Building a Student At-Risk Model: An End-to-End Perspective

Lalitha Agnihotri, Ph.D.
McGraw Hill Education
2 Penn Plaza
New York, NY 10121
1-914-434-2372
lalitha.agnihotri@mheducation.com

Alexander Ott, Ed.D.
New York Institute of Technology
Northern Boulevard
Old Westbury, NY 11568
1-516-686-1037
aott@nyit.edu

ABSTRACT

Poor graduation and retention rates are widespread in higher education, with significant economic and social ramifications for individuals and our society in general. Early intervention with students most at risk of attrition can be effective in improving college student retention. Our research aim was to create a first-year at-risk model using educational data mining and to apply that model at New York Institute of Technology (NYIT). Building the model creates new challenges: (1)the model must be welcomed by counseling staff and the outputs need to be user friendly, and (2)the model needs to work automatically from data collection to processing and prediction in order to eliminate the bottleneck of a human operator which can slow down the process. The result of our effort was an end-to-end solution, including a cost-effective infrastructure, that could be used by student support personnel for early identification and early intervention. The **St**udent **A**t-**R**isk Model (STAR) provides retention risk ratings for each new freshman at NYIT before the start of the fall semester and identifies the key factors that place a student at risk of not returning the following year. The model was built using historical data for the 2011 and 2012 Fall Class and the STAR system went into production at NYIT in Fall 2013.

Keywords

Students At-Risk Model, Ensemble Model, End-to-End system

1. INTRODUCTION

On average less than 60% of full-time students who begin a four-year program of college study graduate in six years [7]. Moreover, the highest rate of attrition occurs during the first year of study — from the student's first fall semester to what would be his or her second fall. Figure 1 shows a box plot of graduation and first year retention rates in the United States for 2006, 2007, 2008, and 2009. (The dataset in the box plots derives from the Delta Cost Project [5], which in turn is based on Integrated Postsecondary Education Data System (IPEDS) data as made available by the National Center for Education Statistics [9].) The graduation rates are around 50% and the first year retention rates are clustered around 70%. Therefore, the logical starting point for improving graduation rates would be to find ways to improve first-year retention.

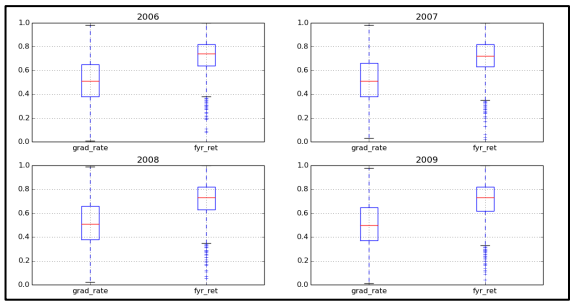

Figure 1. Graduation and First Year Attrition Rates

Research shows that counseling intervention with students at highest risk of attrition can be effective in improving retention [6][10][12]. Essential to this intervention, however, is that it be early in the student's first semester at college [12]. The problem is twofold: (1)how to identify and intervene with these at-risk students before it is too late, and (2)how to identify the key factors putting these students at higher risk of attrition so as to inform the counseling intervention and improve its effectiveness. Given that evidence exists in the literature that it is possible to build a data-mining based model that would address both dimensions of the problem [13], NYIT undertook such an effort, beginning in earnest in the fall of 2012.

However, having the most powerful predictive model would not be useful if either (1)the counseling staff that must ultimately use the model were not willing to do so, either because they did not want the model in the first place or because they were uncomfortable with the model outputs, or (2)the model itself needs intensive manual intervention to produce the output, therefore slowing it down. To overcome these challenges we needed to employ an "end-to-end" approach in user-oriented model creation as well as in building a highly automated model on a technological level.

In terms of building a user-oriented model, the model itself was built in an iterative cycle between the end users and the IT model creator. The problem definition came from the actual users—the retention-focused counseling staff. The data identification was done in collaboration with the solution provider and the users. The IT solution provider did the data gathering and preparation, model building, evaluation, and deployment. Once the knowledge deployment occurred, the users were looped backed in to provide feedback and critique and the process was restarted.

On the technological level, we similarly used an end-to-end solution to build a highly automated model: We used the

Microsoft SQL server as our tool of choice and built the database, prediction models, and the front end used by the counselors all on the same platform. The model was built "in house" at NYIT.

The resulting **St**udent **A**t-**R**isk Model (STAR) provides retention risk ratings for each new freshman at NYIT before the start of the fall semester and identifies the key factors putting a student at risk to not return the following fall. The model was built and used for intervention for the incoming Fall 2013 freshman class.

2. NYIT STAR Model, Version 1.0

NYIT's Student Solutions Center (SSC), which is NYIT's "one-stop-shop" for enrollment services, engages with new students by providing counseling guidance to improve student success and retention. The SSC's counseling intervention is called the 4-3-2-1 Plan, which involves individual counselor-new student meetings occurring early in a new student's first semester at NYIT.

In fall 2011, the SSC attempted to build its own model to identify the most at-risk students, therefore allowing earlier, targeted intervention with these students. This STAR Model, version 1.0, was rather simplistic in its approach. We essentially gathered data on each student from multiple sources and compiled in one Excel sheet. We then used the retention literature and our own inclinations to identify variables and assign each variable with a score of "1" or "0"—with a 1 being a retention risk. The higher the score for each student, the more at risk he or she is.

As one might expect, this approach was highly problematic. On a conceptual level, it was based on student behavior at other institutions (via the retention literature) and not behavior at NYIT. It was also a blunt instrument, in that all variables were weighted equally. On a practical level, compiling the Excel sheet involved significant labor, gathering information from multiple data sources manually.

2.1 NYIT STAR Model, Version 2.0

In order to overcome these "Model 1.0" limitations, we took two key steps: First, we built the dataset in our Data Warehouse where it can be created automatically as soon as a new student registers. Second, we decided to use data mining tools to train machine-learning models to perform the classification task. These models use the variables to predict whether or not a student will return the following year which is then used to flag the risk of new students.

We chose Microsoft SQL Server for the following reasons: All our data exists in the SQL Server; the SQL Server Analysis Services (SSAS) provides capability to use Neural Networks, Naïve Bayes, Logistic Regression, and Decision Tree models for prediction. And finally once a model is trained, SQL Server Reporting Services (SSRS) allows us to query the model on an on-demand basis to populate a report hosted on a SharePoint site that serves as the front end for the counselors to access the data. Figure 2 shows our Microsoft Business Intelligence stack that we used for building the STAR model. The SSAS Modeling part happens only once.

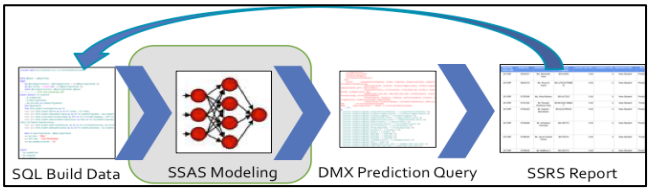

Figure 2. End-to-End Technology Solution

We selected a total of 25 variables after multiple iterations. These variables are from the same three sources as in STAR 1.0—students' admission application data, registration/placement test data at NYIT, or from a survey that the students complete when they take the Compass placement exam at NYIT. However, STAR 2.0 also includes financial data, which research indicates plays a role in retention risk, albeit a complicated one (see, for example, [4][8]).

3. Data Mining Models

Campbell et al. [3] propose the use of analytics to academics, which is what we have attempted in building STAR 2.0. Bayer et al. [1] used student data enriched with data derived from students' social behavior to predict student failure. This works well for longitudinal snapshot data. Romero et al. [11] present data mining methods for classifying students based on their Moodle usage data. They have defined a set of attributes specific to Moodle usage and compared a number of methods and their algorithmic implementations. Taylor and McAleese [15] presented a system that uses data intelligence and analytics for more efficient and effective student success interventions, though they used an analytics company to do the modeling effort. Since we developed our models in house, the data stays in house, so there are no security issues—a significant advantage. Further, having access to the models enables us to drill down into the prediction results to give a detailed picture for each student, as will be demonstrated more fully later. In addition to all the data mining methods, use of ensemble models is growing in popularity as it has the ability to generalize much more than any single method. Yu et al. [16] used Ensemble models in their classifier for 2010 KDD Cup and won the first prize in the challenge.

The process of modeling involves training and testing of multiple models and then selecting the one that works the best for the application. We chose to build four different initial models: Neural Networks, Naïve Bayes, Decision Tree, and Logistic Regression. On top of this we built an Ensemble model that, in addition to taking the variables for each student, takes the output of the initial four models as input and predicts whether a student is at risk or not.

3.1 Model Performance

Key measurements that are used to gauge the power of a predictive classification model are recall and precision. Recall compares the number of students who were predicted as not returned with all those who actually did not return. That is, recall measures the following: Of all those students who actually did not return the following fall, what percentage were correctly predicted by the model as not returning? Precision compares the number of students correctly predicted as not returned with all those who are predicted by the model as not returned. That is, precision measures the following: Of all those students predicted by the model not to return, what percentage of those students actually did not return the following fall?

3.1 Model Selection

Each of the models has a number of parameters that can be changed. We analyzed each of the models and decided to vary the parameters to generate almost 400 variants of the initial models. For example, the reasoning for the number of models that can be built for Naïve Bayes is derived as follows. The Naïve Bayes model has four parameters that can be varied: Maximum Input Attributes, Maximum States, and Minimum Dependency

Probability. We chose not to alter the first parameter that enables feature selection for reasons explained below. We change states from 0 to 250 in steps of 10 for total of 26 different values. Also we vary the minimum dependency probability from 0.1 to 1 in steps of 0.1. In total we get 26*10=260 models of Naïve Bayes.

In order to eliminate feature selection, we ran the 20 possible neural networks with all possible values for parameter Maximum Input Attributes going from 1 to 25 in steps of 1 for a total of 25*20=500 models in all. The same recall and precision was obtained for classifiers with features 21 or more that was the best recall of all the 20 models generated. Further, since all our features are readily available, we decided not to deal with feature selection for our modeling process in order to speed up the development of a model that can be used. In the future, we will revisit the feature selection more rigorously to eliminate variables that may be irrelevant.

Based on analysis for each of the models, we trained a total of 372 models: 20 for NN, 26 for Logistic Regression, 66 for DT, and 260 for Naïve Bayes. SQL Server Analysis Services (SSAS) has a scripting language called DMX that can be used to automatically generate and train models. A DMX script was generated using a SQL query to generate these models automatically. Once the models were generated, their recalls and precisions were computed automatically using another SQL query and stored in a table and the ones with highest recall were selected. If multiple models had the highest recall, we chose the one that provided the highest precision. The model that gave us the highest recall (and precision) for each of the four models was then chosen to generate the ensemble model. The ensemble model takes all of the student data as input as well as the output from the four initial models.

A total of 1453 students who were admitted in fall of 2011 and 2012 were used for the purpose of training and testing the models. Of these 983 students returned to the campus in following fall and 470 did not return. We used 70% of the data for training and 30% for testing. SQL Analysis Services randomly samples the data to help ensure that the testing and training sets are similar.

Following are the models and the selected parameter value chosen for each of the types based on the automatic selection of 372 total models built.

1. Decision Tree: Complexity Penalty = .8, Score Method = Bayesian with K2 Prior, Split Method = Binary
2. Logistic Regression: Maximum States = 10
3. Naïve Bayes: Maximum States = 10, Minimum Dependency Probability = .1
4. Neural Net: Hidden Node Ratio = 19

The Logistic Regression had the best recall and was hence used as the model of choice for the ensemble model. We trained a Logistic Regression model with the same parameters as the chosen initial model to be our final model. This model not only had as input all the student variables, but also the outputs of the four initial models that were chosen automatically as explained above.

3.2 Model Comparison

The model performance using the 2011 and 2012 test data showed a stark contrast between the manual STAR model 1.0 and STAR model 2.0. The recall of the basic four models in version 2.0—Logistic Regression, Neural Network, Naïve Bayes, and Decision Tree—varies from 45% to 62%. This means that the strongest model in version 2.0 in terms of recall was capable of correctly identifying 62% of the not returning population. This represents a major improvement over the 34% recall in version 1.0. In terms of precision, the models in version 2.0 vary from 54% to 70%. This means that in the strongest model in version 2.0 in terms of precision, 70% of those students identified as not returning actually did not return the following fall. This, too, is a major improvement over the 42% precision in version 1.0.

To improve the recall of version 2.0 further, we built an ensemble model that can use the output of the four initial models along with the student data as input and predict whether a student will return or not. This provides the best recall results we have obtained so far, as the model's recall is 74%. This means that the model is able to identify 74% of the students who did not return correctly whereas the other models were able to reach a 62% recall at best. Table 3 summarizes the performance of the models.

Table 3. Performance Comparison of Models

Model Name	Recall	Precision
Manual (STAR 1.0)	34%	42%
Logistic Regression	62%	57%
Neural Network	56%	54%
Naïve Bayes	51%	69%
Decision Trees	45%	70%
Ensemble	74%	55%

In the end, version 2.0 compares very favorably with not only version 1.0 but also with a similar Data Warehouse-based retention modeling effort described at Western Kentucky University [2]. Bogard et al. report that they were able to achieve a recall of only about 30% based on pre-enrollment data, which is also the data space in which our model operates. As noted, we are able to do much better, in fact up to 74% recall for the ensemble model, due in part to our model selection and also due to the inclusion of financial data and a student survey which is missing in the efforts described by Bogard et al.

We did another validation of our method and trained our models as explained above on Fall 2011 data alone (724 students). Then we used the ensemble model to predict the retention risk on Fall 2012 new students (729 students). As can be seen in Table 4, our model was able to generalize well on 2011 students' data and did a comparable job of predicting the retention risk for Fall 2012 new students.

Table 4. Ensemble Model Validation

Model Performance On Training and Testing Data	Recall	Precision
Training Data: 2011 Fall New Students Data	75%	56%
Testing Data: 2012 Fall New Students Data	73%	54%

The answer to the bigger question of how well our model worked as a guide for actual intervention and as a way to change student enrollment behavior will not be known until Fall 2014.

4. STAR MODEL PRODUCT

As mentioned at the start of this paper, all the smart model building and predictions would not be of any use if they cannot be presented in a manner that is easily digestible and pleasing to use for the counseling staff. This is where our end-to-end iterative model building became essential to the project's success. We built an actual "*product*" that the users could use that had the model and its output under the hood.

As soon as the first version of model 2.0 was complete, we built a report using Microsoft's SQL Server Reporting Services to show the prediction output to the counselors. After several cycles of counselor feedback and report revisions, the counselors had a final, user-friendly product that they were comfortable using and that they participated in creating. The final report tells the

counselor which students are at highest risk of not returning the following fall, which allows the counselor to target those students the model is most confident are at risk. Second, the output report lists the reasons for both student risk (e.g., Math placement, affordability) and lack of risk (e.g., high SAT scores, full time enrollment) in the final column. The most recent revision we made to the report was to create a STAR Counselor Log that provides all the output of the model and also allows the counselor to input data about when he or she met with the student, what was discussed, etc.

This STAR Counselor Log is an evolving interface. We have plans to revise this interface further on the basis of feedback from SSC counselors as to how it could be improved. One suggestion is to find a way to categorize a counselor's assessment of the student after the first meeting. For example, we could add a check box that indicates the counselor believes the student is at such high risk that he/she should be followed up with quickly—typically we wait for a second 4-3-2-1 Plan meeting until the following semester.

5. CONCLUSIONS AND CHALLENGES

For colleges considering building an at-risk student model, the key conclusions from the STAR modeling effort are as follows: First, a student at-risk model can be built "in house" if appropriate data is collected and stored in a Data Warehouse. Second, performing data mining is essential to building accurate models in order to weight variables correctly based on student behavior at your institution in particular. Ensemble models can be very useful as the data is rarely clean and each model can only capture so much information on its own. Third, any solution that is provided needs to have an end-to-end perspective in place so that the prediction modeling process is smooth one and the product is user friendly.

As with most attempts to address a complex topic, many challenges remain: First, while the predictive ability of STAR Model 2.0 is quite high, and much higher than STAR Model 1.0, there is still significant room for improvement. For example, the strongest model had a recall of 75%, which means it failed to predict 25% of the students who did not return. Second, assessing whether the STAR-guided intervention is meaningful in a student retention context and, if so, how to demonstrate this. Third, the counseling intervention to at-risk students can affect the model over time. How do we get past the Heisenberg uncertainty principle to build the best model while intervening?

Despite the challenges, the STAR model has been a large step forward at NYIT—it allows NYIT counselors to prioritize intervention with those first-year students most at risk early in the fall semester. This intervention is now based on real student data and is informed by key at-risk variables for each student, allowing the counselor to tailor intervention to the risk factors of each student.

6. ACKNOWLEDGMENTS

Our most sincere thanks to colleagues at NYIT for their contributions to this effort: Mr. Yongxin Ma provided support with Data Warehouse and much wisdom and insight. None of this would have been possible without the support of Mr. Will Wall, SSC Data Guru, who developed version 1.0. And a big shout-out to the counselors in SSC for their invaluable feedback in developing the front end and also their enthusiasm in using the model for intervention. We would like to express our gratitude to Al Essa for his help in editing the paper and also for providing box plots for graduation and first-year retention.

7. REFERENCES

[1] Bayer, J., Bydzovska H., Geryk J., Obsivac T., Popelinsky L. (2012). Predicting drop-out from social behaviour of students. *Educational Data Mining 2012: 5th Intl Conf on Educational Data Mining, Proceedings*. Chania, Greece.

[2] Bogard, M., Helbig, T., Huff, G., & James, C (2011). A comparison of empirical models for predicting student retention. White paper. Office of Institutional Research, Western Kentucky University.

[3] Campbell, J.P., DeBlois, P.B., & Oblinger, D.G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE Review*, *42*(4), 41–57.

[4] Chen, R., & Desjardins, S. L. (2008). Exploring the effects of financial aid on the gap in student dropout risks by income level. *Research in Higher Education, 49*(1), 1–18. doi: 10.1007/s11162-007-9060-9

[5] Delta Cost Project. Retrieved from http://www.deltacostproject.org/

[6] Fowler, P. R., & Boylan, H. R. (2010). Increasing student success and retention: A multidimensional approach. *Journal of Developmental Education, 34*(2), 2–4, 6, 8–10.

[7] Friedman, B. A., & Mandel, R. G. (2009). The prediction of college student academic performance and retention: Application of expectancy and goal setting theories. *Journal of College Student Retention, 11*(2), 227-246.

[8] Kim, D. (2007). The effect of loans on students' degree attainment: Differences by student and Institutional characteristics. *Harvard Educational Review, 77*(1), 64–100, 127.

[9] National Center for Educational Statistics. (2014). Fast facts. Retrieved from http://nces.ed.gov/fastfacts/display.asp?id=40

[10] Pan, W., Guo, S., Alikonis, C., & Bai, H. (2008). Do intervention programs assist students to succeed in college?: A multilevel longitudinal study. *College Student Journal, 42*(1), 90–98

[11] Romero C., Ventura S., Espejo P.G., & Hervás C. (2008). Data Mining Algorithms to Classify Students, *Educational Data Mining 2008: 1st Intl. Conference on Educational Data Mining, Proceedings*. Montreal, Quebec, Canada.

[12] Seidman, A. (2012). Taking action: A retention formula and model for student success. In A. Seidman (Ed.), *College student retention: Formula for student success* (2nd ed.) (pp. 267–284). Lanham, MD: Rowman & Littlefield.

[13] Singell, L. D., & Waddell, G. R. (2010). Modeling retention at a large public university: Can at-risk students be identified early enough to treat? *Research in Higher Education, 51*(6), 546–572. doi: 10.1007/s11162-010-9170-7

[14] Tampke, D. R. (2009). Developing and implementing an early alert system. In R. Hayes (Ed.), *Proceedings of the 5th National Symposium on Student Retention, 2009, Buffalo*. (pp. 143–151). Norman, OK: The University of Oklahoma.

[15] Taylor, L. & McAleese, V. (2012). Beyond retention: Using targeted analytics to improve student success. *EDUCAUSE Review*.

[16] Yu, H.-F., et al. (2010).Feature Engineering and Classifier Ensemble for KDD Cup 2010. *Conference on Knowledge Discovery and Data Mining.* Washington, DC.