Can "at risk" student athletes be identified through predictive analytics?

Presented to the Annual Forum of the

California Association for Institutional Research

Anaheim, CA

November 14-16, 2018

Heidi Carty, Ph.D.,

Research Analyst

Galina Belokurova, Ph.D.,

Research Analyst

University of California, San Diego

http://ir.ucsd.edu/

Overview

- First-time freshman student athletes may need additional support and resources to adjust to the new learning experience in the university environment while meeting the demands of their sport.
- Our predictive model identifies student athletes ``at risk" before they start their academic career using their high school academic preparation, background characteristics and other non-cognitive measures.
- C5.0 Decision Trees, Neural Network, CHAID and Logistic Regression were examined to determine which had the highest precision and accuracy.
- ``Out of the box" C5.0 Decision Trees and CHAID algorithms had 50% accuracy to 57% precision respectively and recall between 15% and 31%. In working with the model we've been able to increase the accuracy to 81% with a new set of data.

Project Purpose

- This presentation should be of interest to institutions that want to utilize SPSS Modeler and predictive analytics with student data to improve the success of their incoming students.
- Our purpose is to illustrate how universities can incorporate predictive modeling into the data analysis routinely performed on applicants' data.
- Students ``at risk'' with similar traits can receive the support with the greatest chance of increasing their success.

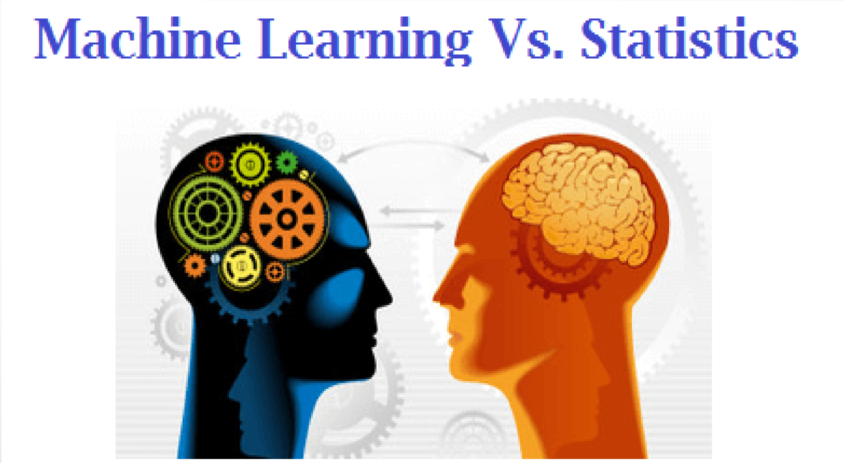
Research Question

- Can we accurately identify “at risk” athletes using predictive analytics?

Heidi Carty, Ph.D. & Galina Belokurova, Ph.D.
Institutional Research, Academic Affairs, UC San Diego

Predictive vs Statistical Modeling

- Causal or explanatory approaches implemented in statistical modeling is a top-down way of thinking based on a theory, from which a researcher generates testable hypotheses. It helps understand the data generating process, yet it does not provide sufficiently detailed individual predictions.
- A data driven approach focuses on discovering patterns in the data that may lead to accurate predictions about individual student outcomes while keeping the mechanisms behind it in the "black box".



Basic Project Characteristics

- **Subjects**: first-time freshmen at a large public university between 2013 and 2017 . Total number of observations equals 21,036. Once the model was identified a group of incoming freshmen athletes were run through the model (N = 158).
- **Approach**: a predictive data-driven one focused more on identifying individuals in need of intervention and less on piecing together the causal mechanism behind it.
- **Target**: Academic Risk (called FALLGPACAT on the SPSS Modeler Canvas) and is designed to capture differences between students in good academic standing (e.g., GPA of 2.6 or higher) and those with poor academic performance (e.g., less than a 2.6 GPA). We do experiment with different definitions of the target.

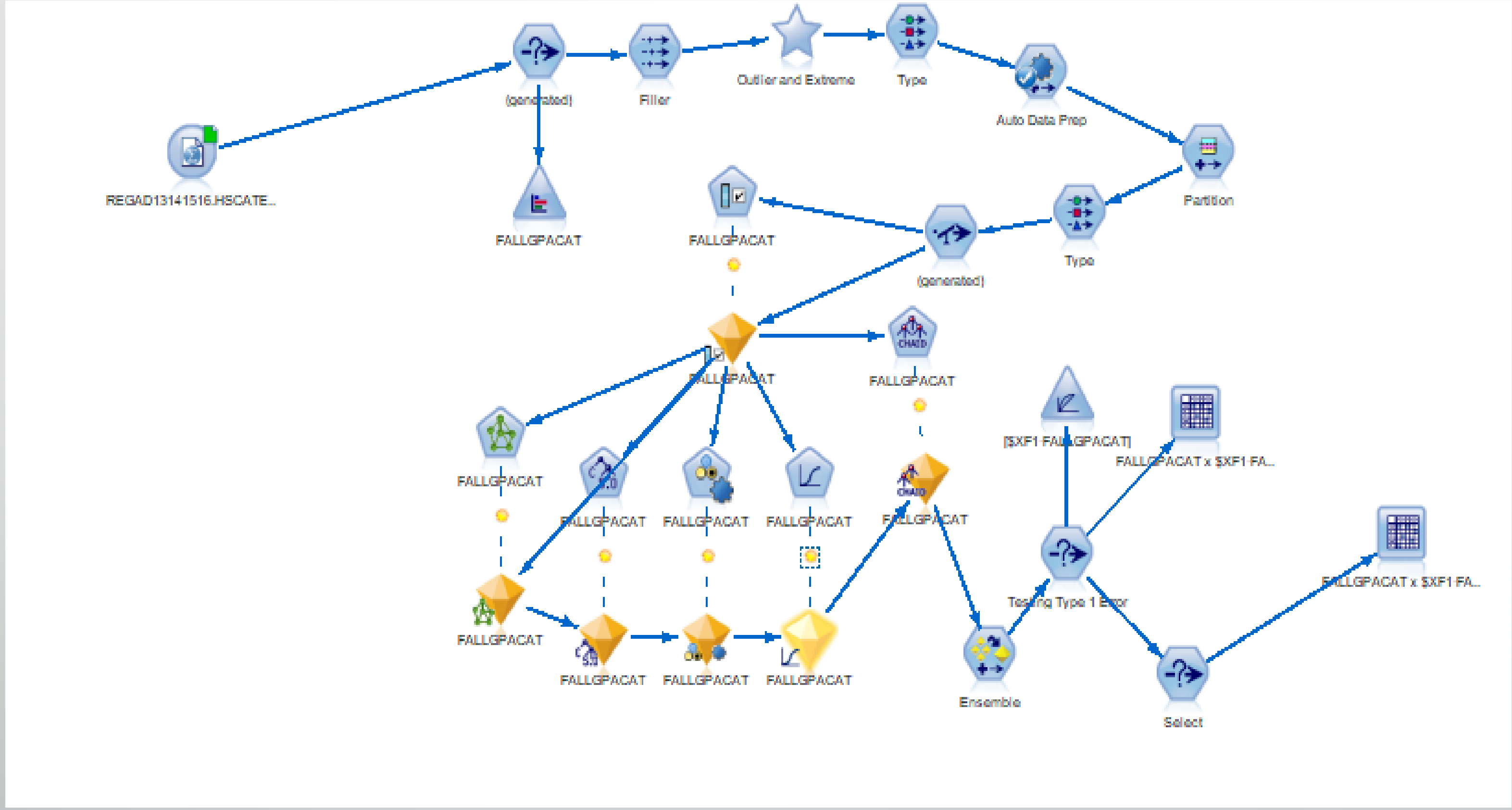
Procedures

- Rescaling/Transforming Data,
- Partitioning,
- Balancing the Training Subset,
- Training the Model, and
- Evaluating the Model using the Testing Subset of the Data.

Rescaling/Transforming, Partitioning & Balancing Training Data

- Re-scaling and transformation help to make sure disparities in the scale of predictors do not affect the classification routines. Outliers potentially capable to bias the estimates are removed.
- Input data randomly divided into two subsets, a training data set and a testing data set. Our model implements 50-50 percent training/testing split.
- The original dataset is unbalanced with a far higher proportion of students in good academic standing. Models perform better if the data have approximately equal numbers of low and high outcomes. To correct for this discrepancy we balanced the dataset by matching the number of low and high outcomes to improve the performance of the training model at detecting at "at risk" cases.

Figure 1. SPSS Modeler Canvas View



Heidi Carty, Ph.D. & Galina Belokurova, Ph.D.
Institutional Research, Academic Affairs, UC San Diego

Groups of Predictors

The table below describes the predictors used in all our models. Not every one attained predictive importance. Encouragingly, the student athlete status did not emerge as an important variable. This indicates no systematic difference between athlete and non-athlete students in terms of the obstacles they face during their first term at the university.

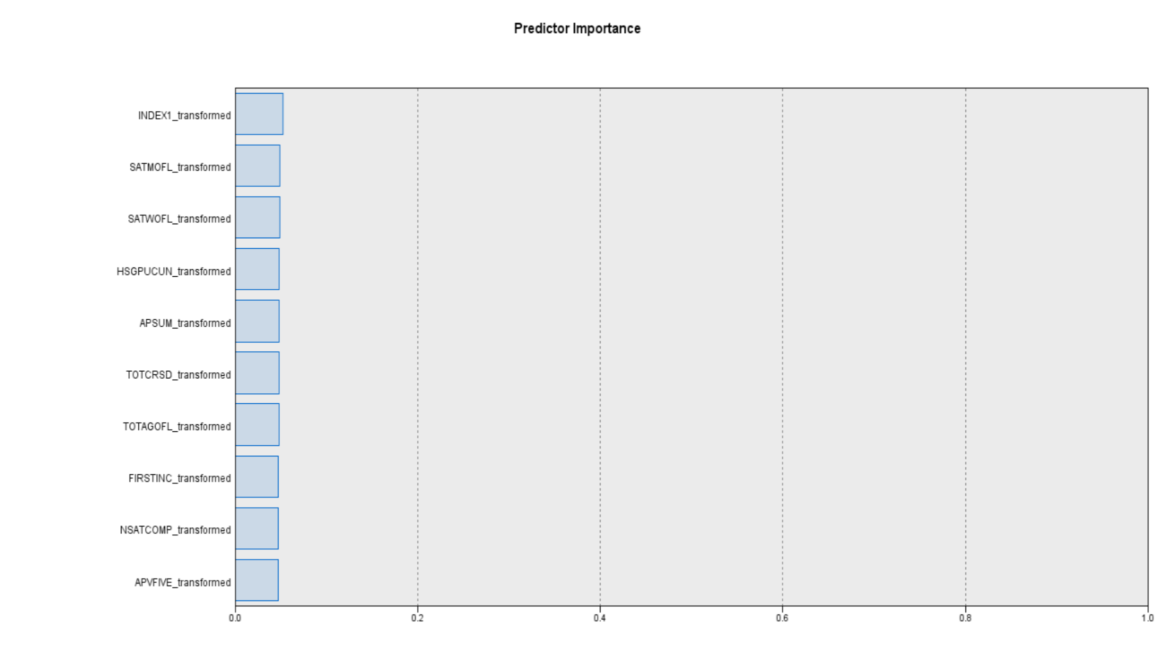
Predictor groups	Definition	Variables
Demographic and socio-economic	Stable individual attributes	Gender, income group (current parents' income), student's home location, first generation status
Institutional	Status and individual's "location" within UC San Diego and feeder institutions	Requested major at UC San Diego, feeder high school state ranking, UC San Diego's college applied to
Course performance in high school	Course Details	Honors courses, AHI requirement, total math courses, total science labs, total number of elective courses, total number of AP courses taken/planned
Aggregate performance	Aggregate measures of academic performance	Cumulative school GPA, academic index score generated by UC San Diego based on grades
Entrance Exams	SAT, ACT, AP Tests	Official SAT math, verbal, written scores, ACT scores, number of AP tests taken with passing score
Non-Academic Factors	Extra-curricular activities, additional skills and experiences	Leadership skills, community service

Heidi Carty, Ph.D. & Galina Belokurova, Ph.D.
Institutional Research, Academic Affairs, UC San Diego

Predictor Importance

- Assessing predictor importance is often the first step in predictive modeling. SPSS Modeler has the capability to make such preliminary evaluation based on the F statistic (how F changes if you drop a predictor) or a p-value when comparing different groups of observations formed during the classification process.
- Figure 2. shows student's admission, SAT scores, and high school GPA are among the most important predictors. There are unfortunately no very strong predictors.

Figure 2. Predictor Importance

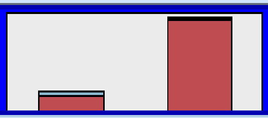
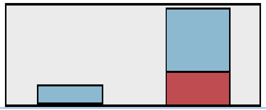
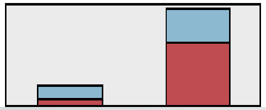


Heidi Carty, Ph.D. & Galina Belokurova, Ph.D.
Institutional Research, Academic Affairs, UC San Diego

Building Predictive Models

- Train different models on the training dataset using a variety of statistical and ML processes: SPSS Modeler automated classifier comes first. The three models with the greatest overall accuracy are CHAID and Discriminant (84% and 66% respectively).

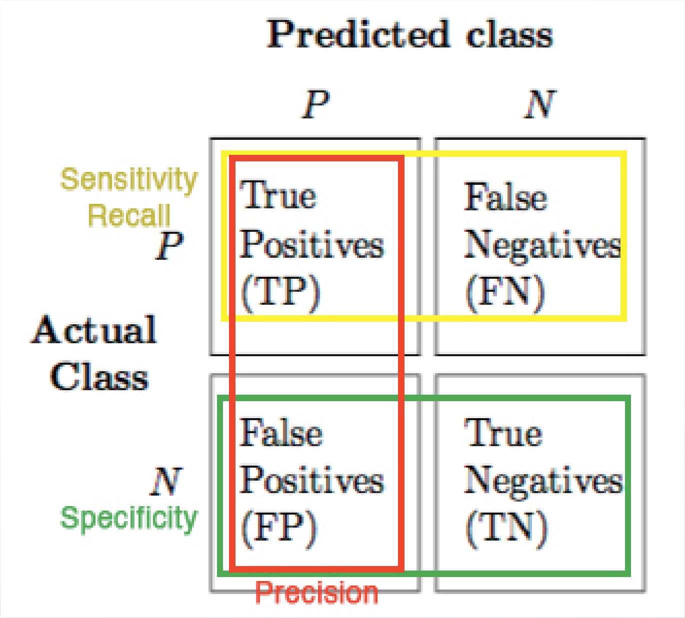
Figure 3. SPSS Modeler Auto Classifier Selection

	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)
☑		CHAID 1	< 1	35817.848	97	1.129	84.011
☑		Decision List 1	< 1	35,680.0	100	1.121	44.781
☑		Discriminant 1	< 1	35,790.0	97	1.114	66.623

We supplemented the in-built routine by running Neural Net, C5.0, Logistic Regression, and CHAID with enhanced model stability (bagging). At the end, we created an ensemble model that combined models on the basis of confidence-weighted voting.

Model Evaluation – Confusion Matrix

- Confusion matrix yields counts of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).
- Sensitivity or recall (TP/(TP+FN)) - the ability of the classifier to detect the “at risk” class; the true positive rate.
- Precision (TP/(TP + FP) – the positive predictive value, i.e., the proportion of relevant cases among retrieved cases.
- The overall accuracy (TP + TN)/(TP+TN+FP+FN) is not always a good metric for evaluating the classifier as it can be dominated by the students in good standing class (TN-FP).



Model Comparison

Figure 4. SPSS Modeler's Auto Classifier Evaluation

Predicted

actual

At risk

Not at risk

$XF-FALLGPACAT

FALLGPACAT		1.000000	2.000000	Total
1.000000	Count	1103	2384	3487
	Row %	31.632	68.368	100
	Column %	42.653	12.921	16.576
	Total %	5.243	11.333	16.576
2.000000	Count	1483	16066	17549
	Row %	8.451	91.549	100
	Column %	57.347	87.079	83.424
	Total %	7.050	76.374	83.424
Total	Count	2586	18450	21036
	Row %	12.293	87.707	100
	Column %	100	100	100
	Total %	12.293	87.707	100

precision

recall

specificity

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 1,449.815, df = 1, probability = 0

Figure 5. Custom Ensemble Model Evaluation

$XF1-FALLGPACAT

FALLGPACAT		1.000000	2.000000	Total
1.000000	Count	259	1461	1720
	Row %	15.058	84.942	100
	Column %	56.674	14.438	16.263
	Total %	2.449	13.814	16.263
2.000000	Count	198	8658	8856
	Row %	2.236	97.764	100
	Column %	43.326	85.562	83.737
	Total %	1.872	81.865	83.737
Total	Count	457	10119	10576
	Row %	4.321	95.679	100
	Column %	100	100	100
	Total %	4.321	95.679	100

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 572.757, df = 1, probability = 0

Model Evaluation

- The overall accuracy of the SPSS Modeler Auto Classifier Ensemble is 87%. The precision is 43%, i.e., those students classified as "at risk" who actually had low GPA (see Figure 4 below). This model performed well in terms of picking 32% of all students who turned out to be "at risk" (this is called recall).
- The overall accuracy of the final custom ensemble model is still 84%. The precision, percent of correctly classified students "at risk", is much higher – 57% (Figure 5). The cost is that only 15% of all students "at risk" were identified as such.

Results

- Of the four algorithms, CHAID and C5.0 decision trees outperform neural networks and logistic regression in their ability to classify students in the "at risk" group.
- The CHAID and C5.0 decision trees together result in a precision of 50% to 57%. About 43% of students identified as "at risk" actually performed well.
- Recall of the custom ensemble model was not as high as the automated routine and fell from 32% (Figure 4) to 15% (Figure 5).

Discussion & Further Research

- With the precision of the model at just over half, the program coordinator wasn't concerned if some students were misidentified. They were more comfortable including some students who may not necessarily need the program, compared to missing students who do.
- Without GPA constraints, we were able to alter the GPA groupings and improve both the accuracy and precision of the model.
- We also ran an additional series of models with a target defined differently. GPA of 3.2 (grade B) splits our student population approximately in half. So, the task was to predict whether student athletes fall below GPA of 3.2. Figure 6 shows the resulting model.

Figure 6. SPSS Modeler's Auto Classifier Evaluation (target is binary – whether GPA is more or less than 3.2)

$XF-FALLGPA_G

FALLGPA_G		1	2
1	Count	5114	4058
	Row %	55.757	44.243
	Column %	67.467	29.821
	Total %	24.136	19.152
2	Count	2466	9550
	Row %	20.523	79.477
	Column %	32.533	70.179
	Total %	11.639	45.073

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 2,810.442, df = 1, probability = 0

This model has lower overall accuracy, but its precision and recall are much better – at 67% and 56% respectively. This is likely the result of the balanced input data, which was artificially created by setting the target GPA threshold at 3.2.

References

- Abdar, M. (2010, January 01) *A Survey and Compare the Performance of IBM SPSS Modeler and Rapid Miner Software for Predicting Liver Disease by Using Various Data Mining Algorithms* Retrieved from https://www.researchgate.net/publication/317633766_A_Survey_and_Compare_the_Performance_of_IBM_SPSS_Modeler_and_Rapid_Miner_Software_for_Predicting_Liver_Disease_by_Using_Various_Data_Mining_Algorithms
- Sayad, S. (2010), *Model Evaluation – Classification* Retrieved from http://www.saedsayad.com/model_evaluation_c.htm
- Grandy, Jeff, Nancy Lough, Chyna Miller (2016, September 01). *Improving Student-Athlete Academic Success: Evaluation of Learning Support Tools Utilized by Academic Advisors for Athletics*. Journal for the Study of Sports and Athletes in Education, 2016. **10**(3): p.199-217
- Hung, J.-L., Y.-C. Hsu, and K. Rice, *Integrating data mining in program evaluation of K-12 online education*. Educational Technology and Society, 2012.**15**(3): p. 27-41.
- Moseley, L.G. and D.M. Mead, *Predicting who will drop out of nursing courses: A machine learning exercise*. Nurse Education Today, 2008. **28**: p. 469-475.

The End

Thank you for attending our presentation. Please fill out an evaluation for this session by clicking on the evaluation link in your CAIR app.

Have questions? Our contact information – Heidi Carty (hcarty@ucsd.edu) & Galina Belokurova (gbelokurova@ucsd.edu)