Statistical Approaches to the Model Comparison Task in Learning Analytics

Josh Gardner[a,*], Christopher Brooks[a]

[a]*School of Information, University of Michigan*

Abstract

Comparing the performance of predictive models of student success has become a central task in the field of learning analytics. In this paper, we argue that research seeking to compare two predictive models requires a sound statistical approach for drawing valid inferences about comparisons between the performance of such models. We present an overview of work from the statistics and machine learning communities and evaluate several methodological approaches, highlighting four approaches that are suitable for the model comparison task. We apply two of these methods to a learning analytics dataset from a MOOC conducted at the University of Michigan, providing open-source code in R for reproducing this analysis on other datasets. We offer several practical considerations for future implementations, as well as suggestions for future research in both the learning analytics community and the broader field of machine learning.

Keywords: Learning analytics, predictive modeling, machine learning, model evaluation

1. Introduction

Identifying effective models for the prediction of student success is one of the cornerstone tasks in the field of learning analytics. For a field concerned with understanding and improving the educational process, identifying students for whom those improvements need to be made is a critical challenge. Additionally, as the fields of learning analytics and educational data mining have matured, so too have the underlying data mining and machine learning techniques, as well as toolkits and hardware which are used to apply these techniques broadly. As a result, it is now possible for even non-experts to leverage machine learning rapidly, and to build predictive models without having to understand details of techniques or implementations thereof – a development that has had an undoubtedly positive overall impact on the study and practice of learning analytics.

Prior learning analytics research comparing the performance of supervised learning models to predict student success has been extensive, particularly in the past four years, but the use of statistical techniques to compare these models has been mixed. In a review of recent research produced by the International Educational Data Mining Society (2013-2016) and the International Conference on Learning Analytics and Knowledge (2015-2016), we uncovered 12 papers specifically focused on using student behavioral data to predict either pass/fail or dropout outcomes. However, of these 12 papers, seven reported simple matrices of model performance metrics using estimates obtained from cross-validation on the training set–an approach which we will argue below is not appropriate for model comparison–with most also using inferential language not appropriate for the methodology of their comparison. This finding was, unfortunately, not surprising, and we find it to be generally representative of work across the field. Many otherwise excellent learning analytics research projects simply present a matrix of accuracies or model performance statistics and conclude that the best or most preferable modeling method is the one with the highest accuracy, AUC, sensitivity/specificity, F-score (or whatever the outcome of interest may be) on their training dataset(s)-entirely passing over the task of evaluating these differences in performance. Correcting this problem by adopting statistically sound

*Corresponding author; jpgard@umich.edu

methods for the comparison of predictive models will allow the learning analytics community to build an enduring, reproducible base of knowledge about methods for predicting student success.

This paper aims to begin this methodological discussion with the broader community by focusing on the two-model comparison task: determining whether two models differ in performance on a given dataset. In Section 2, we situate the challenge of comparing predictive models of student success within the larger strands of research in the learning analytics community and in the larger machine learning and statistics communities, presenting an overview of research on the two-model comparison task. In Section 3, we dive deeper into this task, presenting a framework for understanding the task of comparing supervised machine learning models. We provide an overview of research on methods for statistically valid, high-reproducibility evaluation of performance differences between classifiers, and present methods which are not valid for the two-model task. In Section 4, we continue this discussion, detailing statistically valid methods for the two-model task. In Section 5, we produce a sample application of two of these valid tests to a real-world MOOC dataset, including sample code written in R. In Section 6, we provide practical recommendations and additional considerations for learning analytics researchers implementing these methods in practice. Finally, we conclude and present directions for future research in Section 7.

This paper specifically addresses problems that emerge when researchers have a single dataset, and need to use this dataset to both build models and obtain reliable estimates and comparisons of model performance. While we can easily obtain an unbiased estimate of the mean and variance of the difference if there is a sufficient supply of data and a holdout set is used [1], in practice this approach is often neither practical nor desirable in the context of learning analytics, giving rise to the problems addressed in this paper. Learning analytics researchers often work with datasets which are limited in size, and face a tension between (a) using as much data as possible for model training to avoid starving statistical models of the data necessary for optimal performance, and (b) reserving data for model evaluation to obtain reliable estimates of this performance.

These situations, both characterized by limited data, are particularly prevalent in learning analytics and educational data mining. The privacy restrictions surrounding educational data limit learning analytics and educational data mining researchers more than machine learning researchers in many other fields. Additionally, researchers are often specifically interested in building models for narrow foci - such as admissions to a given university - and not interested in the models' general performance across institutions, leading to datasets which are limited in size with no option for collecting additional data. As such, adopting and consistently implementing a set of statistically robust techniques to evaluate classifiers in the context of learning analytics is especially critical.

2. Prior Research

The broader field of machine learning research, having confronted the task of model comparison and evaluation earlier than the field of learning analytics, has a relatively robust literature of both experimental and theoretical results on methods for obtaining reliable estimates of model performance. Resampling was the accepted method for generating a sample of estimates for applying a t-test throughout much of the 1990s [2]. However, several researchers began to investigate the soundness of such an approach in the 1990s [3, 4, 5]. [6] provided a seminal early analysis of five statistical tests for comparing supervised classification learning algorithms, presenting experimental results from simulated data to estimate the empirical performance of several common statistical tests. This work formed the foundation for future analyses and presented several results that endure today, including the problems with unadjusted resampled t-tests discussed below. [7] extended this analysis, proposing a modification to the t-statistic for resampling (shown below) and demonstrating both a theoretical justification and empirical evidence of its effectiveness. Another important dimension to statistical tests for model comparisons is introduced in [1, 2, 8] by formalizing measurements of replicability (in addition to power and Type I error rate). These papers present further empirical evidence about the performance of several sampling schemes and test statistics on both simulated and real datasets. Several key findings from this literature are discussed in detail in Section 3 below; for additional reading, see references.

Further research extends these conversations to the multi-model comparison task, including [9, 10, 11, 12]. This work provides insight into the issue of comparing multiple models, and even doing so across multiple datasets or domains. Extending the analysis in this paper to multiple models fits the most realistic cases in which researchers encounter these types of comparison tasks; however, the fundamental issues are the same as the two-model case on which we focus herein. While the multiple comparisons case is analogous, it utilizes tests which specifically correct for multiple comparisons (ANOVA, Newman-Keuls, etc.) – and once a statistically rigorous approach to model comparison has been adopted, taking the step from the 2-model case to the n-model case requires only minor methodological changes. The multiple-model-multiple-dataset case, by contrast, faces an entirely different set of considerations than the task evaluated here [11, 9, 13, 12, 14]. A thorough discussion is beyond the scope of this paper, although the authors intend to address it in future work.

3. Statistical Comparisons of Classifiers

3.1. Basic Terminology

In this paper, we address the problem of comparing the performance of two learning algorithms[1], $\mathcal{L}_A$ and $\mathcal{L}_B$, on a single dataset $\mathcal{D}$. Typically, the difference in model performance is estimated using a sample of differences in performance of $\mathcal{L}_A$ and $\mathcal{L}_B$, P_A and P_B, where the observed differences $x_1, ..., x_n = P_{A,1} - P_{B,1}, ..., P_{A,n} - P_{B,n}$ across several iterations of model training and testing, formed on multiple training and testing partitions $\mathcal{D}_{t,1}...\mathcal{D}_{t,n}$ and $\mathcal{D}/\mathcal{D}_{t,1}...\mathcal{D}/\mathcal{D}_{t,n}$ respectively (each training partition $\mathcal{D}_{t,n}$ might be a single random subsample or a set of training folds using k-fold cross-validation; the corresponding test partition $\mathcal{D}/\mathcal{D}_{t,n}$ is the test dataset). Performance can be measured using any metric and loss function which can be calculated from trained models using the testing data, including accuracy, area under the Receiver Operating Characteristic curve (AUC), sensitivity, specificity, etc. Often $x_1, ..., x_n$ is collected using multiple runs, R, of a partitioning method (such as k-fold cross-validation), where the same randomized partitioning method is applied in each of k folds or iterations for each run. This leads to a total of $R \cdot k = n$ samples in $x_1, ..., x_n$.

In a hypothesis test, we use this sample to draw conclusions about the underlying population from which it was drawn. To do so, we calculate a test statistic, referred to in the literature as t or z, and, utilizing assumptions about the population from which our sample was drawn, make a determination as to whether the observed value of the test statistic provides sufficient evidence to reject a null hypothesis, H_0 (which in this case holds that the true average of the differences in performance is zero; that is $H_0 : \mu = 0$). Specifically, we calculate the probability of observing the test statistic under H_0, and reject H_0 if this probability is less than the desired level of significance α. A Type I error occurs when a hypothesis test rejects H_0 when it is in fact true; a Type II error occurs when a hypothesis test accepts H_0 when it is in fact false. The expected frequency of a Type I error is α; this is also referred to as the "acceptable" level of Type I error throughout this paper and the broader literature [6]. The power of a test is related to its Type II error rate, and represents its ability to correctly lead us to reject H_0 when it is truly false (power measures how well we identify true effects when they do indeed exist).

3.2. Model Evaluation as Experiment

When comparing and evaluating supervised machine learning methods, it is useful to think about model evaluation as an experiment. Indeed, much of the research discussed above is grounded in this language, and it provides a familiar framework for understanding the challenges of classifier performance comparison and the methods used to address them. This experiment has two parts:

> (1) a *sampling* component, where we collect a set of samples of differences in the performance of two learners, and
>
> (2) a *hypothesis testing* component where these samples are used to evaluate hypotheses about comparisons between the performance of learning algorithms $\mathcal{L}_A$ and $\mathcal{L}_B$ on $\mathcal{D}$.

[1] These learning algorithms are also referred to as "learners" throughout as is common in machine learning literature, but this is not to be confused with the use of "learners" to refer to students in a course as is common in educational literature. In the context of this paper we always refer to a "learner" as a machine learning algorithm.

The method adopted in (2) typically entails a set of assumptions about the underlying distribution from which the sample in (1) is drawn; often this includes an assumption that the observations are independent and identically distributed draws from a normal distribution.

The challenge of model comparison in the current context primarily occurs at (2), the hypothesis testing stage: in many (but not all) cases, the way model performance data is collected – using a series of overlapping training and testing datasets, such as in k-fold cross validation – renders the analogy to simple random sampling invalid. In particular, the underlying assumption of independence between the observations in the sample relied upon by common testing methods (such as the Student's t-test) is violated. In this case, one must adjust either the approach to (1) so that the sampling more closely meets the assumptions of the hypothesis test used, or adjust (2) to more accurately reflect the distribution of the experimental sample. Below, we discuss a series of techniques for improving both (1) and (2), reviewing the literature to highlight tests which are not appropriate for the model comparison task as well as those which are well-suited to it. While there is no single best approach to this challenge and further research is needed, adopting these methods – and encouraging further conversations within the learning analytics community – will represent tremendous progress as the field attempts to move toward consensus around effective student success modeling techniques.

3.3. The Two-Model Comparison Task; Hypothesis Testing for Model Comparison

While the statistical community has developed a robust set of tools for statistical inference and hypothesis testing, these tools often do not extend directly to the task of model comparison (or, perhaps more accurately, the model comparison task does not conform to these tests). In practice, we often implement hypothesis tests whose assumptions are not entirely met, but are closely approximated by a sample. When the properties of this sample deviate too far from these assumptions, it can produce undesirable results, including elevated probability of rejecting the null hypothesis when it is true (Type I error) or decreased ability to discern a true effect when it exists (low power). Empirical analyses have helped to determine exactly how much various hypothesis tests are impacted by samples of predictive model performance which do not fully match their assumptions, as measured by their Type I and II errors, power, and reproducibility [6, 8, 2, 1].

Throughout this paper, we apply statistical testing in the context of what we will refer to as *the two-model comparison task,* or simply the model comparison task. This is the task that this paper seeks to provide effective tools for evaluating:

> **The two-model comparison task:** Given dataset $\mathcal{D}$ and two statistical learners $\mathcal{L}_A$ and $\mathcal{L}_B$, identify whether the two learners differ in their performance in predicting the class labels of $\mathcal{D}$.

This task is distinct from, but related to, the determination of whether one model performs "better" than another (the details of this distinction are discussed with respect to one-sided and two-sided tests below).

The challenge in this task in the situations described above is that we would like to use as much of our limited dataset $\mathcal{D}$ as possible in each iteration to obtain a realistic estimate of model performance (because withholding available data would either deprive $\mathcal{L}$ of useful training data in that would improve its test performance P, or testing data that could be used to improve the estimate of this performance). As a result, we devise methods to form overlapping training sets $\mathcal{D}_{t,1}...\mathcal{D}_{t,n}$such that each estimate $P_{A,1}, ..., P_{A,n}$ and $P_{B,1}, ..., P_{B,n}$ is not independent from the others (methods for building these training sets include resampling and k-fold cross-validation, discussed below). If we conduct hypothesis tests under the incorrect assumption that our data are independent, we will make errors more often than expected due to chance, and will not have a reliable estimate of the frequency of making such errors. To correct our probability estimates, we can adjust how we build our sample (how each successive $\mathcal{D}_t$ is formed, or how we calculate $\mathcal{P}_{m,n}$from it), or adjust how we evaluate the sample (how t is calculated and tested). Below, we first discuss a series of methods that, in various ways, fail to accurately draw inferences about populations when applied to the model comparison task, and then discuss statistically valid approaches to this task. Specifically, in the following subsection, "Invalid Methods for Model Comparison," we catalogue several problematic approaches to the model comparison test. In the next section, "Valid Methods for Model Comparison," we highlight successful attempts to correct some of the methods considered previously. Finally, we present an application of these valid methods to a MOOC dataset in our case study.

3.4. Invalid Methods for Model Comparison

A model evaluation approach consists of both a sampling method and a hypothesis testing method; these together determine whether the approach is statistically valid for the model comparison task. The approaches evaluated below, and indeed any such approach, might be statistically invalid for a number of reasons, but regardless of the underlying mechanism they will typically display either (a) an inflated Type I error rate, (b) low statistical power, or (c) low reproducibility.[2] Methods are typically evaluated using these performance rates and by some combination of theoretical and experimental evaluation on real or simulated data.

Perhaps the most common approach to the model comparison task, both in the field of learning analytics and in the broader field of applied machine learning research, is using **no statistical test**. Often researchers, having taken the appropriate steps to collect their data, train a set of learners $\mathcal{L}_A$ and $\mathcal{L}_B$, estimate their performance on $\mathcal{D}$, and present the estimates of performance (and their differences) to speak for themselves. Little needs to be said about this approach, which amounts to simply displaying sample averages, other than that it provides insufficient evidence about whether there exist true differences in model performance. This approach leaves entirely unaddressed the question of whether such results may be spurious and caused simply by randomness inherent in the sampling process.

Another common approach to model comparison involves using repeated random sampling to build a successive series of training and testing sets from $\mathcal{D}$ and then applying a hypothesis test to the results of model training and testing on these samples. The intuition behind such approaches is that if the data is shuffled sufficiently between runs and an adequate number of model performance estimates are built on these successive shufflings, the observed differences in model performance $x_1, ..., x_n = P_{A,1} - P_{B,1}, ..., P_{A,n} - P_{B,n}$ will conform to (or closely approximate) the assumptions of the test statistic used to evaluate its distribution. For several randomization methods and the test statistics they are generally evaluated with, however, this is not the case.

One such method is the **resampling** method. Resampling consists of repeatedly (n times) drawing a specified number of observations (typically two thirds of the training data [6]) without replacement to form training sets $\mathcal{D}_{t,1}...\mathcal{D}_{t,n}$ with the remaining data in each iteration used for model evaluation. Both learners $\mathcal{L}_A$ and $\mathcal{L}_B$ are trained and evaluated on the respective training and testing sets, and their differences in performance $x_1, ..., x_n = P_{A,1} - P_{B,1}, ..., P_{A,n} - P_{B,n}$ are used as estimates of the difference in model performance. The sample population $x_1, ..., x_n$ is used to compute a test statistic t with a Student's t-distribution with $n - 1$ degrees of freedom under the assumption that the individual differences $x_1, ..., x_n$ were drawn independently from a normal distribution.

However, these differences will not conform to a normal distribution, for two reasons. First, $P_{A,1}$and $P_{B,1}$ are not independent: they both use identical test/training sets and are correlated with one another. Second, the training and test sets overlap across trials, to varying degrees (depending on the size of each resample). While this approach can be adjusted to more closely match the assumptions of the test statistic, this is not the case with the unadjusted t-test often used with resampled data. This approach is shown to have an extremely high Type I error rate (indeed, the highest of all methods evaluated in [6]). An additional concern with this method is that the difference between the samples can always be made statistically significant by increasing the number of resamples (an undesirable and problematic property for a statistical test). A similar method is the **bootstrap** method. Bootstrapping differs from the resampling approach only in that the observations drawn from $\mathcal{D}$ to form $\mathcal{D}_t$ are drawn with replacement. Similar to the resampling method, then, the bootstrap method produces training datasets with a high degree of overlap, and it has been observed to produce inflated Type I error rates [13].

10-fold cross validation with an unadjusted t-test is another randomization approach that has proven to be problematic for comparing model performance. In 10-fold cross validation, the dataset is partitioned into $k = 10$ random folds, with a series of learners $\mathcal{L}_{A,1}...\mathcal{L}_{A,10}$ and $\mathcal{L}_{B,1}...\mathcal{L}_{B,10}$ trained on a set $\mathcal{D}_{t,k}$ consisting of a subset of $k-1$ folds and evaluated on the kth fold. Due to the fact that each training set $\mathcal{D}_{t,k}$

[2]Replicability is defined as how well the results of an experiment can be reproduced, or the degree to which the outcome of a test depends on the particular random partitioning of the data used to perform it [1, 2].

shares $k-2$ folds (and therefore 80% of the available training data) with each other training set in a given run, this method also produces high rates of Type I error [8, 2, 6]. This is not due to an inherent flaw in the cross-validation approach itself; several authors have offered modifications to both the cross-validation process and its test statistic which will be discussed below.

In contrast to randomization-based approaches, which prescribe an approach to sampling the model performance data, other methods utilize different test statistics in an effort to produce more reliable evaluations of a given sample. One such method is the **difference-in-proportions test**. In this test, a test statistic, z, is calculated based on a simple comparison between the overall error rates P_A and P_B, of the algorithms $\mathcal{L}_A$and $\mathcal{L}_B$, with $p = \frac{P_A - P_B}{2}$ being the average of the two error rates:

$$z = \frac{P_A - P_B}{\sqrt{2p(1-p)/n}} \tag{1}$$

However, this test faces similar (and additional) problems as those discussed previously. First, the unadjusted z-statistic still assumes independence between the measured error rates P_A and P_B, an assumption which is violated by the identical training sets used to estimate P_A and P_B. [6] identifies potential corrections to this statistic, but observes that these corrections are rarely used in practice and does not empirically test the corrected performance. Additionally, even if corrections for non-independence are applied, the difference in proportions test is unable to account for variation resulting from the choice of training sets: it only makes a single comparison between the model error rates P_A and P_B, and thus is structurally unable to measure variation that occurs with multiple successive runs of model training and testing.

4. Valid Methods for Model Comparison

The main goal of this paper is to describe and implement statistically valid methods for the two-model comparison task. While several such methods are described in this section, we select two for detailed description and implementation in the next section of this paper, with a specific emphasis on methods which match the practical conditions under which learning analytics researchers might apply these tests.

The **corrected resampled t-test** is a method that has seen wide adoption in the machine learning community. [15], noting that the inflated Type I Error rate observed by [6] was due to an underestimation of the variance of resamples, proposed a correction to the t-test (shown in Table 1 below). They propose an adjustment to the estimated variance of a traditional t-statistic, from

$$\hat{\sigma}^2 = \frac{1}{n_2} S_L^2 \tag{2}$$

to

$$\hat{\sigma}^2 = \frac{1}{J} + \frac{n_2}{n_1} S_{\mu_j}^2 \tag{3}$$

where the number of samples $J = k \cdot r$, $S_{\mu_j}^2$ is the sample variance of the estimates, n_1 is the number of samples used for training, and n_2 is the number of samples used for testing. This modification is intended to account for the correlation between resamples, which the resampled t-test "boldly" assumes to be zero, by instead estimating that correlation as $\rho = \frac{n_2}{n_1+n_2}$ [14; specifically see sections 3 and 4 for discussion and theoretical justification of this modification]. This modified test statistic is used with the Student's t-distribution and $n-1$ degrees of freedom, in otherwise exactly the same way as an unadjusted Student's t-statistic. This adjustment has been shown to have acceptable levels of Type I error and high replicability, particularly when using many ($R = 100$) runs [1]. Additionally, the correction avoids another critical flaw with the unadjusted resampled t-test, mentioned above, that the significance of the test statistic can be increased without bound simply by increasing k (the number of resamples).

This same corrected test statistic can be used for cross-validation in a general approach we refer to as **corrected cross-validation** [16]. Cross-validation differs from resampling in at least two ways that are relevant to statistical testing. First, the test and training sets are guaranteed not to overlap within each

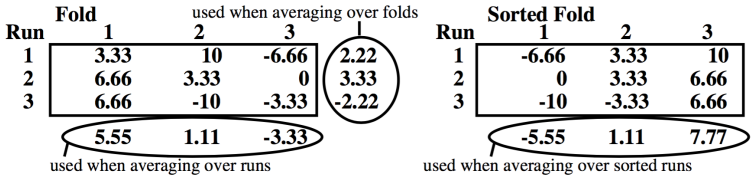

Figure 1: Example from [8] illustrating the data used for averaging over sorted runs (right) compared to averaging over folds or over runs (left).

fold. Second, while training sets do overlap across folds (each training set shares $k-2$ folds of data with the others), they do so in a consistent way. Together, due at least in part to these two factors, it has been empirically demonstrated that a corrected cross-validation approach achieves superior performance relative to the corrected resampled t-test [2, 1, 6].

Several proposals have offered slightly modified versions of the corrected cross-validation approach with different corrections to t, and specific proposed values for R and k. A 5x2 cross-validation approach ($R = 5, k = 2$) was proposed in [6], seeking to avoid correlation between the accuracy measured across folds by using only two folds, which results in nonoverlapping training sets within each run. When this **5x2cv paired t-test** is used in combination with a modified test statistic $\tilde{t}$ with 5 degrees of freedom (shown in table 1), it achieves an acceptable level of Type I error [6]. However, it has been widely recognized that this test achieves low power [15, 6, 8]. Additionally, the test has been shown to have low replicability due to the large random variation inherent in using only 2 folds per run and the fact that the modified test statistic $\tilde{t}$ depends only on the difference $x1, 1$ and not on the full set of differences $x_{1,1}, ..., x_{r,k}$ [8, 1, 6]. Additionally, despite the improved performance, the choice of $R = 5$ is somewhat arbitrary, and has been criticized as ad hoc and lacking in theoretical justification [15, 16].

Another proposal that has been shown to have both better power and higher replicability than the 5x2cv approach is 10x10cv with a variance correction, known as the **corrected repeated k-fold cv test**. This test utilizes a variance correction similar to the corrected resampled t-test, but pairs the modified test statistic with a cross-validation approach. Together, these modifications achieve an acceptable Type I error rate (equal to α, as expected) and better statistical power than 5x2cv [1]. Notably, in a direct comparison between corrected resampling with $R = 100$ and 10x10cv (both of which require 100 iterations of model training and thus the same computational effort), 10x10cv with the corrected repeated k-fold cv test achieves higher replicability [1].

A novel sampling-based approach to the model comparison task, evaluated in [8], shows acceptable Type I Error, high power, and high replicability: the **sorted runs sampling scheme**. This is a modified approach to cross-validation, in which the results $P_{R,1}, ..., P_{R,k}$ for each run are first sorted, and then the averages of the ordered folds are taken according to their respective rank across runs. This results in a sample of k estimates for each learner, where $\overline{P_1} = \frac{1}{R}\sum_{i=1}^{R} min(P_1...k)$ is the average of the lowest scores for a given learner in each of the R runs, P_2 is the average of the second-lowest score for a learner in each run, etc. An illustration of this scheme is shown in Figure 1.

The sorted runs sampling scheme has been shown to achieve a better Type I Error rate than 10x10cv, but at the cost of reduced power [8]. The sorted runs sampling scheme also has the appealing result that it results in a sample for which the independence assumption is not heavily violated, so that no correction in variance or degrees of freedom is required, allowing it to be used with a normal, uncorrected t-statistic and $df = n - 1$.

Below, we present an implementation of the **10x10cv with the corrected repeated k-fold cv test** and the **sorted runs sampling scheme with a standard t-test**. We select the 10x10cv as an example of an approach that can be easily applied to the cross-validation that is already common in most predictive models of student performance, favoring 10x10cv over 5x2cv because of the former's greater statistical power, and over the corrected resampling approach because of its greater reproducibility. Additionally, using 10 folds (instead of 2) substantially reduces the pessimistic bias of the model performance estimates $P_{m,k}$ on each fold, because the models are trained on datasets closer in size to that of the actual dataset we are

Table 1: Overview of statistical tests for the model comparison task with acceptable error rates; (3) and (4) represent methods compared in case study in Section 5.

Test	**Description**	**Sampling Method**	**Test Statistic**	**Distribution**
(1) 5x2cv paired t-test [5]	Adjusted t-test calibrated for use with 5x2cv. Lower power than (2,3,4).	2-fold cross-validation	$t = \frac{x_{11}}{\sqrt{\frac{1}{5}\sum_{j=1}^{5}\hat{\sigma}_j^2}}$	Student's t-distribution, $df = 5$
(2) Corrected resampled t-test [14]	Standard t-test with adjusted variance. Avoids significance inflation for increasing number of resamples. Lower reproducibility than (3).	Random Subsampling	$t = \frac{\frac{1}{n}\sum_{j=1}^{n} x_j}{\sqrt{(\frac{1}{n}+\frac{n_2}{n_1})\hat{\sigma}^2}}$	Student's,t-distribution, $df = n - 1$
(3) Corrected repeated k-fold cv test [3]	$r \cdot k$-fold CV, with the same variance correction as the corrected resampled t-test above.	Cross-validation	$t = \frac{\frac{1}{k \cdot r}\sum_{i=1}^{k}\sum_{j=1}^{r} x_{ij}}{\sqrt{(\frac{1}{k \cdot r}+\frac{n_2}{n_1})\hat{\sigma}^2}}$	Student's,t-distribution, $df = n - 1$ (where $n = k \cdot r$)
(4) Paired t-test with sorted runs sampling scheme [3]	Adjusted sampling scheme with standard t-test.	Sorted runs sampling scheme	$t = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i}{\sqrt{\hat{\sigma}^2}/\sqrt{df+1}}$	Student's t-distribution, $df = r - 1$; note that x_i is summed across sorted runs.

estimating their performance on, $|\mathcal{D}|$. We utilize the sorted runs sampling scheme both because it has been demonstrated to perform at least as well as 10x10cv with the adjusted t-statistic, and because it shows even better reproducibility – one of the key components of an effective model comparison test discussed above. However, there are situations in which one might cautiously utilize an alternative approach (such as the adjusted resampled t-test or 5x2cv) instead of these approaches, particularly if training two learners a total of 100 times each ($R \cdot k = 100$) is not practical. Additional effective methods for evaluating model performance are mentioned in sections 5 and 6 below.

5. Case Study - Evaluating Classifiers on MOOC Data

In this section, we present a case study in model comparison to illustrate the application of the two approaches identified above that are both statistically and practically appropriate for comparing classification algorithms in this case. These are the corrected repeated k-fold cross-validation test with 10x10cv, and the sorted runs sampling scheme with a paired t-test.

Our dataset is compiled from a financial literacy MOOC run on the Coursera platform through the University of Michigan between February and May 2014. From the raw Coursera clickstream logs, a set of features similar to that of [10] was assembled. In particular, an appended featureset was used, where the same set of p features is collected for each week w of the course, leading to a "wider" dataset each successive week. The intuition behind this approach is that it expands the feature space for temporal prediction problems, allowing the learner to exploit more information about each student and their activity patterns as the course proceeds. This particular analysis utilized an appended feature set representing the learner

Table 2: Description of features. The same set of features was collected for both of the first two weeks of the course, and the features were appended (column-wise) to create a set of these features for each of the weeks. For further detail, see [10].

Feature	Description
Forum Views	Count of pageviews of course forum pages.
Active Days	Count of days learner registered any activity in the course (maximum of 7).
Quiz Attempts	Count of attempted quiz questions.
Quiz Exams	Count of attempted exam questions.
Quizzes Human Graded	Count of attempted human-graded quiz questions (0 for all learners; no human-graded quizzes in this course).
Forum Posts	Count of forum posts (this includes both forum posts and comments, which Coursera tracks separately).
Direct Nodes	Number of distinct users a given user responded to on the forums (direct-reply).
Thread Nodes	Number of distinct uses a given user posted in the same forum with (thread-reply).

activity for the first two complete weeks of the course. The original dataset consisted of 51,088 observations of 16 features (8 per week) based on those of [10]. Students who had shown no activity in the first two weeks of the course were removed as inactive students, because including them makes the prediction problem too easy (this is a common approach; as an example, see [17]). 75% of these remaining observations were used for the model training and evaluation (to retain a holdout set for a robustness check comparing hypothesis test results to true test performance or for future research) resulting in a dataset with $|\mathcal{D}| = 23,537$ observations. The feature set was used to predict a binary variable for whether a learner would register any activity in the third week of the course. 4,074 observations were positive (i.e., dropouts), resulting in a dataset with a positive:negative outcome class balance of approximately 1:5.

The third week of activity was chosen as the target for prediction because it appears to be a turning point in the course: many learners register no activity after the first or second week of the course, so the dropout rate is much higher than in subsequent weeks (and we might expect many of those learners to be "explorers" not interested in course completion; see figure 2). Unlike the early dropouts, a learner who shows activity in the first two weeks, persisting to the third week, and then drops out is more likely to be making an honest effort at completion. Additionally, the large decrease in dropouts after week 3 suggests that a higher proportion learners who persist past the third week are able and motivated to successfully complete the course, and that effective predictions (leading to effective interventions) here could substantially impact course completion.

Two models – a multilayer perceptron (also known as a feedforward artificial neural network) and a classification tree – were fit to the data[3]. Hyperparameters for the models are tuned internally by the caret package in R, and the only hyperparameter setting that was adjusted was setting the number of layers for the neural network ($L = 3$). These models were selected because both are used in student success models (the former is gaining widespread use, and the latter is common in previous research) and both make few

[3]The multilayer perceptron used was from the implementation in the RSNNS package in R [18]; the classification tree was an implementation from the rpart package which itself is based on Breiman's orifinal formulation [19]; the caret package was used as a wrapper for the training/testing procedure [7]. We refer the interested reader to these citations for details on the implementations of these models.

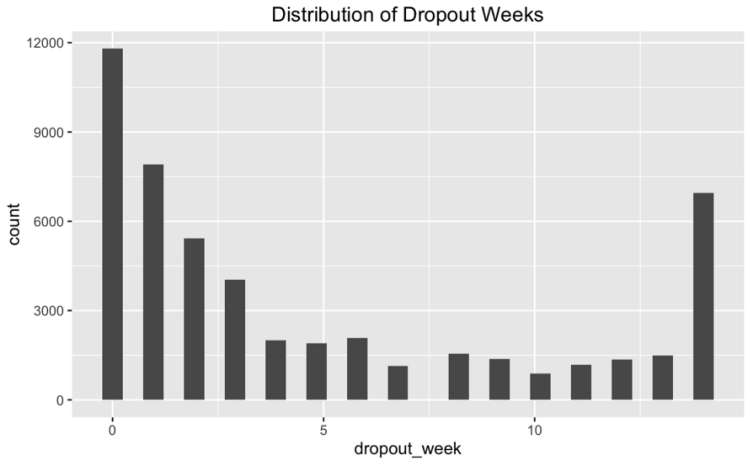

Figure 2: Learner dropout distribution over course weeks.

and weak (if any) assumptions about the underlying data and require no pre-training feature selection. Both models were fit and evaluated using the same subsamples, allowing for valid comparisons between the models (see [7] for details on implementation in R). Caret provides statistics for each model's performance across each individual resample, and our implementation uses identical resamples for both models. Because both of the evaluation approaches identified above could utilize 10x10cv, this was used (the sorted runs sampling scheme can be used with any number of folds, but the adjusted cross-validation approach achieved the performance discussed above specifically with 10x10cv).[4]

The results in Table 3 show the sample averages for three different performance metrics-accuracy, area under the Receiver Operating Characteristic, and specificity – and the resulting comparisons between them. These results are intended to provide a multidimensional view of model performance and to give three points of comparison for the different test statistics used. Additionally, the three performance metrics are used to highlight different approaches to the challenge of the model comparison task in the context of highly-unbalanced student performance datasets where a specific class (such as a dropout or failure) is of particular interest. Note that, in practice, conducting three different comparisons would require a correction for multiple comparisons, such as a Bonferroni correction, and would effectively be moving beyond the two-model comparison task. Since these results are for demonstration purposes only, we will not provide a correction.

A detailed review of the results shows several important practical features of the two-model comparison task and the two test statistics used to evaluate this comparison. In general, the two test statistics (corrected 10x10cv and sorted runs) provide similar p-value estimates. The corrected 10x10cv provides slightly more conservative p-value estimates in each case, which reflects its more conservative variance correction. The increased estimated variance for the 10x10cv approach may be more conservative than necessary (variance too large, shrinking the test statistic), or the uncorrected variance of the sorted runs test statistic may not be conservative enough (variance too small, inflating the test statistic). Interestingly, in this particular application, the two test statistics yielded nearly identical numerators for each of the three comparisons (accuracy, AUC, specificity), despite quite different methods for calculating those sample means (sorted mean across each run, vs. mean across all folds within run). The difference in p-value for the sorted runs test statistic, then, is entirely due to the difference in estimated variance and the difference in degrees of freedom used to calculate p-values from the test statistics. Note that both test statistics use the normal calculation for a Student's t-distribution (with degrees of freedom calculated as shown in Table 2).

[4]Sample code for implementing this model comparison is available in the following repository: `https://bitbucket.org/jpgard/lak-mla-2017/src`

Table 3: Model performance results with p-values for corrected test statistics. Standard deviations shown in parentheses. Note that the models reflected in the Accuracy' column are different than the models reflected in the AUC and Specificity columns – two separate models of each type were trained; one set optimized for accuracy (whose accuracy results are shown), and another set optimized for the more imbalance-robust AUC measure (whose AUC/Specificity results are shown). Note that the p-values of the significance tests used to not always support the conclusions of a simple "choose the best classifier" analysis and indicate the potential that observed differences in performance between the models may be spurious. Additionally, note that the specificity of 0 for the Multi-Layer Perceptron indicates that no positive cases (dropouts) were correctly identified by this model.

	Model	Avg. Performance	Avg. Difference MLP - CART	p-value	
				Corrected 10x10cv	Sorted Runs
Accuracy	Multi-Layer Perceptron	0.826911 (0.000171)	0.000382 (0.000641)	0.089406	0.022840
	Classification Tree	0.826528 (0.000660)			
AUC	Multi-Layer Perceptron	0.596456 (0.015724)	0.029627 (0.034840)	0.016314	0.000196
	Classification Tree	0.566828 (0.034273)			
Specificity	Multi-Layer Perceptron	0.000000 (0.000000)	-0.000736 (0.001328)	0.114304	0.047164
	Classification Tree	0.000736 (0.001328)			

The two test statistics track each other closely, with both yielding their lowest and highest p-values for the same comparisons (AUC and specificity, respectively). Additionally, the different values of these test statistics across metrics clearly reflect the fact that we are not always warranted in concluding that an observed difference is likely to reflect a true difference in model performance. This is a desirable property of null hypothesis statistical testing, which allows us to distinguish between true and potentially spurious results. Particularly in the case of specificity (the proportion of true positives, i.e. dropouts, which are correctly identified as positives), which is a metric of great interest to practitioners given the highly unbalanced nature of many learning analytics datasets and the high cost of correctly identifying the minority class (in this case, dropouts), both test statistics show that a conclusion that the two models had true differences in performance would be weak at best (depending on the chosen α), while a reported result of the simple differences in average might lead readers to conclude otherwise.

The differences observed here almost certainly reflect the peculiarities of this particular dataset, but this dataset is peculiar in one way that is quite similar to most other student performance datasets – it is highly unbalanced, with a dropout:retention ratio of approximately 1:5, as noted previously. While there is little research on how these test statistics perform on unbalanced datasets, [2] notes that the replicability of the sorted runs sampling scheme actually increases as the data becomes more unbalanced, because learners tend to predict the majority class more as that class becomes more prevalent (they do not conduct this evaluation for any other sampling scheme).

6. Using Tests in Practice - Cautions and Caveats

The above application demonstrates an approach to the two-model comparison task in an effort to begin a larger discussion about effective methods for model comparison and evaluation. However, grounding research in statistically valid methodology is not simply about implementing a new test statistic. Several other considerations need to be accounted for in the research and modeling process.

One critical distinction, made at the outset of this paper, is between the two-model comparisons discussed in this work and multiple-model comparisons (or comparisons across multiple datasets). In practice, researchers are likely to encounter the multiple comparison scenario, and methods beyond those described in this paper should be used to avoid invalid inferences. We intend to address such methods in the context of predictive models of student success in future work.

Another potential pitfall that is not new but has received renewed focus is that of obtaining overly optimistic results from about model comparison through multiple, repeated rounds of viewing model comparisons [20, 21]. By repeatedly building different models or feature sets, collecting model performance samples, and testing these various iterations, researchers increase the likelihood of observing spurious correlation and mistaking this for a true difference in model performance. The broad array of approaches to both feature extraction and modeling make this a uniquely challenging element of implementing these procedures in practice, and suggest that a deepened emphasis on pre-registration (specifying hypotheses and methods prior to experimentation) would help diminish such concerns in future evaluation of student performance predictors.

Another consideration – but one which works in the opposite direction of the optimistic bias above – is that approaches which use cross-validation are subject to a pessimistic bias due to smaller datasets being used for training than the actual size of $\mathcal{D}$ a model trained on $k-1$ folds only contains $\frac{k-1}{k}$ of the data (i.e., models trained in a 10-fold cross-validation approach only use 90% of the data). The effective size of this bias depends both on the size of the dataset and the sensitivity of the modeling approach(es) to this size. The multilayer perceptron used above, for example, is likely more sensitive to the training dataset size than the classification tree, and the performance achieved with the perceptron on this relatively small dataset should be seen as a conservative estimate of its performance on datasets of size $|\mathcal{D}|$. Further research on this issue is recommended below, and it reflects a larger set of open questions about whether researchers ought to use different corrections for different types of models in the comparison task.

A final consideration, again familiar to statisticians, is the distinction between two-sided and one-sided significance test, and its implications for conclusions about model performance [13]. The tests discussed in this paper, and in almost all of the literature described above, are generally two-sided tests, although a similar methodology can (cautiously) be used in conjunction with one-sided tests, if such a test is appropriate. Researchers should keep in mind the difference between these two tests: in particular, the two-sided test allows us to evaluate $H_A : \mu \neq 0$ while a one-sided test allows us to evaluate $H_A : \mu > 0$ or $H_A : \mu < 0$.

While the ability to draw a specific conclusion about whether $P_A - P_B \geq 0$ (that is, whether model A performs better than model B) is attractive, using a two-sided test is typically only warranted if the researcher has some *a priori* reason to believe that a given classifier performs better than another. If this is not the case, and a researcher views model performance data before making a determination about the use or direction of a one-sided test, then the results are effectively biased. Using a one-sided test entirely disregards the possibility of a performance comparison going in the opposite direction-but in the general case such an assumption seems unwarranted. Further discussion of this issue in the context of the model comparison task is also recommended below.

As mentioned above, the two tests implemented here – the corrected repeated k-fold cross-validation test with 10x10cv and the sorted runs sampling scheme with a Student's t-test – are not the only acceptable model comparison methods. While we highlight our reasons for favoring these approaches in Table 2 and the analysis above, there are situations in which implementing these two specific tests may not be practical-such as when conducting 100 iterations of model training is not feasible. In these cases, utilizing one of the other approaches, such as adjusted resampling or adjusted 5x2cv, would still be far preferable to not conducting a statistical test. The researcher would need to keep in mind the caveats to these methods mentioned in Table 2 and above (for instance, about the lower replicability or power of certain tests). Other approaches not discussed here may also be useful to researchers, such as the ROC Convex Hull (ROCCH) method proposed by [5]. This method is particularly useful when the learners being compared will be used to predict on new datasets with target conditions (class balances, misclassification cost) that are different and unknown in advance; it is not, however, an approach based on hypothesis testing. [6] also highlights McNemar's test as a potential alternative, but this test has the same inability to measure (and therefore account for) sample variance as the difference-of-proportions test, and showed no substantial performance benefits in that

analysis. Additionally, while the non-parametric McNemar's test may be a more promising alternative to t-testing due to its limited assumptions about the data, it may also suffer when tested on non-independent overlapping data sets [9], though the degree to which this affects its performance is unclear.

7. Conclusion

In this paper, we present an overview of the two-model comparison task, with an emphasis on its application to predictive models of student success. We provide an overview of predictive research in learning analytics, and of statistical and machine learning research directed at adjusting traditional sampling and hypothesis testing approaches to the model comparison task. We present a detailed explication of several of these methods, including invalid methods that researchers should avoid using, and methods that have demonstrated empirical and theoretical support. Finally, we demonstrate an application of two tests, the corrected 10x10 cross-validation and the sorted runs sampling scheme with a t-test, to an actual comparison of two classifiers in the context of MOOC dropout prediction. This application was used to compare and contrast the two methods, both of which are acceptable for use in future research.

7.1. Future Research

As model comparison and evaluation rises as a priority for applied data science in general, and learning analytics in particular, further research is needed on several fronts. Overall, additional studies should evaluate the several different approaches for sampling and hypothesis testing discussed in this paper. A robust foundation of theoretical and applied research, with the latter using both simulated and real data, is an essential prerequisite to effective evaluation of student success models. Future research should revisit the findings of previous predictive modeling experiments and seek to evaluate their results through more statistically rigorous frameworks.[5] Additionally, this research should include discussions of end-to-end methodologies for effective machine learning comparison and evaluation, including developing a consensus around unbiased approaches to feature extraction, model selection, and model evaluation. Such analyses will need to extend beyond the two-model case discussed here, and should be applicable to $n > 2$ classifiers. Given the variety of modeling approaches that continue to emerge, future research also needs to evaluate whether all models should be treated equally in the case of hypothesis testing – as is the current standard – or whether adjustments should be made for different levels of variation across different types of models ([23] presented an early analysis of this challenge). Similarly, researchers should begin more active discussions on how other statistical methods, such as one-sided vs. two-sided hypothesis testing, should be implemented in experiments and under what conditions. Researchers should continue to discuss methods for evaluating classifier performance on highly skewed data, moving the conversation beyond mere accuracy and developing specific methodologies to evaluate classifier performance on unbalanced datasets. While this paper highlights a two-class classification problem, there is a particular need for simple, effective measures of multiclass performance metrics, as many student success prediction problems are multiclass (i.e., grade prediction). Finally, null hypothesis statistical testing is not the only reasonable approach to this problem: Bayesian methods are also attracting increased interest in the machine learning community and these methods have valuable additional insights to add to the model comparison task. Although a comparison ogf null hypothesis statistical testing and Bayesian testing is beyond the scope of this paper, t is our hope that future research further explores Bayesian testing methods and utilizes these insights, such as the concept of practical significance, to evaluate predictive models of student success[14].

References

[1] R. R. Bouckaert, E. Frank, Evaluating the replicability of significance tests for comparing learning algorithms, in: Advances in Knowledge Discovery and Data Mining, Springer, Berlin, Heidelberg, 2004, pp. 3–12.

[5] For a promising effort in this vein, see [22].

[2] R. R. Bouckaert, Estimating replicability of classifier learning experiments, in: Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04, ACM, New York, NY, USA, 2004, pp. 15–.
[3] D. D. Jensen, P. R. Cohen, Multiple comparisons in induction algorithms, Mach. Learn. 38 (2000) 309–338.
[4] A. Feelders, W. Verkooijen, On the statistical comparison of inductive learning methods, in: D. Fisher, H.-J. Lenz (Eds.), Learning from Data, Lecture Notes in Statistics, Springer New York, 1996, pp. 271–279.
[5] F. Provost, T. Fawcett, Robust classification for imprecise environments, Mach. Learn. 42 (3) (2001) 203–231.
[6] T. G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput. 10 (7) (1998) 1895–1923.
[7] M. Kuhn, A short introduction to the caret package.
[8] R. R. Bouckaert, Choosing between two learning algorithms based on calibrated tests.
[9] N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011.
[10] W. Xing, X. Chen, J. Stein, M. Marcinkowski, Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization, Comput. Human Behav. 58 (2016) 119–129.
[11] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30.
[12] S. Garcia, F. Herrera, An extension on"statistical comparisons of classifiers over multiple data sets"for all pairwise comparisons, J. Mach. Learn. Res. 9 (Dec) (2008) 2677–2694.
[13] O. T. Yildiz, E. Alpaydin, Senior Member, Ordering and finding the best of k¿2 supervised learning algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 28 (3).
[14] G. Corani, A. Benavoli, J. Demšar, F. Mangili, M. Zaffalon, Statistical comparison of classifiers through bayesian hierarchical modelling`arXiv:1609.08905`.
[15] C. Nadeau, Y. Bengio, Inference for the generalization error, Mach. Learn. 52 (3) (2003) 239–281.
[16] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.
[17] K. Veeramachaneni, U.-M. O'Reilly, C. Taylor, Towards feature engineering at scale for data from massive open online courses`arXiv:1407.5238`.
[18] C. Bergmeir, J. Benítez, Neural networks in R using the stuttgart neural network simulator: RSNNS, J. Stat. Softw. 46 (1) (2012) 1–26.
[19] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and regression trees, CRC press, 1984.
[20] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning. 2001, NY Springer.
[21] G. Vanwinckelen, H. Blockeel, On estimating model accuracy with repeated cross-validation, in: BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning, 2012, pp. 39–44.
[22] J. M. L. Andres, R. S. Baker, G. Siemens, D. GAŠEVIĆ, C. A. Spann, Replicating 21 findings on student success in online learning.
[23] S. L. Salzberg, On comparing classifiers: Pitfalls to avoid and a recommended approach, Data Min. Knowl. Discov. 1 (3) (1997) 317–328.