Research in Learning Technology
Supplement: ALT-C 2012 Conference Proceedings

Improving student success using predictive models and data visualisations

Alfred Essa* and Hanan Ayad

Desire2Learn Inc, Kitchener, Canada

(Received 12 March 2012; final version revised 13 June 2012)

The need to educate a competitive workforce is a global problem. In the US, for example, despite billions of dollars spent to improve the educational system, approximately 35% of students never finish high school. The drop rate among some demographic groups is as high as 50–60%. At the college level in the US only 30% of students graduate from 2-year colleges in 3 years or less and approximately 50% graduate from 4-year colleges in 5 years or less. A basic challenge in delivering global education, therefore, is improving student success. By student success we mean improving retention, completion and graduation rates. In this paper we describe a Student Success System (S3) that provides a holistic, analytical view of student academic progress.[1] The core of S3 is a flexible predictive modelling engine that uses machine intelligence and statistical techniques to identify at-risk students pre-emptively. S3 also provides a set of advanced data visualisations for reaching diagnostic insights and a case management tool for managing interventions. S3's open modular architecture will also allow integration and plug-ins with both open and proprietary software. Powered by learning analytics, S3 is intended as an *end-to-end solution* for identifying at-risk students, understanding why they are at risk, designing interventions to mitigate that risk and finally closing the feedback look by tracking the efficacy of the applied intervention.

Keywords: predictive models, data visualisation, student performance, risk analytics

1. Introduction

The need to educate a competitive workforce is a global problem. In the US, for example, despite billions of dollars spent to improve the educational system, approximately 35% of students never finish high school. The drop rate among some demographic groups is as high as 50–60%. At the college level in the US only 30% of students graduate from 2-year colleges in 3 years or less and approximately 50% graduate from 4-year colleges in 5 years or less (Bill and M. G. Foundation 2010). A basic challenge in delivering global education, therefore, is improving student success. By student success we mean improving retention, completion and graduation rates. In this paper we describe a Student Success System (S3) that provides a holistic, analytical view of student academic progress.[1] The core of S3 is a flexible predictive modelling engine that uses machine intelligence and statistical techniques to identify at-risk students pre-emptively. S3 also provides a set of advanced data visualisations

*Corresponding author. Email: Alfred.Essa@Desire2Learn.com

ISBN 978-91-977071-4-5 (print), 978-91-977071-5-2 (online)

http://dx.doi.org/10.3402/rlt.v20i0.19191

for reaching diagnostic insights and a case management tool for managing interventions. S3's open modular architecture will also allow integration and plug-ins with both open and proprietary software. Powered by learning analytics, S3 is intended as an *end-to-end solution* for identifying at-risk students, understanding why they are at risk, designing interventions to mitigate that risk and finally closing the feedback look by tracking the efficacy of the applied intervention.

2. Related work

Student Success System (S3) draws and builds upon work in risk analytics in education and health care. In this section we begin by describing how predictive modeling has been applied in health care and education. We also describe methodological limitations to current risk modelling approaches (e.g. Signals Project at Purdue University) in learning analytics. Current approaches to building predictive models for identifying at-risk students are stymied by two serious limitations. First, the predictive models are one-off and, therefore, cannot be extended easily from one context to another. We cannot simply assume that a predictive model developed for a particular course at a particular institution is valid for other courses. Can we devise a flexible and scalable methodology for generating predictive models that can accommodate the considerable variability in learning contexts across different courses and different institutions? Secondly, current modelling approaches, even if they generate valid predictions, tend to be black boxes from the standpoint of practitioners. The mere generation of a risk signal (e.g. green, yellow, red) does not convey enough information for designing meaningful personalised interventions. The design of S3, both from an application and research perspective, is intended to overcome these limitations.

2.1. Risk analytics in health care

The use of risk analytics in health care provides an instructive example of how segmentation strategies and statistical models can lead to substantial cost savings and Return on Investment (ROI). Risk analytics is also the first step in designing personalised interventions and therefore optimizing the quality of health care delivery.

The starting point of predictive models in health care delivery is a well-known phenomenon: a small percentage of subscribers in health care plans account for a disproportionately large percentage of health care costs. Typically, 20% of the population accounts for 80% of the costs. The middle 60% of the population accounts for another 15% of the costs. Finally, the remaining 20% of the population accounts for only 5% of the costs.

Predictive models gained popularity in health care delivery initially as a strategy for managing costs. Since then analytics is being used to deliver personalised care, thereby improving the quality of health outcomes. Subscribers are first segmented into risk pools using predictive analytics. Then each pool is managed using a different intervention strategy.

- **Risk Pool 1:** the chronically ill, who need personalised and well-integrated care services
- **Risk Pool 2:** the newly diagnosed, who have an immediate need for disease specific information and timely and cost-effective options

- **Risk Pool 3:** the well, who need information on well-being, staying well and avoiding disease

Because of the cost dynamics, an accurate predictive model alongside a well-designed intervention strategy can translate into substantial ROI. As members of Risk Pool 3 in effect pay for the claims of other members and are the major source of profitability for health care plans, retaining them is critical for financial success and viability. Similarly, as members of Risk Pool 1 require constant access to services and treatment across multiple modalities, assigning a case-worker, with the ability and knowledge to negotiate and navigate on the patients behalf, can turn out to be an effective financial strategy despite the initial cost overhead.

Predictive modelling of health populations has traditionally focused on identifying the determinants of disease within anonymous populations using large-scale models. The models are then applied to segment subscribers into risk pools. Members in each risk pool receive their appropriate set of health care services or interventions. Health care is evolving now so that analytics is beginning to be used to deliver a "just-in-time" personalised approach, where predictive models and interventions are customised not just to the group but also individually for each patient. With S3 we offer a similar model of risk stratification and personalised intervention.

2.2. Predictive models in education

In education predictive models for identifying at-risk students were pioneered by John Campbell and the Signals Project at Purdue University (Campbell, DeBlois, and Oblinger 2007). Similar work has been underway at a variety of institutions, including Capella University, University of Phoenix and Rio Salado College (Gilfus Education Group).

The Course Signals system and recent research studies provide strong evidence that student e-learning activities (i.e. behaviours in online environments) are predictive of course success. Regression modelling such as logistic regression has been applied to build course-based predictive models. Such models incorporate the most significant Learning Management Systems (LMS) variables such as total number of discussion messages posted, total number of mail messages sent and total number of assessments completed. The models are also supplemented with Student Information System (SIS) data, such as whether a student is taking other courses at the same time, their grades in previous courses and their current Grade Point Average (GPA).

Macfadyen and Shane have discussed the *limitations* of this approach in terms of its overall generalisability and interpretation. In particular, the generalisability of such models can be limited by the sample courses used for model fitting, or by focusing on fully online courses within one institution (Macfadyen and Dawson 2010). In addition, these approaches generate a risk indicator (high-risk, medium-risk and low-risk) or prediction but fail to provide any additional data that would allow the practitioner to devise a meaningful intervention. Imagine a physician having access to a health prediction system that indicates that a particular patient is at high-risk, medium-risk or low-risk of illness and that the prediction is 70% accurate. Such a system would be useful but highly limited because the "black box" predictive model does not provide practitioners the ability to take action in terms of designing personalised interventions.

The limitations of this modelling strategy, in terms of generalisability and interpretability, critically hinder the wide-ranging deployment of discovered models to educational institutions in a meaningful way. Hence, it limits the potential benefits that institutions can draw from their data through the development of predictive analytics capabilities for modeling learner success. S3 relies on a predictive modelling strategy that aims at closing this gap. We focus on providing a generalisable modelling strategy that is well suited for supporting the wide-ranging needs of educational institutions and for taking full advantage of predictive analytics. S3 provides an adaptive framework and a stacked-generalisation modelling strategy, whereby intelligent data analysis can be applied at all levels and graciously combined to express higher-level generalisations.

2.3. Challenges and opportunities

A core problem in current approaches, as applied in Course Signals-type systems, is that a single hypothesis/model that best fits a collection of course data is chosen from the space of all possible hypotheses, and then applied to make predictions across different courses in different programmes and institutions. There are potential sources of bias in this solution. This methodology is expected to work well when courses on which the model is applied have a relatively consistent instructional model with the courses used to discover the best-fit model, but otherwise lead to a risk of systematic errors in predictions, i.e. relatively high bias.

A second key problem is that current predictive modelling systems do not provide diagnostic information. For example, Course Signals generates a prediction that indicates the identified level of risk; however, there is no direct insight into the specific causes, thus making a recommended remediation difficult to specify. Furthermore, the system does not incorporate human insight that can be leveraged via model tuning, if needed.

To enable an effective synthesis of machine intelligence and human insight, S3 provides an interpretable model and data visualisations. Another issue with a Signals-type model is that ignores or takes a narrow treatment of a key aspect of learning, namely social learning or learning network analysis. Research is beginning to show that a student's peer interactions, in the right context, can stimulate and accelerate learning. Social network analysis can be utilised to provide insights into the student learning community and the patterns of peer interactions. In S3, a social network analysis and visualisation is incorporated to capture and explain the social learning aspect.

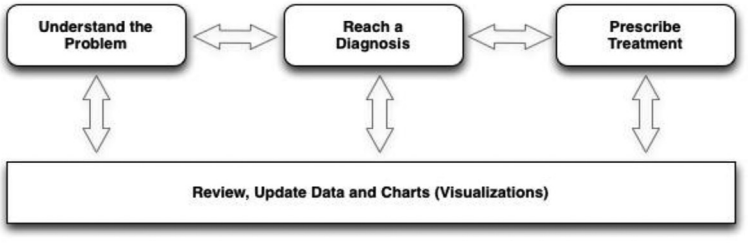
In S3, to overcome these limitations, we have implemented an ensemble strategy whereby a domain-specific decomposition allows for the development and integration of specialised models and algorithms that are best suited for different aspects of learning. In particular, in S3, the proposed decomposition provides an abstraction of learning behaviour into semantically meaningful units.

The ensemble idea provides an adaptive framework that integrates multiple models. Prediction ensembles provide a powerful and flexible paradigm for enhancing the relevance and generalisability of predictive analytics. It can also be viewed as enabling a collaborative platform, whereby institutions can plug their own proprietary model as part of the ensemble. Thus, it enables an open, community-driven R&D platform for the application of predictive models to advance learning analytics as well as institutional analytics capabilities.

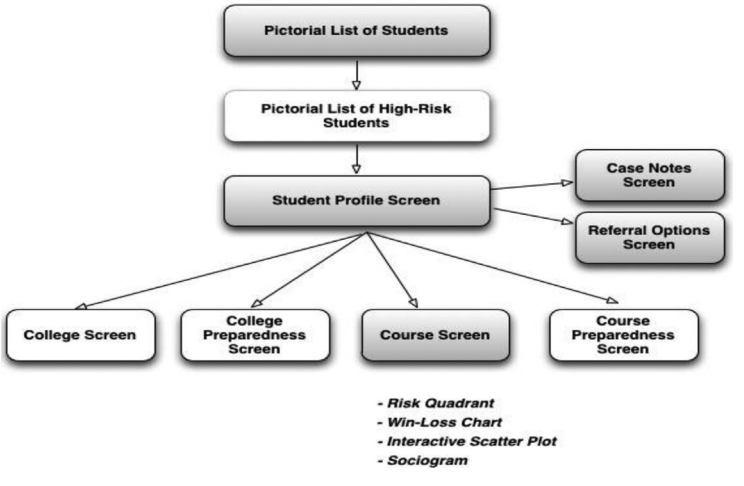
3. Student Success System

3.1. S3 functional overview

The workflow for S3 is analogous to the workflow in a typical patient–physician relationship. As a first step, the physician tries to understand the problem. Next, the physician tries to reach a diagnosis. Next, the physician prescribes a course of treatment or makes a referral. At any point, the physician might review additional data or "charts" to round out the picture of the patient. Finally, the system behind the scenes records the intervention with the purpose of establishing a feedback loop.

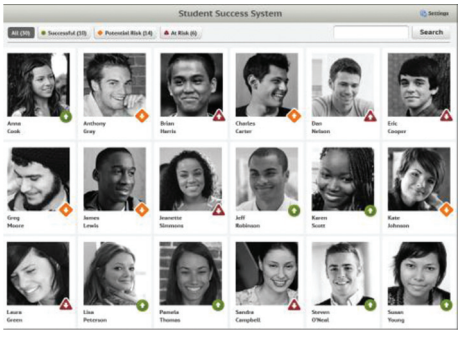


The basic screens in S3 provide a similar workflow. First, upon login to S3 an advisor or instructor (roles in S3) is presented with a pictorial list of his/her students. Associated with each student is a risk indicator: green indicates not at-risk, yellow indicates possibly at-risk and red means at-risk. The advisor or instructor can immediately click on a particular student or view the screen showing the list of students in a particular category (e.g. high risk).



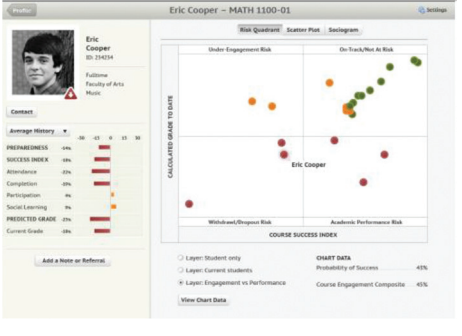
Associated with each student is his/her Student Profile Screen. The Student Profile Screen serves as a gateway to other screens, which collectively provide a comprehensive view of student academic progress and risk factors. The Notes Screen provides case notes associated with the student while Referral Screen provides all the relevant referral options available at the institution.

Pictorial List of Students: The primary screen of S3 displays a pictorial list of students along with a risk indicator: green, yellow and red. An up or down arrow indicates the projected trajectory of either improvement or decline.

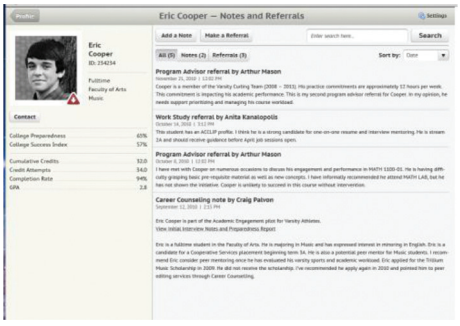


Student Profile Screen: The Student Profile Screen is the primary screen associated with each student in S3. It is intended to provide an at-a-glance view of a student's profile and risk factors.

Course Screen: By clicking on Math in the Student Profile Screen we pull up Eric Cooper's performance charts and predictions for Mathematics. An explanation of the visualisations is provided in the Section 3.2.



Notes Screen: The Notes Screen provides a running case history of the student's interactions with various advisors, faculty and counselors. It can be regarded as a case management tool. We can imagine scalable versions of S3 would integrate with an enterprise CRM tool to provide deeper case management functionality.



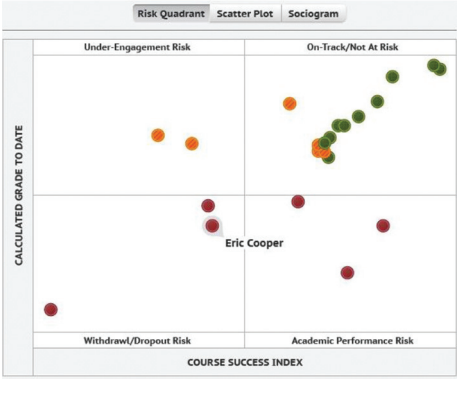
Referral Screen: The Referral Screen lists all relevant referral options at the institution. In addition, a communication pathway (e.g. email) is provided from within the screen rather than having to step outside the context.

In summary, the basic screens of S3 provide a synoptic view of a student's academic progress and the ability through single click interactions to isolate areas of risk or potential risk. Once we have seen that a student is projected to be at-risk, what kind of insights can we derive from the data and patterns as a basis for designing an intervention strategy? A key feature of S3 is the set of data visualisations.

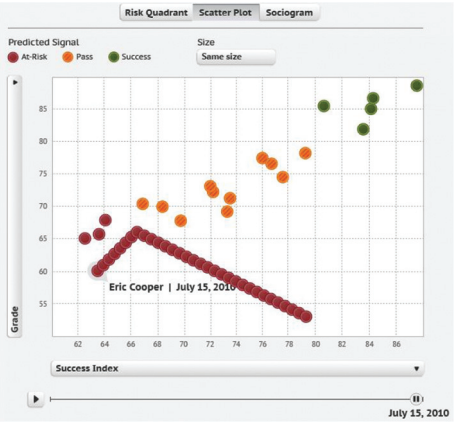
3.2. S3 Visualisations

As the user of the S3 navigates through the various success indicators, the underlying models and data are presented in an intuitive and interpretable manner, going from one level of aggregation to another. Furthermore, at the course level we present dynamic and interactive chart that allow the user of the S3 to interact with the data and to explore and understand its patterns and characteristics. Some sample visualisations in S3 are displayed below:

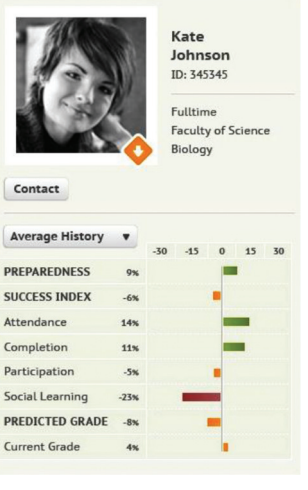
Risk Quadrant. At a course level each point represents a student in the class. The top right quadrant contains all students who are on-track and not at-risk. The bottom right quadrant contains students who are academically at– risk, meaning that they are projected to receive a D or F in the course. The bottom left quadrant contains students who are likely to Withdraw or Dropout. Finally, the top left quadrant contains students who are under-engaged, meaning that the students are projected to succeed in the course but their pattern of under-engagement might be a cause of concern for other reasons.



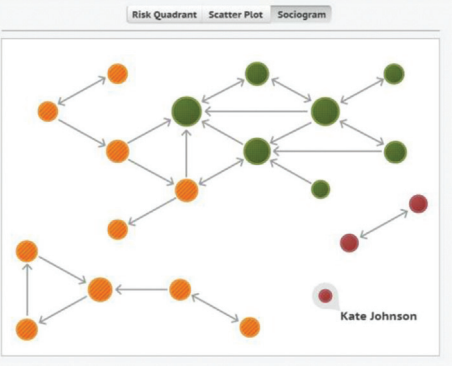
Interactive Scatter Plot. A user of S3 is able to explore the data that make up the predictive model by selecting the success indicators associated with each domain and visualise patterns such as cluster structures and relations between different indicators and measures of performance. The chart is also dynamic in the sense that data can be animated to visualise paths/trails depicting changes in learner behaviours and performance over time.



Win-Loss Chart. This chart shows at a glance how the learner compares to peers along the overall indicator and each of the sub-indicators. Options to compare to learner own history along these indicators are also presented.

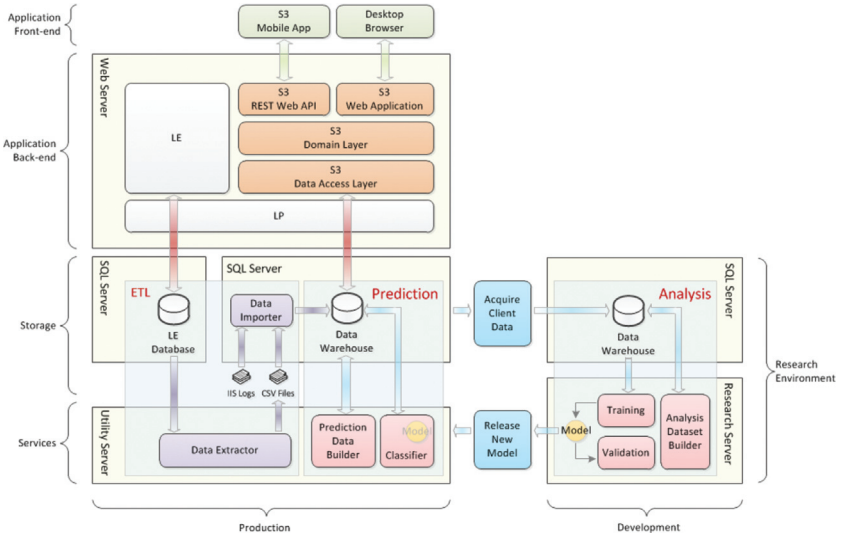


Sociogram. The chart shows patterns of communication or collaborations among learners. It is depicted as a network with nodes representing learners and link representing interactions. Size, colours and link width are used to indicate relevant variables. Furthermore, statistical and topological measures are used to describe patterns, cluster structures and other characteristics, and to evaluate the health of individual social learning and of the overall learning community. As part of the analysis of this domain, text mining, cognitive and learning theory are applied to extract relevant factors of learning success and to identify at-risk learners.



4. S3 architecture

The practical implementation of a system such as S3 in an enterprise requires multiple layers of integration and infrastructure. The overall design is consistent with a standard Business Intelligence (BI) or Analytics environment where data from operational source systems are aggregated and stored using Extraction, Transformation, Lead (ETL) processes into a data warehouse or series of Data Marts. The primary sources of data for S3 are LMS, Web Logs and SIS. But the architecture of S3 allows for ingestion of data from other data sources. Although it is beyond the scope of this paper to describe, S3 architecture follows an open, modular approach to allow maximum flexibility and integration of components.



The S3 architecture involves multiple components that serve different purposes. The above diagram illustrates the various components, their dependencies and interactions, and the data flow between them. The application itself is a web application that has a typical layered architecture. In addition to providing access to desktop browsers, it also provides access to mobile devices through a REST API.

The application uses the Analytics data warehouse for storage of its data. It integrates with the rest of the Analytics architecture, which involves synchronising data from the Learning Environment through an ETL process. In production, a classifier service will be used to make predictions of student success based on live student data. The classifier service relies on a predictive model that has been produced in development based on historical data. In order to produce this predictive model, a process by which historical data are acquired from clients is employed. An analysis process is performed on the historical data, in which a training algorithm produces a predictive model capable of predicting student success.

The application front-end is offered in two versions: a web browser and a native mobile version. The web browser version utilises an Model-View-Controller (MVC) web framework, including standard MVC controls. The native mobile version will be offered initially for the iPad. The mobile app will communicate with the S3 back-end through a REST API.

S3 visualisations will utilise the same mechanism for rendering charts on the client in both the web version as well as the mobile version. The client in both cases will host the chart on a web page (web view in case of mobile). Client-side JavaScript representation of the chart will be sent down from the server to the client, where the client will invoke a function to render the chart inside the web page/view.

The application back-end has a typical layered architecture. The front-end facing layer consists of two components: an MVC web application layer for serving the desktop web version of the application, and a REST API layer for serving the mobile app. Both the MVC web application and the REST API depend directly on the S3 domain layer.

The domain layer is where the domain entities and business logic lies. The domain layer is also responsible for enforcing security through authorisation rules. The domain layer depends directly on the S3 data access layer for storage and retrieval of data. The domain layer manages translation between data layer Data Transfer Objects (DTOs) and domain entities. The data access layer is responsible for Create, Read, Update and Delete (CRUD) operations against the database. This layer depends directly on LP data access framework, as well as stored procedures defined in the database.

The predictions made by the S3 classifier rely on student data collected from LE. LE data are synchronised to the Analytics data warehouse on a nightly basis through an ETL process. This part of the system is already established and is being used for reporting purposes. A data extraction service extracts relevant data from the LE database and stores them into Comma-Separated Values (CSV) files in a predefined location on the filesystem. A data importer component then imports the extracted data, along with IIS web logs, into the data warehouse.

Predictions are made based on a classification model that has been generated in development. A Prediction Data Builder service builds the input data used for prediction by transforming existing data in the data warehouse into a format suitable for classification. The prediction data are stored back in the data warehouse. A classifier service then goes through the prediction input data and produces the predictions. The classifier service uses a model that has been generated during development.

Data analysis is performed at development time, not in production. Analysis is done on historical data acquired from certain clients. An Analysis Dataset Builder component builds the input data used for training and validation by transforming historical data in the data warehouse into a format suitable for analysis. The analysis dataset is stored back in the data warehouse. A training component then performs

predictive modelling by learning the association of the input data to the actual output data. The output of the training component is a predictive model. A validation component then validates the model by evaluating the accuracy of predictions made on test data. The purpose of the validation component is to make sure that prediction accuracy is suitable for use in production. Once the predictive model is produced and validated, it is incorporated into the classifier component to be released in the next version of S3.

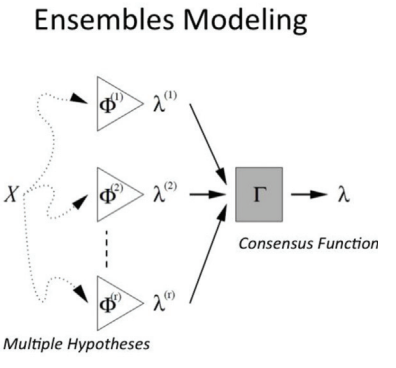
5. S3 modelling strategy

It is beyond the scope of the paper to describe our modelling techniques in detail. Here we provide an outline. Current approaches to building predictive models in learning analytics are faced with two serious limitations. The first limitation is the ability to generalise across different learning contexts. It is erroneous to assume that a predictive model developed for a particular course at a particular institution will be valid for a different course at the same institution, let alone at another institution. Current predictive models in learning analytics are mostly one-off and cannot be extended easily from one context to the next. The second limitation in current approaches is the ability of practitioners to interpret the results of a prediction with the aim of deriving insights or designing interventions. Current models are either black boxes or obscured in technical jargon.

5.1. Ensemble modeling

Our proposed solution is to apply an ensemble method for predictive modelling using a strategy of decomposition into semantic units, where each unit has meaning in a learning context. Decomposition provides a flexible technique for generalising predictive models across different contexts. Decomposition into interpretable semantic units, when coupled with data visualisations and case management tools, allows practitioners to extend predictions towards designing personalised interventions, thereby building the missing bridge between prediction and action.

Ensemble methods are designed to boost the predictive generalisability by blending the predictions of multiple models. For example, stacking, also referred to as blending, is a technique in which the predictions of a collection of base models are given to a second-level predictive modelling algorithm, also referred to as a meta-model. The second level algorithm is trained to combine the input predictions optimally into a final set of predictions.

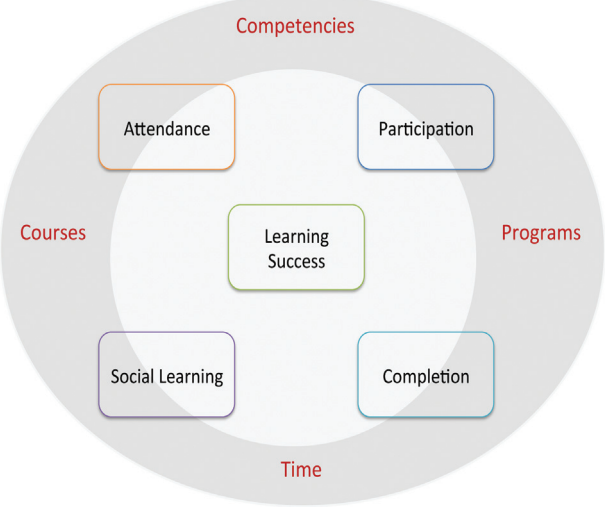


In general, ensuring that a predictive modelling algorithm matches the properties of the data is crucial in providing meaningful results that meet the needs of the particular application scenario. One way in which the impact of this algorithm-to-application match can be alleviated is by using ensembles of predictive models, where a variety of models (either different types of models or different instantiations of the same model) are pooled before a final prediction is made. Intuitively, ensembles allow the different needs of a difficult problem to be handled by models suited to those particular needs. Mathematically, classifier ensembles provide an extra degree of freedom in the classical bias/variance trade-off, allowing solutions that would be difficult (if not impossible) to reach with only a single model (Oza and Tumer 2008).

Stacking, data fusion, adaptive boosting and related ensemble techniques have successfully been applied in many fields to boost prediction accuracy beyond the level obtained by any single model (Polikar 2006). S3 represents a particular instance of the ensemble paradigm. It employs aspects of data fusion to build base models for different learning domains. Furthermore, the system utilises a stacked generalisation strategy. A best fit meta-model takes as input predictors the output of the base models and optimally combine them into an aggregated predictor, referred to as a success indicator/index. In this type of stacked generalisation, optimisation is typically achieved by applying Expectation Maximization (EM) algorithm.

A large data arising from learner-produced data trails, ubiquitous learning and networks of social interactions are giving rise to the new research area of learning analytics. These diverse and abundant sources of learner data are not sufficiently analysed via a single best-fit predictive model, as in the Course Signals system. Instead, the discovery and blending of multiple models to effectively express and manage complex and diverse patterns of the e-learning process is required.

The idea is that data from each learning modality, context or level of aggregation across the institution can be used to train base predictive models, whose output can then be combined to form an overall success or risk-level prediction. Applications in which data from different sources with different input variables are combined to make a more informed decision are generally referred to as data fusion applications.



Hence, the data fusion model is useful for building individual predictive models that are well suited for sub-domains of an application. In the context of S3 these models correspond to each data tracking domain and represent different aspects of the learning process. That is, each model is designed for a particular domain of learning behaviour. An initial set of domains are defined as: Attendance, Completion, Participation and Social Learning.

6. Conclusion

In this paper we have outlined a holistic ensemble-based analytical system for tracking student academic success. The core idea of the S3 synthesises several strands of risk analytics: the use of predictive models and segmentation to identify academically at-risk students, the creation of data visualisations for reaching diagnostic insights and the application of a case-based approach for managing interventions.

There are several fundamental limitations in current approaches to building predictive models in learning analytics. The first limitation is the ability to generalise across different learning contexts: how can we build predictive models that generalise across different courses, different institutions, different pedagogical models, different teaching styles and different learning designs? A second limitation is the ability to interpret the results of a prediction for the purpose of decision and action: how can a non-technical practitioner (e.g. an advisor) design meaningful interventions on behalf of an individual learner when the underlying mechanism of prediction is either a black box or obscure?

S3 applies an ensemble method for predictive modelling using a strategy of decomposition. The units of decomposition have the added property that they are semantically significant in a learning context. Decomposition provides a flexible mechanism for building predictive models for application in multiple contexts. Decomposition into semantic units provides an added bonus, namely the ability to extend our predictions towards reaching diagnostic insights and designing personalised interventions.

Note

1. S3 is in development by Desire2Learn Inc. A beta version of the software will be demonstrated at the Alt-C conference. The production version will be available in January 2013.

References

Bill & Melinda Gates Foundation. (2010) Next generation learning (pdf, 8 pages). *Technical report, Bill & Melinda Gates Foundation*, Seattle, USA.

Campbell, J. P., DeBlois, P. B. & Oblinger, D. G. (2010) 'Academic analytics: a new tool for a new era', *Educause Review*, vol. 42, no. 4, pp. 40–57.

Gilfus Education Group. (2012) Academic analytics – New eLearning diagnostics, [online] Available at: http://www.gilfuseducationgroup.com/academic-analytics-new-elearning-diagnostics

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

Oza, N. C. & Tumer, K. (2008) 'Classifier ensembles: select real-world applications', *Information Fusion*, vol. 9, no. 1, pp. 4–20.

Polikar, R. (2006) 'Ensemble based systems in decision making', *IEEE Circuits and Systems Magazine*, Third Quarter, pp. 21–45.