

Predictive Modeling for Student Retention

From Application to Graduation

Jose R. Bautista

The University of Arizona

July 21st, 2016

Outline

- 1 Introduction
 - Research Focus and Goals
 - MAUI
- 2 Methods
 - Statistical Learning
 - Decision Trees
- 3 R Tools
 - randomForest
 - Boosting and gbm
- 4 Results and Conclusions

Research Focus and Goals

- The University of Iowa has been attempting to find a way to increase enrollment, retention, and academic performance among first year undergraduate students.
- What if we can cater visits, interactions, and advising to each student individually? What determines whether a student will come back for their sophomore year? Can we develop reasonable academic expectations for students simply based off application information?

The Data

MAUI

Made At the University of Iowa

- Database that contains diverse information regarding students that are enrolled, graduated, or had some other interaction with the University of Iowa.
- E.g. Date of Application, Language Proficiency, High School GPA, etc.
- Many different tables to choose from. Lack of data is not the issue, however differentiating between what could be a useful predictor and what may be noise is a challenge within itself.
- **Snapshot:** a tool in MAUI that allows you to extract data from specific date ranges. I.e. We are able to select admission data from 2014 independent from any future data.

Structure of MAUI and Preparation of Data

- MAUI's primary function is storage of student's information, contains many different tables.
- A data set with information for every student does not exist, it needs to be compiled from smaller tables and organized in a way to prepare for analysis and model building.

Compiling a Master Data Frame

Specific use of the R package: dplyr

- **inner_join**: Multiple tables have information that would be useful to place in a master data set.. `inner_join` allows us to join two sets by a common factor. "MASTER_ID" is in all MAUI data sets, which made it perfect to join with.
- Other MAUI data sets did not have observations for all IDs. For example, when looking at student's relatives that are alumni, not every student has an alumnus relative. **left_join** allows you to join two sets, however if there are not any observations for an ID, the resulting data is given a value of NA
- Missing information proves to be quite useful in building classification and regression models. It is important to keep them as "MISSING" instead of simply deleting the values.

Predictive Focus

Predictor Variables

High school four-pt GPA, ACT score, number of years of second language, major, home state, honors status, application date, number of contacts with UI (campus visits, phone calls, etc.), high school class size, and the number of ACT/SAT tests a student has taken, etc. In total, there are 25.

- Lots of factors could potentially effect whether a student comes back for their second year and their first year GPA at the University of Iowa.
- These predictor variables will be the focus in building regression and classification models to predict both 2015 GPA and retention.

Predictive Modeling

- Also know as: statistical learning, machine learning, informatics, etc.
- Generally, a problem is presented and the goal is to achieve the ability to predict unknown future outcomes.
- Predictive models use known events and their outcomes in order to understand the product of a set of new events.
- Strong models follow trends of the initial data, but adapt when introduced to new data.

Traditional Regression vs. Statistical Learning

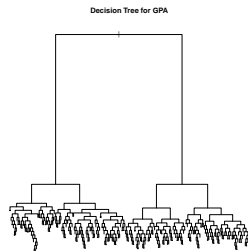
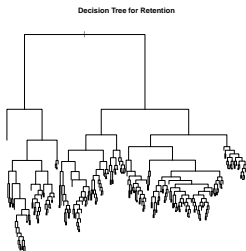
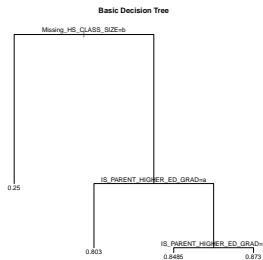
- Traditional regression models such as linear regression and logistic classification focus on gaining an understanding of a linear relationship between a set of input variables and a response variable. E.g.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Statistical learning methods attempt to estimate the functional form of the relationship between inputs and outputs.
- The models are fit in order to maximize predictive power. Because of this, the models lose interpretability and are sometimes referred to as "black box techniques".

Decision Trees

- Decision trees may be used for both classification and regression problems.
- Divide observations into groups based on predictor variables. They go through predictors one by one and split the data in two based on each predictor variable.



RandomForest

randomForest

A package in R that utilizes bootstrap re-sampling and decision trees.

- Bootstrapping takes samples from the initial sample, with replacement. The goal is to mimic the population distribution, by repeatedly sampling many times.
- The random forest algorithm builds decision trees on each bootstrapped sample. Each tree uses a random subset of predictor variables to grow the tree.
- The product is an average of the trees which has low variance.

Boosting

Definition

A family of machine learning algorithms which convert weak learners to strong ones.

- Boosting is a technique that is similar to random forests in that it uses bootstrapping techniques. However, it focuses on reducing error in a stepwise manner than relying on overall averaging like random forests.

Boosting with Trees

- In the context of trees, a tree is grown based off a single predictor. At the next step prediction risk is minimized, given a particular loss function. Then a new tree is grown from the previous tree and the process repeats.
- It is because of this sequential minimization of loss that boosting is a good technique, but the model also becomes subject to over fitting the data.
- K-fold cross validation helps with this problem. Each bootstrapped sample is split into k equal subsets, $k - 1$ subsets are used to build the model and the k th is used to test the model. This repeats for all $\binom{k}{k-1}$ iterations.
- The R tool for building boosted tree models is **gbm**.

Selected Models

- Both randomForest and gbm models were built using the defined predictors.
- Gbm models proved to have more accurate predictive power than randomForest models.
- A classification gbm model was built for predicting retention and a regression gbm model for predicting college GPA.

Influence of Predictors

Table : Enrollment

Predictor	Relative Influence
PGMS_PROGRAM_DESCR	38.2738530
HS_FOUR_PT_GPA	35.4288658
HOME_STATE_KEY	14.1636557
APPLICATION_DT	4.4703879
NUM_REFERRALS	2.9500166
IS_PARENT_HIGHER_ED_GRAD	2.1336947
BEST_CONCORDANT_ACT_SCORE	0.8589485
HS_CLASS_SIZE	0.7175438
APPL_HONORS_STATUS_EN	0.6372876
SADV_ACAD_ORG_UNIT_KEY	0.1938790
NUM_RELATIONS	0.1039076
TEST_ACT_CNT	0.0338477
IS_FAFSA_SUBMITTED	0.0192004
SADV_ADVISOR_ID	0.0058601
IS_2YEARS_ONE_LANG	0.0053226
Missing_HS_FOUR_PT_GPA	0.0037293
PROGRAM_COLLEGE_ACAD_NAME	0.0000000
PGMS_OBJECTIVE_KEY	0.0000000
SELF_REPORTED_RESIDENCY_DESCR	0.0000000
IS_ATHLETE	0.0000000
IS_SECOND_MAJOR	0.0000000
ENGLISH_PROF_EXAM_DESCR	0.0000000
IS_2YEARS_EACH_TWO_LANGS	0.0000000
TEST_SAT_CNT	0.0000000
Missing_HS_CLASS_SIZE	0.0000000

Table : GPA

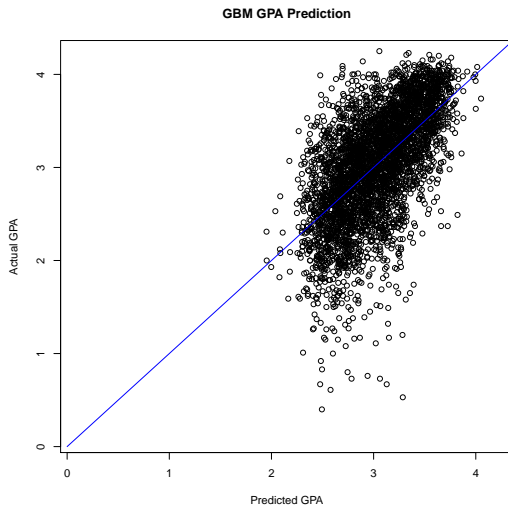
Predictor	Relative Influence
HS_FOUR_PT_GPA	48.9638741
PGMS_PROGRAM_DESCR	22.8213196
APPL_HONORS_STATUS_EN	9.7029570
HOME_STATE_KEY	6.8500780
BEST_CONCORDANT_ACT_SCORE	5.5747054
HS_CLASS_SIZE	2.8318357
APPLICATION_DT	1.8863117
NUM_REFERRALS	0.5950821
IS_PARENT_HIGHER_ED_GRAD	0.3934433
NUM_RELATIONS	0.2627024
ENGLISH_PROF_EXAM_DESCR	0.0485402
SADV_ADVISOR_ID	0.0304009
IS_FAFSA_SUBMITTED	0.0153109
TEST_ACT_CNT	0.0094728
TEST_SAT_CNT	0.0076099
SADV_ACAD_ORG_UNIT_KEY	0.0063559
PROGRAM_COLLEGE_ACAD_NAME	0.0000000
PGMS_OBJECTIVE_KEY	0.0000000
SELF_REPORTED_RESIDENCY_DESCR	0.0000000
IS_ATHLETE	0.0000000
IS_SECOND_MAJOR	0.0000000
IS_2YEARS_ONE_LANG	0.0000000
IS_2YEARS_EACH_TWO_LANGS	0.0000000
Missing_HS_FOUR_PT_GPA	0.0000000
Missing_HS_CLASS_SIZE	0.0000000

Predictions of Enrollment

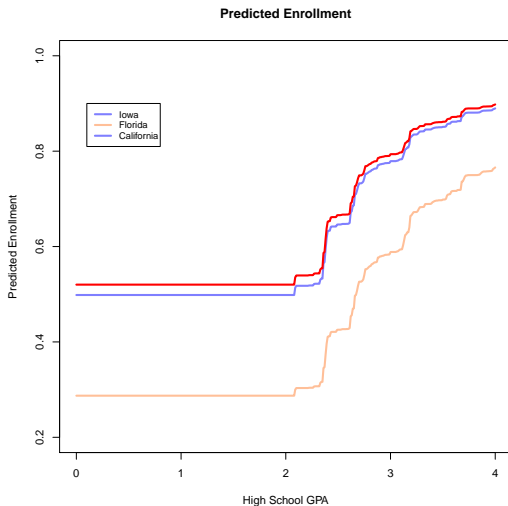
Table : Enrollment Model Performance

	Correct	Incorrect
Retained	0.9966726	0.0033274
Not Retained	0.0158730	0.9841270

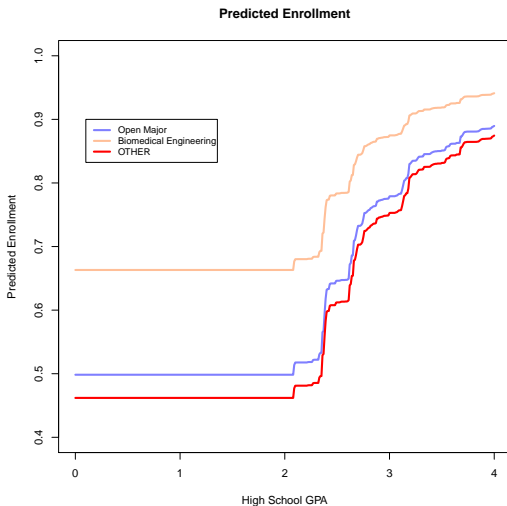
GPA Model Performance



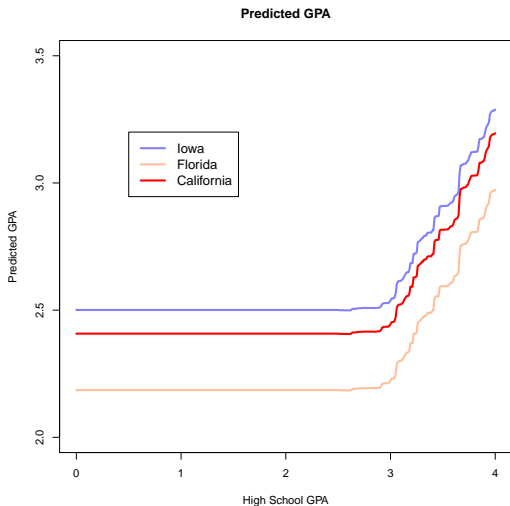
Home State Influence



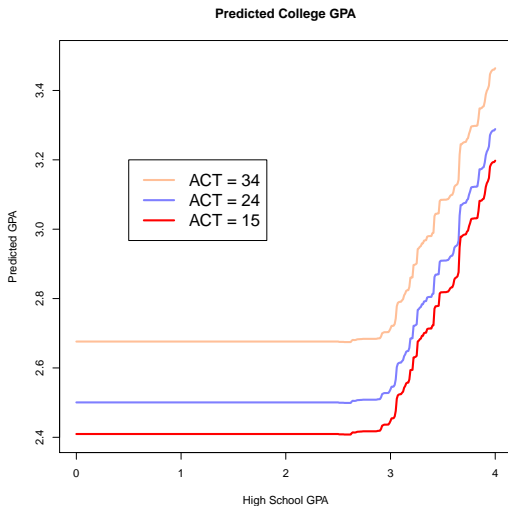
Major Influence



Home State Influence






ACT Score Influence



Conclusions

- The 25 predictor variables may not be completely sufficient in determining first year GPA and whether a student will be retained.
- High school GPA and major appear to have an effect on both retention status and first year college GPA.
- Ideally, we could continue this research and include more predictors from the MAUI database and attempt to gain a more viable predictive model.

References

-  Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani *An Introduction to Statistical Learning: With Applications in R.* ebook.
-  Stanford Online - Lagunita *Statistical Learning.* Web. Online Course.
-  Grant Brown, Knute Carter *Predictive Modelling Techniques with Application to Enrollment Management at UI.* Center for Public Health Statistics; Talk; December 14th, 2015

Acknowledgements

- Iowa Summer Institute in Biostatistics Mentor Staff
- National Institute of Health (NIH)
- The National Heart, Lung, and Blood Institute (NHLBI)
- K.D. Zamba
- Dr. Grant Brown
- Terry Kirk, Miles Dietz, Ann Weber
- Lauren Sager