Paper 044-2013

A Data Driven Analytic Strategy for Increasing Yield and Retention at Western Kentucky University Using SAS Enterprise BI and SAS Enterprise Miner

Matt Bogard, Western Kentucky University, Bowling Green, KY

ABSTRACT

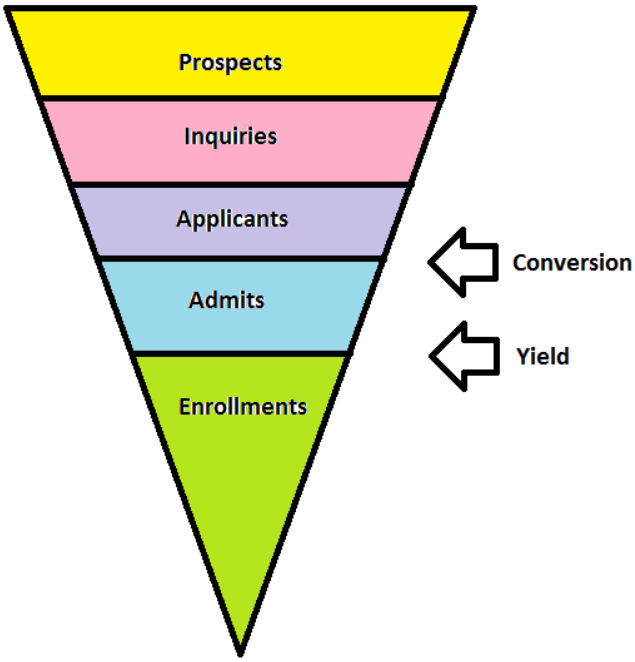
As many Universities face the constraints of declining enrollment demographics, pressure from state governments for increased student success, as well as declining revenues, the costs of utilizing anecdotal evidence and intuition based on 'gut' feelings to make time and resource allocation decisions become significant. However, grasping advanced statistical methods and analytics for data driven decision making can be overwhelming to some staff making buy in difficult. This paper describes how we are using SAS® Enterprise Miner to develop a model to score university students based on their probability of enrollment and retention early in the enrollment funnel so that staff and administrators can work to recruit students that not only have an average or better chance of enrolling but also succeeding once they enroll. Incorporating these results into SAS® EBI will allow us to deliver easy-to-understand results to university personnel.

INTRODUCTION

As many Universities face the constraints of declining enrollment demographics, pressure from state governments for increased student success, as well as declining revenues, the costs of utilizing anecdotal evidence and intuition based on 'gut' feelings to make time and resource allocation decisions become significant. Each year at WKU less than 50% of admitted students actually enroll. Of those that enroll, only 75% retain after 1 year. One way to improve retention may be to have more stringent admissions standards that admit only students with the highest probability of graduation. However, social costs and competitive pressures make this solution nearly impractical. With a limited budget and other resource constraints, how can you best utilize your resources to recruit students that have an average or better chance of enrolling that also have the best chance of succeeding once they enroll?

ISSUES TO ADDRESS

In higher education, students' enrollment decisions are often visualized in terms of an enrollment funnel.



Conversion is the percentage applicants that are admitted. Yield is the percentage of admits that actually enroll. As previously mentioned, our yield rate is only about 50%. Less than 75% of our yielding students are still enrolled one year later and only about 50% actually graduate. Previously at WKU we have utilized data at the 5th week of enrollment and SAS® Enterprise Miner to identify students at risk, and have delivered these analytics to decision makers using SAS® EBI. (Bogard, James, Helbig & Huff, 2012). However, until now, we have not developed analytical tools to address admissions and enrollment.

The first point in the enrollment funnel for which we have sufficient data populated in our data warehouse related to student characteristics is at the applicant stage. The question becomes, are there opportunities at the applicant stage to improve our yield, implement cost savings, and shape our freshmen class to maximize retention? Is there a way of knowing which applicants are most likely to enroll and succeed?

A challenge we face with applicant data is that it is transactional and accumulates over time. We don't have all of the applicants for the entire cohort until almost the start of term. However, by February we typically have as much as 80% of our applicant data. It is also important to emphasize the transactional nature of the data implies that variable values that appear on the application in February may change by the start of term. (admissions decisions, flags for financial aid awards, etc. are applied as the applications are processed by various offices across campus at different times). Without snapped data files that are warehoused for each point in time, important information can be overwritten. In 2008 we began warehousing our applicant files on a weekly basis. This is valuable because it enables us to build a model based on historical data actually available in February (or any point in time) as opposed to the data that's housed in our live warehouse files (which for past years would reflect the latest instance of that record). February seems like a great starting point for model development because it is

early enough in the recruiting season to still influence applied and admitted students' decision to enroll.

MODELING STUDENT ENROLLMENT AND PERSISTENCE: METHODS FROM THE LITERATURE

The vast majority of the literature related to the empirical estimation of retention models includes a discussion of the theoretical retention framework established by authors such as Bean (1980), Braxton (2000), Braxton, Hirschy, & McClendon (2004), Chapman & Pascarella (1983), Pascarella &Terenzini (1978) and Tinto (1975). Most empirical models have utilized some form of logistic regression to predict retention (Herzog, 2005; Miller, 2007; Miller and Herried,2008; Ronco and Cahill, 2006; Stratton et al 2008; Dey and Astin, 1993). Most empirical models for the decision to enroll are also based on logistic regression.(Goenner and Pauls,2006;DesJardins,2002;Curs and Singell,2002;Thomas, Dawes, & Reznik,2001;. Bruggink and Gambhir,1996). Literature indicates that data mining or algorithmic approaches to prediction can provide superior results vis-à-vis traditional statistical modeling approaches (Delen, Walker, & Kadam, 2004; Delen, Sharda, & Kumar, 2007; Kiang, 2003; Li, Nsofor, & Song, 2009). Because we are interested in making accurate predictions for decision support, as opposed to obtaining specific parameter estimates to make inferences, algorithmic approaches are preferred if they outperform traditional approaches of statistical inference in terms of predictive accuracy and generalization error (Brieman,2001). Previously at WKU we utilized algorithmic approaches to predict student retention at the 5th week using a decision tree implemented in SAS Enterprise Miner® (Bogard, James, Helbig & Huff, 2012).

DATA AND METHODS

Decision to Enroll Model

Starting in the fall of 2008 we began warehousing our applicant data files on a weekly basis. These weekly data snaps have provided a rich data source for decision support and predictive modeling. Utilizing data from the first applicant snapshot in February for 2009-2010 fall applicants we developed a model for student enrollment using a decision tree. Ultimately our goal was to strike a balance between developing powerful analytics that would ultimately be acceptable and interpretable to our end users. Decision trees provide a very clear picture of the relationships between variables and the target. Our end-users typically are not interested in interpreting regression co-efficeints or odds ratios associated with methods like logistic regression. We implemented our decision trees using SAS® Enterprise Miner, with optimization based on average square error, limiting leaf size to 50 and the number of rules to 3 to improve interpretation.

Retention Model

With regard to retention, our concern was more with predictive accuracy than model clarity because in our previous work we have already characterized the determinants of retention using decision trees (Bogard, James, Helbig & Huff, 2012). In this endeavor, we compared the results of four different algorithms using SAS® Enterprise Miner.

Neural Networks

SAS® Enterprise Miner's default settings for a multilayer perceptron architecture were utilized. Neural networks are complex nonlinear models composted of multiple hidden layers. Multilayer perceptrons can be thought of as a weighted average of logits (Kennedy, 2003).

Decision Trees

The decision tree algorithm we implemented (using SAS® Enterprise Miner's - similar to CART and CHAID) searches through the input space and finds values of the input variables (split values) that maximize the differences in the target value between groups created by the split. The final model is characterized by the split values for each explanatory variable and creates a set of rules for classifying new cases.

Gradient Boosting

The gradient boosting algorithm used by SAS® Enterprise Miner is based on the stochastic gradient boosting algorithm developed by Freidman (Friedman,2001). Boosting algorithms are ensemble methods that make predictions based on the average results of a series of weak learners. Gradient boosting involves fitting a series of trees, with each successive tree being fit to a resampled training set that is weighted according to the classification accuracy of the previously fit tree. The original training data is resampled several times and the combined series of trees form a single predictive model. This differs from other ensemble methods using trees, such as random forests. Random forests are a modified type of bootstrap aggregation or bagging estimator (Freidman et al,2009). With random forests, we get a predictor that is an average of a series of trees grown on a bootstrap sample of the training data with only a random subset of the available inputs from the training data used to fit each tree (De Ville, 2006). SAS® Enterprise Miner does not have an implementation of Random forests, however gradient boosting can perform similarly to random forests and boosting may tend to dominate bagging methods on most problems. (Freidman et al,2009).

Double Scoring

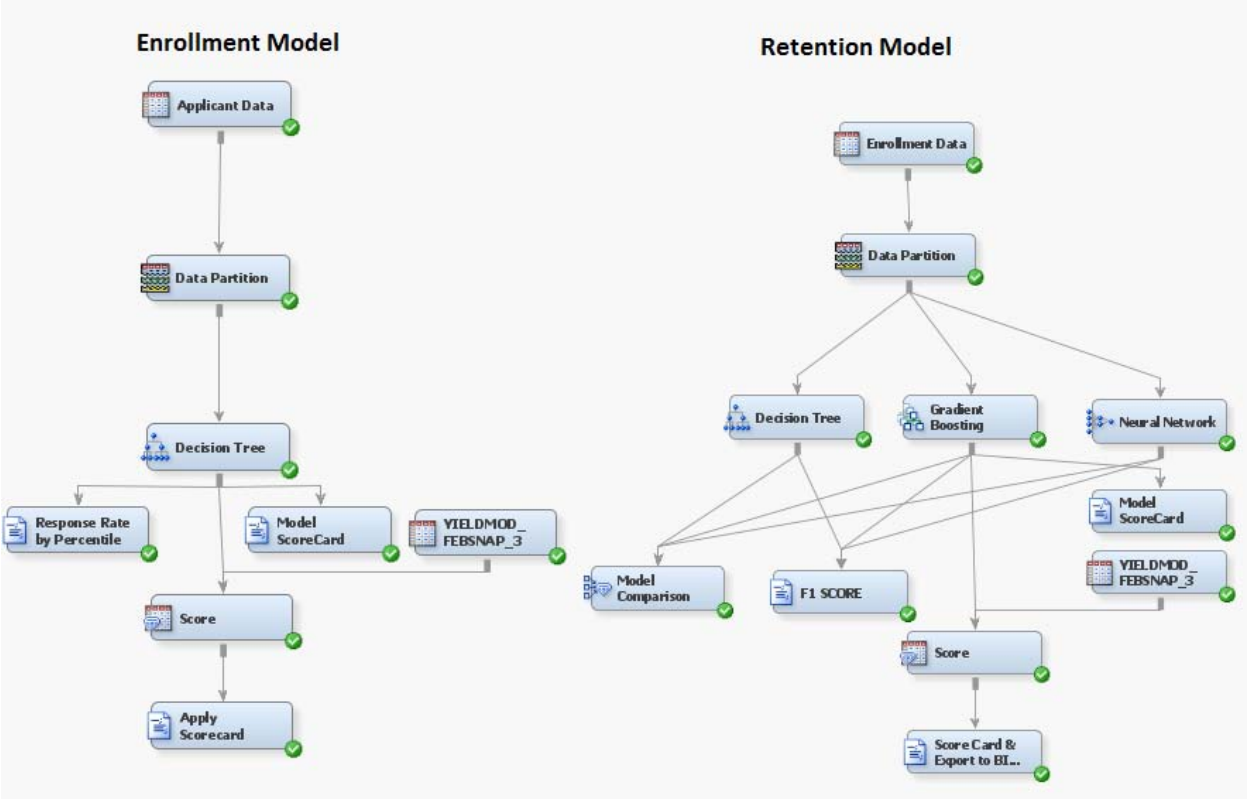
The models are implemented in SAS® Enterprise Miner as follows. First we fit the model for the decision to enroll and score our 'new student applicant' data set for probability of enrollment. Next we fit the models for retention based only on the data available at the time of application. The best models are chosen from each stage and new applicants are scored using both models.

Model Performance

Models were compared based on the area under the ROC curve. This area is equivalent to the probability that our model will correctly rank a randomly chosen training example that chooses to enroll (or retain in the case of our retention model) higher than a randomly chosen example that does not enroll (or does not retain). Alternatively the ROC curve can be interpreted to measure the tradeoff between true positives and false positives (Provost,1998). This method is used increasingly in the machine learning community and is preferred over other measures of fit like precision or the F1-Score because it evaluates model performance across all considered

cutoff values vs. an arbitrarily chosen cutoff (Caruana and Niculescu-Mizil,2006). This is important to us because we are not interested in identifying a single cutoff to make a classification. We are interested in the entire distribution of predicted probabilities, and want to create a score card that steps across the distribution and groups students into relative categories related to their respective probabilities of enrolling and retaining.

Figure 1 - SAS® Enterprise Miner Process Flow



RESULTS

You can find the results in terms of model performance based on various cutoffs below.

Enrollment Model Results

Model	Cutoff	Precision	Recall	F1-Score	ROC
Decision Tree	0.01	0.4	1	0.57	0.702
Decision Tree	0.23	0.48	0.88	0.62	0.702
Decision Tree	0.58	0.65	0.38	0.49	0.702
Decision Tree	0.66	0.78	0.14	0.23	0.702

Retention Model Results

Model	Cutoff	Precision	Recall	F1-Score	ROC
Decision Tree	0.74	0.83	0.68	0.75	0.675
Neural Network	0.74	0.75	0.78	0.76	0.6
Gradient Boosting	0.74	0.83	0.62	0.71	0.698

Based on these results the best algorithm for predicting retention utilizing the area under the ROC curve is gradient boosting. We would like this metric to be higher, but the true test of our model will be its ability to let us segregate our student populations based on their probabilities and improve the efficiency of resource allocation. An imperfect model may still outperform adhoc guesses based on gut feelings. After scoring students with each of these models, we developed a scorecard that segments them by risk category. For enrollment predictions, students were classified into four categories: Least Likely, Unlikely, Average, and Most Likely. For retention, students were classified (from most likely to drop out to least likely) into four different categories: Double Red, Red, Yellow, Green. Based on actual historical enrollment and retention data, it turns out that these classifications do a good job discriminating between the groups of students.

Probability of Enrollment	Actual Enrollment
Least Likely	0.07792
Unlikely	0.23324
Average	0.39263
Most Likely	0.65362

Probability of Retention	Actual Retention
Double Red	0.48039
Red	0.65
Yellow	0.72303
Green	0.90511

MODEL IMPLEMENTATION

After deriving the models and scoring students, recruitment efforts can then be prioritized based on each student's likelihood of enrolling and retaining. For instance, we may want to reduce the amount of resources expended on recruiting the students highlighted in red below and concentrate efforts on those students depicted in green for which we can make the greatest marginal difference in terms of enrollment and retention. A unique strategy can be tailored for each type of student classified into each cell based on the model results.

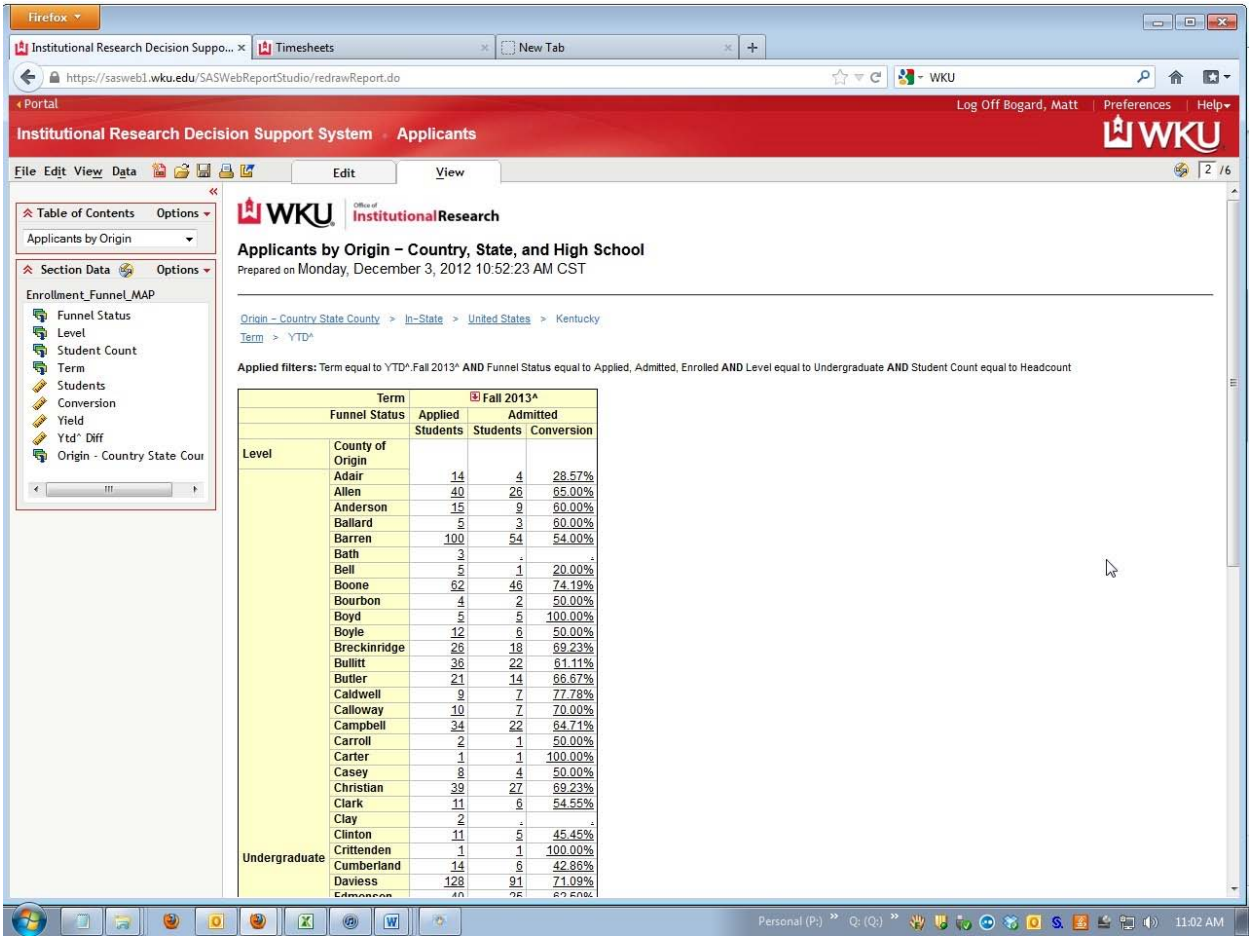
Propensity to Enroll	Attrition Risk				All
	Green	Yellow	Red	Double Red	
Most Likely	668	441	126	75	1310
Average	488	1056	591	502	2637
Unlikely	220	525	149	57	951
Least Likely	89	197	79	37	402
All	1465	2219	945	671	5300

THE POWER OF BUSINESS INTELLIGENCE

Using the score code generated by SAS® Enterprise Miner, these analytics can be incorporated into the SAS® EBI production environment via an ETL process in SAS® Data Integration Studio (see Bogard, James, Helbig & Huff, 2012) . This will allow administrators, faculty, and professional staff at the institution to easily incorporate advanced analytics into their recruitment and retention strategies on a regular basis, without having to rely on manually generated lists of at-risk students, or less precise ad hoc reports generated solely on the basis of intuition. Previously WKU has implemented an online decision support system using SAS® EBI that includes flexible drill down reporting and dashboards incorporating analytics for student retention (for currently enrolled students). Figure 2 is a snapshot displaying student applicants by county of origin currently accessible to University personnel through our Decision Support System implemented using SAS® EBI. As depicted in figure 3, with the new analytics presented in this paper, university personnel can identify applicants at risk for not retaining on a county by county basis (or whatever criteria meets their needs). In addition they will also be able to assess the probability of enrollment based on the indicators mentioned above.

To dynamically implement the program generator and score code into the SAS® EBI production environment, an ETL will be created in SAS® Data Integration Studio including both components. This ETL will be submitted in batch every Monday morning before the open of business allowing the updated list of students to be scored for the week. The SAS® data set generated by the ETL can then be used to create an OLAP cube with student level information. This OLAP cube consists of variables necessary for the statistical model and other demographic and academic variables useful to the WKU user community.

Figure 2: Applicant Report in SAS® EBI

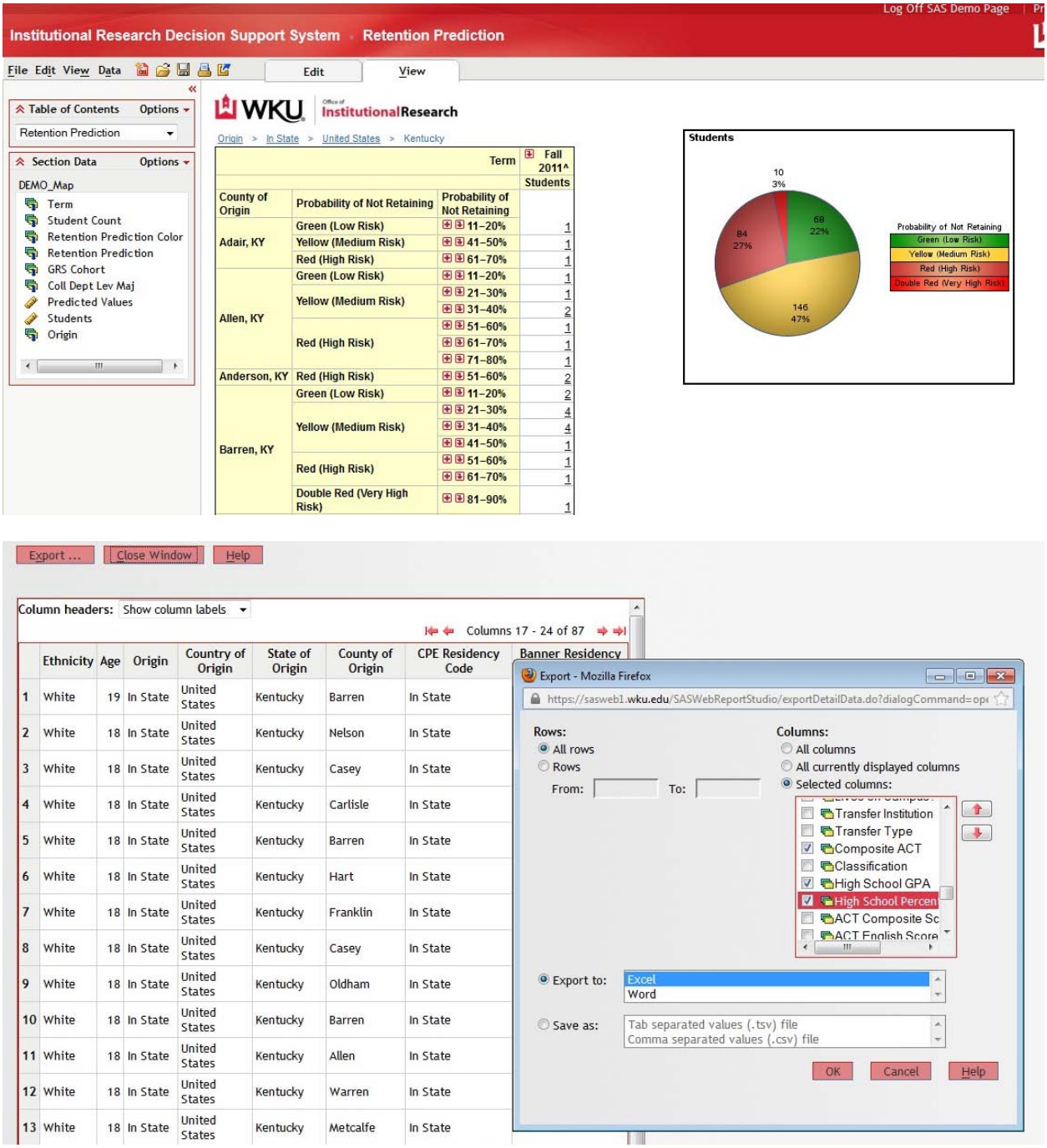


Term		Fall 2013^		
	Funnel Status	Applied	Admitted	
		Students	Students	Conversion
Level	County of Origin			
Undergraduate	Adair	14	4	28.57%
	Allen	40	26	65.00%
	Anderson	15	9	60.00%
	Ballard	5	3	60.00%
	Barren	100	54	54.00%
	Bath	3	.	.
	Bell	5	1	20.00%
	Boone	62	46	74.19%
	Bourbon	4	2	50.00%
	Boyd	5	5	100.00%
	Boyle	12	6	50.00%
	Breckinridge	26	18	69.23%
	Bullitt	36	22	61.11%
	Butler	21	14	66.67%
	Caldwell	9	7	77.78%
	Calloway	10	7	70.00%
	Campbell	34	22	64.71%
	Carroll	2	1	50.00%
	Carter	1	1	100.00%
	Casey	8	4	50.00%
	Christian	39	27	69.23%
	Clark	11	6	54.55%
	Clay	2	.	.
	Clinton	11	5	45.45%
	Crittenden	1	1	100.00%
	Cumberland	14	6	42.86%
	Daviess	128	91	71.09%

DRILL-THROUGH TO DETAIL

While many of the end users may not fully appreciate the complexity of the models used to drive the analytics employed, they find that the to drill-through to detail capability made possible with SAS® OLAP to be one of the most powerful aspects of our system. This capability dramatically expands the utility of our deployed models. Instead of requiring staff in the Office of Institutional Research to send multiple reports via hardcopy, pdf, or email, end users can create custom reports with the exact variables, analytics, and level (hierarchy) of details that they require. They can also export the final report to excel, word, or get the raw data in a csv file.

Figure 3: Incorporation of Analytics and Drill-Through to Detail Table



CONCLUSION

With WKU's increased focus on retention, in the fall of 2011 we implemented a Student Attrition Risk Analysis tool using models developed in SAS® Enterprise Miner and implemented in SAS® EBI. This project is the next step of our ongoing mission to deliver advanced analytics

and business intelligence to a wide community of faculty and staff. The automated delivery made possible with SAS® EBI saves a tremendous amount of resources that otherwise would have been tied up in more manual and error prone reporting efforts. This enables us to focus more resources to developing more advanced analytics applications for our end-users.

REFERENCES

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, *12*(2), 155–187.

Bogard, M.T., James, C., Helbig, T., & Huff,G. (2012). Using SAS® Enterprise BI and SAS® Enterprise Miner™ to reduce student attrition. Paper 031-2012. *SAS Global Forum 2012 Proceedings*. North Carolina; SAS® Institute

Braxton, J. M. (Ed.) (2000). *Reworking the student departure puzzle.* Nashville, TN: Vanderbilt University Press.

Braxton, J. M., Hirschy, A. S., & McClendon, S. A. (2004). Understanding and reducing college student departure,(*ASHE-ERIC Higher Education Report* No. 30.3), San Francisco: Jossey-Bass.

Breiman,L. (2001). Statistical Modeling:The Two Cultures(with comments and a rejoinder by the author). *Statistical Science*, Volume 16, Issue 3,199-231.

Bruggink, T.H. and Gambhir, V. (1996). Statistical Models for College Admission and Enrollment: A Case Study for a Selective Liberal Arts College. *Research in Higher Education,* 37:2. 221 – 240.

Caruana ,R. & Niculescu-Mizil ,A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *The Proceedings of the 23rd International Conference on Machine Learning (ICML2006)*. June 2006, pp. 161-168.

Chapman, D. & Pascarella, E. (1983). Predictors of academic and social integration of college students. *Research in Higher Education*, *19*, 295-322.

Curs, B & Singell Jr., L. D. (2002) An analysis of the application and enrollment processes for in-state and out-of-state students at a large public university. *Economics of Education Review*, 21 111–124

Dey, E. L. & Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education*, *34*(5).

Delen, D., Walker, G. & Kadam, A. (2004). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine, 34*(2) 113–127.

Delen, D., Sharda, R., & Kumar, P. (2007). Movie forecast guru: a web-based DSS for Hollywood managers, *Decision Support Systems, 43*(4) 1151–1170.

DesJardins, S. L.; (2002). "An Analytic Strategy to Assist Institutional Recruitment and Marketing Efforts." *Research in Higher Education,* 43 (5): 531-553.

DeVille, B. (2006). *Decision Trees for Business Intelligence and Data Mining Using SAS® Enterprise Miner*. SAS® Institute. Carey, North Carolina.

Friedman, J. H. (2001), Greedy function approximation: A gradient boosting machine. *The Annals of Statistics,* 29, 1189-1232.

Goenner, C. F. & Pauls, K.(2006). A Predictive Model of Inquiry to Enrollment *Research in Higher Education*, Vol. 47, No. 8. 935-956

Hasti,T.,Tibshirani, R. & Friedman, J. (2009)*Elements of Statistical Learning: Data Mining, Inference, and Prediction.* (2nd ed.). Springer-Verlag.

Herzog , S. (2005).Measuring Determinants of Student Return vs. Dropout/Stopout vs. transfer: A First to Second Year Analysis of New Freshmen. *Research in Higher Education,* Vol 46 No 8 Dec.

Kennedy,Peter.(2003). *A Guide to Econometrics.* (5th ed.). MIT Press.

Kiang, M.Y. (2003). A comparative assessment of classification algorithms, *Decision Support Systems, 35*, pp. 441–454.

Li, X., Nsofor, G.C., Song, L. (2009). A comparative analysis of predictive data mining techniques. *International Journal of Rapid Manufacturing, 1* (2) 150–172.

Pascarella, E.T. & Terenzini, P.T. (1978). The relation of students' precollege characteristics and freshman year experience to voluntary attrition. *Research in Higher Education, 9*, 347-366.

Provost, F. J., Fawcett, T.,& Kohavi, R. (1998). The Case against Accuracy Estimation for Comparing Induction Algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp.445-453)(ICML '98), Jude W. Shavlik (Ed.). Morgan Kaufmann Publishers Inc.,San Francisco, CA, USA.

Ronco, S. ,& Cahill, J. (2006). Does it Matter Who's in the Classroom? Effect of Instructor Type on Student Retention, Achievement and Satisfaction. *AIR PRofessional File*. Number 100, Summer.

Stratton, L.S., O'Toole, D. M., & Wetzel.,J. N. (2008). A multinomial logit model of college stopout and dropout behavior.*Economics of Education Review,* 27 319–331.

Thomas, E., Reznik, G., and Dawes, W. (2001).Using Predictive Modeling to Target Student Recruitment: Theory and Practice. *AIR Professional File.* 78.

Miller, T. E. (2007). Will They Stay or Will They Go?: Predicting the Risk of Attrition at a Large Public University. *College & University,* v83 n2 2-4, 6-7.

Miller, T.E. & Herreid, C.H. (2008). Analysis of variables to predict first year persistence using logistic regression analysis at the University of South Florida. *College & University, 83*(3).

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125.

CONTACT INFORMATION

Matt Bogard
Institutional Research, WKU
1906 College Heights Blvd. #11011
Bowling Green, KY 42101-1011
Work Phone: (270) 745-3250
Fax: (270) 745-5442
Email: Matt.Bogard@wku.edu