data were missing. This model used the default 50% missing value cutoff in order to implement data imputation.

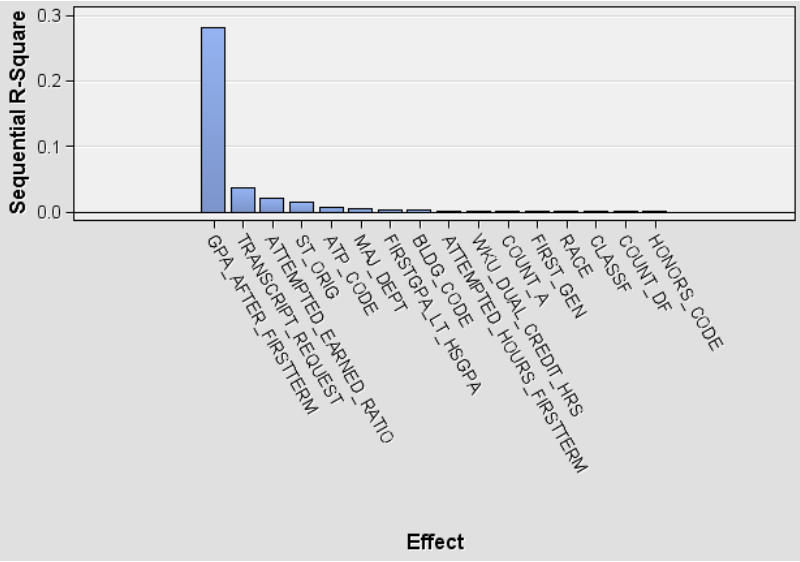

Figure 6. R^2 values for input variables

Default settings on the modeling nodes provided the basis for an easy-to-implement predictive model; however, some adjustments were made. In this model, the binary retention prediction included five algorithms: a stepwise logistic regression, gradient boosting, a decision tree, a neural network, and an ensemble method. The stepwise logistic regression used validation error as the selection criterion, which minimized error and overfitting (Berry & Linoff, 2008). The ensemble node generated an average of the posterior probabilities based on the input models, which provided stronger prediction and classification of the target variable (Maldonado, Dean, Czika, & Haller, 2014).

Assessment of this model included the model comparison node and the cutoff node. The model comparison node was set to generate a receiver-operating characteristics curve (ROC) and calculate misclassification rates for each algorithm, assessing each one and determining the strongest method. The cutoff node examined the predicted probabilities for the chosen algorithm and adjusted the decision classification based on either a user-input value or a user-determined method (see Figure 7).

Score	
Cutoff Method	Event Precision Equal Recall
Cutoff User Input	0.5

Figure 7. Cutoff input methods

The cutoff node analyzed the “balance between true positive and false positive predictions” (Shah, 2012). The retention model used the event precision equals event recall method, which applied a cutoff based on the best ratio of true positive and true negative rates (see Figure 8).

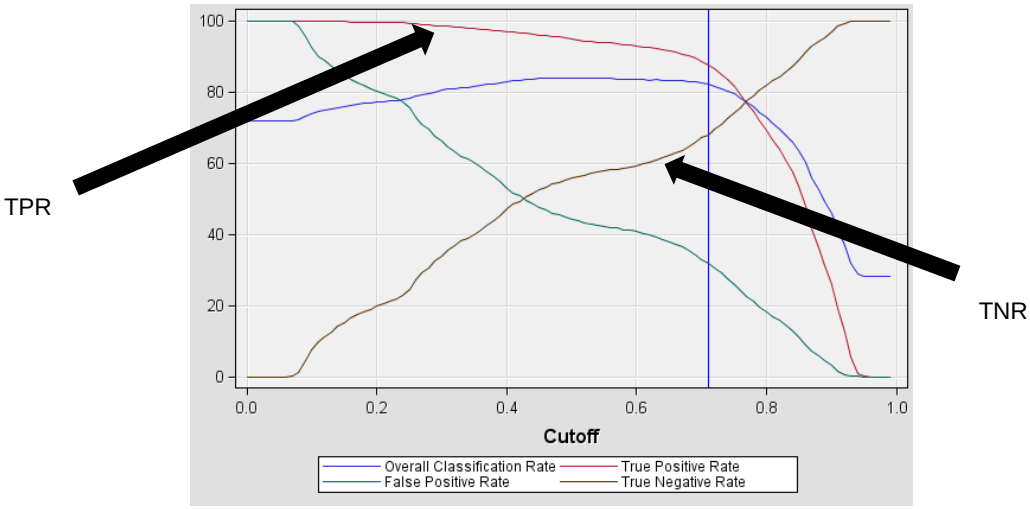

Figure 8. Probability cutoff at ~ .72

The utility tab atop the SAS® Enterprise Miner™ screen provides a SAS® Code node, which allows the user to apply the determined cutoff decision programmatically. Application of the cutoff probability involved writing the following line of code into the score code tab of the editor (see Figure 9):

I_<INSERT TARGET VARIABLE> = EM_CUTOFF;

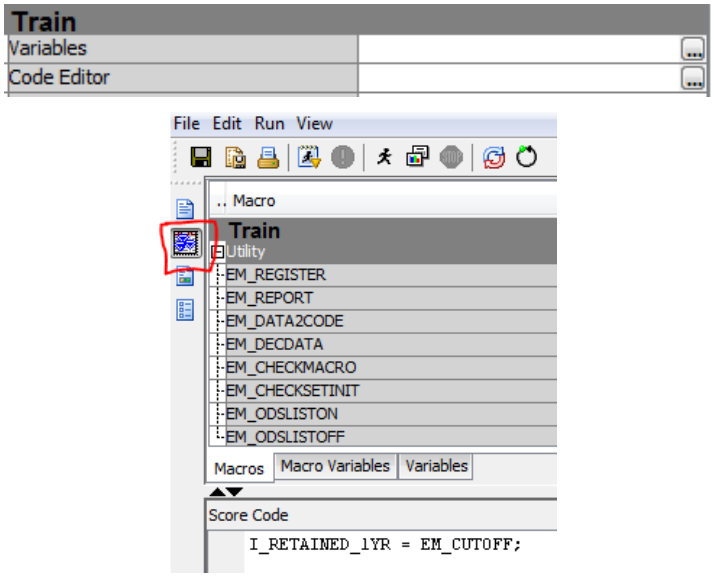

Figure 9. Code input window

SIX-YEAR GRADUATION RATE MODEL

This model applied the same approach as the retention model and predicted the binary target variable "TARGET_GRAD_SIX." However, this model implemented an additional modification technique (see Figure 10). The target variable was either a "Y" or "N" variable based on whether the student graduated within six years of their starting term. Exploration of this sample included the use of both Graph Explore and Stat Explore.

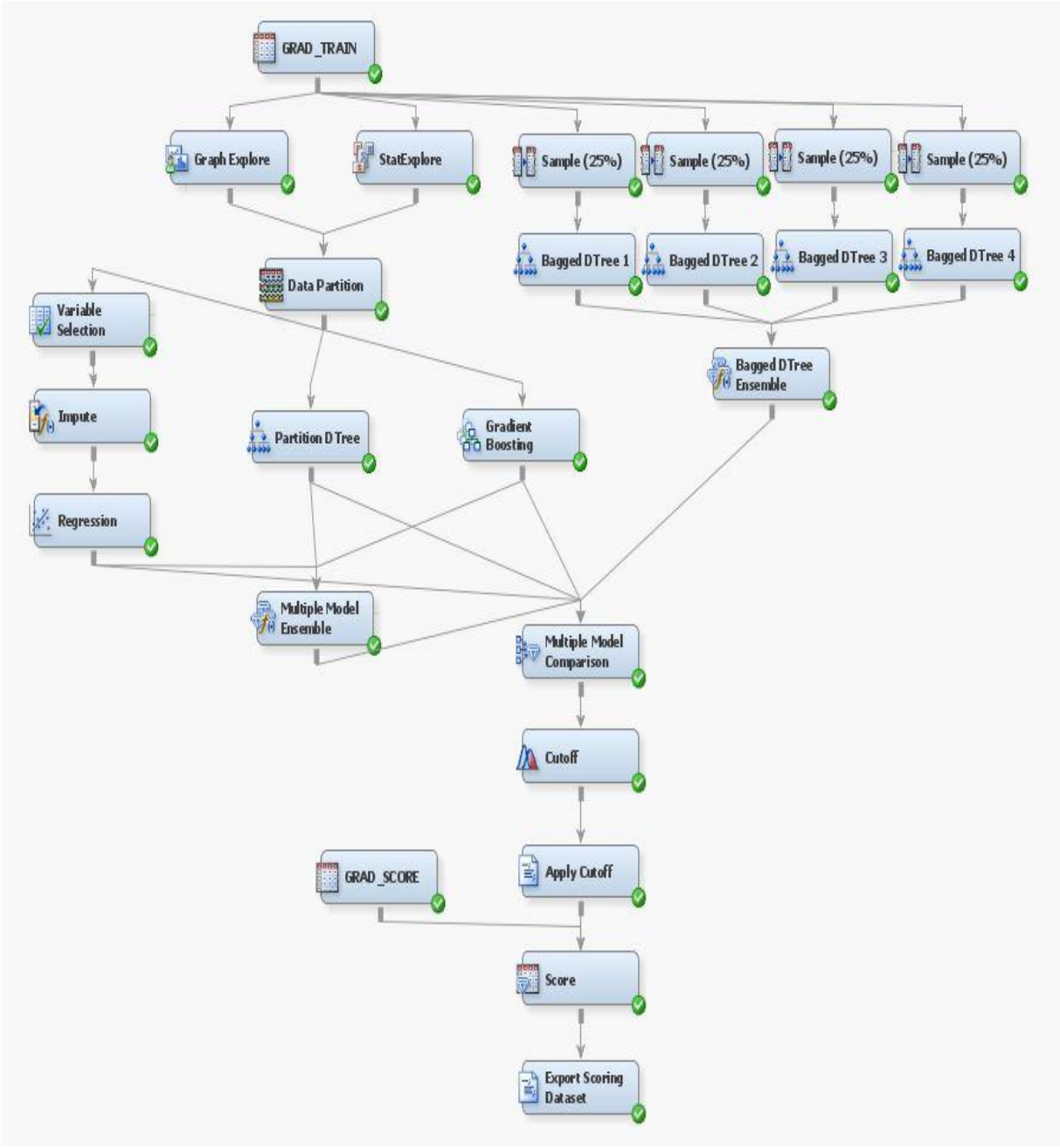

Figure 10. Screen capture of the graduation model

The graduation model introduced a new modification technique: bootstrap aggregating (or “bagging”). The bagging method trained multiple samples and then assessed the average predicted probabilities of those samples (for a summary of this method, see Maldonado et al., 2014). The bagging approach divided the training sample into four 25% samples, each generated with a different random seed (see Table 2). In addition to bagging, data were also partitioned identical to the retention model (i.e., a 60% training partition and a 40% validation partition). The variable selection node fed to the imputation node, followed by the logistic regression node. The imputation parameters were identical to the retention model.

Sample	Random Seed
1 (25%)	12345
2 (25%)	13456
3 (25%)	14965
4 (25%)	17324

Table 2. Random seeds for bagging samples

For the bagging component of this model, default decision trees were run on the validation table of each sample and the ensemble node calculated the average predictive possibilities based on the results of the four decision trees. For the standard partitioning component of this model, a default logistic regression, decision tree, and gradient boost were used. The bagged and multi-model ensemble nodes calculated the posterior probabilities, which fed into the model comparison containing the other algorithm nodes.

PREDICTIVE MODEL RESULTS

Identical to before, model assessment included the comparison and cutoff nodes. The model comparison node assessed the strongest model using an assortment of metrics. The cutoff node used the same event precision equals recall parameter, which the utility code node applied to the scoring dataset. Of specific interest were both the ROC curves and indices (see Figure 11), and misclassification rates.

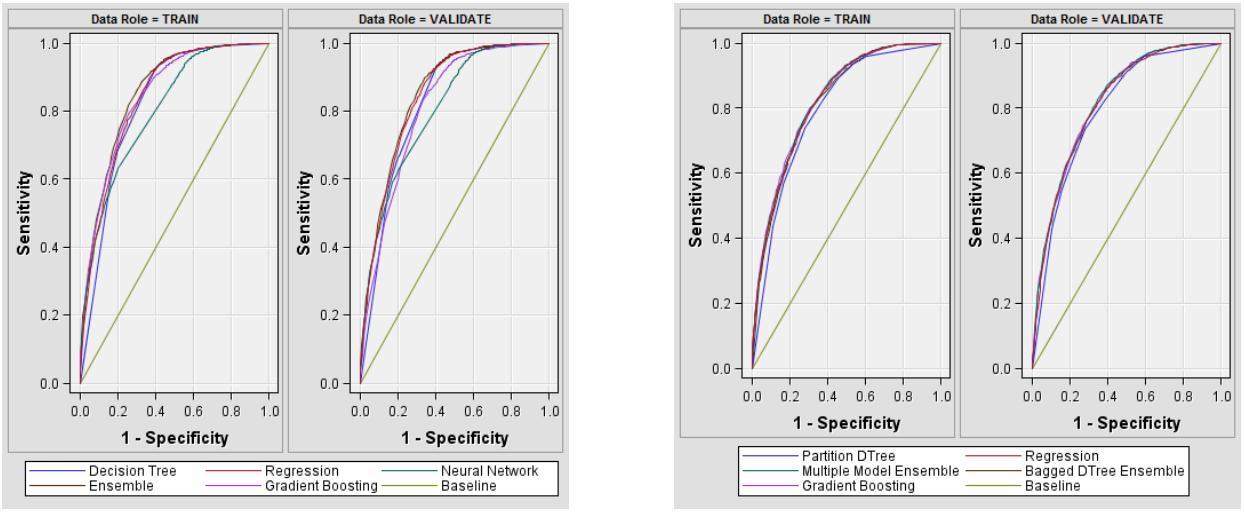

Figure 11. Model ROC curves for retention (left) and graduation (right) targets

The output of the model comparison node provided a ranking of the predictive models (see Figure 12). For the retention model, the ensemble approach was the strongest method and gradient boosting was the best method for the graduation model. According to SAS® documentation, gradient boosting "is an approach that resamples the analysis data several times to generate results that form a weighted average of the resampled data set." Further, "boosting makes no assumptions about the distribution of the data…and is less prone to over fit the data than a single decision tree."

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Roc Index
Y	Ensmbl	Ensmbl	Ensemble	RETAINED_1YR		0.847
	Reg	Reg	Regression	RETAINED_1YR		0.842
	Tree	Tree	Decision Tree	RETAINED_1YR		0.821
	Boost	Boost	Gradient Boosting	RETAINED_1YR		0.809
	Neural	Neural	Neural Network	RETAINED_1YR		0.801

Figure 12. Model comparison output for the RETAINED_1YR model

According to Gönen (2006), ROC curves and misclassification rates provide simple methods "…for assessing the accuracy of predictions." ROC curves visualize the relationship between the true and false positive rates, whereas misclassification rates (MR) provide the ratio of false negatives and positives to the sum of all possible outcomes (i.e., TP, FN, FP, TN). Simply put, the model accuracy can be calculated using one minus MR.

For both models, the ROC curves had roughly the same area under the curve (AUC), represented by the ROC index. Table 3 summarizes the results. Interestingly, the six-year graduation model appeared to be slightly less accurate than the one-year retention model. However, this discrepancy could be due to the graduation model encapsulating a much longer duration of time than the one-year retention model.

Target Variable	Best Model	ROC Index / AUC	Misclassification Rate
RETAINED_1YR	Ensemble	.847	.16 (~84% predicted accuracy)
TARGET_GRAD_SIX	Ensemble	.815	.24 (~76% predicted accuracy)

Table 3. Summary of predictive models

APPLYING THE PREDICTED RESULTS USING SCORING

Scoring the new dataset based on training data provided the predicted results. This step required (1) the score node, which was found under the assess tab, and (2) the dataset that was to be scored. When the score code executed, several new variables were created. Out of the new variables, this project focused on two: EM_CLASSIFICATION and EM_EVENTPROBABILITY. The classification variable represented the predicted binary value for the target and was populated with either a “0” or a “1” (i.e., 0 referred to a “No” and 1 referred to “Yes”). The EM_EVENTPROBABILITY variable defined the probability of obtaining a “1” classification for the target variable. Exporting the scored data involved the use of a second utility code node and an export assignment to a SAS® library.

The scored data examination occurred in Base SAS® and used PROC FREQ to display the prediction percentages and PROC SGPLOT to view a histogram of the predicted probabilities. Simple histograms helped to visualize probability distributions and determine thresholds. Tables 4 and 5 display the PROC FREQ prediction results for each target respectively.

Prediction for RETAINED_1YR		
EM_CLASSIFICATION	**Frequency**	**Percent**
N	1025	35.04
Y	1900	64.96

Table 4. Binary target prediction for RETAINED_1YR; historical WKU retention is ~72%

Prediction for TARGET_GRAD_SIX		
EM_CLASSIFICATION	**Frequency**	**Percent**
N	1507	51.52
Y	1418	48.48

Table 5. Binary target prediction for TARGET_GRAD_SIX; historical WKU six-year rate is ~50%

The probability distributions showed clusters that allowed for risk category assignment. This project used the following thresholds and code to assign risk:

```
/*RETENTION*/

DATA SASGF.VA;
        SET SASGF.RET_SCORED;
        LENGTH RISK $ 25;
        IF EM_EVENTPROBABILITY GE .87 THEN RISK = "VERY LOW";
        IF .72 LE EM_EVENTPROBABILITY LT .87 THEN RISK = "LOW";
        IF .48 LE EM_EVENTPROBABILITY LT .72 THEN RISK = “MODERATE";
        IF .24 LE EM_EVENTPROBABILITY LT .48 THEN RISK = “HIGH";
        IF EM_EVENTPROBABILITY LE .24 THEN RISK = "VERY HIGH";
RUN;
```

```
/*GRADUATION*/

DATA SASGF.VA2;
        SET SASGF.GRAD_SCORED;
        LENGTH RISK $ 25;
        IF EM_EVENTPROBABILITY GE .76 THEN RISK = "VERY LOW";
        IF .60 LE EM_EVENTPROBABILITY LT .76 THEN RISK = "LOW";
        IF .40 LE EM_EVENTPROBABILITY LT .60 THEN RISK = "MODERATE";
        IF .20 LE EM_EVENTPROBABILITY LT .40 THEN RISK = "HIGH";
        IF EM_EVENTPROBABILITY LE .20 THEN RISK = "VERY HIGH";
RUN;
```

DATA VISUALIZATION USING SAS® VISUAL ANALYTICS™

Visualizing data is more effective than mindlessly staring at a raw data file. These reports provided the end-user with several options: the ability to view overall prediction percentages, the ability to sort by college and risk category, and the ability to access and export raw student information for contact purposes. Figure 12 shows the basic dashboard for the RETAINED_1YR prediction, which was composed of a risk category bar chart, an overall prediction percentage pie chart, and a student information table. The raw table provided the prediction (i.e., 1 for retained and 0 for not retained), and data for the key variables identified in the model.

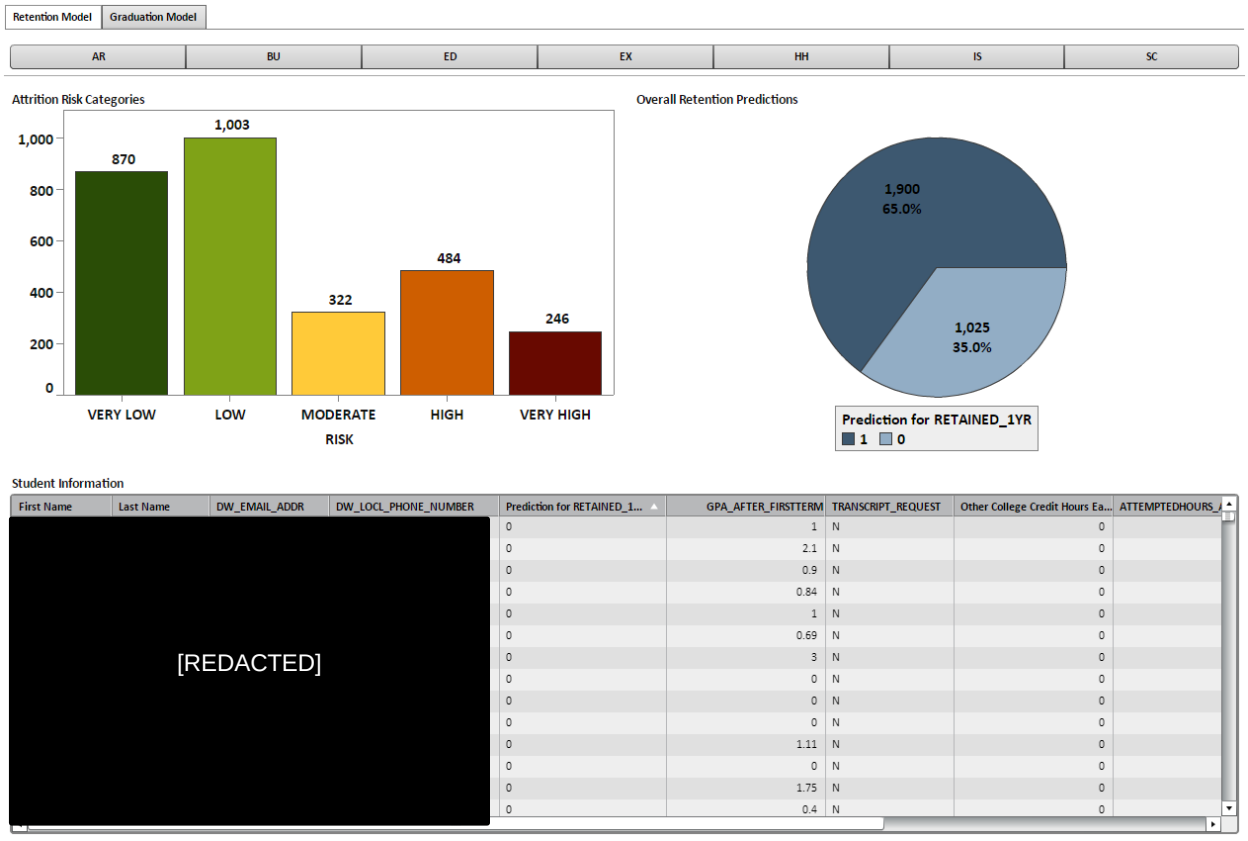

Figure 12. Retention dashboard

The SAS® Visual Analytics ™ reports provided users with the ability to examine interactions between report components. As Figure 13 shows, the selection of the College of Science and Engineering (i.e., SC) changed the pie and risk category charts, and adjusted the raw data table to reflect students enrolled in the specific college and at the chosen risk level. The report showed that 54 students in this college had a very high chance of not retaining to the next fall.

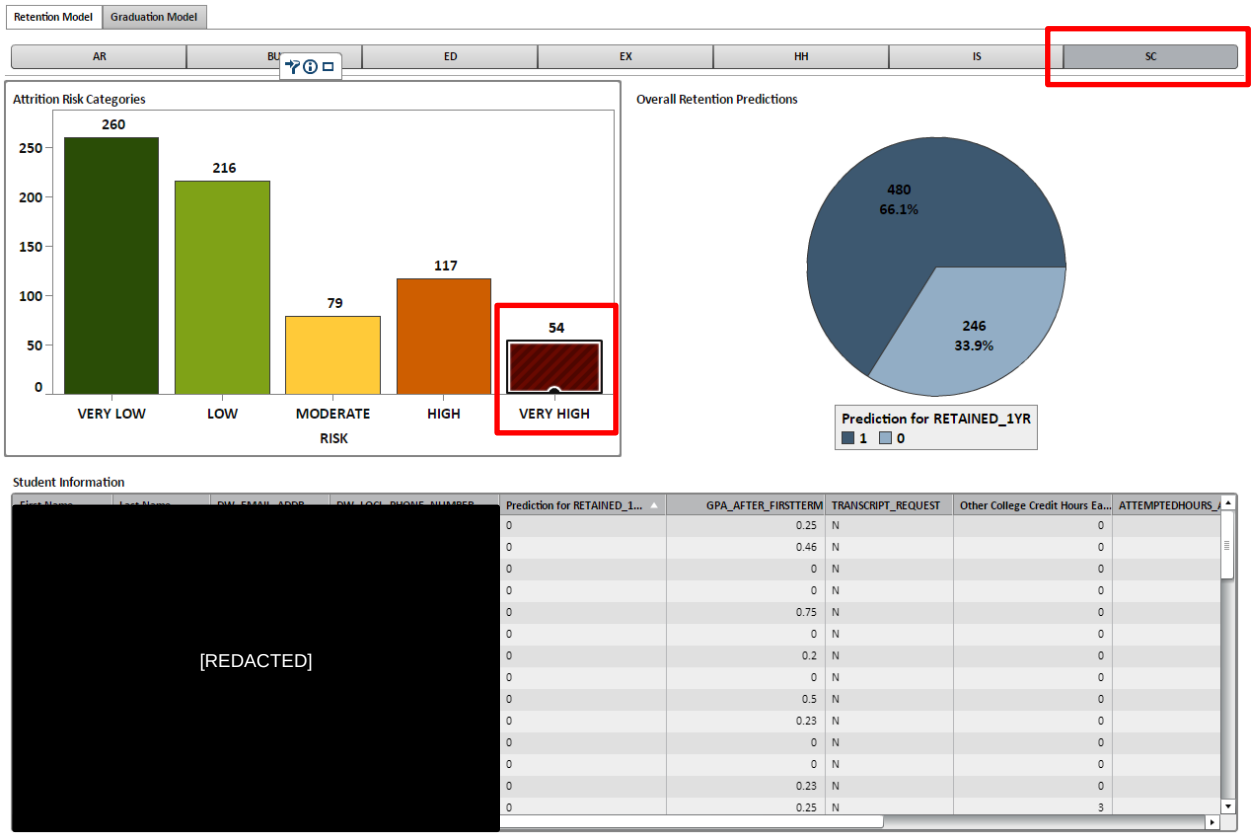

Figure 13. College and risk options.

The list table included important variables identified in the models. After viewing the report, the end user could then export data by right clicking the list table and selecting “export <table name>”; in this case “Student Information” (see Figure 14).

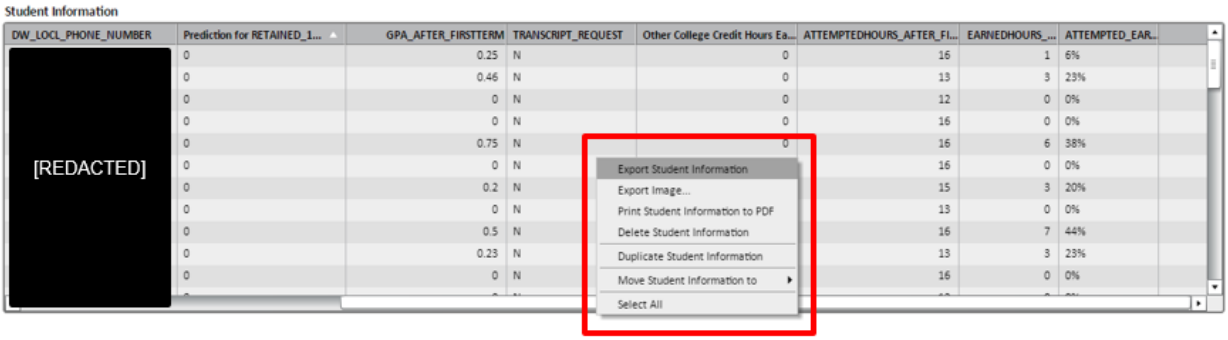

Figure 14. Exporting student information

CONCLUSION

SAS® Enterprise Miner ™ and SAS® Visual Analytics™ are powerful tools for analyzing higher education data. Previous research has shown that data mining and machine-learning techniques can be very useful in the realm of higher education (Luan, 2002; Lykourentzou, Giannoukos, Nikolopoulos, Mpardis, & Loumos, 2009). By using these principles and machine-learning techniques, the WKU Office of Institutional Research identified important variables that predicted both retention and graduation. More directly, the techniques used in this paper were implemented in order to identify at-risk students.

SAS® Visual Analytics™ has been heavily used at Western Kentucky University. An internal usage report showed that between January 1, 2016 and January 1, 2017, the WKU Director of Enrollment Management was the most frequent user of SAS® Visual Analytics™. The predicted data from the models in this paper could provide the perfect tool for university administrators. Similarly, academic advisors could use these reports to identify varying levels of at-risk students and allocate their resources appropriately.

Given the upcoming importance of performance funding in Kentucky, these tools and reports will prove to be highly effective for increasing the fall-to-fall retention and six-year graduation rates.

REFERENCES

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.

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.

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.

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

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.

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.

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.

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

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.

Spalding, A. (2016). *Performance funding in Kentucky should promote successful outcomes for all students.* Accessed January 2017. Retrieved from http://kypolicy.org/performance-funding-kentucky-promote-successful-outcomes-students/.

ACKNOWLEDGMENTS

I would like to acknowledge Matt Bogard, who was the first author on my primary reference paper. As my paper was designed as an update to his, I strived to remain consistent in both theoretical and methodological ideologies.

CONTACT INFORMATION

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

Taylor Blaetz, M.S.
Research Analyst
Western Kentucky University
taylor.blaetz@wku.edu

Tuesdi Helbig, Ph.D.
Director of Institutional Research
Western Kentucky University
tuesdi.helbig@wku.edu

Gina Huff
Senior Applications Programmer Analyst
Western Kentucky University
gina.huff@wku.edu

Matt Bogard
Adjunct, Department of Economics
Western Kentucky University
matt.bogard2@gmail.com