From a Prediction Model to Meaningful Reports in School

Lief Esbenshade
Google
liefesbenshade@google.com

Jonathan Vitale
Google
jonvitale@google.com

Ryan S. Baker
University of Pennsylvania
ryanshaunbaker@gmail.com

ABSTRACT
Predictive analytics and prediction modeling has emerged as one of the largest applications of educational data mining methods, with particularly widespread application in dropout prediction. In this demo, we present reports from the BrightBytes Early Warning Solution, a widely-used platform providing information on which students are at risk of dropout and why. We show reports for teachers on individual students, the intervention management system they are integrated into, and more aggregated reports for school and district leaders. These reports and integration with intervention management are designed to help districts move from knowing which students are at risk to taking action to reduce student risk and improve outcomes.

Keywords
Prediction modeling, predictive analytics, dropout, teacher reports, school leader reports, early warning systems

1. INTRODUCTION
Predictive analytics and prediction modeling has emerged as one of the largest applications of educational data mining methods. Prominent since the beginning of our field [2], prediction models for education have scaled to application on millions of learners [7, 6]. Their predictions are increasingly used to drive decision-making in schools, from the decisions made by teachers to the decisions made by school administrators. However, simply predicting an important learner outcome is not sufficient to answer the question of what we should do with that information [21]. It is also essential to provide actionable and meaningful information about those predictions to the people using them to make decisions.

Perhaps the widest application of prediction models in education has been the prediction of student dropout -- defined as when a student withdraws from or simply fails to complete a formal educational program such as high school [18] -- and student stopout -- defined as when a student ceases to participate in a learning activity such as a MOOC course [20] or a problem set [3].

Before modern machine learning techniques were available to school systems, districts were already experimenting with using data to identify students at risk of dropping out and targeting supports to those students [19]. However, early methods were not particularly accurate at predicting dropouts [11]. Machine learning methods are now able to predict student dropouts with more accuracy, which returns school systems to the question of how to intervene.

The first widely-used prediction model for dropout was the Chicago Model [1]. Named because of the data set used in its creation and its first location of wide-scale deployment, the Chicago model used a small set of manually derived indicators to predict if a student was at risk of high-school dropout. Taken outside of its original context, it performed relatively poorly [5, 6]. Later work used more advanced machine learning techniques to predict if a student would drop out of high school [e.g. 14, 16, 6, 7]. This work achieved substantially more accurate and precise prediction than the Chicago model [e.g. 6, 7] but at the cost of lower interpretability. Related work predicting stopout in MOOCs [20, 10], higher education [13, 23], and homework assignments [3] has leveraged similar machine learning techniques, producing successful prediction on new or held-out data, but again at some cost to interpretability.

In this work, there has often been a focus on comparing predictive accuracy between different models [5, 7, 23]. This approach has led to increases in our field's ability to successfully identify students at risk of dropping out.

To promote effective action, the next step is then to provide reports about why students are at risk of dropping out to teachers, school counselors, and school leaders. Such reports must support natural processes of inquiry and sense-making among their users [22] and provide information that supports taking actions, whether for a group of students or individual students [22]. There is increasing interest and readiness for reports of this nature among teachers [12], particularly data that goes beyond just test scores [15].

2. BRIGHTBYTES: PREDICTIVE ANALYTICS ON DROPOUT AND POST-SECONDARY READINESS
The BrightBytes platform offers several types of actionable information to K-12 teachers and administrators, including reports on educational app usage, summaries of student surveys, and formative assessment reports. The platform is designed with the goal of supporting educators, administrators, and stakeholders invested in improving the academic and overall well-being of all K-12 students. BrightBytes's platform is designed in partnership with researchers at the American Institutes for Research (AIR ®).

One of the key areas of functionality offered by BrightBytes is its Early Warning Solution, which provides leading indicators and early warnings to stakeholders across K-12 to identify students who meet or exceed school district-specific risk thresholds. The system helps a principal or teacher to determine which students need intervention immediately and allows stakeholders to assess the performance of their school. Models are used to predict both high school dropout and post-secondary readiness (operationalized as enrollment in higher education, post-high school).

For large school districts with a history of high-quality data, a Random Forest model is trained using the past data from the school district where it is applied. However, not all districts have sufficient data available. For those districts, BrightBytes's Early Warning Solution uses an ensemble of Random Forest models, each trained using districts with particularly large data sets and high data quality. These "pillar models" are then ensembled using weightings derived from measures of the similarity between the district that contributed data to the model and the district the model is being used in [7].

Coleman et al. [7] studied the effectiveness of the BrightBytes model when applied to unseen school districts. They found that the model achieved an AUC ROC of 0.813 when applied to 9th-12th grade students and an AUC ROC of 0.740 when applied to 6th-8th grade students. They also found that the BrightBytes model performed substantially better than the widely-used Chicago Model.

Further analysis in Coleman [8] investigated algorithmic bias in this model. The analysis determined that the model performed acceptably for a range of demographic groups, but also noted that there was some variation in predictive accuracy across groups, performing 0.043 worse (AUC ROC) for the least-effectively predicted group than the median across groups.

3. FROM A PREDICTION MODEL TO MEANINGFUL REPORTS

A key step for the use of predictive models in education is the move from simply offering prediction of overall student outcomes to providing insights into why a specific student is at risk [22]. In this case, we are able to distill explanations for specific students' predictions using information from the feature engineering process. Using semantically meaningful features as predictors makes it easier to then communicate the factors placing an individual student at risk (see, for instance, [17]). By contrast, using auto-encoders or other neural network based approaches to distill complex features makes it considerably harder to inspect specific cases and understand why a decision was made. In this case, the use of a theoretically-aligned set of predictors based on past research makes it easier to then translate predictions back to those theories and leverage decades of practice on dropout intervention based on those theories.

Specifically, we analyzed the data and our models to determine simplified relationships and risk thresholds for each of the predictors with high feature importance (which can be distilled automatically for the Random Forest algorithm, a benefit to using this algorithm). We then present the individual indicators and their risk levels along with the overall prediction. For example, a teacher using the system can click on a student identified to be in the highest risk bracket to see a dashboard that shows which features in the student's data are contributing most to their risk label.

The visual, interactive data analytics tool allows administrators and educators to analyze current student data through dashboards. These dashboards provide important context about students' academic, attendance, and behavioral risk, so that users can identify students at risk of falling behind and address each student's areas of need (Figure 1). Our solution provides a holistic view by showing the interventions previously assigned to a student and allowing teachers to share notes and documentation of qualitative concerns about a student.

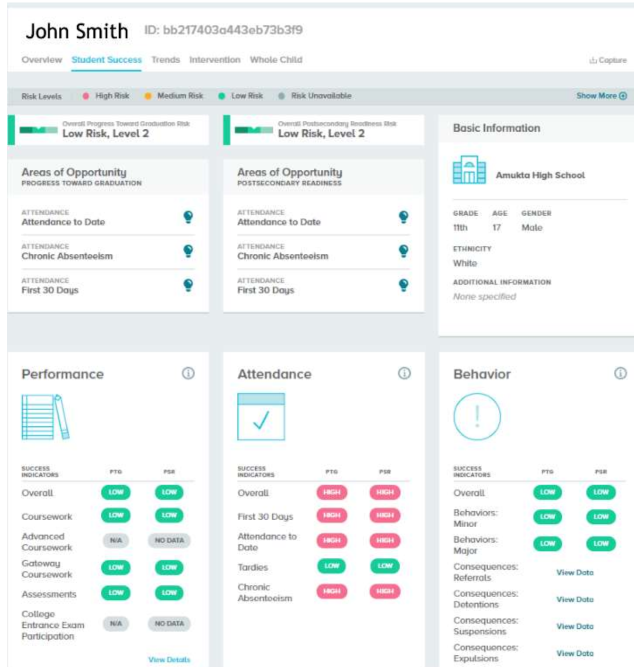

Figure 1. BrightBytes report on a specific student's risk factors

A school district in the midwestern USA had access to a state-provided student risk prediction, but that binary model provided no gradations in risk or information as to what services a student needed. The district was able to use the information provided in the BrightBytes risk predictions to target different kinds of intervention to students using their existing multi-tiered support system (MTSS) [9]. The differentiation provided in the BrightBytes risk model allowed the district to intervene earlier, and in a more cost effective manner, than they could previously.

Those interventions can be managed using intervention reports, which are designed so that teachers, principals, and district leaders can see key information in one place. The reports provide information in a timely fashion, so that educators can intervene early with specific programs like small group targeted reading sessions, one-on-one behavior counseling, and attendance management. They can track change over time with monthly refreshes of student risk predictions and daily updates of student data. The system also provides recommendations for action with specific students, based on the research literature. Principals can identify areas of need across their schools and district staff can view district and school challenges to distribute support resources in a targeted and efficient manner.

The intervention management solution is designed with the intention of bridging the gap between identification and intervention. Intervention management features a cumulative view by showing all student intervention records and documented resources across school years. The Intervention Student Profile allows teachers, principals, and leaders to see all intervention-related data for the student on a single page and is linked directly from the student's risk profile. In one school district in the southern

USA, district school social workers reported that the BrightBytes platform replaced an onerous collection of disparate excel spreadsheets. This allowed them to spend less time identifying which students needed interventions, and more time getting students the help they needed. In one instance they were able to identify and intervene quickly with a student who had a sudden increase in behavior issues at school, helping the student to get back on track to graduate before his grades or attendance faltered.

Additionally, our solution includes overview dashboards that show the district-specific features that are contributing most to students' risk of not graduating. The district overview - which can be filtered by school - disaggregates why students are at risk into three categories for intervention: Performance, Attendance and Behavior. This disaggregation enables school staff to understand which areas are driving risk and what types of interventions should be developed.

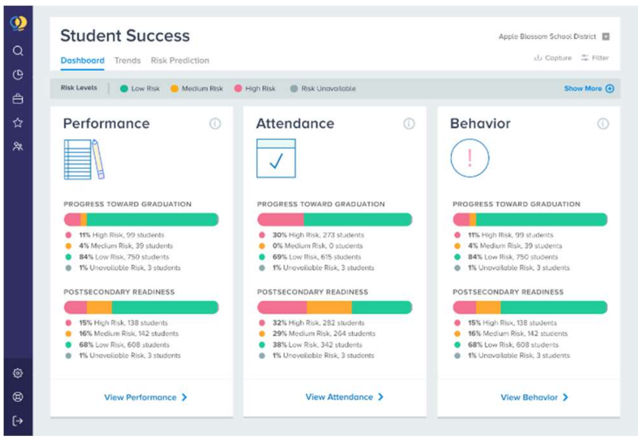

Figure 2. BrightBytes report on a district's overall risk factors

Several districts across the USA have reported using the risk models' school aggregate views to track student transitions into 6th and 9th grade. The districts used this data to allocate counseling and other support resources based on the number of entering students who were predicted to be at higher risk.

4. CONCLUSIONS

Much of the published work on predictive modeling in education (and in data science more generally) focuses on improving and reporting improvement on goodness metrics such as AUC (and RMSE, log likelihood, F1, and other metrics). While this work is important -- we don't want to present a teacher with a prediction barely better than chance -- predictive analytics of dropout is not just about achieving the best overall prediction. It is first and foremost about communicating what a student needs to someone who can do something to help them.

At this point, the challenge of being able to predict key outcomes like student dropout has largely been solved. We certainly can do more to optimize our models -- though much progress has already been achieved (see reviews in [4, 8]) better performance is still feasible. But overall, EDM as a field is now able to produce models that effectively predict student dropout.

The remaining challenge is one of human-computer interaction. We need to improve how we communicate not just whether a student is at risk, but why. And, from there, we need to study and refine our strategies for intervening based on the reports. The reports presented in this paper are a step towards the first of these goals. We are also working, as are many others, on the second of these challenges. Through our design efforts on these two goals, we can learn to better understand students and to better support them in successful achievement in this key phase of their lives.

5. ACKNOWLEDGEMENTS

We would like to thank the teams at BrightBytes and the American Institutes for Research for their support and participation in these efforts, and the teachers and school leaders who have used the BrightBytes platform and given us their feedback and suggestions on improving our designs.

6. REFERENCES

[1] Allensworth, E., Easton, J.Q. 2005. *The On-Track Indicator as a Predictor of High School Graduation*. Technical Report. Consortium on Chicago School Research at the University of Chicago, Chicago, IL.

[2] Baker, R. S., & Yacef, K. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of educational data mining*, 1(1), 3-17. DOI= https://doi.org/10.5281/zenodo.3554657

[3] Botelho, A. F., Varatharaj, A., Patikorn, T., Doherty, D., Adjei, S. A., & Beck, J. E. 2019. Developing early detectors of student attrition and wheel spinning using deep learning. *IEEE Transactions on Learning Technologies*, 12(2), 158-170. DOI= https://doi.org/10.1109/tlt.2019.2912162

[4] Bowers, A. J. 2021. Early warning systems and indicators of dropping out of upper secondary school: the emerging role of digital technologies. *OECD Digital Education Outlook 2021 Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, 173. DOI= https://doi.org/10.1787/c8e57e15-en

[5] Bowers, A. J., Sprott, R., & Taff, S. A. 2012. Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 77-100. DOI= https://doi.org/10.1353/hsj.2013.0000

[6] Christie, S. T., Jarratt, D. C., Olson, L. A., & Taijala, T. T. 2019. Machine-Learned School Dropout Early Warning at Scale. *Proceedings of the International Conference on Educational Data Mining*.

[7] Coleman, C., Baker, R., Stephenson, S. 2019 A Better Cold-Start for Early Prediction of Student At-Risk Status in New School Districts. In *Proceedings of the 12th International Conference on Educational Data Mining*, 732-737.

[8] Coleman, C. 2021. *Exploring a Generalizable Machine Learned Solution for Early Prediction of Student At-Risk Status*. Unpublished Ph.D. Dissertation. Teachers College Columbia University, New York City, NY.

[9] Eagle, J. W., Dowd-Eagle, S. E., Snyder, A., & Holtzman, E. G. 2015. Implementing a multi-tiered system of support (MTSS): Collaboration between school psychologists and administrators to promote systems-level change. *Journal of Educational and Psychological Consultation*, 25(2-3), 160-177. DOI= https://doi.org/10.1080/10474412.2014.929960

[10] Gardner, J., & Brooks, C. 2018. Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*, 28(2), 127-203. DOI= https://doi.org/10.1007/s11257-018-9203-z

[11] Gleason, P., & Dynarski, M. 2002. Do we know whom to serve? Issues in using risk factors to identify dropouts.

Journal of Education for Students Students Placed at Risk, 7(1), 25-41. DOI= https://doi.org/10.1207/s15327671espr0701_3

[12] Holstein, K., McLaren, B. M., & Aleven, V. 2017. Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)*, March 13-17, 2017, Vancouver, British Columbia, Canada. ACM Inc., New York, NY, 257-266. DOI= https://doi.org/10.1145/3027385.3027451

[13] 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. *Journal of Learning Analytics* 1, 1, 6–47. DOI= https://doi.org/10.18608/jla.2014.11.3

[14] Knowles, J. E. 2015. Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin. *Journal of Educational Data Mining*, 7(3), 18-67.

[15] Mandinach, E. B., Honey, M., & Light, D. 2006. A theoretical framework for data-driven decision making. Presentation at the *Annual Meeting of the American Educational Research Association, San Francisco, CA* (Apr. 2006).

[16] Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. 2016. Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1), 107-124. DOI= https://doi.org/10.1111/exsy.12135

[17] Ocumpaugh, J., Baker, R., Slater, S., San Pedro, M.O., Heffernan, N, Heffernan, C., Hawn, A. 2017. Guidance Counselor Reports of the Assistments College Prediction Model (ACPM). In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)*, March 13-17, 2017, Vancouver, British Columbia, Canada. ACM Inc., New York, NY, 479-488. DOI= https://doi.org/10.1145/3027385.3027435

[18] Rumberger, R. W. 1987. High school dropouts: A review of issues and evidence. *Review of educational research*, 57(2), 101-121. DOI= https://doi.org/10.3102%2F00346543057002101

[19] Stanford, J., & Simons, R. 1999. *Victory in our schools*. Bantam.

[20] Taylor, C., Veeramachaneni, K., & O'Reilly, U. M. 2014. Likely to stop? predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382*.

[21] Verbert, K., Ochoa, X., De Croon, R., Dourado, R. A., & De Laet, T. 2020. Learning analytics dashboards: the past, the present and the future. In *Proceedings of the Tenth International Learning Analytics & Knowledge Conference,* (Mar. 2020), 35-40. DOI= https://doi.org/10.1145/3375462.3375504

[22] Wise, A. F., & Jung, Y. 2019. Teaching with Analytics: Towards a Situated Model of Instructional Decision-Making. *Journal of Learning Analytics*, 6(2), 53-69. DOI= https://doi.org/10.18608/jla.2019.62.4

[23] Yu, R., Lee, H., & Kizilcec, R. F. 2021. Should College Dropout Prediction Models Include Protected Attributes?. In *Proceedings of the Eighth ACM Conference on Learning@ Scale (L@S '21),* June 22-25, 2021, Virtual Event Germany. ACM Inc., New York, NY, 91-100. DOI= https://doi.org/10.1145/3430895.3460139