Predicting Enrollment with Stacked Models: Bringing Together Theory and Empirics

Haroon Atcha

6/4/2021

# Introduction

Enrollment forecasting is one of the most consequential tasks a given Institutional Research department will undertake. The pressures of getting it ‘right’ are especially pronounced in the community college setting. As open enrollment institutions that are uniquely reliant on public funding, accurate enrollment forecasts are critical for guaranteeing adequate state-level funding and proper planning. Moreover, community colleges operate under unique data availability constraints and admission processes that make the task difficult.

Despite the importance of this task, the extant literature on enrollment forecasting - especially in the community college context - is slim and relatively disconnected from theoretical advances in the study of the predictors of retention and enrollment. This manifests itself in two main problems.

First, practitioners operating in open enrollment contexts have little recourse to literature that speaks to their specific circumstances. Much of the literature has been generated by practitioners in selective-enrollment institutions and speaks to issues particular to that context. (Aksenova, Zhang, and Lu 2006), (Chen 2008), (Nandeshwar and Chaudhari 2009), and (Slim et al. 2018) for instance, exclusively apply their models to the selective enrollment context. Indeed, X’s models are built to specifically predict the likelihood of enrollment *after* acceptance; a process that has no direct analogue in the open enrollment context. Work addressing the problem in the community college context is largely dated. While a highly useful survey of forecasting method used at community colleges, (Bender 1981)’s work is now several decades old and understandably is not reflective of modern forecasting tools and approaches. Likewise, (Lawrence 1980) is useful as a snapshot of best forecasting practices of the time period, but is similarly dated. (Pennington, McGinty, and Williams 2002) represents the most recent and comprehensive project addressing this issue. Numerous dissertations (see: x, y, and z) have reproduced and expanded this work in varying local contexts, adding important contextual knowledge but few explicitly addressing modeling choices. In all, the literature that speaks to predictive modeling of enrollment in the community college context is limited.

Second, predictive models of enrollment do not reflect our theoretical understanding of the processes that are being modeled. Much of the modeling in the extant literature falls into one of two camps: time series approaches at the aggregate level (ARIMA) or linear modeling techniques (OLS, SVG, etc.). While such approaches produce perfectly acceptable predictions, such models are rarely reflective of the theoretical processes they attempt to capture. (Chen 2008), for instance, compares predictions from both, ARIMA and linear models though uses the aggregate number of enrollments in a given semester as the target variable despite noting that ‘[…] the portion of a university’s student enrollees (freshmen) depends on the number of high school graduates within the state […] [and] the same pool of eligible enrolled or returning students (one-year lagged OSU enrollment).’ The recognition that two distinct sub-populations constitute the population of enrollees is not reflected in the modeling which treats each semester’s cohort uniformly[[1]](#footnote-20). By contrast, (Trusheim and Rylee 2011) does model returning and new enrollees separately, and while this serves as a good example of using theoretical knowledge as a foundation for in empirical exercises, the methods used to achieve predictions are fairly unsophisticated.

Here, I argue that, in the interest of predictive accuracy and knowledge accumulation, our statistical models ought to be theoretically informed. To demonstrate this, I present a hypothetical example with generated data, an approach widely used in the study of statistics. I show that bringing theoretical knowledge to bear on predictive problems - even if that knowledge is tentative or incomplete - is a fruitful avenue for improving forecast accuracy.

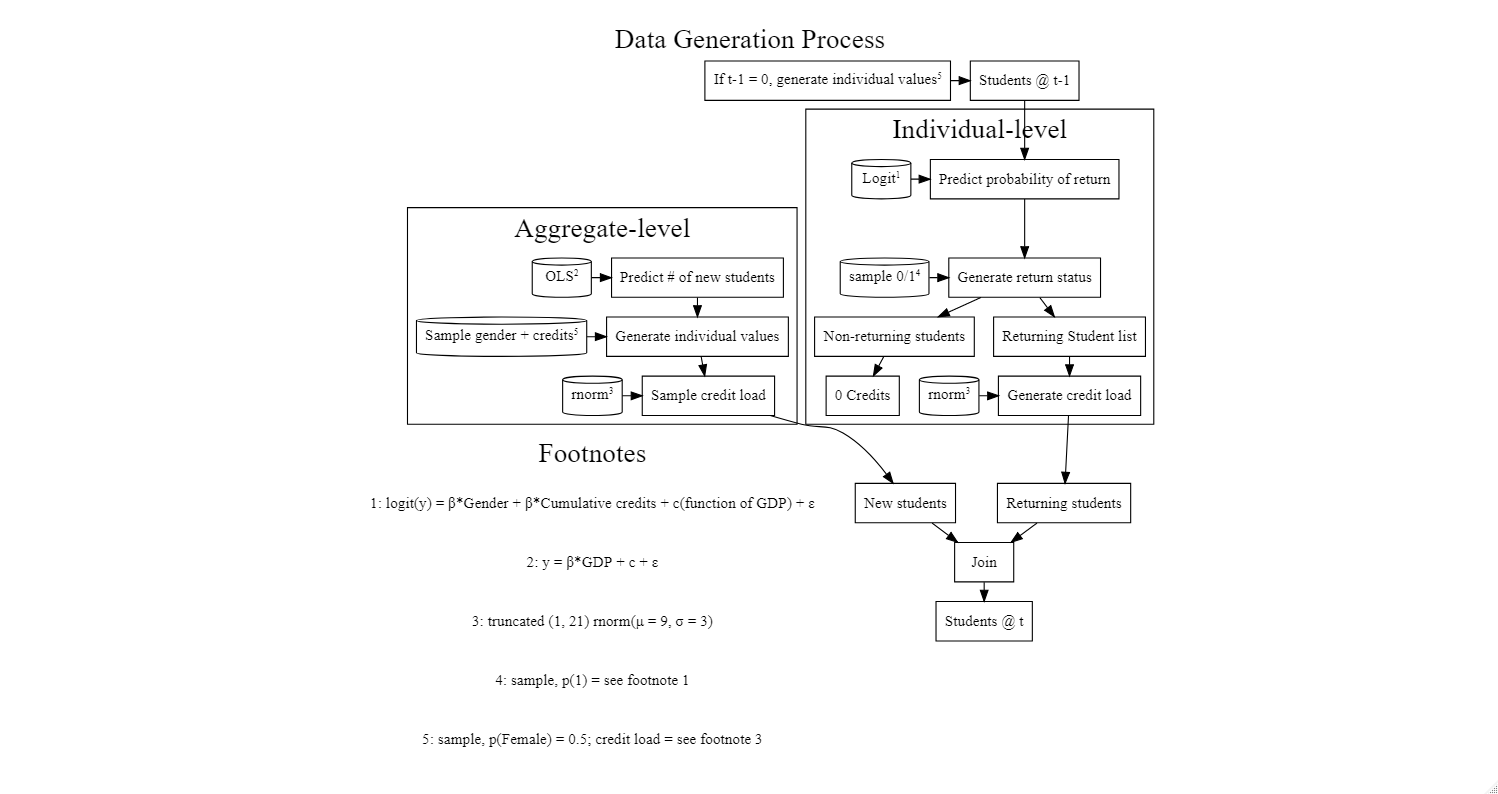
## Simulating Data

The data used in this paper was generated specifically for the purpose and was **not** gathered at any specific community college. This was done for several reasons:

1. Simulated data are generated with *known* parameters, allowing us to determine how accurately models recover *true* estimates
2. Simulated data allow for generalizability. Data generating processes may differ dramatically across institutions.
3. Simulated data allow us to demonstrate theoretically important concepts by assuring these concepts are the *only* relevant ones in our model.

## Data Generating Process

The data used in this study was generated through two processes. First, the number of new enrollees in a given semester was generated at the aggregate level (i.e. total number of new students). Second, the retention status of enrolled students in a given semester was generated at the individual level (i.e. did student n return at t + 1). Finally, the individual level predictions were aggregated and summed with the number of new enrollees to produce total headcount for a given semester. This process is visualized in chart 1.



### Returning Enrollees

The retention status of currently enrolled students was generated at the individual level. Three variables are observed for each student: Gender, current semester credit load, and cumulative credits taken. Each student’s likelihood of returning in the subsequent semester is a linear function of their gender and curvilinear function of cumulative credits taken such that students are *more* likely to be retained as they approach graduation and *less* likely thereafter. Constant and error terms are also included.

The formula used to generate each student’s probability of returning is given below; the coefficients associated with each term as well as the values for the constant and error parameters can be found in the Appendix. As noted above, I chose to generate data using a simple model for the purpose of demonstration; Gender and cumulative credit load are widely cited predictors of retention.

logit(y) = + + + *C* +

### New Enrollees

The number of new enrollees in a given semester was generated at the aggregate level and is given in heuristic form below. The number of new enrollees in a given semester is a linear function of change in GDP and semester such that higher GDP growth is associated with greater enrollment. Moreover, Fall, Spring, and Summer semester terms each add a constant number of new enrollees in decreasing order. Results from the model were then de-normalized and rounded to generate an integer number of new enrollees. The coefficients associated with each term as well as the values for the constant and error parameters can be found in the Appendix. As with the individual level data, local GDP change and semester effects are widely documented in the literature[[2]](#footnote-24).

y = + + + + *C* +

New Students = ( \* y) +

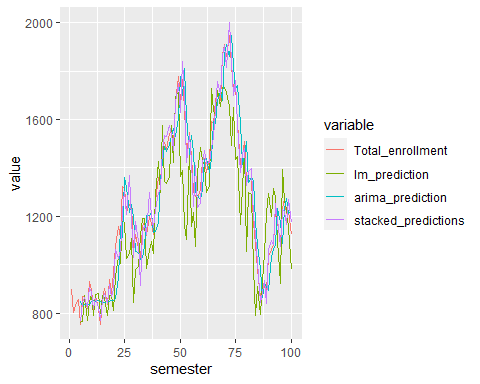
# Methods

Modeling choices ought to reflect our underlying knowledge about the data generating process. In this example, we know that the processes that generated our observations vary across sub-populations. Different generative models determine the number of new enrollees and returning students in a given semester. Rather than fitting a single model to the aggregate count of enrollees, which would entail a misalignment between theory and practice and require us to throw away a large amount of information at the individual level, we can estimate a ‘stacked’ model instead, fitting different models to different sub-populations and aggregating those predictions.

To demonstrate the value in building theory-informed predictive models, I fit and compare several such models. I test autoregressive integrated moving average (ARIMA) and simple linear models at the aggregate level, as these are the most common approaches in the extent literature[[3]](#footnote-28) and compare them to three stacked models. For the first stacked model, I predict the number of returning students at the individual level using logistic regression and use that number as a linear predictor of the *total* number of enrolling students at semester t + 1. Second, I fit a theoretically misspecified model, predicting the number of returning students at the individual level using *only* lower order terms in a logistic regression and the number of new students using ARIMA. Finally, I fit a correctly specified model which differs only insofar as it uses a correct higher order term in the individual prediction section.

I compare the results of these models using the Mean Absolute Percent Error (MAPE), Mean Absolute Standard Error (MASE), and Root Mean Squared Error (RMSE), three widely used measures of model accuracy.

# Results



|  |  |  |  |
| --- | --- | --- | --- |
| Model | MAE | MAPE | RMSE |
| Linear Model | 120.44 | 9.36% | 155.95 |
| ARIMA | 76.16 | 6.12% | 95.54 |
| Stacked Model | 33.91 | 2.72% | 44.38 |
| Stacked Model w/ Polynomial | 34.68 | 2.78% | 44.94 |

# Discussion

A skeptical reader may justifiably question the value in the above exercise. After all, of course the better specified model predicted more accurately, we *knew* the true parameters and data generating process before starting. This is, however, exactly the point. Often, practitioners approach enrollment prediction from a point of imposed ignorance. Brute force and automated methods of feature selection are treated as substitutes for theoretically informed modeling choices[[4]](#footnote-32). In the community college context, where data availability can be limited and the process that generates enrollment data is idiosyncratic in known ways, discarding such knowledge has practical consequences for the accuracy of our enrollment forecasts and consequently, on resource availability. Practitioners ought to take pains to incorporate theoretical knowledge in predictive forecasts. Employing stacked models, as shown above, can be a simple manner of doing so.

# Appendix A: Coefficients, Parameters, and Variable Values

## Coefficients and Model Parameters

|  |  |  |
| --- | --- | --- |
| Term | Model | Value |
|  | New | 2 |
|  | New | 3 |
|  | New | 0 |
|  | New | 6 |
|  | New | 0 |
|  | New |  |
|  | Returning | 0.1 |
|  | Returning | 0.02 |
|  | Returning | 0.9 |
|  | Returning |  |

## Individual Variable Summary Statistics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Level | Min | Max | Mean | Generation | Parameters |
| Gender | Binary | 0 | 1 | 0.5 | Sample (0:1) | P(1) = 0.5 |
| Credits | Interval | 1 | 21 | 9 | Sample Truncated Normal | = 6, = 9 |
| Cumulative Credits | Interval | 1 | 121 | 135 |  | - |
| Likelihood of Return | Ratio | 0.02 | 0.83 | 0.72 | Linear Function | See Equation 1 |
| Return | Binary | 0 | 1 | 0.72 | Sample (0:1) | P(1) = Likelihood of Return |

# Appendix B: Cut Parts

In this paper, I have demonstrated the value in bringing theoretical knowledge to bear on predictive models of enrollment. Community colleges construct such models under unique constraints. Moreover, the financial consequences of inaccurate models is more acutely felt than at other types of institutions of higher education. While the extant literature has sought improvements to predictive accuracy by using ‘brute-force’ methods - e.g. increasing the number of features in models, implementing more model types, exhaustively searching across model hyper-parameters - I argue that these efforts essentially re-invent the wheel insofar as they ‘rediscover’ relationships that are already well established. Our theoretical knowledge of the predictors of enrollment and persistence is robust. That knowledge should be reflected in our empirical models.

# Appendix C: Summary Statistics

Aksenova, Svetlana S., Du Zhang, and Meiliu Lu. 2006. “Enrollment Prediction Through Data Mining.” In *2006 IEEE International Conference on Information Reuse Integration*, 510–15. <https://doi.org/10.1109/IRI.2006.252466>.

Bender, Louis W. 1981. “Community College Enrollment Projection Study: A National Survey of Approaches Used by State Agencies for Community/Junior Colleges,” March. <https://eric.ed.gov/?id=ED208930>.

Chen, Chau-Kuang. 2008. *An Integrated Enrollment Forecast Model. IR Applications, Volume 15, January 18, 2008*. Association for Institutional Research. <https://eric.ed.gov/?id=ED504328>.

Lawrence, Sue. 1980. “Forecasting Enrollments in a Virginia Community College.” *Dissertations, Theses, and Masters Projects*, January. https://doi.org/<https://dx.doi.org/doi:10.25774/w4-gd66-bc63>.

Nandeshwar, Ashutosh, and Subodh Chaudhari. 2009. “Enrollment Prediction Models Using Data Mining,” April.

Pennington, Kevin L., Dixie McGinty, and Mitchell R. Williams. 2002. “Community College Enrollment as a Function of Economic Indicators.” *Community College Journal of Research and Practice* 26 (5): 431–37. <https://doi.org/10.1080/02776770290041783>.

Pfitzner, Charles Barry. 1987. “A Short‐Term Forecasting Procedure for Institution Enrollments.” *Community Junior College Research Quarterly of Research and Practice* 11 (3): 141–52. <https://doi.org/10.1080/0361697870110301>.

Slim, Ahmad, Don Hush, Tushar Ojah, and Terry Babbitt. 2018. *Predicting Student Enrollment Based on Student and College Characteristics*. International Educational Data Mining Society. <https://eric.ed.gov/?id=ED593221>.

Trusheim, Dale, and Carol Rylee. 2011. “Predictive Modeling: Linking Enrollment and Budgeting.” *Planning for Higher Education* 40 (1): 12–21. <https://www.proquest.com/docview/906339461/abstract/36613B5AD823484CPQ/1>.

1. See also (Pfitzner 1987) for an early implementation of time series modeling in the community college context that approaches the problem at the aggregate level [↑](#footnote-ref-20)
2. Note that, since GDP data were *not* gathered from actual sources and were instead generated using general sampling process, it could just as easily be interpreted as the percent change in number of graduating high school students. For ease of interpretation and generalizabiilty, I choose to present the variable as change in GDP. [↑](#footnote-ref-24)
3. Exponential smoothing approaches are also common early in the enrollment forecasting literature; X, for instance, lists such an approach as one of the most common ways community college practitioners approach modeling. However, this approach sees little use - at least in its simple form - among contemporary practitioners who prefer other methods of time series modeling. For this reason, I elect not to include an exponential smoothing model in my comparison [↑](#footnote-ref-28)
4. This is not to argue that automated and brute force methods of feature selection are inherently bad. Note that the ARIMA model fit in the stacked version above iterates through all combinations of parameters to find the best fit. Such methods are powerful tools but they are best leveraged when informed by theoretically grounded modeling choices. [↑](#footnote-ref-32)