Jeremy Miller
November 5, 2018

Machine Learning for Learning Success

Goals

My project is to develop a machine learning model (or ensemble of models) to more rapidly and effectively predict success in online college level courses. The first target of my model will be to predict the probability that a given student will successfully complete a given online course. My metrics for success will the ROC-AUC score, recall, precision, and accuracy. Time permitting I will seek a creative and relevant way to develop a Profit analysis, since success in this case is measured not solely in monetary terms. The second target will be to predict that student's final grade in that particular course. Here success will be measured by the Root-Mean-Squared-Error. A third aim may be to identify groups of student via clustering. This could potentially increase the understanding of student online behavior and be helpful for devising economical methods of intervention. Success for this clustering will be defined by calculating the Within-Cluster Sum of Squares and the Gap statistic.

Motivation

In 2012 Sebastian Thrun prophesied that in 50 years there would be only 10 higher education institutions in the world due to the proliferation of the Massive Open Online Courses. Though they have yet to take over higher education on this scale, the proportion of college courses taken online continues to grow. It is reasonable to assume that going forward, online modalities will be used for a significant portion of college level coursework. Relative to traditional "live" courses, non-completion rates in online learning are high and highly varied. Factors that contribute to completion and success need greater definition. Many colleges have implemented early warning systems which can trigger interventions for struggling students.

Though early alert has been shown in several studies to be correlated with higher grades and completion rates, the success of these systems is mixed and there are no industry standards.

Data and Process

I will be using the the Open University Learning Analytics dataset published by the Open University in the United Kingdom in 2015 and updated in 2017. Time permitting I will attempt to integrate the HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset. A breakdown of project steps is as follows:

- Acquire and explore data and documentation
- Clean and join data as appropriate
- Create initial baseline models; use these to determine modeling direction
- Engineer features where possible / practical
- Tune and develop model for predicting completion
- Determine how features contribute to classification
- Tune and develop model for predicting final grades (if possible)
- Determine how features contribute to regression
- Wrangle and integrate Harvard-MIT data set (time permitting)
- Write formal analysis of model results
- Identify possible next steps and directions
- Build web app for visualization and demonstration
- Present project to public amazement

Tools

I will use Python as my programming language and make use of its data science libraries such as Numpy, Pandas, Matplotlib, Sciki-Learn, and Statsmodels. I plan to develop my pipeline on a small scale on my local machine and then execute my model fitting and tuning on an Amazon EC2 instance. The data are not large in terms of file size, but some of the tables contain a large number of rows (from several thousand up to about ten million), suggesting that the transformations, fitting, and prediction will require substantial computing

power. If necessary I will use the Spark Machine Learning Library across an Amazon EMR cluster, though I do not expect this to be the case. I will then use Flask and other tools as necessary to create a web web app for visualizing the results of my model(s). Time permitting, I will make interactive visualizations using a library such as Plotly or Bokeh. I strongly believe in power of well-designed, interactive visualizations to convey complex concepts, particularly to a non-technical audience.

Conclusions

I will conclude my analysis with specific recommendations regarding the results and future development and / or use of my model(s).

References

Aragon, Steven and Elaine Johnson. *Factors Influencing Completion and Noncompletion of Community College Online Courses*. The American Journal of Distance Education, 22: 146-159, 2008.

Choi, Hee Jun and Ji-Hye Park. *Factors Influencing Adult Learners' Decision to Drop Out or Persist in Online Learning*. Education Technology and Society, October 2009.

Glenn, Ivonne. *The Role of Self-efficacy in Self-Regulation Learning in Online College Courses.* Dissertation: Northcentral University. San Diego, CA; April 2018.

Leckart, Steven. *The Stanford Education Experiment the Could Change Higher Learning Forever*. Wired: Science, March 20, 2012.

Siegel, David M. *Should You Bother Reaching Out? Performance Effects of Early Direct Outreach to Low-Performing Students.* The Impact of Formative Assessment: Emphasizing Outcome Measures in Legal Education, March 3, 2017.