# A Learning Analytics Approach to Assessing Student Risk in Active Learning

MOHSEN DORODCHI, MOHAMMAD J. MAHZOON, MARY LOU MAHER, AND AILEEN BENEDICT

## Introduction

Learning analytics is an emerging discipline within data science. It is analytics that is concerned with developing methods for exploring the unique and increasingly large-scale datasets collected from educational settings, including the collection, analysis, and visualization of such educational data. The goal of the analyses and visualizations is to understand and improve students' learning and their learning environments. These methods are developed and applied in the same way as general data analytics, including exploiting statistical and machine learning for prediction, clustering, outlier detection, knowledge discovery with models, text mining, knowledge tracing, relationship mining, etc. to search for unobserved patterns and underlying information in learning processes (Agudo-Peregrina, Iglesias-Pradas, Conde-Gonzáles, & Hernandez-García, 2014).

Learning analytics of a course includes the gathering and analysis of data about a course and its students with the goal of improving its learning environment. The tracking of a student's progress and potential factors for success and failure can be valuable for the evaluation of the course. Coursework could then be redesigned by exploring these factors and learning more about student patterns, such as examining how student attitude and motivation can affect their success. These insights could then help create a better learning environment for the students over time. By using computer science techniques and creating visualizations for these factors, we are able to discover and understand patterns more easily, something that could be much more difficult to accomplish by simply looking at the raw data itself.

Learning analytics has been used in many situations. For example, Cherenkova, Zingaro, and Petersen (2014) explored which student difficulties arise within beginning computer science courses by mining data from CodeLab, a "web-based interactive programming problem system," finding that conditions and loops are the main challenges for students. They also encourage the use of large data from many institutions to lead to greater insight. Agudo-Peregrina et al. (2014) have applied learning analytics, specifically bivariate correlation anal-

ysis, to find the correlation between interactions (i.e., student-to-student interactions within the learning management system [LMS], student interactions with LMS content, and student interactions with the professor) and student performance. In another study, learning analytics was used to identify significant behavioral indicators of learning. Results showed that students' regular study, times of assignment submission, number of login sessions, and proof of reading course information were all significant factors in predicting course achievement (You, 2016).

Agudo-Peregrina et al. (2014) extracted different data from the LMS, such as student-to-student interaction inside the LMS, student interactions with LMS and content of the LMS, and student interactions with the professor within the LMS. Furthermore; they applied statistical methods, such as bivariate correlation analysis, to find the correlation between such interactions and student performance for an online as well as a face-to-face class. The results showed some correlation from mid-to-strong for the online class; however, for the face-to-face class, they found a no-to-weak correlation. The differences between the course structures and LMS structures are not very clear.

## Learning Analytics for Student Risk Analysis

Research is drawn from various areas that view analytics in different perspectives or dimensions. Gašević, Dawson, and Siemens (2015) identified three common dimensions in learning analytics research: design, theory, and data science. For example, action research (McNiff & Whitehead, 2011) and personalized adaptive learning best fit into design or theory categories because their research focuses on improving teaching practice. However, our research contributes to learning analytics from a data science perspective and uses research in theory and design dimensions to make sense of data.

From the data science perspective, we focus on the issue of predicting student success/risk using analytical methods. Research in student risk analytics helps instructors keep track of student performance, and given the prediction results, policymakers can plan for improving retention by helping at-risk students.

Depending on the goal of the research, student risk analytics employs one of two general ways to define success and risk. The first method is to consider a student's final grade. For example, a student with an acceptable final grade for a course (usually C and above) is deemed successful in the course while others are considered at-risk. Other studies use course completion rather than the final grade to determine success. This definition of success is often best used when analytics is done at a micro level by looking at individual key courses and focusing on student success in those courses. Analytics in this area contributes to better student performances by identifying issues that students may have while taking courses and by providing insights to create interventions to help fix those issues. The second method is to look at student graduation. For example, a student graduating "on-time" is successful, while one who does not is at-risk (of dropping out). This definition of success and risk can be useful for academic leaders and executives who need to check the health of the education system from the macro

level. Analytics in this domain analyzes student behaviors to identify issues, such as flaws in curriculum design.

Regardless of how success and risk are defined, research in student risk analytics needs to be confirmed with some data source, such as the LMS, to obtain granular and meaningful data from students. For example, Macfadyen and Dawson (2010) used Blackboard Vista LMS to extract 15 features correlating with students' final grades. These features include the total number of discussion messages posted, mail messages sent, and assessments completed. Macfadyen and Dawson (2010) used logistic regression to classify students as successful or at-risk with 81% accuracy. As another example, Wolff, Zdrahal, Nikolov, and Pantucek (2013) used click behaviors in the virtual learning environments as the data source to identify students at-risk using a decision tree model. Moreover, Jayaprakash, Moody, Lauría, Regan, and Baron (2014) combined the log data from student interactions within Sakai Learning Management System with student demographics, aptitude data, course grades, course-related data, and partial contributions to students' final grades such as individual assignment grades.

In terms of the models used in student risk analytics, we refer to surveys done by Romero and Ventura (2007) and Romero, Ventura, and García (2008) showing different approaches taken in the learning community to discern student behavior using machine learning or statistical methods. Generally, the data mining approaches discussed in their surveys used statistics or machine learning techniques operating on a feature vector representation of each student having data such as demographic information, course grades, and LMS logs. Several others, such as Mohamad and Tasir (2013) and Peña-Ayala (2014), review approaches that used different analytics with similar feature sets for their vector representations.

When it comes to the analyses of student learning, the major question is what student-related features can be used to accurately analyze performance, such as study patterns, exhibited emotions, and temporal features. By analyzing these features, it is possible to extract crucial information, such as identifying at-risk students to improve the course or to intervene on their behalf. In a study of 350 college students, a learning analytics model was used to predict course achievement as measured by their activities inside a LMS (You, 2016). The study demonstrated that their pattern of study, late submissions, and whether they reviewed the materials was predictive of performance. In another study, students' emotional reactions were correlated with student performance on programming assignments (Lishinski, Yadav, & Enbody, 2017). This work influences our use of sentiment to identify risk.

In this chapter, we look at learning analytics methods, particularly the sequence analytics method, a temporal approach to analyzing data. In particular, we look into course-related data that can be extracted from the LMS and/or student reflections pointing directly or indirectly toward their learning in the classroom.

## Learning Analytics Using Time

During the last decade, increasing research in the data mining and machine learning communities have produced many approaches to analyze time-related raw data to identify trends and

unexpected behaviors over time. However, these approaches still have not been widely adapted for learning analytics, and state-of-the-art approaches in student success and risk analysis do not consider temporal aspects of data.

Molenaar (2014) argues that temporal aspects of student data deserve more attention, and temporal analysis yields a paradigm shift addressing new research questions in learning analytics. Similarly, previous work in computer-supported collaborative learning (CSCL) and self-regulated learning (SRL) emphasizes the importance of temporal features in student data (Kapur, 2011; Bannert, Reimann, & Sonnenberg, 2014).

There are potentials in time series analysis, data stream mining, and sequence pattern mining that can contribute to analyzing student data while preserving the temporal dependencies. However, for each of these approaches to be used with student data, there are potential obstacles as described below.

### *Time Series*

Time series analysis aims to arrive at a mathematical or statistical model to describe a series of observations over time, and it has applications from the stock market to weather forecasting. Various methods have been proposed in time series analysis literature to solve prediction, classification, and regression problems. All of these models were built on the same assumptions that (a) the data are in a numerical format, and (b) a significant number of data samples are available. Neither of these assumptions is necessarily true for student data because it is highly heterogeneous, containing ordinal and categorical features in addition to numerical features. Even though some data items such as grades and other performance features can be converted to numerical data, many features such as reflection data and quiz answers cannot be represented in numbers while preserving meaning.

The data we have for each student are limited and uniquely different from that of other students. The data about a single student cannot be generalized to a format that reconciles it with the data on all students without significant information being lost. The amount of data available for each student is also unique and can vary widely. Additionally, time series analysis usually looks for recurring patterns or regularities within a time period. In contrast, student data are temporal but not periodic. Students' progress each week as they acquire knowledge and prepare for the next activity. While time series can still be applied to student data to identify periodic patterns for numerical features, our sequence data facilitate detecting trends and irregularities in sequences having heterogeneous and variable length data items.

### *Data Stream Mining*

Data stream mining is a subdomain of data mining that presents methods to efficiently process continuous massive sequences of data items called streams. These methods can watch for "concept drift" (Widmer & Kubat, 1996) when the general statistical properties of the target prediction change. Methods in data stream mining adapt to the changes in the stream to produce better prediction for new instances of data. For example, Hulten, Spencer, and Domingos (2001) present a model to maintain and update a decision tree for concept-drifting data streams. The

model is always up-to-date with the latest instances of the stream while discarding old concepts that were changed over time. Adapting data stream mining ideas to the student data analytics faces several challenges. In student data analytics, we are not dealing with massive continuous data streams. Student sequences have a clear starting point and a duration of several weeks, and therefore the streams are not massive and content is sparse. Furthermore, data stream algorithms do not keep track of changes in data since they discard the changed concepts to account for the newest ones. To interpret students' behavior and investigate what it means to be at risk, we need to capture changes in trends and identify unexpected patterns.

*Sequence Pattern Mining*

Another subdomain of data mining that works with sequences is sequence pattern mining used to identify frequent sets of items or patterns in data or strings (Agrawal, Imielinski, & Swami, 1993). This domain is generally used for identifying behavior patterns of consumers in the business domain. One such approach detects frequent items bought together from a dataset of all transactions. For example, Padmanabhan and Tuzhilin (1999) propose an interestingness measure to filter all frequent items to obtain interesting items that happen to be unexpected transactions, contradicting beliefs.

We can make an analogy to transfer ideas from sequence pattern mining to student sequence data mining. If we treat each student sequence as a transaction, then the task becomes frequent events happening together in student sequences. However, there are certain assumptions in sequence pattern mining that make it hard to continue the analogy further. For instance, in sequence pattern mining, it is assumed that we know beforehand about all potential items in transactions (i.e., all items being sold in a store). This assumption holds in business and marketing because the number of items is finite and known. However, student data sequences have a wide range of possibilities such as quizzes taken, assignment grades, forum participation, and other academic and nonacademic activities. It is a daunting task to generate all potential events for a student sequence.

## Sequence Data Model

We define a student sequence as data items that are grouped into temporally ordered structures called "nodes." For example, a node may represent a semester and may contain a student's data items related to that semester: courses taken, grades received, extracurricular activities, and so on. This grouping gives context to the data items and allows for analysis at the level of both data items and nodes.

Figure 7.1 illustrates the structure of the sequence data model in which information about a student is grouped by semester. The sequence starts with an initial node that captures attributes outside of the node-based temporal sequence, such as demographics and prior academic achievement. A node is then included for each semester the student is enrolled and finishes with an outcome node. The properties that characterize a sequence data model include time dependency, contextualization, segmentation, and storytelling.

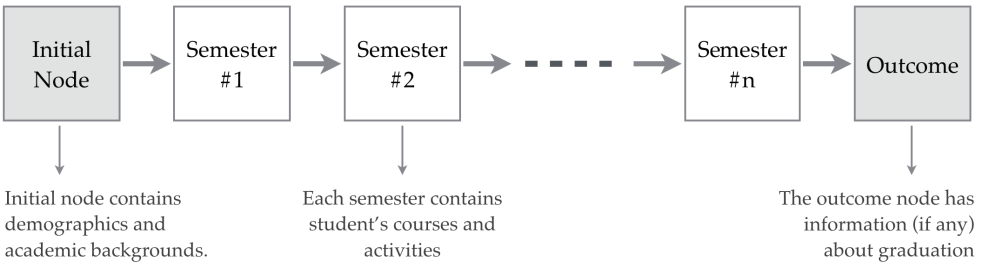


Figure 7.1. The structure of the sequence data model in which information about a student is grouped by semester.

**Time Dependency.** The sequence data model explicitly represents that the later data items can depend on former data items. This allows the explicit representation of temporal dependencies, such as the correlation between final grade and student assignment grades. In comparison, a vector representation assumes that data points are independent of each other, and features (independent variables) do not have a correlation with each other.

**Contextualization.** The grouping of data items into nodes gives context to salient features that are selected for analysis. For example, if each node groups different information throughout the course mostly coming from the LMS for a given week within the semester, then data can be identified as salient features within each node, such as grades of the activities, while other features, such as student activities, are the context of the salient feature.

**Segmentation.** The nodes in a sequence allow us to represent the data in segments. Different choices for the beginning and end of each node define a principle for a window of time and allow the data model to capture a different granularity for the segments, for example, semesters versus weeks. Access to LMS data makes finer-grained node segmentation possible, which may lead to more timely assessments of academic risk.

**Storytelling.** A sequence of information expresses a student's learning events throughout a particular course. This property enables us to view each node as a collection of student data from course events happening during a particular week in a semester. Moreover, there is an opportunity to infer a narrative from the nodes to tell a story about a specific or typical student in order to hypothesize about success or risk.

## Applying Learning Analytics to a Course

To verify the effectiveness of our active learning course, learning analytics methods are applied to a course as explained in the following sections (Dorodchi et al., 2018).

### *Data Collection*

The data were collected from the LMS from 91 university students enrolled in the Introduction to Computer Science (CS1) course in the spring semester of 2017. The course demographics

consisted of 21.9% female students and 72.5% computer science (CS) majors. The data have three main categories per student:

- Student background information
- Student performance scores
- Student reflections and self-assessments

Each of the above three categories includes specific attributes used in our algorithm as features. More specifically, student background includes attributes such as age group, gender, and major. Student performance scores are numerous as is typical in an active learning classroom in which students are submitting items for preparation, in-class and in-lab activities, and assignments outside the classroom. In the example in this chapter, we include grades for all quizzes (18 total), pre- and postlabs (16), long assignments (four), lecture tests (four total, including three during the semester and one final), lab tests (four), lab/lecture activities (37), and extra credit activities (four) for a total of 87 different columns per student. Reflections are informal surveys students take regularly after class (both labs and lectures) activities, assignments, quizzes, and tests for a total of 23 reflections per student. Students reflected on their learning of different course topics, as well as on the learning processes, group activities, or the tests and assignments. Some of the reflections, therefore, were mandatory as a part of the activity while others were optional extra credit activities. We ended up with a heterogeneous dataset for different reasons: (a) we have both numerical and textual data; (b) data items' frequency of occurrences are different such as weekly, biweekly, or monthly; and (c) the data included objective and subjective measures, as well as self-assessment or group assessment by students.

All the 110 different grades and quantified reflections are spread over our dataset, based on the date of the activity, which highlights their strong temporal dependencies with each other. Therefore, these data are a good candidate for using temporal data analysis models. It should be emphasized that the temporal dependency of the data items comes from the fact that students must do different types of activities in the lab and lecture as explained while providing reflections over time. The activities are all dependent on and build on each other. In addition, students were reflecting on their learning and outcomes of activities that suggest the strong dependency as shown in Figure 7.2.

In other words, it is possible that a student who received low grades for the first few weeks of the course might change their study pattern to make up for the low performance. Consequently, we have dependencies in activities themselves as well as dependencies in the time between reflections and activities.

### *Student Data Model*

Our goal was to discover the trends of students' activities throughout the semester, predict the student outcome (success or fail), and discover the impact of reflections on the prediction. To do so, we built a temporal data model, called the "student sequence model" (Mahzoon, Maher, Eltayeby, Dou, & Grace, 2018). In this model, we put all the data for one week into one node as shown in Figure 7.2. Next, we connected the nodes to form a sequence. The sequence was then

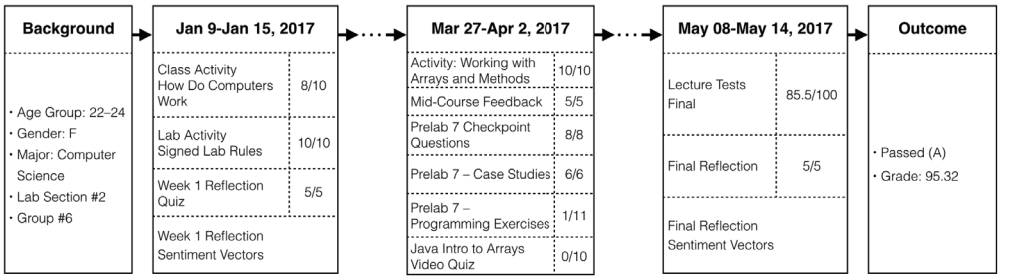


Figure 7.2.

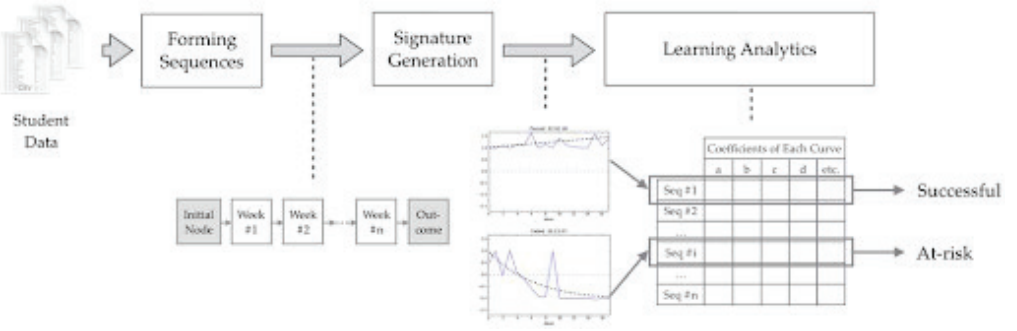


Figure 7.3.

passed to a signature generation submodule followed by the learning analytics submodule for final determination, as shown in Figure 7.3.

### *Analysis: Sequence Model versus Feature Vector Model*

While the student sequence model uses nodes to sort and group data items temporally, a more common method uses feature vectors to represent data items. Feature vector representation in knowledge discovery and data mining constructs features vectors for data items in which each data item is represented by one vector with a fixed set of features or dimensions. For example, in student data, each data item could contain a vector of one student's performance in a certain course or program. The features of the vector could then include the student's background information (e.g., demographics and course information) and the student's performance (e.g., grades, assignments, and activities).

Feature vector representation makes strict assumptions about data dependencies that enable the use of conventional machine learning tools. This representation assumes that vectors are independent of each other and features are independent of other features. These independency assumptions, as well as the fixed length of the vectors (i.e., the number of features), make the application of machine learning tools widely available.

However, having that strict assumption for data dependencies in vector representations ignores the temporal correlations in student data—something we wanted to emphasize. A typical example of such temporal correlation is the correlation between the final grade and different types of grades (e.g., class activities, lecture tests, assignments) over time for the same course. The order in which these grades occur provides important information for predicting success or risk. However, that order is discarded in feature vector representation due to its inability to represent temporal correlations.

### *Structure of the Sequence Model*

Our sequence model consists of 19 nodes: one node at the beginning of the semester for student background data, 17 intermediary weekly nodes that include grades and reflection responses, and one outcome node containing the overall course grade. There are four background features in the first node. The 110 grade scores and maximum of 23 reflection responses (depending on the individual student) are then spread out over the intermediary weekly nodes. We converted reflective surveys from text to numbers using the commercial linguistic sentiment analytics tool called Linguistic Inquiry and Word Count, or LIWC (Tausczik & Pennebaker, 2010). LIWC generates 93 sentiment features as numbers for every input text. Many of these features were highly correlated with each other. For this reason, we only chose 18 sentiment features with the least correlation to each other. This also improves computational efficiency. Therefore, each reflective survey's text was converted to a vector of 18 sentiment features.

### *Analysis*

One of the benefits of our sequence data model and analytics is its capability to repeat the analysis with different salient features in the data model to identify the predictive impact of each data category. Based on the model by Mahzoon et al. (2018), salient features are involved in the analytics while contextual features are used for interpretability after the analysis. Our three main salient features were tests, activities, and reflections. We then experimented with these features both individually and together to evaluate their relative predictive impact to help us understand the effectiveness and importance of each feature as predictors of success. For each salient feature, we ran sequence analytics to classify students at risk of obtaining at-risk grades of D, F, or W in the course. Figure 7.4 shows two examples of individual student signatures that were generated for successful grades of A, B, or C and at-risk (DFW) students.

Figure 7.5 shows the averages of all the student signatures in the class grouped by final grade category.

Figure 7.6 shows the averages grouped by successful (ABC) or at-risk (DFW). In all three figures, the data include tests, activities, and reflections used to generate the signatures.

The classification is performed in two phases: training and validation. In the training phase of classification, we use the 10-fold cross-validation method (Kohavi, 1995) to split our data into training and validation sets. After training, the system output will be validated by repeating the validation set 10 different times. The performance measures of the analytics were then averaged over the 10-fold cross-validation.

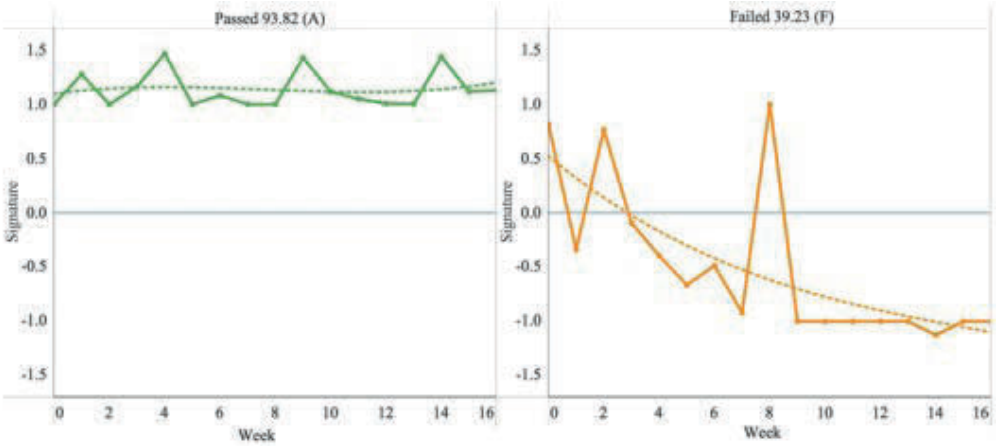


Figure 7.4. Two examples of individual student signatures that were generated for successful grades of A, B, or C and at-risk (DFW) students.

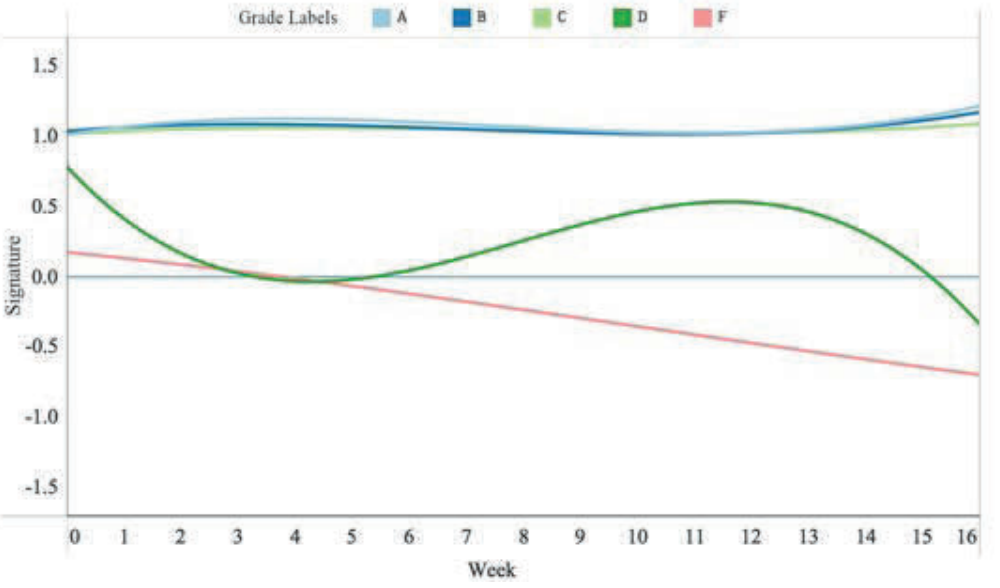


Figure 7.5. The averages of all the student signatures in the class grouped by final grade category.

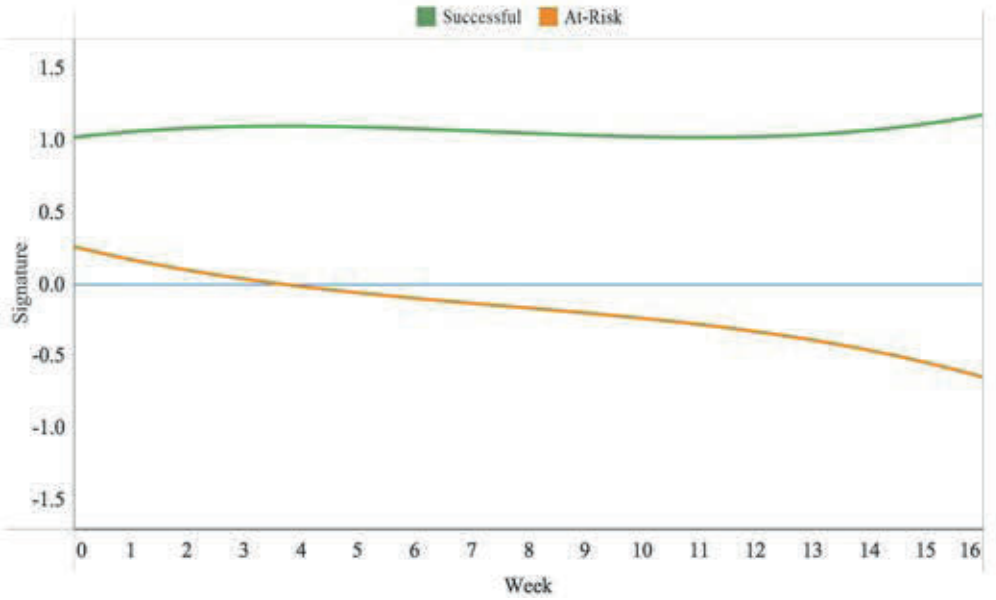


Figure 7.6. The averages grouped by successful (ABC) or at-risk (DFW).

We evaluated the sequence model incrementally at multiple points in time to assess how the temporal model's accuracy changes over time. Figure 7.7 reports the model's accuracy for the following three salient features: tests, activities, and reflections.

These features were plotted over time to show how the accuracy improved as more data were included. In this figure, the horizontal axis shows the number of weekly nodes included in the model, and the vertical axis shows the accuracy of the model as a percentage. For example, from Figure 7.7 we can conclude that if we only use one week of the data (e.g., tests, activities, or reflections), we are able to accurately classify the risk status of 70% of students. This accuracy increases as we include more nodes (i.e., more weeks into the semester) in the sequence model. However, the trend of increasing accuracy is not the same for different salient features. For instance, having tests as the only salient feature will improve accuracy but only up to the four-week point in the semester; on the other hand, having activities as the only salient feature produces models with higher accuracies in comparison with tests after five weeks of activities. Reflections as the salient feature perform even better than activities or tests and can predict at-risk students with 90% accuracy even after having only five weeks of reflections. In all cases, the additional benefit for including more information diminishes at the halfway point of the course. It is important to try to maximize earlier prediction rather than overall accuracy after the middle of the semester. At this early point, there are still opportunities to intervene on behalf of the student. Some examples of interventions include: helping the student understand patterns of the active learning class and how to prepare, working with other students, and learning materials by practicing. It is worth noting that the

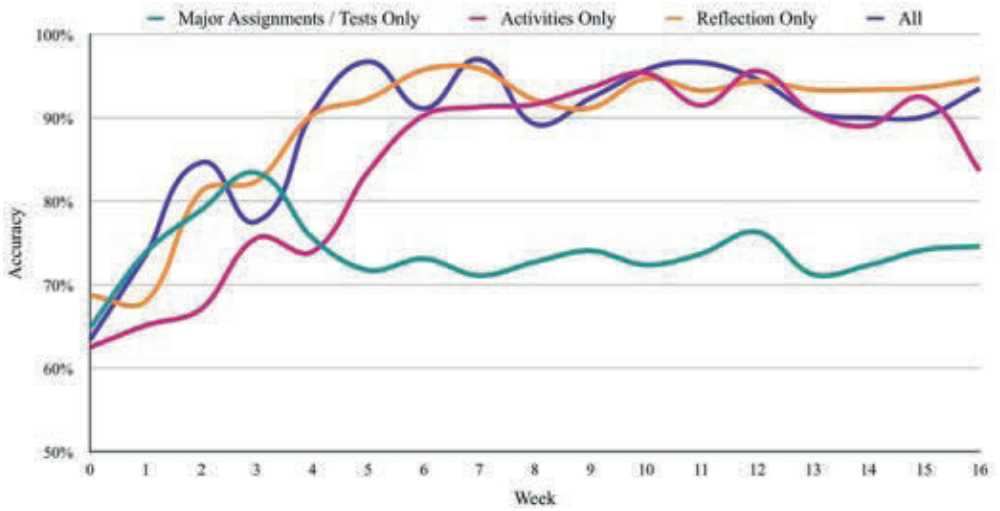


Figure 7.7. The model's accuracy for the following three salient features: tests, activities, and reflections.

closest individual salient feature compared to including all features together in Figure 7.7 is a reflection-only plot.

Based on our results, we observed that including reflections as a feature improves the accuracy of our risk classification model. This shows that including student reflection in a course proves to be useful, as we can use them in a predictive model to improve both accuracy and time-to-classify student success and risk. Having an improved time-to-classify is important, as interventions need to be made early enough to help the at-risk students adjust and make improvements toward success. Thus, using student reflections provides additional motivation for instructors as they not only improve our risk classification model but can serve as an effective learning tool for students.

Our findings are encouraging for integrating reflections into the curriculum. Previous research has investigated reflections as a tool for learning and has cited many different potential benefits, such as the development of metacognitive skills (Turns, Sattler, Yasuhara, Borgford-Parnell, & Atman, 2014). What we have shown in this work is that, in addition to previously explored benefits affecting students, there are also benefits for instructors and administrators. For example, they will have the ability to predict the students who may be at-risk early on. With that knowledge, instructors can intervene to aid the at-risk students. Furthermore, it is crucial that these predictions be accurately made early on so that it is not too late for the student to make improvements when those interventions are implemented.

Although our results suggest that reflections were predictive of student success on their own, they were most effective when used with traditional features such as tests, activities, and assignments. While most features performed well after the first six weeks of the course,

reflections served as the earliest predictors of success for students. Hence, it also suggests that infusing student reflective practices between activities and throughout the course is effective as an early predictor of success. Reflection in CS has the ability to help students think more deeply about the course material and make broader connections to other courses and aspects of computing. Our work has shown that in addition to these benefits, there are also administrative benefits that help instructors and teaching staff identify at-risk students sooner and more accurately. Our results provide another compelling reason to integrate reflections into engineering and CS courses.

## Conclusions

Learning analytics provides a broad range of tools to analyze and predict student progress as a whole and individually. This provides an opportunity for the course instructors to detect the at-risk students in the early weeks of the semester and to intervene in different forms before it is too late. Accuracy of learning analytics algorithms significantly increases by infusing more feedback points from students. This is in line with the notion of activity-based active learning that provides students with many different forms of activities throughout a week. By analyzing such data, we are able to predict with some level of accuracy the students who are at-risk of failing the course.

### Acknowledgments

This work is supported by NSF-1820862: EAGER: An Interactive Learning Analytics Framework Based on a Student Sequence Model for Understanding Students, Retention, and Time to Graduation.

### References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Database mining: A performance perspective. IEEE transactions on knowledge and data engineering, 5(6), 914-925. http://dx.doi.org/ 10.1109 /69.250074

Agudo-Peregrina, A., Iglesias-Pradas, S., Conde-Gonzáles, M. A., & Hernandez-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior, 31*, 542–550. https://doi.org/10.1016/j.chb.2013.05.031

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, *9*(2), 161–185. http://dx.doi.org/10.1007/s11409-013-9107-6

Cherenkova, Y., Zingaro, D., & Petersen, A. (2014). Identifying challenging CS1 concepts in a large problem dataset. *Proceedings of the 45th ACM Technical Symposium on Computer Science Education, 14*, 695–700. https://doi.org/10.1145/2538862.2538966

Dorodchi, M., Benedict, A., Desai, D., Mahzoon, M. J., MacNeil, S., & Dehbozorgi, N. (2018). Design and implementation of an activity-based introductory computer science course (CS1) with peri-

odic reflections validated by learning analytics. *Proceedings of Frontiers in Education, 1*, 1–8. https://dx.doi.org/10.1109/FIE.2018.8659196

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71. http://dx.doi.org/10.1007/s11528-014-0822-x

Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 97–106. https://doi.org/10.1145/502512.502529

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics, 1*(1), 6–47.

Kapur, M. (2011). Temporality matters: Advancing a method for analyzing problem-solving processes in a computer-supported collaborative environment. *International Journal of Computer-Supported Collaborative Learning, 6*, 39–56. https://dx.doi.org/10.1007/s11412-011-9109-9

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings* of *International Joint Conference on Artificial Intelligence, 14*, 1137–1145.

Lishinski, A., Yadav, A., & Enbody, R. (2017). Students' emotional reactions to programming projects in introduction to programming: Measurement approach and influence on learning outcomes. *Proceedings of the 2017 ACM Conference on International Computing Education Research, 17*, 30–38. https://dx.doi.org/10.1145/3105726.3106187

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*, 588–599.

Mahzoon, M. J., Maher, M. L., Eltayeby, O., Dou, W., & Grace, K. (2018). A sequence data model for analyzing temporal patterns of student data. *Journal of Learning Analytics, 5*(1), 55–74. https://doi.org/10.18608/jla.2018.51.5

McNiff, J., & Whitehead, J. (2011). *All you need to know about action research* (2nd ed.). London, England: Sage Publications.

Mohamad, S. K., & Tasir, Z. (2013). Educational data mining: A review. *Proceedings of Procedia—Social and Behavioral Sciences, the 9th International Conference on Cognitive Science*, *97*, 320–324. http://dx.doi.org/10.1016/j.sbspro.2013.10.240

Molenaar, I. (2014). Advances in temporal analysis in learning and instruction. *Frontline Learning Research*, *2*(4), 15–24. https://doi.org/10.14786/flr.v2i4.118

Padmanabhan, B., & Tuzhilin, A. (1999). Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support System, 27*(3), 303–318. http://doi.org/10.1016/S0167-9236(99)00053-6

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*, 1432–1462. http://dx.doi.org/10.1016/j.eswa.2013.08.042

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*, 135–146. http://dx.doi.org/10.1016/j.eswa.2006.04.005

Romero, C., Ventura, S., & Garcia, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*, 368–384. https://doi.org/10.1016/j.compedu.2007.05.016

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24–54. https://doi.org/10.1177/0261927X09351676

Turns, J. A., Sattler, B., Yasuhara, K., Borgford-Parnell, J. L., & Atman, C. J. (2014, June). *Integrating*

*reflection into engineering education. Proceedings Presented at ASEE Annual Conference & Exposition, Indianapolis, Indiana*. Retrieved from https://peer.asee.org/20668

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning, 23*, 69–101. https://dx.doi.org/10.1007/BF00116900

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. *Proceedings of the Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium*, 145–149. http://dx.doi.org/10.1145/2460296.2460324

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education, 29*, 23–30. https://doi.org/10.1016/j.iheduc.2015.11.003