Meta-Predictive Retention Risk Modeling: Risk Model Readiness Assessment at Scale with X-Ray Learning Analytics

Aleksander Dietrichson[*,b], Diego Forteza[b], John Whitmer[a]

[a] *500 Act Drive, Iowa City, IA, 52244*
[b] *40 East Main Street, Newark, DE, 19711*

Abstract

Deploying *X-Ray Learning Analytics* (Blackboard Inc 2015) at scale presented the challenge of deploying customized retention risk models to a host of new clients. Prior findings made the researchers believe that it was necessary to create customized risk models for each institution, but this was a challenge to do with the limited resources at their disposal. It quickly became clear that usage patterns detected in the Learning Management System (LMS) were predictive of the later success of the risk model deployments. This paper describes how a meta-predictive model to assess clients' readiness for a retention risk model deployment was developed. The application of this model avoids deployment where not appropriate. It is also shown how significance tests applied to density distributions can be used in order to automate this assessment. A case study is presented with data from two current clients to demonstrate the methodology.

Introduction

X-Ray Learning Analytics (Blackboard Inc 2015) is a learning analytics package offered as an add-on to Moodlerooms'[1] clients as well as to institutions that use Moodle on a self-hosted. Among X-Ray's features is a retention risk report usually based in its entirety on data endogenic to the Learning Management System (LMS). The retention risk is assessed with statistical models which are trained and fitted for each institution individually. Back-testing (Dietrichson 2016; and Forteza 2016) have shown accuracies in the 90s and have typically been along the lines of the researchers' expectations during modeling. Other cases have been less fortunate –in several cases the recommendation has been *not* to deploy a risk model at all, since its utility would likely be insignificant or even counter-productive. These cases quickly became a source of some embarrassment since the analytics team was already

[*]Corresponding Author
Email addresses: dietrichson@gmail.com (Aleksander Dietrichson), diegoforteza@gmail.com (Diego Forteza), jcwhitmer@gmail.com (John Whitmer)

[1] *Moodleroooms* is Blackboards managed open-source offering.

in meetings with clients at this point in the process. Consequently it became apparent there was a need for a procedure to assess readiness prior to engaging with the client, a model to predict performance of the risk models, in other words: *a meta-predictive model*.

Methodological Bases

This research is based on some notions that have emerged from prior experience, both in the form of formal research and by ad-hoc observation. This section briefly describes some of the concepts that guided our efforts.

Customized Models. Multiple research studies on *individual courses* have found a significant relationship between frequency of use of the LMS and student grades (Rafaeli and Ravid 1997; Morris, Finnegan, and Wu 2005; McWilliam, Dawson, and Pei-Ling Tan 2008; Macfadyen and Dawson 2010; Fritz 2011; Ryabov 2012; Whitmer, Fernandes, and Allen 2012). The value of LMS data has been far more important than what is found in conventional demographic or academic experience variables in explaining variation in course grades. However, when analysis is expanded to all courses at an institution, several studies have found no relationship or an extremely weak relationship (Campbell 2007; Lauria 2015). These findings were in line with the researchers' experience, and congruent with the view that risk models not only need to be customized on a per-institution basis, and also that a likely outcome of a thorough modeling exercise is the deployment valid for only a subset of courses and even several different models for distinct and distinguishable groups of courses.

Course Archetypes. Previous work (Forteza and Nuñez 2016) on *course archetypes* demonstrated that online courses can be classified into five categories:

1. Supplemental – high in content but with very little student interaction
2. Complementary – used primarily for one-way teacher-student communication
3. Social – high peer-to peer interaction through discussion boards
4. Evaluative – heavy use of assessments to facilitate content mastery
5. Holistic – high LMS activity with a balances use of assessments, content, and discussion

While it may be immediately intuitive that developing a single model (or even model template) to cover these five use-cases, and that use cases (1) and (2) will likely always result in non-performant models, we still wanted to operationalize this distinction and its implication through empirical evaluation. It is also clear that these categories represent a multi-dimensional continuum, and that the named categories refer to the centroids of each cluster. As such there is clearly going to be some overlap and modeling may be possible for courses that straddle one of more of these categories. Real life experience has also indicated that each institution comes with a unique mix of these archetypes as well as other characteristics relevant to the modeling effort.

Risk Model Performance. The term *Model Performance* is used loosely to refer to the potential *usefulness* of a model, rather than as a weighted (or not) proportion of model precision or recall. While several algorithms –for example: Lopez-Raton et al. (2014)– exist for optimizing this relationship. The exact balance point will to a large degree depend on each client's needs: the degree to which interventions are planned as a result of predictions made by X-Ray, the cost of those interventions, institutional policy and practical considerations regarding each institutions' ability to act on the information generated by the system.

Population Parameters. The *outcome variable*, i.e. that which we are trying to predict, is typically a dichotomized course pass/fail, although cases with *qualified pass* are also encountered. In either case, a successful modeling exercise necessitates some variance in this variable. This fact allows us to immediately discard institutions with extremely high or extremely low passing rates. For example, an institution which graduates 95% of its students is not a candidate for risk modeling: Simply predicting *success for all students* would already result in a .95 precision rate. We thus only consider institutions whose population parameters fall within a certain heuristically defined range.

Risk Model Readiness Assessment

In order to determine the likelihood of a successful modeling exercise some global course-level measures are considered. These include: passing rate, proportion of students who have accessed the course (in the LMS), number of graded items, number of quizzes, number of assignments, correlation between quiz grades and final grades, correlation between assignment grades and final grades, mean number of access-log entries (clicks) per student and correlation between clicks and final grades. These measures are constructed based on the historical LMS activity. *Final grades* refer to the course-grades in the LMS or, if the institution does not use the course-level evaluation in the LMS, from an external source, typically the institutional SIS. When substantial use of discussion fora is detected, linguistic variables are also extracted and included.

The courses into are then divided into three categories a) courses that can be used for training a model, b) courses to which the trained model would be applicable and c) discarded courses. The criteria for the second category (b) is somewhat softer than the training data. This gives us an initial estimate of whether we have enough data to train a risk model (a), and an estimate of the proportion of courses in which we would be able to deploy a performant risk model for the client in question.

In order for a course to be useful as part of the training set is needs to have have relevant activity, and this activity must be related to the outcome variable (pass/fail or final grade), i.e. it must have *discriminatory value*. Courses where this is clearly not the case are immediately removed from consideration. For example, courses in which the proportion of passing students is greater that the proportion of students who have accessed the course are not considered, because

it is clear that students' access to the course is not relevant for determining the outcome variable.

Additional restrictions are applied and courses further filtered. The filters applied to the training and application categories are summarized in Table 1

Table 1: Summary of Restrictions and Filters

Filter	Training	Application
Proportion of Students' Access	> 0	> 0
Proportion of Students' Access	$>$ pass-rate	—
Pass-Rate	not 0 and not 1	—
Standard Deviation Final Grade	> 0	$>$ pass-rate
Quizzes or Assignments	> 5	> 5
Graded Items	> 10	> 5
Correlation between Assignment and Final Grades	$> .5$	$> .25$
Correlation between Quiz and Final Grades	$> .5$	$> .25$
Clicks per Student	> 500	> 100
Correlation between Clicks and Final Grades	$>.5$	$> .25$

All correlations for these filters are calculated using the *point biserial correlation* (Glass and Hopkins 1995) since the outcome variable has been dichotomized into pass/fall.

Reference Institution. The measures found in Table 1 were also been calculated for institutions where a successful modeling exercise had already taken place. These measures were collapsed into a *reference institution* and used for comparison with *candidate institions*. The process is best described by means of an example, or case-study, presented in the next section.

Case Study. In this section we present anonymized data from two real candidate institutions, both North American Higher Education Institutions. In the following we will refer to them as *Candidate I* and *Candidate II*

Let us first consider the two types of graded items that have shown to be of most importance for predicting the outcome variable in our reference data, namely: grades on quizzes and grades on assignments. Table 2 and Table 3 show the proportion of these two types of graded items for the two candidate institutions as well as the reference.

Table 2: Proportion of Courses with Assignments at Different Levels

	Candidate I	Candidate II	Reference
>1	76 %	13 %	90 %
>5	61 %	7 %	71 %
>10	32 %	3 %	55 %

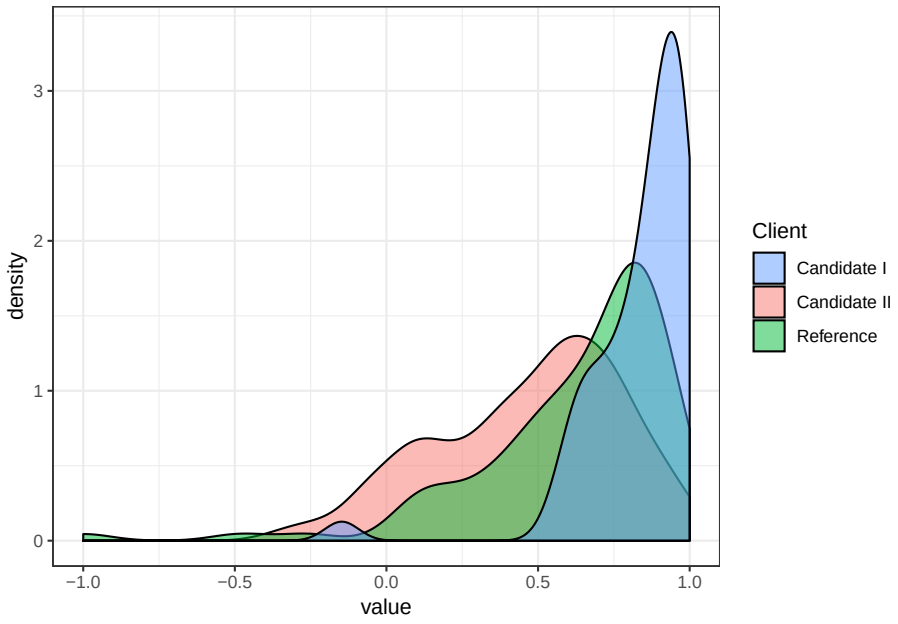

Figure 1: Correlation between Assignment Grade and Final Grade

Table 3: Proportion of Quizzes in Courses at Different Levels

	Candidate I	Candidate II	Reference
>1	88 %	8 %	78 %
> 5	54 %	4 %	50 %
>10	38 %	2 %	23 %

We see that Candidate I has a solid performance on these metrics, in terms of quizzes per course even higher than the reference client while Candidate II shows significantly lower use of these platform features.

The presence of quizzes and/or graded assignments is, however, not enough to be able to fit a risk model. These grades need to show some variance as well as some correlation to the final grades or other outcome variable. To ascertain if such a pattern exists we generate a density plot of *point biserial correlation* calculated between these variables of a course by course basis, for each of the clients as well as the reference data. Figure 1 shows the density of correlation between assignment grades and the outcome variable.

We see that Candidate I has an even higher density of strong correlation between the variables than the reference institution. Candidate II shows a lower correlation overall, and, interestingly, a non-trivial portion of the density curve is found below the zero midpoint, i.e. indicates some systematic portion of negative correlation between the variables. These cases, where systematic

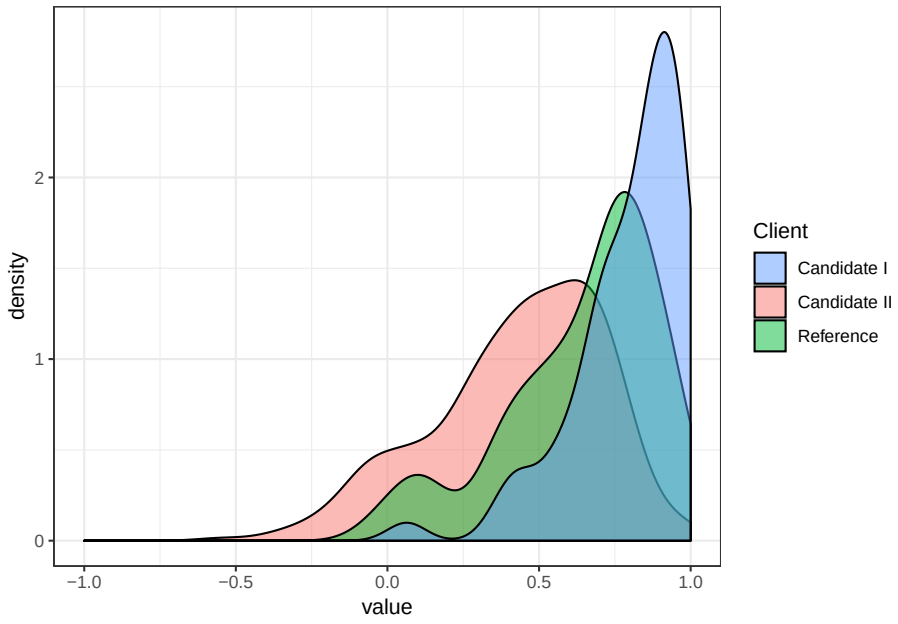

Figure 2: Density of Correlations between Quiz and Final Grades

negative correlations are found, are assumed to be invalid and are discarded for modeling purposes.

Figure 2 shows the same density plot for the correlations between quiz-grades and final grades. A pattern similar to that in Figure 1 can be appreciated, albeit with slightly lower incidence of negative correlation.

We also analyze the density of correlation between platform access (clicks) and final grades. Table 4 shows a summary of the observations for the institutions in question and Figure 3 shows the density of correlations.

Table 4: Proportion of Clicks per Student in Courses

	Candidate I	Candidate II	Reference
>100	99 %	43 %	76 %
> 500	74 %	6 %	36 %
> 1000	4 %	1 %	3 %

We see that Candidate I fares well. Candidate II, however, shows higher density of lower correlations (and even a substantial density of negative correlations) between the two variables, meaning that overall activity-level is not a stable predictor of success. This pattern is typically found when the institution has a higher proportion *Supplemental* and/or *Complementary*, as per the archetypes discussed previously.

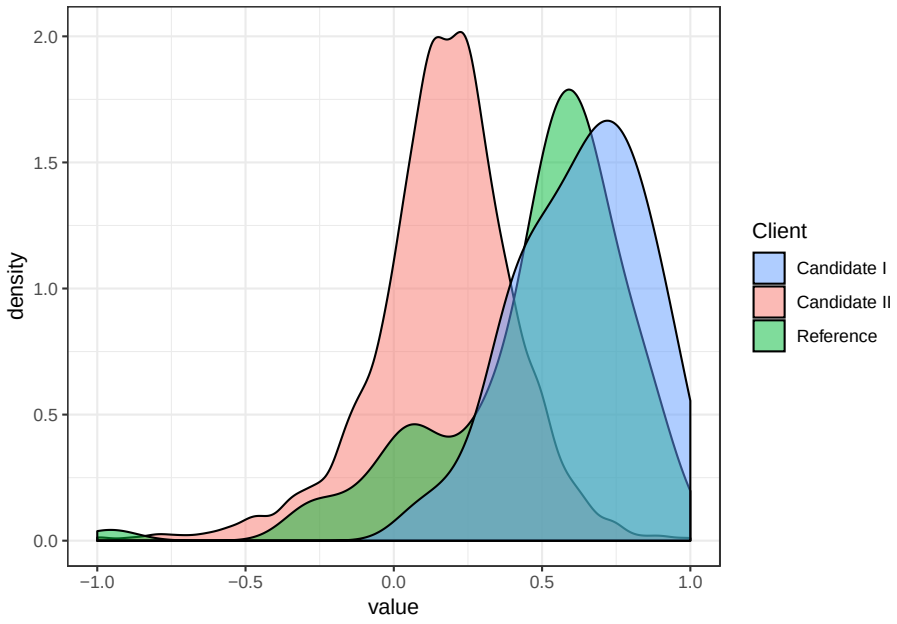

Figure 3: Density of Correlation Between Number of Clicks per Student and Final Grade

Based on these observations we draw the conclusion that fitting a risk model for Candidate I is likely to be successful, while Candidate II does not have enough meaningful use of the LMS for this to be the case.

Automatization. The example explored in the previous section shows that it is possible to predict the usefulness of a predictive retention risk model starting from the parameters and variables we chose. The decision to proceed or not with modeling is still, however, left up to the researchers, i.e. the very last step is still a manual one. For this procedure to become a scalable solution we need to be able to automate all the steps in the process. The application of filters as per Table 1 is trivial, but the determination of conformity of the density distribution to a reference is a bit more involved. Analysis of the data in R (R Core Team 2016) with the *fitdistrplus* package (Delignette-Muller and Dutang 2015) found that the density distribution can be modeled as a *beta distribution*[2] (with α=1.732 and β=0.952). Having a theoretical distribution to test against allows us to use the *Kolmogorov-Smirnov* (Kolmogorov 1933; and Smirnov 1939) statistic as a significance test, where the null-hypothesis in that the probability density of the correlations is not significantly different from the theoretical distribution. For practical purposes it does not matter, indeed it is beneficial, if these the densities

[2] PDF $= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$, where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, and $\Gamma(n) = (n-1)!$.

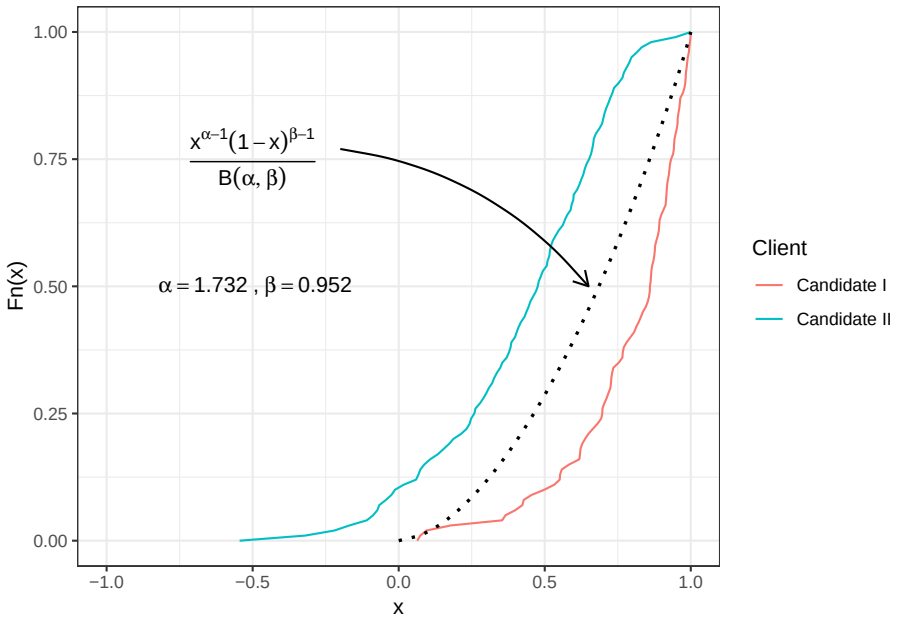

Figure 4: Cumulative Density Distribution - Theoretical and Observed“

are concentrated close to the 1.0 mare, so a one-way test is appropriate. One way to visualize this is by plotting the *cumulative* density of each distribution alongside the theoretical one. An example of this is shown in Figure 4, where we see that the totality of the of the cumulative densities for each candidate are found on either side of the theoretical reference.

The results from the one-way Kolmogorov-Smirnov test for the two candidate institutions are shown in Table 5, and are congruent with the researchers' intuition. These results show that the procedure can be set up as a completely automated system.

Table 5: Kolmogorov-Smirnov Test for Each of the Candidates

Client	D	p.value
Candidate I	0.0181265	0.9395385
Candidate II	0.3225494	0.0000000

Conclusions and Outcomes

This study shows that is it possible to quantify and predict the likelihood of a successful risk-modeling exercise based on historical data. By applying both heuristic filters and empirically extracted parameters we can avoid deploying under-performing retention risk models as well as target deployments where likelihood of success is higher.

As a result of this research, processes were put in place to pre-screen clients for risk-modeling. The X-Ray Learning Analytics product is now offered *without* risk modeling by default, and risk modeling is only offered where the pre-screening shows that a deployment is likely to be successful. We thus drastically reduce or even eliminate the deployment of under-performing models. At the same time we are now to identify clients for whom a deployment might be appropriate even if they are not currently using X-Ray.

Limitations and Next Steps

The initial filters both for institutions (population parameters) as well as course-level filters were applied based on the researchers intuition. This constitutes a limitation of the study since these precepts can and ideally should be empirically tested. The same is true for the cases where negative correlation was found between potential predictors and outcome variables. At present these are unceremoniously discarded as invalid, but it is clear that further inquiry into these *marginal* cases is merited as it may result in a more complete understanding of the patterns that govern and predict success in the modeling of retention risk.

References

Blackboard Inc. 2015. *X-Ray Learning Analytics*.

Campbell, J. P. 2007. "Utilizing Student Data Within the Course Management System to Determine Undergraduate Student Academic Success: An Exploratory Study." In *Educational Studies*, edited by Educational Studies. Indiana: A. G. Rud.

Delignette-Muller, Marie Laure, and Christophe Dutang. 2015. "fitdistrplus: An R Package for Fitting Distributions." *Journal of Statistical Software* 64 (4): 1–34. http://www.jstatsoft.org/v64/i04/.

Dietrichson, Aleksander. 2016. "Backtest of Ucem Risk Models." Internal Report. Blackboard Inc.

Forteza, Diego. 2016. "Backtest of Oshu Risk Models." Blackboard Inc.

Forteza, Diego, and Nicolas Nuñez. 2016. "Patterns in Course Design: How Instructors Actually Use the Lms." http://blog.blackboard.com/patterns-in-course-design-how-instructors-actually-use-the-lms/.

Fritz, John. 2011. "Classroom Walls That Talk: Using Online Course Activity Data of Successful Students to Raise Self-Awareness of Underperforming Peers." *The Internet and Higher Education* 14 (2): 89–97.

Glass, Gene V., and Kenneth D. Hopkins. 1995. *Statistical Methods in Education and Psychology*. Allan & Bacon.

Kolmogorov, A.N. 1933. "Sulla Determinazione Empirica Di Una Legge Di Distribuzione." *Giornale Dell'Instituto Italiano Degli Attuari* 4 (6.1): 83–91.

Lauria, Joshua, E. J. M. B. 2015. *Mining Sakai to Measure Student Performance: Opportunities and Challenges in Academic Analytics*. Academic

Conferences; Publishing International Limited.

Lopez-Raton, M., M.X Rodriguez-Alvarez, C. Cadarso-Suarez, and F. Gude-Sampedro. 2014. "OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests." *Journal of Statistical Software* 61 (8): 1–36.

Macfadyen, Leah P, and Shane Dawson. 2010. "'Mining LMS Data to Develop an "Early Warning System" for Educators: A Proof of Concept'." *Computers & Education* 54 (2): 588–99.

McWilliam, Erica, Shane Dawson, and Jen Pei-Ling Tan. 2008. "Teaching Smarter: How Mining Ict Data Can Inform and Improve Learning and Teaching Practice." *Graduate School of Medicine - Papers*.

Morris, Libby V, Catherine Finnegan, and Sz-Shyan Wu. 2005. "Tracking student behavior, persistence, and achievement in online courses." *The Internet and Higher Education* 8 (3): 221–31.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rafaeli, Sheizaf, and Gilad Ravid. 1997. "OnLine, Web Based Learning Environment for an Information Systems Course: Access Logs, Linearity and Performance." *ISECON*.

Ryabov, I. 2012. "The Effect of Time Online on Grades in Online Sociology Courses." *MERLOT Journal of Online Learning and Teaching* 8 (1): 13–23.

Smirnov, N.V. 1939. "Estimate of Deviation Between Empirical Distribution Functions in Two Independent Samples." *Bulletin of Moscow University* 2: 3–16.

Whitmer, J., K. Fernandes, and W Allen. 2012. "Analytics in Progress: Technology Use, Student Characteristics, and Student Achievement." *EDUCAUSE Review*.