Of Stacks and Muses: Adventures in Learning Analytics at Marist College

Eitel J. M. Lauría[(*)] [1], Edward Presutti [(**)], Maria Kapogiannis[(**)]
[(*)] *School of Computer Science & Mathematics, Marist College, Poughkeepsie, NY, USA*
[(**)] *Data Science & Analytics Group, Marist College, Poughkeepsie, NY, USA*
{Eitel.Lauria, Edward.Presutti, Maria.Kapogiannis}@marist.edu

Abstract: We describe the development of MUSE, an early-alert system of academically at-risk students implemented at Marist College, based on a stacked ensemble of classifiers. Experimental tests are carried out to demonstrate the predictive performance of the stack when making predictions on college-wide data.

Keywords: stack ensembles, machine learning, learning management systems, early alert systems, higher education

1. Introduction

The numbers are appalling: according to The Chronicle of Higher Education College Completion website, in the United States the average six-year degree completion across all four-year institutions, of those students starting bachelor degree programs, stands at 58% for public institutions to 65% for private institutions, with percentages plummeting when considering black student populations (40%; 44%) or Hispanic student populations (50%; 61%). Four-year graduation rates are even more worrying (see College Completion, https://collegecompletion.chronicle.com, for more details).

Will artificial intelligence come to the rescue? At Marist College we think so. And we are working hard to that effect. MUSE—aka Marist Universal Student Experience—is the code name of the learning analytics project recently implemented at Marist College. MUSE is a predictive modeling application based on machine learning techniques that provides early alert and detection of academically at-risk students. The system uses data from previous semesters to train statistical learning models that then make predictions of student academic performance a few weeks into the semester.

In the last decade, a number of research projects and initiatives have flourished that tackle the issue of monitoring student academic performance (Pistilli and Arnold, 2010; Smith et al, 2012; Jayaprakash et al., 2014; Lauría et al, 2016). Generally, researchers use machine learning algorithms (e.g. linear models, Bayesian learners, maximum margin algorithms, and decision trees) to develop predictive models that try to identify academically at-risk students.

2. Materials and methods

Our current research (Lauría et al, 2018) introduces a stacked ensemble architecture to build early detection models. This machine learning method is mostly absent and minimally referenced in the learning analytics literature. Hence, this work makes two relevant contributions: 1) it provides a methodology for building a stacked ensemble architecture learnt from data; 2) it provides proof of concept of how stacked ensembles can be applied in the context of early detection of academically at-risk students.

---

[1] Corresponding author. Eitel Lauría is a professor and director of graduate programs at the School of Computer Science & Mathematics at Marist College, where he also serves as the senior data scientist of the Learning Analytics Initiative. He was born and raised in Buenos Aires, Argentina.

MUSE uses multiple sources of data to build its models, including learning management system data, course grades and course related data, student aptitude data and student characteristics, including age and sex. MUSE does not use race to make predictions. As for learning management system data, a learning management system is a type of web-based portal implemented by most colleges and universities that manages and records the educational delivery, as well as the interaction of students and instructors in a given course. Although initially developed for online learning, these platforms have gained widespread use in face-to-face settings. MUSE tracks the student interaction with Sakai, the learning management system used by Marist[2]. This student interaction with the learning management system can be regarded as a surrogate for student effort. MUSE also tracks partial contributions to the final grade coming from assignments, tests and other gradable activities recorded by the learning management system. All these data elements extracted from different sources are condensed into one bucket which is subsequently used to build statistical models for student performance analysis and prediction. The models are applied to new incoming student data to make predictions early on in the semester as to how the students will perform.

MUSE has three main components: the predictive modeling engine—the software that trains statistical models and makes predictions on new student data; the data preparation layer which pulls data from multiple sources and feeds them to the predictive modeling engine; and the presentation layer, which uses charts, graphs, and dashboards to display the prediction outcomes. In this presentation we focus on the implementation of the predictive modeling engine. As mentioned before, the predictive modeling engine is built around a stacked ensemble of classifiers, a machine learning technique that uses multiple classifiers to produce predictions, which are in turn fed to a second-stage classifier that delivers the final predictions on student performance. The 'stack' implemented at Marist uses the XGBtree algorithm (XB; Chen and Guerin, 2016), Random Forests (RF; Breiman, 2001), feed-forward neural networks (NN) and Bayesian learners (NB) as base learners. The second-stage model is trained using regularized logistic regression using the the LibLinear library (LOG; Fan et all, 2008) and Logistic model trees (LMT; Landwehr et al, 2005). The MUSE engine implemented at Marist detects at least 85% of the actual at-risk students, when trained with good-quality data (see Table 1).

Stage 1	Stage 1	Stage 2	Mean (AUC)				Stack AUC		paired t-test	Recall	Specif.
Classifiers	Predictors	Classifiers	XB	NN	RF	NB	Mean	SE	pvalue		
XB+NNET+RF	ALL	LMT	0.928	0.920	0.925		0.934	0.003	0.020	0.865	0.873
XB+NNET+RF	ALL	LOG	0.928	0.920	0.925		0.936	0.002	0.000	0.867	0.875
XB+NNET+RF	noCS	LMT	0.846	0.833	0.834		0.855	0.001	0.001	0.792	0.767
XB+NNET+RF	noCS	LOG	0.846	0.833	0.834		0.855	0.001	0.000	0.792	0.770
XB+NNET+NB	ALL	LMT	0.928	0.920		0.858	0.933	0.002	0.057	0.869	0.865
XB+NNET+NB	ALL	LOG	0.928	0.920		0.858	0.933	0.002	0.026	0.870	0.865
XB+NNET+NB	noCS	LMT	0.846	0.833		0.775	0.851	0.002	0.115	0.794	0.757
XB+NNET+NB	noCS	LOG	0.846	0.833		0.775	0.852	0.001	0.021	0.791	0.761

Table 1. Stack Predictive Performance Results

[2] Marist College is a long-time user of Sakai, a popular open source learning management and collaboration system. Other examples of world-renowned learning management systems are Moodle and Blackboard (a commercial system).

Predictive performance improvement is moderate but consistent, and particularly relevant considering the high AUC values displayed by all base classifiers. The stack seems to extract additional predictive performance from its component classifiers. Also, the stacked ensemble cushions weaker performances of its components, promoting more stable predictions when faced with varying characteristics of the data.

3. Conclusion

This work provides first-time insight of the use of a stacked ensemble architecture in the domain of learning analytics and early detection of academically at-risk students. We recognize that much more could be written about each of these topics. However, we will provide more complete coverage of these topics at the workshop. We hope that, as with the muses of Greek mythology, MUSE will be a source of inspiration, in this case to researchers and practitioners in the field of learning analytics, and with it help provide a better path for student success.

References

- Breiman, L., 2001. Random Forests. Mach Learn 45, 5–32.
- College Completion, The Chronicle of Higher Education (https://collegecompletion.chronicle.com/)
- Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. CoRR abs/1603.02754.
- Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008. LIBLINEAR: A Library for Large Linear Classification. J Mach Learn Res 9, 1871–1874.
- Jayaprakash, S.M., Moody, E.W., Lauría, E.J.M., Regan, J.R., Baron, J.D., 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. J. Learn. Anal. 1, 42.
- Landwehr, N., Hall, M., Frank, E., 2005. Logistic Model Trees. Mach. Learn. 59, 161–205.
- Lauría, E.J., Presutti, E., Sokoloff, M., Guarino, M., 2016. Crossing the Chasm to Big Data: Early Detection of at-Risk Students in a Cluster Computing Environment. Presented at the Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK'17)- Practitioner Track. Vancouver, Canada.
- Lauría E., Presutti E, Kapogiannis M., Kamath A., "Stacking Classifiers for Early Detection of Students at Risk", Proceedings of the 10th International Conference of Computer Supported Education (CSEDU 2018), Funchal Madeira, Portugal, March 15-17, 2018, vol 1, 390-397
- Pistilli, M.D., Arnold, K.E., 2010. Purdue Signals: Mining Real-time Academic Data to Enhance Student Success. Campus 22–24.
- Smith, V.C., Lange, A., Huston, D.R., 2012. Predictive Modeling to Forecast Student Outcomes and Drive Effective Interventions in Online Community College Courses. J. Asynchronous Learn. Netw. 16, 51–61.