COMMUNITY COLLEGE JOURNAL OF RESEARCH AND PRACTICE

Volume 34
Number 8
August 2010

Predicting Enrollment with Stacked Models: Bringing Together Theory and Empirics

Journal:	*Community College Journal of Research and Practice*
Manuscript ID	Draft
Manuscript Type:	Exchange Article
Keywords:	Quantitative, Research, Student Affairs, Students

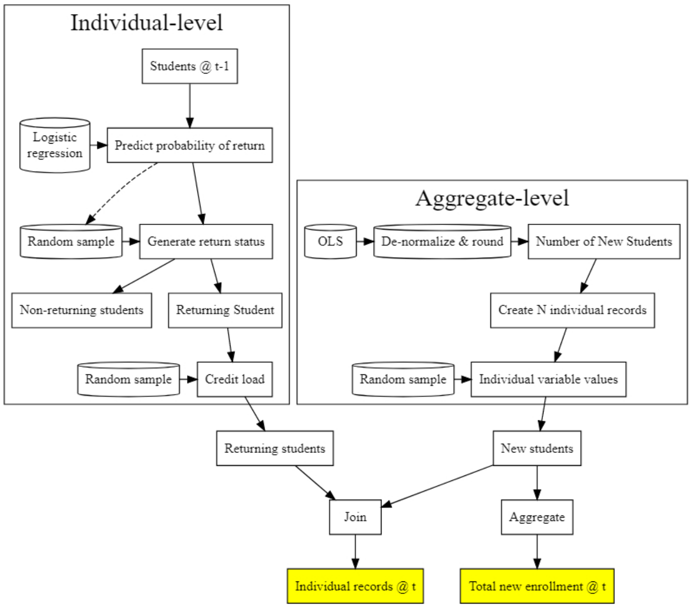


Figure 1

800x510mm (38 x 38 DPI)

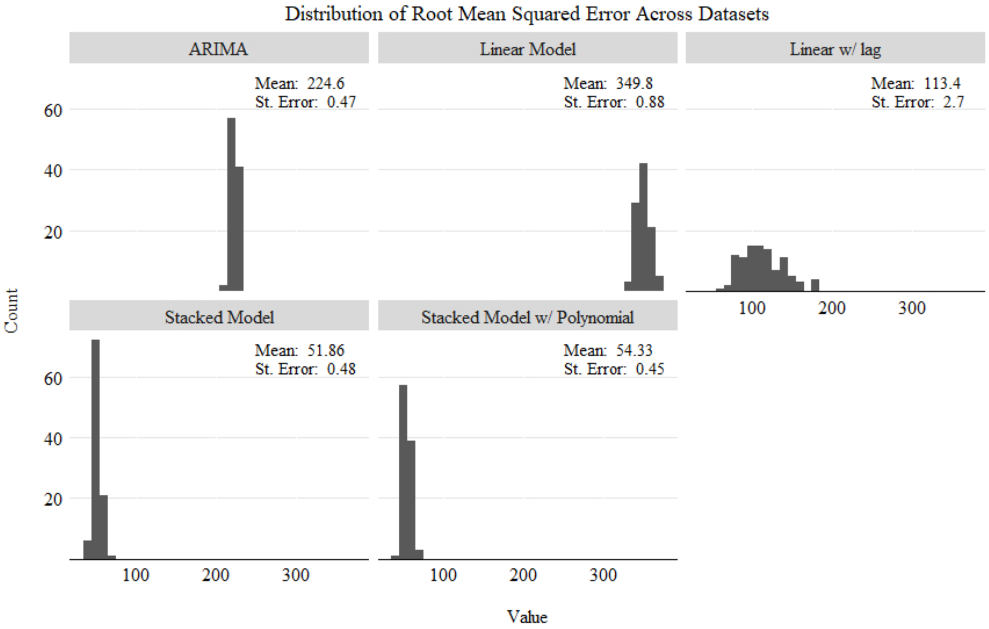
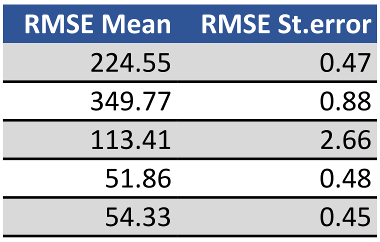


Figure 2

576x367mm (38 x 38 DPI)

Model	MAE Mean	MAE St.error	MAPE Mean	MAPE St.error
ARIMA	109.84	0.48	6.47%	0.03%
Linear Model	246.73	0.59	14.48%	0.04%
Linear w/ lag	43.78	0.46	2.47%	0.02%
Stacked Model	40.5	0.42	2.43%	0.03%
Stacked Model w/ Polynomial	42.26	0.39	2.53%	0.02%



RMSE Mean	RMSE St.error
224.55	0.47
349.77	0.88
113.41	2.66
51.86	0.48
54.33	0.45

Introduction

Proper institutional planning requires accurate enrollment forecasts. This is especially true in the community college context given open enrollment policies and relative reliance on public funds. Despite the importance of this task, the extant literature is slim. Much of the literature has been generated by practitioners at selective-enrollment institutions and speaks to issues particular to that context[1]. Moreover, the work is largely disconnected from theoretical advances in the study of retention and enrollment. Predictive models rarely reflect our knowledge about the processes being modeled[2].

One way to address these shortcomings is to incorporate theory into predictive modeling by building 'stacked' models. As opposed a single model which predicts enrollment at the aggregate level, a stacked model is an aggregation of models that predict sub-populations. Put simply: new and re-enrolling students follow different paths to enrollment and, more importantly, we have data at different levels of aggregation about them. Modeling these differences can help us make better use of available data and produce more accurate and less volatile forecasts.

Making the Case

To make the case for stacked models, I compare them to forecasting approaches commonly cited in the literature. The data used in this paper was simulated specifically for the purpose and was **not** gathered at any specific community college. This approach allows us to:

[1] See (Aksenova et al., 2006), (Chen, 2008), (Nandeshwar & Chaudhari, 2009), and (Slim et al., 2018) for recent literature in the selective-enrollment context and (Bender, 1981), (Lawrence, 1980), and (Pennington et al., 2002) for work in the community college setting.

[2] Much of the extent literature falls into one of two camps: time series approaches at the aggregate level (ARIMA) or linear modeling techniques (OLS). See (Chen, 2008) for work that compares both and (Trusheim & Rylee, 2011) for work that approaches modeling distinct processes separately albeit with fairly unsophisticated methods.

1) know the *true* theoretical data generating process
2) generalize findings; Enrollment processes may differ dramatically across institutions
3) demonstrate important concepts in a controlled environment

This simulated enrollment data was generated through two processes. The number of new enrollees in a semester was generated at the aggregate level (i.e. total number of new students given the semester and GDP change) and the retention status of enrolled students was generated at the individual level (i.e. did student n return at t + 1 given gender and polynomial term for cumulative credits). Finally, the individual level predictions were aggregated and summed with the number of new enrollees to produce total enrollment. This process is visualized below. Nodes in yellow are 'outputs' of the generative process which are used in the model fitting section.

[FIGURE 1]

Method

Modeling choices ought to reflect our knowledge about the underlying data generating process. In this example, we know that the processes that generated our observations vary across sub-populations. 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 individual information about potentially returning students. Instead, we can estimate a 'stacked' model, fitting different models to different sub-populations and aggregating those predictions. For each semester, I train my models on all data from the first semester through the current semester. I then predict enrollment one semester ahead. I repeat this 100 semesters out.

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 and compare them to two stacked models. For the first stacked model, I fit a theoretically mis-specified model, predicting the number of returning students at the individual level using a non-polynomial 'Cumulative Credits' term in a logistic regression and the number of new students using ARIMA. For the second, I fit a correctly specified model with the correct polynomial term.

I compare the results of these models using the Mean Absolute Error (MAE), Mean Absolute Percent Error (MAPE), and Root Mean Squared Error (RMSE), three widely used measures of predictive accuracy. To ensure results are not the product of random chance, I fit each model to 100 uniquely generated data sets and present the mean value and standard error for each measure of model accuracy.

Results

While all tested models predict enrollment with some degree of accuracy, there are large differences between models. Simple linear models perform significantly worse across all measures of predictive accuracy, with large mean MAE, MAPE, and RMSE scores. Simple ARIMA models are only a moderate improvement over linear models.

[TABLE 1]

Comparisons between linear models with lag terms and stacked models illustrates the critical point of this exercise. While linear models with lag terms perform essentially on par with the stacked models in terms of MAE and MAPE, linear models perform significantly worse on RMSE and are more volatile in this regard. In essence, when linear models with lag terms get it wrong, they get it much 'wronger' than stacked models. Additionally, across datasets, the prevalence of highly 'wrong' values varies considerably with linear models that include lag terms; hence its volatility. Only stacked models predict enrollment precisely and accurately.

[FIGURE 2]

Finally, it should be noted that the correctly specified stacked model does not outperform the parsimonious stacked model though differences are statistically negligible across all measures of predictive accuracy. Neither are volatile in the same way that linear models with lag terms are.

Implications

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[3]. 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 and effective manner of doing so.

[3] 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

Works Cited

Aksenova, S. S., Zhang, D., & Lu, M. (2006). Enrollment Prediction through Data Mining. *2006 IEEE International Conference on Information Reuse Integration*, 510–515. https://doi.org/10.1109/IRI.2006.252466

Bender, L. W. (1981). *Community College Enrollment Projection Study: A National Survey of Approaches Used by State Agencies for Community/Junior Colleges*. https://eric.ed.gov/?id=ED208930

Chen, C.-K. (2008). An Integrated Enrollment Forecast Model. IR Applications, Volume 15, January 18, 2008. In *Association for Institutional Research (NJ1)*. Association for Institutional Research. https://eric.ed.gov/?id=ED504328

Lawrence, S. (1980). Forecasting enrollments in a Virginia community college. *Dissertations, Theses, and Masters Projects*. https://dx.doi.org/doi:10.25774/w4-gd66-bc63

Nandeshwar, A., & Chaudhari, S. (2009). *Enrollment Prediction Models Using Data Mining*.

Pennington, K. L., McGinty, D., & Williams, M. R. (2002). Community College Enrollment as a Function of Economic Indicators. *Community College Journal of Research and Practice*, *26*(5), 431–437. https://doi.org/10.1080/02776770290041783

Slim, A., Hush, D., Ojah, T., & Babbitt, T. (2018). Predicting Student Enrollment Based on Student and College Characteristics. In *International Educational Data Mining Society*. International Educational Data Mining Society. https://eric.ed.gov/?id=ED593221

Trusheim, D., & Rylee, C. (2011). Predictive Modeling: Linking Enrollment and Budgeting. *Planning for Higher Education*, *40*(1), 12–21.