Identifying At-Risk Students in Massive Open Online Courses

Jiazhen He[†⋆] **James Bailey**[†⋆] **Benjamin I. P. Rubinstein**[†] **Rui Zhang**[†]
[†]Department of Computing and Information Systems, The University of Melbourne, Australia
[⋆]Victoria Research Laboratory, National ICT Australia
jiazhenh@student.unimelb.edu.au, {baileyj, brubinstein, rui.zhang}@unimelb.edu.au

Abstract

Massive Open Online Courses (MOOCs) have received widespread attention for their potential to scale higher education, with multiple platforms such as Coursera, edX and Udacity recently appearing. Despite their successes, a major problem faced by MOOCs is low completion rates. In this paper, we explore the accurate early identification of students who are at risk of not completing courses. We build predictive models weekly, over multiple offerings of a course. Furthermore, we envision student interventions that present meaningful probabilities of failure, enacted only for marginal students. To be effective, predicted probabilities must be both well-calibrated and smoothed across weeks. Based on logistic regression, we propose two transfer learning algorithms to trade-off smoothness and accuracy by adding a regularization term to minimize the difference of failure probabilities between consecutive weeks. Experimental results on two offerings of a Coursera MOOC establish the effectiveness of our algorithms.

Introduction

With the booming popularity of Massive Open Online Courses (MOOCs), such as Coursera, edX and Udacity, MOOCs have attracted the attention of educators, computer scientists and the general public. MOOCs aim to make higher education accessible to the world, by offering online courses from universities for free, and have attracted a diverse population of students from a variety of age groups, educational backgrounds and nationalities. Despite these successes, MOOCs face a major problem: low completion rates. For example, Table 1 shows the student participation in the first offering of a Coursera MOOC *Discrete Optimization* by The University of Melbourne in 2013, which illustrates low completion rates seen by other MOOCs. Of 51,306 students enrolled, only 795 students completed: a completion rate of 1.5%. And only 27,679 (about 54%) students ever engaged in lectures and quizzes/assignments; even restricted to this group, the completion rate was a mere 2.9%.

In this paper, we explore the accurate and early identification of students who are at risk of not completing courses.

Table 1: Student participation in the first offering of *Discrete Optimization (DisOpt)* launched in 2013; actions are measured in terms of viewing/downloading lectures and completing quizzes/assignments.

	DisOpt MOOC
Number of students enrolled	51,306
Number of students with actions	27,679
Number of students completed	795

Early prediction can help instructors design interventions to encourage course completion before a student falls too far behind. We focus on Coursera MOOCs, which often last for several weeks with students engaging in activities such as watching/downloading lectures, attempting assignments/quizzes, and posting to/viewing discussion forums. To obtain early predictions, we build models weekly and leverage multiple offerings of a course to obtain ground truth to supervise the training of our models. Exploration of predictive analysis on MOOCs across multiple offerings has been limited thus far, but is nonetheless important, since data distributions across offerings is likely non-stationary: *e.g.*, different cohorts of students enroll in offerings, and course materials (lectures and assignments) are refined over time. It is not clear *a priori* whether a model trained on previous offerings will serve a new offering well.

A key aspect of our approach is a plan for interventions that involve presenting at-risk students with meaningful probabilities of failure. We hypothesize that such carefully crafted interventions could help students become aware of their progress and potentially persist. However a necessary condition for such an approach to be effective, is to have probabilities that are well calibrated. By focusing on intervening with only those students near the pass/fail borderline, we aim for students who could be motivated by being 'nearly there' in succeeding in the class. Our intervention plan expressly avoids displaying failure probabilities for high-risk students, for fear of discouraging them from further participation in the course. Therefore calibration is not necessary across the entire unit interval, only near 0.5.

By examining individual students' failure-probability trajectories, we observe huge fluctuations across weeks, which is undesirable for a number of reasons, such as confusing students or undermining credibility of the intervention system. Therefore, we impose a requirement of smoothed probabilities across consecutive weeks. Towards this end, we propose two transfer learning algorithms—Sequentially Smoothed Logistic Regression (LR-SEQ) and Simultaneously Smoothed Logistic Regression (LR-SIM)—to balance accuracy with smoothness. These algorithms add a regularization term, which takes the probabilities in consecutive weeks into account, so as to minimize their difference. While LR-SEQ uses knowledge from the previous week to smooth the current week in a sequential fashion, LR-SIM learns across weeks simultaneously.

Contributions. The main contributions of this paper are:

- The first exploration of early and accurate prediction of students at risk of not completing a MOOC, with evaluation on multiple offerings, under potentially non-stationary data;
- An intervention that presents marginal students with meaningful failure probabilities: to the best of our knowledge a novel approach to completion rates;
- Two transfer learning logistic regression algorithms which would be practical for deployment in MOOCs, for balancing accuracy & inter-week smoothness. Training converges quickly to a global optimum in both cases; and
- Experiments on two offerings of a Coursera MOOC that establish the effectiveness of our algorithms in terms of accuracy, inter-week smoothness and calibration.

Related Work

Low completion rates is a major problem in MOOCs. One way to address this problem is to identify at-risk students early and deliver timely intervention. A few studies have focused on predicting students' success/failure in MOOCs. Jiang et al. (2014) use students' Week 1 assignment performance and social interaction to predict their final performance in the course. Ramesh et al. (2013) analyze students' online behavior and identify two types of engagement, which is then used as a latent feature to help predict final performance. The same methodology is then used to predict a similar task for whether students submitted their final quizzes/assignments (Ramesh et al. 2014). However, these predictions are not studied for intervention. Instead, we propose to intervene students by presenting meaningful predicted probabilities, with only those who are on the pass/fail borderline targeted. Stages of targeted interventions have parallels to epidemiological approaches to education (Lodge 2011) and are conceptually similar to defense-in-depth and perimeter defenses in computer security (Kaufman, Perlman, and Speciner 2002).

A task similar to ours is dropout prediction, where the class label is whether or not a student will dropout instead of fail. While we have not focused on dropout prediction, our techniques should readily apply to this setting. Most studies focus on developing features from students' behaviors and engagement patters to help prediction. Kloft et al. (2014) predict dropout from only click-stream data using a Support Vector Machine (SVM). Taylor, Veeramachaneni, and O'Reilly (2014) utilize crowd-sourced feature engineering (Veeramachaneni, O'Reilly, and Taylor 2014) to predict dropout based on logistic regression. Balakrishnan (2013) extracts features mainly from discussion forums and video lectures, and employs Hidden Markov Models (HMMs) to predict student retention. Halawa, Greene, and Mitchell (2014) study accurate and early dropout prediction using student activity features capturing lack of ability or interest.

Previous work has concentrated on using different data sources, carrying out feature engineering and using off-the-shelf classifiers evaluated only within one offering of a course. However to the best of our knowledge, none have recognized the importance of calibrated prediction probabilities for predicting failure or dropout; explored and motivated the need for temporally smooth prediction probabilities in the context of education and interventions; applied transfer learning for this purpose; and shown that a model trained a previous MOOC offering can be used effectively for predicting within a future offering.

Another research area exploring low completion rates is correlative analysis to understand factors influencing success/failure or dropout/retention. Various factors have been investigated, such as demographics (DeBoer et al. 2013b; 2013a), student behavior and social positioning in forums (Yang et al. 2013), sentiment in forums (Wen, Yang, and Rosé 2014) and peer influence (Yang, Wen, and Rose 2014). This can help better understand the reason for success/failure or dropout/retention and potentially help devise detailed feedback, but it is not our focus in this paper. However in the future we plan to combine it with predictive analysis to provide accurate prediction and effective intervention.

Problem Statement

We explore the accurate and early prediction of students who are at risk of failing, which we cast as a supervised binary classification task where possible class labels are whether or not a student will fail a course.

Predicted probabilities can serve a dual purpose, both for the identification of at-risk students and within subsequent intervention. We hypothesize that carefully employing the predicted probabilities as part of an intervention message could incentivize students to invest further in the course. Specifically, we propose to intervene with those who are on the pass/fail borderline rather than high-risk students. For example, given a 0.45 predicted probability, a hypothetical intervention message might resemble the following.

> Great work on your efforts so far—you're nearly there! In fact our statistical models suggest your profile matches students with a 55% chance of passing. This is based mainly on your lecture downloads this week. We'd like to encourage you to watch lecture 4 and post to the board. Doing just these 2 activities have greatly improved outcomes for students like you!

By targeting only those students near the pass/fail border, we are focusing on the part of the cohort that with an incremental investment could most personally benefit and increase the course pass rate.

Our application motivates 4 requirements of the learner.

- **Early & accurate predictions** enable timely interventions for at-risk students, with minimal unfounded and missing interventions;
- **Well-calibrated probabilities** allow proper targeting of interventions to those students who are truly near the classifier's decision boundary and to supply meaningful interventions: *e.g.*, approximately 60% of students with a risk prediction of 0.6 should eventually fail the course;
- **Smoothed probabilities** across consecutive weeks mitigate large fluctuations from slight changes in activities. Such fluctuations (*cf.* Figure 1) may undermine the credibility of intervention messages—we opt for consistent feedback. Moreover smoothing admits a principled approach to learning from the entire course when distributions change and even feature spaces change (*i.e.*, a form of regularization through transfer learning); and
- **Interpretable models** suggest additions to intervention messages such as explanations for the current prediction and possible areas for improvement. Moreover such models can be useful in informing instructors on the profiles of successful vs. struggling students.

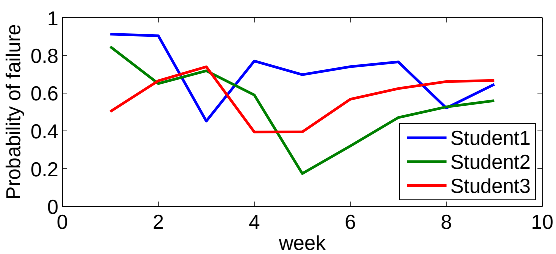

Figure 1: Failure-probability trajectories for three students across nine weeks produced by logistic regression with cross-validation performed weekly on *DisOpt* launched in 2014.

Algorithms

In initial experiments we explored a variety of supervised binary classifiers for predicting failure weekly: regularized logistic regression, SVM (LibSVM), random forest, decision tree (J48), naive Bayes, and BayesNet (in weka with default parameters used). Results (omitted due to space) indicate that regularized logistic regression performs best in terms of Area Under the ROC Curve (AUC), followed by BayesNet, naive Bayes, random forest, decision tree and SVM. Only BayesNet is comparable to logistic regression, whilst SVM performs worst. In addition to the advantage of outperforming other classifiers, logistic regression: produces interpretable linear classifiers with weights indicating relative importance (under certain assumptions); is naturally well-calibrated (Niculescu-Mizil and Caruana 2005b); and is a technique widely appreciated by researchers in the education community. Therefore in the sequel we focus our attention on approaches based on logistic regression.

To address smoothness, we propose two adaptations to logistic regression. To aid their development, we first review basic regularized logistic regression. A glossary of symbols used in this paper is given in Table 2.

Table 2: Glossary of symbols

Symbol	Description
n	The number of weeks
n_i	The number of students by week i
$n_{i,i-1}$	The number of extant students by both week i and week i-1
$\boldsymbol{x}_i$	The set of students by week i
$\boldsymbol{x}_{ij}$	The jth student by week i
d_i	The number of features for student $\boldsymbol{x}_{ij}$
$\boldsymbol{x}_i^{(i-1,i)}$	The set of students in week i also existing in week i-1
$\boldsymbol{x}_{ij}^{'}$	The jth student with extended feature space by week i
$\boldsymbol{x}_i^{'(i-1,i)}$	The set of students with extended feature space by week i also existing in week i-1
$\boldsymbol{w}_i$	The weight vector for week i
$\boldsymbol{w}$	The weight vector for all weeks
$\boldsymbol{y}_i$	The set of labels for students by week i
$\boldsymbol{y}_{ij}$	The label of jth student by week i
λ_1	Regularization parameter for overfitting
λ_2	Regularization parameter for smoothness

Logistic Regression (LR)

Let n be the number of weeks that a course lasts for. We have n_i students by the end of week i ($1 \leq i \leq n$). $\boldsymbol{x}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{in_i}\}$ is the set of students in week i. Each student $\boldsymbol{x}_{ij}$ is described by d_i features. Note that the number of students by the end of each week i can be different, since students can enter a course at any time while it is running.

Logistic regression predicts label y (fail for y=1 and pass for y=-1) for input vector $\boldsymbol{x}_{ij}$ (a student) according to,

$$p(y|\boldsymbol{x}_{ij}, \boldsymbol{w}_i) = \sigma(y\boldsymbol{w}_i^T\boldsymbol{x}_{ij}) = \frac{1}{1+\exp(-y\boldsymbol{w}_i^T\boldsymbol{x}_{ij})} \quad (1)$$

where $\boldsymbol{w}_i = [w_{i1}, w_{i2}, \cdots, w_{id_i}]^T$ is the weight vector to be learned.

From a data set by week i, given by $(\boldsymbol{x}_i, \boldsymbol{y}_i) = [(\boldsymbol{x}_{i1}, y_{i1}), (\boldsymbol{x}_{i2}, y_{i2}), \cdots, (\boldsymbol{x}_{in_i}, y_{in_i})]$, we wish to find $\boldsymbol{w}_i$ by L_2-regularized maximum likelihood estimation: minimizing with regularization parameter $\lambda_1 > 0$,

$$\mathcal{L}(\boldsymbol{w}_i) = \sum_{j=1}^{n_i} \log(1+\exp(-y_{ij}\boldsymbol{w}_i^T x_{ij})) + \frac{\lambda_1}{2}\|\boldsymbol{w}_i\|^2 \quad (2)$$

This convex problem can be solved by Newton-Raphson which produces the update equations known as iteratively-reweighted least squares. The result is n logistic regression models learned separately by the end of each week.

Sequentially Smoothed LR (LR-SEQ)

In order to smooth probabilities across weeks, we propose a transfer learning algorithm, *Sequentially Smoothed Logistic Regression (LR-SEQ)*. Transfer learning leverages the knowledge learned in related tasks to better learn a new task. In our setting, the previous week's knowledge is used to help learn smoothed probabilities for the current week.

A natural approach is to follow existing transfer learning approaches to linear classifiers (Ando and Zhang 2005): add a regularization term minimizing the difference between $\boldsymbol{w}_i$ and $\boldsymbol{w}_{i-1}$. However, the data distribution across weeks can be non-stationary as engagement varies and prescribed activities evolve. Moreover the number of features might change ($d_i \neq d_{i-1}$). Instead we seek to minimize the difference between predicted probabilities between two weeks directly. Unfortunately this leads to a non-convex objective. Therefore we minimize a surrogate: the difference between $\boldsymbol{w}_i \boldsymbol{x}_i^{(i-1,i)}$ and $\boldsymbol{w}_{i-1}\boldsymbol{x}_{i-1}^{(i-1,i)}$, where $\boldsymbol{x}_i^{(i-1,i)}$ denotes the set of students in week i that also exist in week $i-1$, and similarly $\boldsymbol{x}_{i-1}^{(i-1,i)}$ denotes the set of students in week $i-1$ that also exist in week i. The objective function[1] for week i is

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}_i) &= \sum_{j=1}^{n_i} \log(1+\exp(-y_j \boldsymbol{w}_i^T \boldsymbol{x}_{ij})) + \frac{\lambda_1}{2}\|\boldsymbol{w}_i\|^2 \\ &+ \lambda_2 \sum_{j=1}^{n_{i,i-1}} \left\| \boldsymbol{w}_i^T \boldsymbol{x}_{ij}^{(i,i-1)} - \boldsymbol{w}_{i-1}^T \boldsymbol{x}_{i-1j}^{(i,i-1)} \right\|^2 \quad (3)\end{aligned}$$

where parameter $\lambda_2 > 0$ controls smoothness and the level of transfer. This surrogate objective function is convex therefore efficiently solved by Newton-Raphson to a guaranteed global optimum. To recap: n weekly logistic regression models are learned sequentially such that week i's model cannot be built until model for week $i-1$ is obtained.

Simultaneously Smoothed LR (LR-SIM)

The drawback of LR-SEQ is that early inaccurate predictions cannot benefit from the knowledge learned in later weeks (where data is closer to the end of the course), in-turn undermining models learned later. To combat this effect, we propose *Simultaneously Smoothed Logistic Regression (LR-SIM)* that simultaneously learns models for all weeks. LR-SIM allows early and later prediction to be correlated and to influence each other, which we expect should yield improved prediction due to inter-task regularization but also good smoothness.

We first extend the feature space for each student $\boldsymbol{x}_{ij}$ to a new space with n components. The student $\boldsymbol{x}'_{ij}$ with new feature space has $\sum_{i=1}^n d_i$ dimensions, with the ith component having d_i features corresponding to the features in the original feature space by the end of week i, and others zero. For example, for a student at the end of week 2, $\boldsymbol{x}_{2j}$, we extend to a new point $\boldsymbol{x}'_{2j}$, where the 2nd component with d_2 features are actually the same as $\boldsymbol{x}_{2j}$, and others being zero. Hence we encode the same information by the end of week 2 that contributes to the outcome. We must learn a single $\boldsymbol{w}$, which also has $\sum_{i=1}^n d_i$ dimensions corresponding to $\boldsymbol{x}'_{ij}$. But only the ith component—the ith model—contributes to the prediction by the end of week i, due to the zero values of other dimensions of $\boldsymbol{x}'_{ij}$.

$$\begin{array}{c} \\ \boldsymbol{x}'_{1j} \\ \boldsymbol{x}'_{2j} \\ \vdots \\ \boldsymbol{x}'_{nj} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & n \end{array} \\ \begin{pmatrix} \boldsymbol{x}_{1j} & [0,\cdots,0] & \cdots & [0,\cdots,0] \\ [0,\cdots,0] & \boldsymbol{x}_{2j} & \cdots & [0,\cdots,0] \\ \vdots & \vdots & \ddots & \vdots \\ [0,\cdots,0] & [0,\cdots,0] & \cdots & \boldsymbol{x}_{nj} \end{pmatrix} \end{array}$$

Based on the extended $\boldsymbol{x}'_{ij}$ and $\boldsymbol{w}$, we can minimize the difference of probabilities predicted for week i and week $i-1$ for i ($i \geq 2$) together, via a simple expression, as shown in Eq. (4). Again the objective function is convex and efficiently minimized.

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) = &\sum_{i=1}^{n}\sum_{j=1}^{n_i} \log(1+\exp(-y_j \boldsymbol{w}^T \boldsymbol{x}'_{ij})) + \frac{\lambda_1}{2}\|\boldsymbol{w}\|^2 \\ &+\lambda_2 \sum_{i=2}^{n}\sum_{j=1}^{n_{i,i-1}} \left\| \boldsymbol{w}^T \boldsymbol{x}'^{(i,i-1)}_{ij} - \boldsymbol{w}^T \boldsymbol{x}'^{(i,i-1)}_{i-1j} \right\|^2 \quad (4)\end{aligned}$$

Our algorithms can operate for tasks with differing feature spaces and feature dimensions. For example, one might use individual-level features for each lecture and assignment, which might be released weekly, to help understand and interpret student performance.

Experimental Results

We conduct experiments to evaluate the effectiveness of our algorithms on real MOOCs.

Dataset Preparation

Discrete Optimization* Dataset** The first offering of *Discrete Optimization* (DisOpt1***) launched in 2013 by The University of Melbourne lasted for nine weeks, with 51,306 students enrolled, of which 795 students received a certificate of completion for the course. This course has an open course curriculum with all the videos and assignments released at the beginning of the course, enabling students to study at their own pace. There are 57 video lectures and 7 assignments in total. Students can watch/download video lectures, and attempt assignments multiple times. Their final grade is assessed by the total score on 7 assignments.

The second offering of *Discrete Optimization* (***DisOpt2***) launched in 2014 also lasted for nine weeks, attracting 33,975 students to enroll, of which 322 students completed. There are 4 fewer video lectures compared to *DisOpt1*, with 43 video lectures. The number of assignments remain but some of the assignment contents differ to those of *DisOpt1*. The total score of all assignments differs between offerings. An overview of the two offerings is shown in Table 3.

[1] For $i \geq 2$; the objective for week 1 is identical to LR in Eq. (2).

Table 3: Overview on two offerings for *DisOpt*

	DisOpt1	*DisOpt2*
Duration	9 weeks	9 weeks
Number of students enrolled	51,306	33,975
Number of students completed	795	322
Number of video lectures	57	53
Number of assignments	7	7
Total score of all assignments	396	382

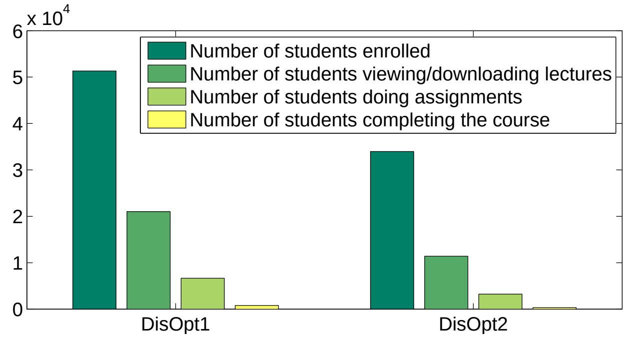

Figure 2: Student participation in the first and second offering of *Discrete Optimization*

Cohorts Among all the students enrolled, only a tiny fraction complete, which makes the data extremely imbalanced. Figure 2 shows the number of students in different course activities. In *DisOpt1*, among all the students enrolled, only around 41%, 13% and 2% of the students watch/download videos, do assignments and complete the course respectively. The same thing happens in *DisOpt2* with low completion rate, and *DisOpt2* had fewer students enrolled. Students enroll for various reasons. For example, some treat MOOCs like traditional courses by taking lectures and assignments at a fixed pace, while others treat MOOCs as online references without doing any assignments (Anderson et al. 2014). In this paper, we are interested in helping those who intend to pass. Therefore, we focus on students who are active in assignments/quizzes, which indicates an intention to pass. In particular, at the end of each week, we retain the students who did at least one assignment by that week.

Features Used We extract features from student engagement with video lectures and assignments, and performance on assignments by the end of each week to predict their performance at the end of the course. The features are shown in Table 4. In order to easily apply the model trained on previous offerings to a new offering, we extract features present across offerings.

Performance Measure

To evaluate the effectiveness of our proposed algorithms, we train prediction models on *DisOpt1*, and test on *DisOpt2*. Due to the class imbalance where high proportion of students fail, we prefer area under the ROC curve (AUC), which is invariant to imbalance. To measure the smoothness for week i, we compute the difference of probabilities between

Table 4: Features for each week i for *DisOpt*

Features
Percentage of lectures viewed/downloaded by week i
Percentage of lectures viewed/downloaded in week i
Percentage of assignments done by week i
Percentage of assignments done in week i
Average attempts on each assignment done by week i
Average attempts on each assignment done in week i
Percentage of score on assignments done by week i, to total score on all assignments

week i and week i-1 for each active student (in terms of our rule for maintaining students) in week i and i-1, and obtain the averaged difference for all students, and standard deviation (stdev).

Smoothness and AUC

To evaluate the effectiveness of our proposed algorithms LR-SEQ and LR-SIM, we compare them with two baselines, LR and a simple method using moving averages, denoted LR-MOV. LR-MOV predicts as final probability for week i an average of LR's week i and i-1 probabilities, $(i \geq 2)$. The prediction for week 1 is the same as LR. We train models using the above four algorithms on $DisOpt1$, where $\lambda_1 = 10$ and $\lambda_2 = 1$, and apply them to $DisOpt2$. Figure 3 and Table 5 show the smoothness and AUC across weeks respectively.

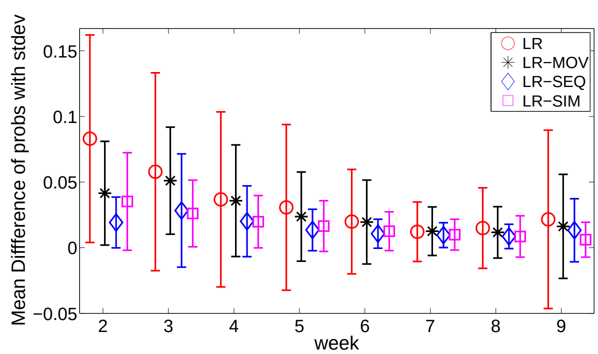

Figure 3: Comparison of LR, LR-MOV, LR-SEQ and LR-SIM on smoothness across weeks. Mean difference of probabilities across students plus/minus standard deviation. Closer to zero difference is better.

As we can see from Figure 3, LR-SEQ and LR-SIM achieve better smoothness (average difference) and low standard deviation, especially in the first five weeks where early intervention is most critical. LR attains smooth probabilities in the last few weeks, but with high standard deviation, when intervention is less impactful. LR-MOV achieves the same

Table 5: Comparison of LR, LR-MOV, LR-SEQ and LR-SIM on AUC across weeks.

Week	LR	LR-MOV	LR-SEQ	LR-SIM
1	0.788	0.788	0.788	**0.800**
2	0.867	0.856	0.849	**0.872**
3	0.901	0.890	0.867	0.892
4	0.928	0.923	0.907	0.923
5	0.947	0.944	0.934	0.944
6	0.962	0.958	0.953	0.960
7	0.970	0.968	0.963	0.969
8	0.984	0.981	0.981	0.986
9	0.996	0.997	0.997	0.995

smoothness as LR with reduced standard deviation, demonstrating the need for performing some kind of smoothing.

From Table 5, we can see that LR-SIM and LR-MOV are comparable to LR in terms of AUC, while LR-SEQ decreases slightly. (Note: LR-MOV cannot achieve better smoothness as shown in Figure 3.) LR-SIM does outperform LR in the first two weeks (in bold): one reason might be that the reduced model complexity due to transfer learning helps to mitigate overfitting; another reason might be that later, more accurate predictions improve early predictions via transfer learning in *DisOpt1* and the data distributions over *DisOpt1* and *DisOpt2* do not significantly vary. On the other hand, LR-SEQ gets continually worse in the first three weeks: LR-SEQ only uses the previous week's knowledge to constrain the present week, but early predictions might be inaccurate, which undermine models learned later (*cf.* week 3, with the worst AUC). Later, LR-SEQ catches up with LR as data closer to the end of the course becomes available.

Overall, LR-SIM and LR-SEQ outperform LR consistently in terms of smoothness. And LR-SIM maintains or even improves on LR's AUC in early weeks, while LR-SEQ suffers slightly inferior AUC in the first few weeks, and is comparable to LR in the last few weeks. Notably, using the data collected by the end of early weeks we can achieve quite good AUC: about 0.87 by week 2 and 0.9 by week 3, *establishing the efficacy of early identification of at-risk students*. Furthermore, this demonstrates that a model trained on the first offering works well on the second offering.

Parameter Analysis

We compare the performance of LR-SIM, LR-SEQ and LR in terms of smoothness and AUC varying λ_1 and λ_2. Figure 4 shows results for week 2. We choose week 2 to emphasize early intervention. The curves from right to left show varying λ_2 from 10^{-4} to 10^4. The smoothness is computed between week 2 and week 1, and AUC is for week 2. It can be seen that LR achieves good AUC but poor smoothness. LR-SIM dominates LR-SEQ. As λ_2 increases, LR-SEQ and LR-SIM get smoother. But LR-SIM can achieve better AUC while LR-SEQ gets worse. Overall, LR-SIM clearly outperforms LR-SEQ and LR.

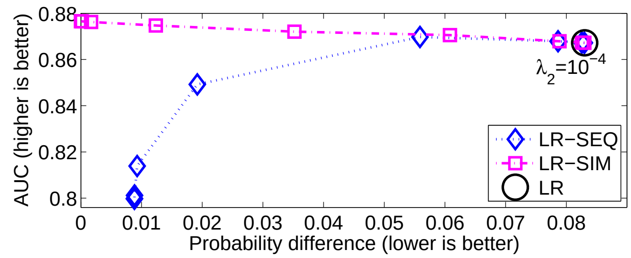

Figure 4: Smoothness versus AUC for LR, LR-SEQ and LR-SIM for week 2 when $\lambda_1 = 10$, and $\lambda_2 = 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}$.

Calibration

Given an instance, it is not possible to know what the true underlying probability is, therefore some approximations are often used. A common way is to group instances based on the ranked predicted probability into deciles of risk with approximately equal number of instances in each group, and compare the predicted probability with observed probability within each group. A reliability diagram plotting the predicted probability with observed probability, is commonly used for calibration (Niculescu-Mizil and Caruana 2005a; Zhong and Kwok 2013).

Figure 5 shows the reliability diagram using LR-SIM for week 2. Our predicted probabilities agree closely with the observed probability in the gray region of marginal at-risk students for whom we wish to intervene.

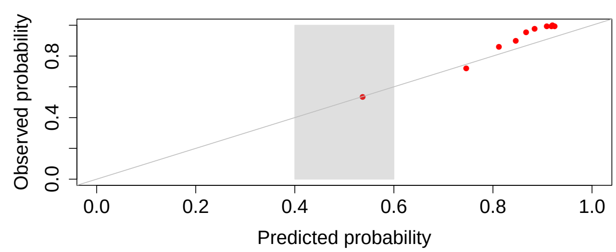

Figure 5: Reliability diagram for class *fail* using LR-SIM week 2. Grey area shows an intervention interval [0.4,0.6], which could be varied according to educational advice.

Conclusion

We have taken an initial step towards early and accurately identifying at-risk students, which can help instructors design interventions. We have compared different prediction models, with regularized logistic regression preferred due to its good performance, calibration and interpretability. Based on the predicted probabilities, we envision an intervention that presents students meaningful probabilities to help them realize their progress. We developed two novel transfer learning algorithms LR-SEQ and LR-SIM based on regularized logistic regression. Our experiments on Coursera MOOC data indicate that LR-SEQ and LR-SIM can produce smoothed probabilities while maintaining AUC, with LR-SIM outperforming LR-SEQ. LR-SIM has exceptional AUC in the first few weeks, which is promising for early prediction. Our experiments leveraging the two offerings of a Coursera MOOC demonstrate that the prediction models trained on a first offering work well on a second offering.

Model interpretability is important in learning analytics, where detailed feedback may be favored over generic feedback like 'how's it going?'. Such specifics can shed light on why a student is failing, and also what strategies other students follow to succeed. In particular, within logistic regression, the learned weight vectors can be used for explaining the contribution of each feature—albeit under certain assumptions on feature correlation. In these cases, features are not only important for prediction, but also for interpretability.

In the future, we will collaborate with course instructors to deploy our identification models and subsequent interventions in a MOOC for A/B testing to determine efficacy.

References

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2014. Engaging with massive online courses. In *Proceedings of the 23rd International Conference on World Wide Web*, 687–698. International World Wide Web Conferences Steering Committee.

Ando, R. K., and Zhang, T. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research* 6:1817–1853.

Balakrishnan, G. 2013. Predicting student retention in massive open online courses using hidden Markov models. Master's thesis, EECS Department, University of California, Berkeley.

DeBoer, J.; Ho, A.; Stump, G. S.; Pritchard, D. E.; Seaton, D.; and Breslow, L. 2013a. Bringing student backgrounds online: MOOC user demographics, site usage, and online learning. *Engineer* 2:0–81.

DeBoer, J.; Stump, G.; Seaton, D.; and Breslow, L. 2013b. Diversity in MOOC students' backgrounds and behaviors in relationship to performance in 6.002 x. In *Proceedings of the Sixth Learning International Networks Consortium Conference*.

Halawa, S.; Greene, D.; and Mitchell, J. 2014. Dropout prediction in MOOCs using learner activity features. In *Proceedings of the European MOOC Summit*.

Jiang, S.; Warschauer, M.; Williams, A. E.; ODowd, D.; and Schenke, K. 2014. Predicting MOOC performance with week 1 behavior. In *Proceedings of the 7th International Conference on Educational Data Mining*.

Kaufman, C.; Perlman, R.; and Speciner, M. 2002. *Network Security: Private Communication in a Public World*. Prentice Hall, 2nd edition.

Kloft, M.; Stiehler, F.; Zheng, Z.; and Pinkwart, N. 2014. Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses*.

Lodge, J. M. 2011. What if student attrition was treated like an illness? an epidemiological model for learning analytics. In Williams, G.; Statham, P.; Brown, N.; and Cleland, B., eds., *Changing Demands, Changing Directions. Proceedings ascilite Hobart 2011*, 822–825.

Niculescu-Mizil, A., and Caruana, R. 2005a. Obtaining calibrated probabilities from boosting. In *UAI*, 413.

Niculescu-Mizil, A., and Caruana, R. 2005b. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine learning*, 625–632.

Ramesh, A.; Goldwasser, D.; Huang, B.; Daumé III, H.; and Getoor, L. 2013. Modeling learner engagement in MOOCs using probabilistic soft logic. In *NIPS Workshop on Data Driven Education*.

Ramesh, A.; Goldwasser, D.; Huang, B.; Daume III, H.; and Getoor, L. 2014. Learning latent engagement patterns of students in online courses. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press.

Taylor, C.; Veeramachaneni, K.; and O'Reilly, U.-M. 2014. Likely to stop? predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382* 273–275.

Veeramachaneni, K.; O'Reilly, U.; and Taylor, C. 2014. Towards feature engineering at scale for data from massive open online courses. *CoRR* abs/1407.5238.

Wen, M.; Yang, D.; and Rosé, C. P. 2014. Sentiment analysis in MOOC discussion forums: What does it tell us? *Proceedings of Educational Data Mining*.

Yang, D.; Sinha, T.; Adamson, D.; and Rosé, C. P. 2013. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*.

Yang, D.; Wen, M.; and Rose, C. 2014. Peer influence on attrition in massive open online courses. *Proceedings of Educational Data Mining*.

Zhong, L. W., and Kwok, J. T. 2013. Accurate probability calibration for multiple classifiers. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 1939–1945. AAAI Press.