Replicating MOOC Predictive Models at Scale

Josh Gardner
University of Michigan
Ann Arbor MI, USA
jpgard@umich.edu

Christopher Brooks
University of Michigan
Ann Arbor MI, USA
brooksch@umich.edu

Juan Miguel Andres
University of Pennsylvania
Philadelphia PA, USA
andresju@upenn.edu

Ryan Baker
University of Pennsylvania
Philadelphia PA, USA
rybaker@upenn.edu

ABSTRACT

We present a case study in predictive model replication for student dropout in Massive Open Online Courses (MOOCs) using a large and diverse dataset (133 sessions of 28 unique courses offered by two institutions). This experiment was run on the MOOC Replication Framework (MORF), which makes it feasible to fully replicate complex machine learned models, from raw data to model evaluation. We provide an overview of the MORF platform architecture and functionality, and demonstrate its use through a case study. In this replication of [41], we contextualize and evaluate the results of the previous work using statistical tests and a more effective model evaluation scheme. We find that only some of the original findings replicate across this larger and more diverse sample of MOOCs, with others replicating significantly in the opposite direction. Our analysis also reveals results which are highly relevant to the prediction task which were not reported in the original experiment. This work demonstrates the importance of replication of predictive modeling research in MOOCs using large and diverse datasets, illuminates the challenges of doing so, and describes our freely available, open-source software framework to overcome barriers to replication.

INTRODUCTION

The aim of public science is to build a shared body of knowledge. The repeated verification of scientific results on new data – *replication* – is necessary in order to solidify scientific knowledge, guard against spurious results, discover the potential limitations of findings, and use experimental results to inform theory. However, several fields are undergoing what has been referred to as a *replication crisis*, wherein canonical research findings have been found to be difficult, if not impossible, to replicate [7]. This has prompted concerns about both the theory based on these findings and the methodology that produced them. The learning sciences field has also been criticized for poor efforts toward replicability [30]. Unfortunately, the professional incentives for scientists do not typically support replication [32], and several technical barriers to replication exist [4, 11].

In this work, we argue that there is a need for a replication platform that can support the unique technical, methodological, and data requirements for replicating predictive modeling research in online learning data, such as the data now emerging from Massive Online Open Courses (MOOCs). We are motivated in part by the rise in the number of scientific works focused on early warning systems for student success, and present a platform aimed at addressing replication challenges. We demonstrate the capabilities of this platform through a large-scale replication. The conclusions we draw from this replication demonstrate the issues of relying upon a single study, and provide further evidence for the need for replication in the educational data sciences.

Replication of Predictive MOOC Models

There has been considerable research on predicting student success, particularly in MOOCs (see Section 2), but relatively little assessment of whether the models produced are general across courses, platforms, or student cohorts. This lack of replication is problematic for three reasons.

Novelty: While there has been some work to predict student outcomes in traditional schooling [37], the applications of these methods to MOOCs is relatively novel. MOOCs are distinct from several other superficially similar contexts, such as distance education or for-credit online learning, because of the large breadth of intrinsic and extrinsic motivators, demographics, and outcomes involved. This makes it difficult to draw directly from these other fields. Existing models of engagement (e.g., [36]) have been of relatively low influence on predictive modeling efforts. The field has had little consensus and

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

theory to guide investigation, and this lack of consensus extends to the types of features and algorithms that best predict MOOC learner outcomes [12, 21, 38]. Conducting research in a new and emerging domain – particularly one where the artifacts under study, MOOCs, are evolving over time – virtually guarantees that some results will not be robust, and without replication, we have no way of knowing which those are.

Different Experimental Subpopulations: MOOC research is often conducted on student populations which vary substantially across studies. Furthermore, there are large differences in the subpopulations used in different studies, with many restricting their analysis to between 50% and 5% of the remaining students in a course due to the features of interest [23]. Together, these differences (in the original student population, and the subpopulation selected from it) can produce substantial differences in observed predictive performance [21], also casting doubt on the generalizability of findings.

Researcher Degrees of Freedom and Overfitting: Repeated randomization on small datasets, with unlimited "researcher degrees of freedom" to investigate results without controls for the many comparisons conducted, virtually guarantees statistically significant results [25]. Even when researchers are not directly attempting to probe for statistically significant results, these comparisons may be performed by automatic model tuning libraries such as `caret` or `auto-WEKA`. Such practices, performed on small datasets that are not publicly available, can lead to results which fail to generalize to new data.

These considerations collectively make a strong argument for the creation of infrastructure for large-scale replication, experimentation, and analysis of machine learned models in MOOCs. The field's ability to construct reliable, methodologically sound knowledge and shared practices will be enhanced by the existence and use of such an infrastructure.

Challenges and Tractability of Predictive Model Replication in MOOCs

In prior work, we identify key challenges to predictive model replication in MOOCs [24], which we briefly summarize here. The primary factors which we identify as hindrances to replication in the current landscape are:

Technical Complexity: Replicating machine-learned models for prediction of student success depends on minute implementation details typically not reported in journals or conference proceedings, such as the methods for feature extraction from raw data, model-building, and model evaluation. Complete replication of a machine-learned model on new data requires (i) a complete encapsulation of the experimental procedures (i.e., code) and its execution environment; (ii) new data, ideally with the same schema, and large enough to be informative about the model's generalizability over a broad population of courses and learners; (iii) the computational resources to execute this experiment across n courses by repeating the *extract-train-test* cycle n times.

Methodological Discrepancy: Currently, there is no standard practice for building and evaluating predictive models in MOOCs, making it often impossible to compare or reconcile the results of different experiments. Models are often claimed to be useful to student learning interventions broadly, but experiments use varying procedures to identify subpopulations of interest, extract features from raw data, and train and evaluate models [23].

Data Scarcity: Many universities interpret FERPA, the IRB Common Rule, or other regulations in ways that severely limit access to MOOC data. This has limited researchers' access to data, leaving many able to access data only from individual or small numbers of courses [23, 39]. This risks the field building theories of learner success on small and non-representative subsets of MOOC learners or specific content areas and disciplines. The few openly-available MOOC datasets that do exist, such as the HarvardX-MITx Person-Course Dataset [27] and the 2015 KDD Cup dataset from XuetangX (no longer publicly available), have been utilized extensively in predictive modeling research, suggesting that many interested researchers face barriers to access and that there is a need for more access to MOOC data. Unfortunately, the anonymization techniques applied to the currently open datasets limit the scope of potential research, for example, by aggregating data to only summary statistics (as with XuetangX data), eliminating information such as the text of discussion forum posts (e.g. DataStage) or removing traces from learners if unless they form a homogeneous subpopulation (e.g. [13]).

Despite these challenges, we also optimistically observe that large-scale replication is a more tractable problem in predictive modeling in MOOCs than in many other contexts, as noted previously in [24]. First, the raw data formats used to generate these models are largely consistent, and their schema are thoroughly and publicly documented. This ensures that researchers engaging in replication can conduct analyses on new and even unseen data from the same platform with a clear knowledge of its schema, and analyses executed on any one course can be replicated across *all* courses from the same platform. Second, the tools for building predictive models from raw data – typically, code and software built using open-source programming languages and even operating systems – are highly portable and replicable. Using containerization, described in Section 3, replicating an experiment can become as straightforward as re-running a containerized experiment against new data [9, 31].

Together, these two features of MOOC data and analyses suggest that efforts for replication are relatively feasible in this context. This stands in contrast to the replication challenges in other fields, such as experimental psychology, where replicating laboratory conditions exactly is

nearly impossible, and researchers have instead resorted to developing criteria for replications that are "as close as possible" [7]. Further, the ability to generalize *through replication* to huge populations of diverse learners is a potential opportunity for MOOC researchers that is not generally available within other fields [28]. However, attempts at replicating sophisticated models are limited to guesswork and partial replication in the absence of a platform for conducting and sharing such research.

PRIOR RESEARCH

Predictive Dropout Modeling in MOOCs

There is a large and growing body of research on predictive modeling in MOOCs, particularly modeling of whether a student will drop out, stop out, or otherwise fail to complete a MOOC. Prior research has attempted to predict this student outcome using a variety of features extracted from clickstream data and natural language in discussion forum posts [12], social networks [42], and assignment grades and activity [38]. This work has also explored a diverse model space, including survival models [42] and a range of machine learning algorithms [3, 8, 18, 29]. For a detailed overview of prior work on predictive modeling in MOOCs, including both feature extraction and statistical modeling techniques, see [23].

Scientific Replication At Scale

While there is less research involving replication than original research, what does exist generally supports the value of replication – and paints a bleak picture of the replicability of much published work. A study of over 100 experimental and correlational results in experimental, social, and cognitive psychology found that only 39% of previously-reported results replicated, despite using high-powered designs and materials provided by the original authors [10]. A recent survey of the top 100 education research journals [30] notes (i) a dearth of replication research, with only 0.13% of education articles being replications (a replication rate eight times lower than in the field of psychology); and (ii) of those attempted, 67.4% replicated the original results fully, 19.5% replicated some, but not all, findings, and 13.1% failed to replicate any original findings. The results were significantly more likely to replicate when there was overlap between the authors of the two studies. Both [10] and [30] note the technical challenges of replication when relying solely upon published descriptions of experimental and analysis methods. The NSF Committee on Social, Behavioral, and Economic Sciences notes [5] that researchers who study "the causes of human behaviors and the effectiveness of strategies meant to change behavior" must exercise particular caution to ensure their research is robust and replicable (pp.1).

Frameworks to support replication have been slow to appear; the existing frameworks are often incomplete or unsuitable for MOOC predictive model replication, for a variety of reasons: they are either limited to code- or data-sharing, but not both; domain-specific; or lack the privacy controls required for the underlying data. General replication tools include the Open Science Framework[1], a cloud-based reproducible preregistration, code, and data hosting platform geared toward science researchers; and Codalab[2], which allows for the construction of directed acyclic graphs to model the procedures used to process data. Both platforms represent progress toward replication and open research; however, neither framework supports the privacy-protecting measures required for working with and sharing MOOC data.

MOOC-specific computational research platforms exist, but none are specifically designed to support replication. LearnSphere[3] is a community infrastructure to support the sharing and analysis of online learning data, and unifies several individual projects (DataShop, MOOCdb, DataStage, DiscourseDB). However, the analytical capabilities on this platform are limited to specific plugins and therefore restricted in their capabilities (users cannot, for example, upload their own code or software). DataStage[4] provides access to the largest raw MOOC dataset currently available to researchers (94 Coursera sessions, 57 edX sessions); however, access to this data is restricted, and no computational resources are provided to conduct analyses needed for effective replication. MOOCdb [14], which provides shared data schema and access controls across MOOC platforms, does not itself provide either data or computational resources for analysis, and therefore only resolves one of the many barriers to replication in MOOCs.

MORF 1.0

The MOOC Replication Framework was introduced in its first form (1.0) in [1, 2]. This initial implementation enabled the replication of findings as production rules. The first published study using MORF attempted to replicate 21 findings from eight different studies within the context of a single MOOC. Only nine of the original findings replicated, with two other findings found to be statistically significant in the *opposite* direction from the published effect [2]. A second study attempted to replicate 15 previously published findings from five different studies across 29 sessions of 17 MOOCs. 12 of the 15 findings replicated significantly across the datasets, with two replicating significantly in the opposite direction [1].

The simplicity and interpretability of MORF 1.0 come at the cost of being unable to replicate more complex findings and models. Much MOOC research – particularly predictive modeling research – is far more complex than can be represented efficiently using production rules. Findings which attempt to control for even a single continuous predictor – i.e., *students whose posts spanned a duration of 1 standard deviation higher than average were 60% less likely to drop out, holding their authority score constant* [42] – are difficult to represent

[1]https://osf.io/
[2]https://worksheets.codalab.org/
[3]http://learnsphere.org/
[4]https://datastage.stanford.edu/

in this manner. MORF 1.0's success in examining several findings across many MOOCs demonstrates the feasibility of replication in assessing the robustness of published results on MOOCs, and the opportunity created by utilizing large, multi-course datasets to conduct those replications. The production rule framework, however, limits the extent to which previous findings can be replicated, and is unable to replicate a substantial portion of predictive modeling research.

THE MORF 2.0 PLATFORM

To address the need for a predictive model replication framework we redeveloped the MORF software and created a new computational infrastructure for replication which instead follows the end-to-end workflow of a predictive modeling experiment. This software is open-source and freely available, and its architecture is described in detail in [24] and at https://educational-technology-collective.github.io/morf/. We provide a brief overview here prior to discussing the replication experiment conducted on the MORF platform.

Infrastructure and Workflow

MORF utilizes the Docker containerization service and a cloud computing environment to support fully reproducible analyses, addressing the technical challenges to MOOC replication outlined above. MORF provides a Python API which allows users to fully specify the extract-train-test-evaluate workflow for their experiment. This allows users to run scalable, parallelized analyses across these courses, where the underlying code can be written in any language that can be installed in a Linux-based Docker container. For example, the analysis presented in Section 4 utilized Python, R, and Java in an Ubuntu-based environment.

MORF addresses the challenge of researcher degrees of freedom in predictive modeling by providing support for better model evaluation, using an effective model evaluation procedure as a default. Cross-validation is overwhelmingly the most common choice for model evaluation in MOOC experimentation [22]. However, research specific to MOOCs has shown that evaluating models using data from students in the same session results in higher estimates of model performance than data from new sessions of the same course [8, 38, 39], suggesting that there is an issue of over-fitting even to a specific run of a course and that cross-validation within sessions produces optimistically biased model performance data. To build models which generalize well to new data (e.g. for use in student early warning systems), it is important to use evaluation methods which provide accurate estimates of how well models will perform on new sessions of a course, when this is the aim of prediction. As a result, MORF requires experiments to predict on the most recent session of a course, which is held-out from the model at training time. We refer to this approach as *future session* prediction [24]; this is a special case of transfer learning [6]).

Data

The MORF software system was deployed at the Universities of Michigan and Pennsylvania as a platform-as-a-service (called the "MORF Platform"). The MORF Platform provides access to the complete raw data for 209 sessions of 77 unique MOOCs, including clickstream, discussion forum, video watching and navigation, demographic, and assignment data, as well as course metadata (information about individual lecture videos, assignments, exams, etc.). Overall, nearly one million unique students and nearly three million unique interactions are stored in the MORF Platform. For reasons of security, privacy, and data ownership, the data available in MORF is not available for export or download, but instead is available for analysis through a secure platform which is currently governed by a data use agreement.

CASE STUDY: REPLICATION WITH MORF

Study for Replication

We replicate a dropout modeling experiment by Xing et al. [41]. As we describe below, this study serves as an example of "standard" practice for predictive modeling research in MOOCs for several reasons, and replication across the large, diverse MOOC dataset in MORF has the potential to contextualize these results, as we will show. Our goal is two-fold: we seek (1) to demonstrate the benefits and challenges to conducting end-to-end replication, but also (2) to evaluate the generality of the original findings. In [41], both feature extraction and predictive modeling techniques were varied to provide fine-grained temporal dropout predictions in a single MOOC. The original authors employ three different feature types (*week-only*, *summed*, and *appended*), which represent different methods for aggregating features over time, in combination with three different predictive modeling techniques (Bayesian network, C4.5 trees, and ensembles. These feature/model combinations are used to predict, at each week of a single seven-week MOOC, which active students will drop out the following week. This prediction is performed with the intention of identifying much smaller and temporally-specific, and therefore more actionable, groups of students at risk of dropping out of a MOOC.

This study is a representative example of predictive modeling research in MOOCs, and thus a useful candidate for replication, for several reasons. First, the original study evaluates only a single course, which was offered on the Canvas platform. Second, this study uses a *post hoc* prediction architecture, evaluating the experimental models' performance on new data from the same course used for model training, a common practice in the field, as mentioned above. Third, the study does not provide a statistical comparison of the differences between models it reports on, simply concluding that one of the model and feature extraction methods has better average performance – also a standard practice in the field. Fourth, the study considers a large number of comparisons, evaluating three different models across three different fea-

ture types using two evaluation metrics over seven weeks of the course. With so many different comparisons, the possibility of observing spurious differences in model performance becomes likely. Finally, the operationalization of many aspects of the models, including the feature definitions, algorithms, and training/hyperparameter tuning techniques used, are ambiguous; we were unable to obtain clarification on these questions even after multiple attempts to correspond with the study's authors. This ambiguity, detailed in Section 4.4 demonstrates the many challenges of replication in computational science [34] and without author cooperation [7], although such issues are challenging even with collaboration from study authors.

Replication Methodology

The original study presents a variety of descriptive performance data for each model. In this replication we seek to investigate its two core findings from [41]:

- **F1: A stacked ensemble of Bayesian Network + C4.5 Decision Tree classifiers performs better than either base learner**.
- **F2: Appended features perform better and with greater stability (less variability over weeks) than week-only or summed features**.

We do not attempt to replicate a third finding, in which a final model is formed by 'switching' from appended to week-only features in later weeks of the course based on PCA, because the methodology (which was largely exploratory, manual, and qualitative) and avenues for replicating the approach across multiple courses was not clear enough to us to be confident that we were correctly replicating the authors' original intent. We evaluate **F1** and **F2** using statistical testing; no statistical testing was used in the original work.

We used the MORF Platform and followed these steps: (1) For each session of each course ($n = 133$), separately **extract** all three feature sets (week-only, summed, appended) for all weeks.(2) For each course, **train** all three models (C4.5, Bayesian Network, stacked ensemble) for every $week \times feature$ combination. (3) **Test** models on the most recent session of each course. (4) **Evaluate F1** and **F2** using statistical testing.

These steps are also the four components of the MORF API workflow [24], and match the general structure of almost all end-to-end predictive modeling experiments. This replication required extracting three sets of features (week-only, summed, appended) and training nine models (C4.5, GBN, and ensemble, with one of the three feature types) per week of each course, a total of 1,588 models. Due to the concerns outlined above about overfitting to specific training runs, particularly when a model might be trained on the first session of a course and tested on the second session (prior research suggests that the differences between course sessions are largest between the first session and subsequent sessions [6, 17]), we only use courses for which at least two training runs were available. This limits our course population to the 28 courses in MORF with at least three sessions (two training + one testing) – a total of 133 sessions overall.

This replication attempts to meet the five criteria for a "convincing close replication" from [7]. In particular, our replication attempts to demonstrate (1) definition of the effects and methods intended for replication; (2) following as exactly as possible the methods of the original study; (3) having high statistical power; (4) making complete details about the replication available[5]; and (5) evaluating the results of the replication, comparing them critically to the original study results.

Replication Results

For both **F1** and **F2**, we find that the authors' original findings do not fully replicate, and that some findings replicate significantly in the opposite direction. Each finding is evaluated using the statistical testing procedure from [19], which utilizes the equivalence of the Area Under the Receiver Operating Characteristic Curve (AUC, or A') statistic and the Wilcoxon statistic to generate a test statistic Z to evaluate a null hypothesis of equivalent performance between two predictive models (the reader is directed to [19, 26] for the details and theoretical justification of this approach):

$$Z_{course} = \frac{A'_1 - A'_2}{\sqrt{SE(A'_1)^2 + SE(A'_2)^2}} \tag{1}$$

Z-scores for each condition were averaged across weeks within a course due to non-independence, and then aggregated across courses using Stouffer's method [35]:

$$Z \sim \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{2}$$

where k is the number of courses, and Z_i is the z-score for the comparison in course i averaged across all weeks of the course. This test evaluates the null hypothesis of whether two models have equivalent average dropout prediction AUC performance across the n weeks of a course. We reject the null hypothesis of equivalent performance using the rejection threshold for a standard Z test at level α ($\alpha = 0.05$ in this experiment), using the `auctestr` R package for testing.

F1: Ensemble vs. Base Classifiers

In order to test **F1: A stacked ensemble of Bayesian network + C4.5 decision tree classifiers performs better than either base learner**, we apply the statistical testing method outlined in Equations 1 and 2 above to comparisons of *ensemble vs. Bayesian network* and *ensemble vs. C4.5* for each of the three feature methods. Results are shown in Table 1, and show only limited

[5]Code to replicate this analysis is available at https://github.com/educational-technology-collective/xing-replication.

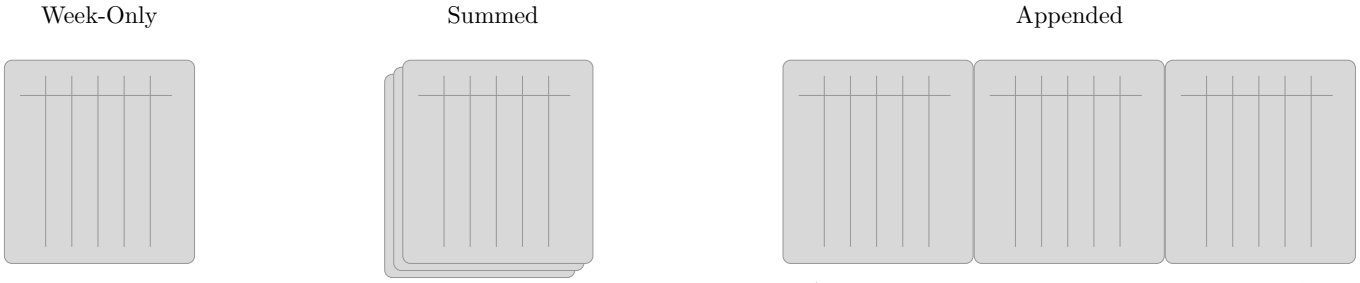


Figure 1. Temporal feature aggregation methods replicated from [41] and evaluated in F2. Each method represents a different approach to aggregating a set of seven base features.

Models		Features	Z	p-value
Ensemble	BN	Appended	15.7	$\ll 0.0001^{\star}$
Ensemble	BN	Sum	-47.2	$\ll 0.0001^{\dagger}$
Ensemble	BN	Only	-26.5	$\ll 0.0001^{\dagger}$
Ensemble	C4.5	Appended	-2.3	$0.0227^{\dagger}$
Ensemble	C4.5	Sum	2.7	$0.0073^{\star}$
Ensemble	C4.5	Only	-2.6	$0.0186^{\dagger}$

Table 1. Replication results for F1 (comparison of ensemble vs. base learner AUC performance). $\star$: statistically significant replication of original finding at $\alpha = 0.05$; $\dagger$: statistically significant replication of opposite finding from original. See also Fig. 2.

support for the authors' original conclusions. Recall that in the original experiment no statistical testing was performed, and only a cross-validation with *post hoc* prediction architecture was used to evaluate results. Our results show that when appropriate statistical testing is used to evaluate these findings at scale across a large and diverse sample of MOOCs, the ensemble only significantly outperforms the base learners in two of the six possible model configurations, and the ensemble actually performs significantly *worse* than the base learners in the remaining four configurations.

While ensembling is theoretically guaranteed to improve the performance of uncorrelated base learners under mild conditions [15], the ensemble fails to do so in four of the six cases evaluated here. This suggests that the significant computational expense of training additional classifiers and then ensembling them may not be justified. In particular, the ensemble performs worse than C4.5 alone in two of the three cases tested. This finding is relevant because of the substantial computational complexity of learning the structure and parameters of a Bayesian network for a large dataset, particularly for the case of the relatively high-dimensional data of the appended features: the number of possible network configurations when learning the structure of a Bayesian network is super-exponential in the number of features [20], making structure learning on such high-dimensional datasets computationally expensive. The results of our replication suggest that effective dropout models can be constructed without computationally-intensive Bayesian network structure and parameter learning. Furthermore, these results demonstrate that modeling the behavioral features used here as independent (as a tree-based model does) instead of attempting to learn dependencies between features (as a Bayesian network does) can achieve better predictive performance. If there is no discernible dependence between features (i.e., between forum views, active days, social network degree), this has important implications for theory driven by these models.

F2: Appended vs. Other Features

We also evaluate **F2: Appended features perform (a) better and (b) with greater stability than summed or week-only features**. The original work states: "[T]he performance of modeling based on summed features fluctuates over weeks and has several curves. By contrast, modeling which relies on appended features is more stable across time. The possible reason might be due to more historical data concerning the features being available for model building" ([41] pp.126). Results from our evaluation of **F2a** are shown in Table 2, which shows performance comparisons by feature type, and results from **F2b** are shown in Figure 2, which shows the stability of both AUC and precision over time (both metrics were used to evaluate **F2b** in the original work). We discuss each below.

F2a: Table 2 shows the results of the feature comparisons averaged over weeks following the method outlined above. In the original work, comparisons were only given for the base models (not the ensemble) and only for appended vs. summed (not for week-only features); we report the results of all possible comparisons but only evaluate replication on those reported in the original study. In both of these comparisons, **F2a** replicated significantly. However, this original finding only evaluated two of the nine potential comparisons shown in Table 2, with no justification for why the other comparisons were not evaluated (those for ensemble models, and for summed vs. week-only features). We believe these comparisons are relevant to the larger questions about feature-algorithm synergy raised by this work. In particular, the additional comparisons in Table 2 suggest that week-only features can achieve good performance (out-

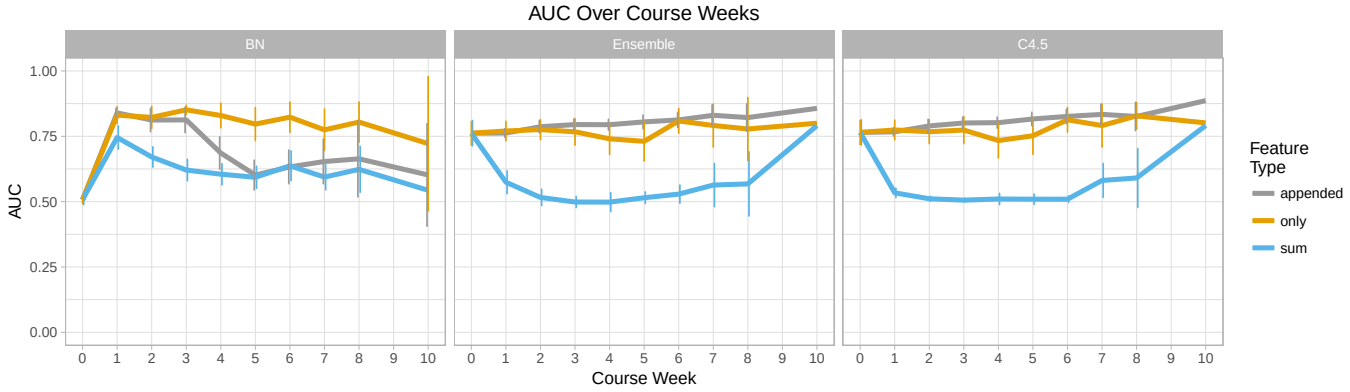
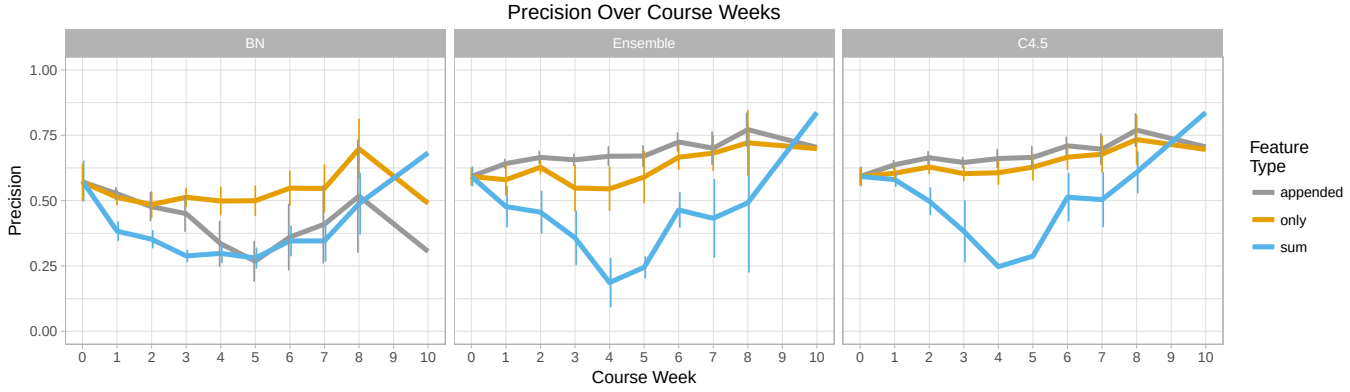


Figure 2. Above: AUC of feature types over course weeks. Below: Precision of feature types over course weeks. Only weeks 0-10 are shown due to small sample of courses with length > 10 weeks. See also Table 2.



performing summed features in all specifications tested, and outperforming appended features when used with a BN algorithm). This finding is relevant to the original results in at least two ways. First, it has practical implications for effective model building, implying that effective predictive models can be built using less data (only the previous week, not all previous weeks), which reduces the data processed for both feature extraction and model training. Second, it has theoretical implications, suggesting that only more recent behavior (the previous week, from week-only features) or aggregate behavior over all weeks (summed features) may be relevant to students' decision to drop out, with earlier behavior alone being less relevant (in appended features). Similar findings regarding the limited performance benefits of appended vs. week-only features are reported in [33].

F2b: This finding concerns the *stability* of performance with appended features relative to all other feature sets. Because the testing method used for F2a only applies to AUC (due to its equivalence to the Wilcoxon statistic, enabling statistical inference), and not to the other metric reported in [41], precision, we evaluate the variability of both AUC and precision over course weeks by estimating 95% confidence intervals for each feature type and model, shown in Figure 2. This figure shows that appended features do indeed demonstrate greater stability than summed features. This matches the original finding

Features		Model	Z	p-value
Appended	Only	C4.5	9.7	$\ll 0.0001$
Appended	Only	BN	-32.4	$\ll 0.0001$
Appended	Only	Ensemble	9.8	$\ll 0.0001$
Appended	Sum	C4.5	100.2	$\ll 0.0001^\star$
Appended	Sum	BN	32.0	$\ll 0.0001^\star$
Appended	Sum	Ensemble	95.5	$\ll 0.0001$
Sum	Only	C4.5	-90.2	$\ll 0.0001$
Sum	Only	BN	-64.7	$\ll 0.0001$
Sum	Only	Ensemble	-85.5	$\ll 0.0001$

Table 2. Results for F2a (comparison of AUC performance by feature types). ⋆: statistically significant replication of original finding. Unmarked comparisons were not reported in [41]. See also Fig. 2.

(again, results for **F2** were only provided for appended vs. summed features on the base learners in the original work). However, week-only features achieve nearly identical stability to appended features as measured by both AUC and precision across all model types; week-only features display *greater* stability than appended features when used with Bayesian networks, shown in the left panel of Figure 2. This is likely due to instability in the Bayesian networks' structure-learning algorithm when applied to relatively high-dimensional appended feature sets. This comparison was not reported in the

original work, but is relevant to the interpretation of **F2**: if week-only features can achieve both comparable performance and stability to appended feature sets, there are situations in which week-only features may be preferable to appended features. This includes when modeling algorithms do not scale well with the number of features, such that training models on the substantially wider appended feature sets incurs great computational cost; or in later weeks of a course, when reviewing a user's entire previous history requires processing much more data than only processing the previous week.

Collectively, our evaluation of F1 and F2 have relevance both to the replication of the original work, and to future modeling efforts. Collecting appended features and fitting models across the considerably larger feature space that accrues over many weeks requires substantially more computation time, especially for Bayesian network structure learning, and means processing much more historical data. If this approach achieves neither better generalization performance nor smaller variability compared to using the features from only the current week, then using the smaller feature space of week-only features could be preferable. While the original findings of **F2** generally replicated in our study, we find unexamined (or at least unreported) patterns which add important context to the original results, and which may adjust the original conclusion to favor ensembled models and appended features far less strongly, instead favoring week-only features and the base learners.

Ambiguity and Good-Faith Replication

The case study presented here highlights several challenges to conducting predictive model replication in MOOC research. Our aim is not to single out the authors of this particular study. Instead, it is to demonstrate these challenges, and focusing on the details of a specific study is necessary in order to illuminate these challenges. In this section, we describe potential gaps between our replication and the original work, in order to make clear the difficult choices that researchers may confront in conducting the best possible implementation of the original work when firsthand knowledge of the original procedure is unavailable.

The first critical area where our replication efforts required inferring the authors' methods was in feature definitions. The original work utilized a "social network degree" feature based on co-posting in discussion fora, but the method for defining edges in this network was unclear. In particular, the authors do not define whether they build edges only between students who posted adjacent to each other on a thread, or between all students on a thread. We implemented both as separate features (*reply-node* and *thread-node*, respectively). The definition of 'active' students which are used to train the model is also not directly defined in the original work. In our analysis, we assume this means students who did not drop out of the course in weeks prior to the target week (equivalently, those without one or more consecutive weeks of inactivity immediately prior to the target week). This challenge highlights the need for clear feature and methodology descriptions in published work or, ideally, open-source and reproducible code for the full modeling workflow.

A second set of challenges in replicating the original work relate to model-fitting. A Bayesian Network is a particularly complex model to fit, and requires learning (a) the network structure and (b) the network parameters. Neither of these are discussed in [41]: while the authors are explicit that they did *not* use a Naïve Bayes model, the method for learning the network structure is not mentioned. Not knowing which structure-learning algorithm was originally implemented (e.g. a score-based algorithm such as hill climbing; a constraint-based algorithm such as grow-shrink; or expert specification), we utilized a hill-climbing algorithm in consultation with other experts in the field[6]. The method for learning the network *parameters* is also omitted from [41]. We adopted a standard Maximum Likelihood Estimation method. Responses to inquiries sent to the original article's authors or open source code would have prevented this challenge.

There is also no discussion of the meta-learner used to build the ensemble that forms the core of the analysis in [41] and is the focal finding **F1**, which led to additional model replication difficulties. Both the procedure used to collect predictions of base learners in a way that avoids data leakage (we assume and implement a cross-validation approach, following [40], and utilize 3-fold cross-validation to limit the number of iterations of model fitting required), and the type of meta-learner used, are not described. We used a logistic regression meta-learner, because (a) the outcome was binary, (b) a low-variance model seemed suitable for the low-dimensional input data from the base learners (only four predictors: predicted class probabilities from 2 models $\times$ 2 outcome classes), and (c) prior work demonstrated success using a logistic meta-learner in MOOCs [3]. Finally, many hyperparameters could be tuned for both of the base algorithms and the meta-learner. Again, because the authors provided no discussion of these decisions and were unresponsive to requests for clarification, we used default or standard settings where possible. Clear explication of modeling parameters in published work or open-source code would have prevented this issue.

Each of these decisions has a potential to affect the outcome of the replication. Additionally, each would be avoidable if the Learning at Scale community utilized a shared replication framework for predictive modeling in MOOCs. We are attempting to build such an infrastructure with MORF, where researchers can leverage, interrogate, and build off of each others' experiments. As David Donoho notes regarding replication in modern data science research, "[a]s computations have become more ambitious, the gap between what readers know

[6]We would like to thank Dr. Christina Conati for guidance on effective Bayesian network approaches to this task.

about what authors did has become immense" [16]. It is our intent that MORF reduces – or eliminates – this gap for MOOC researchers.

CONCLUSION AND FUTURE WORK

This paper describes the need for replication of research and practice in predictive modeling in MOOCs, introduces an open-source framework to address this need, and demonstrates the replication process with a case study where we replicate a previously published set of findings in the context of a much larger data set. This case study highlights the many challenges of conducting a replication, but also demonstrates the continuous process of refining scientific knowledge that is central to the replication task. With this work and the MORF platform it introduces, we hope to encourage future replication and to provide a foundation for open replication which is accessible and robust. Additionally, we hope that MORF can provide a common foundation for researchers interested in replicating – or, indeed, conducting – research across large, representative MOOC datasets and that future work utilizes MORF data as a resource or even a publicly-available benchmark for such modeling tasks.

Several further replication experiments on the MORF platform are either planned or underway, and expansions to the MORF platform to extend its capabilities are also in development. This includes making extracted features available to other users and natively conducting more robust model evaluation tests to encourage the adoption of valid and state-of-the-art model inference techniques.

REFERENCES

1. J. M. L. Andres, R. S. Baker, G. Siemens, D. Gašević, and S. Crossley. Studying MOOC completion at scale using the MOOC replication framework. In *Proceedings of the International Conference on Learning Analytics and Knowledge*, pages 71–78, Mar. 2018.
2. J. M. L. Andres, R. S. Baker, G. Siemens, D. Gašević, and C. A. Spann. Replicating 21 findings on student success in online learning. *Technology, Instruction, Cognition, and Learning*, pages 313–333, 2016.
3. G. Balakrishnan and D. Coetzee. Predicting student retention in massive open online courses using hidden markov models. Technical report, Univ. Calif. at Berkeley EECS Dept., 2013.
4. C. Boettiger. An introduction to docker for reproducible research. *Oper. Syst. Rev.*, 49(1):71–79, Jan. 2015.
5. K. Bollen, J. T. Cacioppo, R. M. Kaplan, J. A. Krosnick, J. L. Olds, and H. Dean. Social, behavioral, and economic sciences perspectives on robust and reliable science. Technical report, NSF Subcommittee on Replicability in Science, 2015.
6. S. Boyer and K. Veeramachaneni. Transfer learning for predictive models in massive open online courses. In *Artificial Intelligence in Education*, pages 54–63. Springer, Cham, June 2015.
7. M. J. Brandt, H. IJzerman, A. Dijksterhuis, F. J. Farach, J. Geller, R. Giner-Sorolla, J. A. Grange, M. Perugini, J. R. Spies, and A. van 't Veer. The replication recipe: What makes for a convincing replication? *J. Exp. Soc. Psych.*, 50:217–224, 2014.
8. C. Brooks, C. Thompson, and S. Teasley. A time series interaction analysis method for building predictive models of learners using log data. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 126–135. ACM, Mar. 2015.
9. J. Cito, V. Ferme, and H. C. Gall. Using docker containers to improve reproducibility in software and web engineering research. In *Web Engineering*, Lecture Notes in Computer Science, pages 609–612. Springer, Cham, June 2016.
10. O. S. Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, Aug. 2015.
11. C. Collberg, T. Proebsting, G. Moraila, A. Shankaran, Z. Shi, and A. M. Warren. Measuring reproducibility in computer systems research. Technical report, Univ. Arizona Dept. of Comp. Sci., 2014.
12. S. Crossley, L. Paquette, M. Dascalu, D. S. McNamara, and R. S. Baker. Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 6–14, 2016.
13. J. P. Daries, J. Reich, J. Waldo, E. M. Young, J. Whittinghill, A. D. Ho, D. T. Seaton, and I. Chuang. Privacy, anonymity, and big data in the social sciences. *Commun. ACM*, 57(9):56–63, 2014.
14. F. Dernoncourt, C. Taylor, K. Veeramachaneni, and U. O. Reilly. Moocdb: Developing standards and systems for mooc data science. Technical report, Technical Report, MIT, 2013.
15. T. G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15. Springer, Berlin, Heidelberg, June 2000.
16. D. Donoho. 50 years of data science. In *Princeton NJ, Tukey Centennial Workshop*, pages 1–41, 2015.
17. B. J. Evans, R. B. Baker, and T. S. Dee. Persistence patterns in massive open online courses (MOOCs). *J. Higher Educ.*, 87(2):206–242, Mar. 2016.
18. M. Fei and D. Y. Yeung. Temporal models for predicting student dropout in massive open online courses. In *Intl. Conf. on Data Mining Workshop (ICDMW)*, pages 256–263, 2015.

19. J. Fogarty, R. S. Baker, and S. E. Hudson. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface 2005*, pages 129–136, 2005.

20. J. A. Gámez, J. L. Mateo, and J. M. Puerta. Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Min. Knowl. Discov.*, 22(1-2):106–148, Jan. 2011.

21. J. Gardner and C. Brooks. Dropout model evaluation in MOOCs. In *Proceedings of the Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*.

22. J. Gardner and C. Brooks. Evaluating predictive models of student success: Closing the methodological gap. *The Journal of Learning Analytics*, 2018. In press.

23. J. Gardner and C. Brooks. Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*, 2018.

24. J. Gardner, C. Brooks, J. M. L. Andres, and R. Baker. MORF: A framework for MOOC predictive modeling and replication at scale. 2018.

25. A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 2013.

26. J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, Apr. 1982.

27. M. A. HarvardX. HarvardX-MITx Person-Course academic year 2013 De-Identified dataset, version 2.0, May 2014. Title of the publication associated with this dataset: HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0.

28. R. F. Kizilcec and C. Brooks. Diverse big data and randomized field experiments in MOOCs. In C. Lang, G. Siemens, A. Wise, and D. Gašević, editors, *Handbook of Learning Analytics*, pages 211–222. Society for Learning Analytics Research, 2017.

29. R. F. Kizilcec and S. Halawa. Attrition and achievement gaps in online learning. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, pages 57–66, 2015.

30. M. C. Makel and J. A. Plucker. Facts are more important than novelty: Replication in the education sciences. *Educ. Res.*, 43(6):304–316, 2014.

31. D. Merkel. Docker: Lightweight linux containers for consistent development and deployment. *Linux J.*, 2014(239), Mar. 2014.

32. B. A. Nosek, J. R. Spies, and M. Motyl. Scientific utopia: II. restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.*, 7(6):615–631, Nov. 2012.

33. T. Sinha, N. Li, P. Jermann, and P. Dillenbourg. Capturing "attrition intensifying" structural traits from didactic interaction sequences of MOOC learners. Sept. 2014.

34. V. Stodden and S. Miguez. Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *Journal of Open Research Software*, 2(1):1–6, 2013.

35. S. A. Stouffer. *Adjustment during army life*. Princeton University Press, 1949.

36. V. Tinto. Research and practice of student retention: What next? *J. Coll. Stud. Ret.*, 8(1):1–19, 2006.

37. T. J. Tobin and G. M. Sugai. Using Sixth-Grade school records to predict school violence, chronic discipline problems, and high school outcomes. *J. Emot. Behav. Disord.*, 7(1):40–53, Jan. 1999.

38. K. Veeramachaneni, U.-M. O'Reilly, and C. Taylor. Towards feature engineering at scale for data from massive open online courses. July 2014.

39. J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley. Delving deeper into MOOC student dropout prediction. Feb. 2017.

40. D. H. Wolpert. Stacked generalization. *Neural Netw.*, 5(2):241–259, 1992.

41. W. Xing, X. Chen, J. Stein, and M. Marcinkowski. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Comput. Human Behav.*, 58:119–129, 2016.

42. D. Yang, T. Sinha, D. Adamson, and C. P. Rosé. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*, volume 11, page 14, 2013.