August 2020

Predicting Incoming Union College Class Profiles using Machine Learning

Jason D'Amico
Union College - Schenectady, NY, damicoj@union.edu

Predicting Incoming Union College Class Profiles using Machine Learning

By Jason D'Amico (Computer Engineering '23) with Professor Tomas Dvorak (Economics) and David Glasser (Admissions)

Introduction

Predictive analytics is increasingly influential on a wide variety of fields and industries. In college admissions, predictive analytics is used to forecast enrollment, tuition revenue, financial aid and other characteristics of the incoming class. In this study, we explore a variety of machine learning models to predict class characteristics at Union College.

Admissions Cycle Timeline

(dates approximated and could experience slight variance in practice)

Before app. Prospective applicants interact with the College (campus events, interviews, etc.)

Jan. 15 **Applications submitted for regular decision.** Basic data received (SAT score, transcript, etc.)

March 1 Initial admission decisions are made, **class profile is generated by model.**

March 22 Final admission decisions released

May 1 Enrollment decisions due

Literature review

Machine Learning Sources:

Burkov, Andriy. *The Hundred-Page Machine Learning Book*. USA: Creative Commons, 2019.

Jolly, Kevin. *Machine Learning with Scikit-Learn Quick Start Guide*. Birmingham, England: Packt, 2018.

Klosterman, Stephen. *Data Science Projects with Python*. Birmingham, England: Packt, 2019.

Predictive Analytics in Higher Education:

Aulck, Lovenoor, et al. "Using Machine Learning and Genetic Algorithms to Optimize Scholarship Allocation for Student Yield." ACM, 2019, http://ml4ed.cc/attachments/Aulck.pdf.

Hayes, J. Bryan, et al. "A Simple Model for Estimating Enrollment Yield from a List of Freshman Prospects." *Academy of Educational Leadership Journal*, vol. 17, no. 2, 2013, pp. 61–68.

Data

We use data on all applicants to Union College over the past 8 years. This includes information on each applicant's demographic, academic and financial characteristics, their admission and enrollment outcomes as well as some measures on their engagement with Union (e.g. campus visits).

The majority of the data in its original form is not correctly formatted to be used in predictive analysis, and thus needs to be restructured. The data manipulation/feature engineering methods include:

- **Binning**, which was applied to fields where the actual value that an applicant held in that field was not as important as the general range it was in relative to the rest of the applicants (ex: SAT score, events attended).
- **Scaling** to better interpret model coefficients.
- **Filling empty values**, either by:
  a. Removing a field from consideration as the concentration of applicants with empty values in said field is too great.
  b. Making an estimate of the empty value based on other field values.

In the final model, 34 features were used.

Modeling

Splitting data into train, test and validation sets

To train the model, only regular decision (RD) admits from 2013-2018 are considered: 2019 admits are used as a holdout set. The 2013-2018 data contained over 12,000 applicants. From this pool, two subsets are created at random: a larger subset to use for model training, with the remainder used for testing and evaluating the performance of the model.

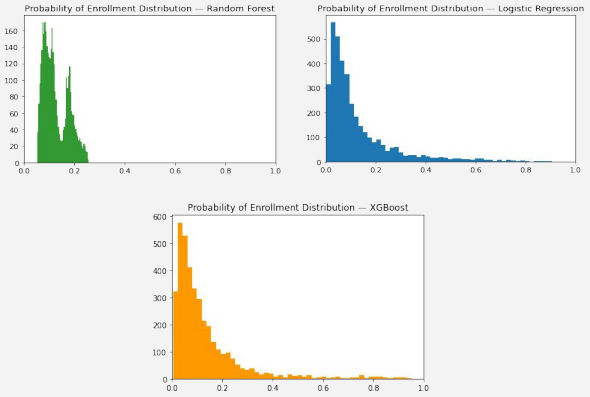
Estimating models

Only classification models are considered in this study due to the binary nature of the problem. Three separate classification models are explored:

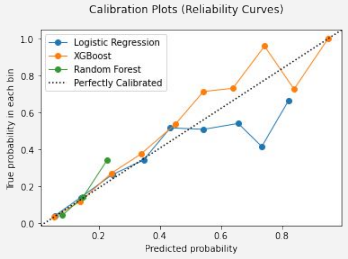
- Logistic regression
- Random forest classification
- XGBoost

For each applicant, the models produce a probability between 0-1 indicating to how likely it is for the applicant to enroll.

The ***histograms*** below show the distribution of the predicted probabilities of enrollment across the training data.

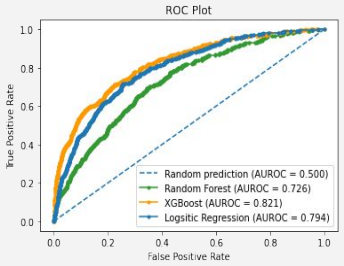


Evaluating Models



A calibration plot, or reliability curve, measures how accurate the predictions made by the model really are. This is accomplished by binning predictions within a given range, determining the actual rate of enrollment within that bin, and comparing it to the predicted rate of enrollment that the bin represents. From the above histograms, we note that the probabilities of enrollment are concentrated primarily towards 0, therefore, the bins corresponding to lesser predicted values are the most indicative of model performance.

Evaluating Models (cont.)



The ***ROC Curve*** is a measure of the model's effectiveness of predicting whether or not someone enrolls. The larger the area under the curve, or AUC, the better or more accurate the algorithm. The dotted blue line corresponds to a random prediction (AUC of 0.5), and a line that perfectly hugs the top left corner of the plot corresponds to a perfect prediction (AUC of 1.0). Generally speaking, an AUC within the range of 0.8-0.9 is considered to be an exceptional predictor.

Creating Class Profile

Given the estimated probabilities of enrollment for each admitted student, we create an expected class profile for the test data set. This is accomplished by multiplying the probability of enrollment for an applicant by a certain metric: for instance, if the applicant has a 20% chance of enrolling, then that counts for 0.2 of a person in the expected class profile. For each predicted characteristic of the class, a confidence interval is calculated. An example of an XGBoost-created class profile for the test set (approximately 3,000 applicants) is shown below:

```
TOTAL ENROLLED
Predicted: 499.53918
Actual: 508

FEMALE
Predicted: 246.08022890496068
Actual: 249

INTERNATIONAL
Predicted: 62.678400413598865
Actual: 78

MINORITY
Predicted: 82.19842721452005
Actual: 90

COLLEGE COST
Predicted: 9391621.137821859
Actual: 10183282.0

COLLEGE RECEIVES
Predicted: 22359899.133240294
Actual: 22125158.0

TESTING CONSIDERED
Predicted: 378.5473284062464
Actual: 373

SAT TOTAL (AVG)
Predicted: 1367.510504150508
Actual: 1360.212765957447

ACT TOTAL (AVG)
Predicted: 29.76716688521226
Actual: 29.60355029585799
```

Conclusion

In mid-August, 2020, the model and its performance on the 2019 holdout data will be presented to the Admissions office to determine whether this model can replace existing methods provided by an external vendor. Given the performance metrics shown by the calibration plot and the ROC curve, we expect that the XGBoost model will be an improvement over existing methods, saving the College money while improving its ability to forecast the enrollment and financial impacts of its admission decisions.