The Internet and Higher Education 37 (2018) 66–75

Contents lists available at ScienceDirect

The Internet and Higher Education

journal homepage: www.elsevier.com/locate/iheduc

Contrasting prediction methods for early warning systems at undergraduate level

Emma Howard[a,*], Maria Meehan[a], Andrew Parnell[a,b]

[a] School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland
[b] Insight Centre for Data Analytics, University College Dublin, Belfield, Dublin 4, Ireland

ARTICLE INFO

Keywords:
Learning analytics
Early warning systems
Undergraduate education
Prediction modelling

ABSTRACT

Recent studies have provided evidence in favour of adopting early warning systems as a means of identifying at-risk students. Our study examines eight prediction methods, and investigates the optimal time in a course to apply such a system. We present findings from a statistics university course which has weekly continuous assessment and a large proportion of resources on the Learning Management System Blackboard. We identify weeks 5–6 (half way through the semester) as an optimal time to implement an early warning system, as it allows time for the students to make changes to their study patterns while retaining reasonable prediction accuracy. Using detailed variables, clustering and our final prediction method of BART (Bayesian Additive Regressive Trees) we can predict students' final mark by week 6 based on mean absolute error to 6.5 percentage points. We provide our R code for implementation of the prediction methods used in a GitHub repository[1].

Abbreviations: Bayesian Additive Regressive Trees (BART); Random Forests (RF); Principal Components Regression (PCR); Multivariate Adaptive Regression Splines (Splines); K-Nearest Neighbours (KNN); Neural Networks (NN) and; Support Vector Machine (SVM)

1. Introduction

Early warning systems to identify at-risk students (of dropping out or failing) are in practical use in large classes and online courses (Corrigan, Smeaton, Glynn, & Smyth, 2015; Pistilli & Arnold, 2012; Wolff, Zdrahal, Herrmannova, Kuzilek, & Hlosta, 2014). We provide findings from a large first year statistics course in which most of the learning materials are available online and therefore student engagement with them can be measured via the Learning Management System (LMS) Blackboard. We acknowledge the impact course design, in particular weekly continuous assessment, has on developing early warning systems. We contrast results from eight prediction methods (Random Forest; BART; XGBoost; Principal Components Regression; Support Vector Machine; Neural Network; Multivariate Adaptive Regression Splines; and K-Nearest Neighbours) and the impact of cluster membership (based on student engagement) on reducing prediction error. We reasonably predict a student's final grade as early as week 5 of a 12-week teaching semester. This study was completed using R software and we have provided our R code on GitHub at https://github.com/ehoward1/Early-Warning-System, and in Appendix A.

This study forms part of a larger goal to use the predictions we create to allow for more precisely targeted interventions for poorly performing students. Determining the timing at which these interventions should occur is one of the key goals of this study. We would like to intervene as early as possible, but with little information from the LMS and necessarily limited continuous assessment at the start of the semester, the predictions are inaccurate. This accuracy increases as we move through the semester but at the price of intervening later and so lessening the impact of any interventions. We monitor the performance of the predictive models on a week by week cumulative basis. For each week, we aim to predict the final percentage mark of the student based on all current information. We do not dichotomise students' performance to pass/fail unlike many other studies (Azcona & Casey, 2015; Marbouti, Heidi, & Madhavan, 2016) which would lessen the accuracy. At week 6 (of a 12-week semester) we obtain a mean absolute error (MAE) of approximately 6.5 percentage points.

The structure of our paper is as follows: in Section 2 we discuss the rationale and prediction methods behind current early warning

Abbreviations: BART, Bayesian Additive Regressive Trees; RF, Random Forests; PCR, Principal Components Regression; Splines, Multivariate Adaptive Regression Splines; KNN, K-Nearest Neighbours; NN, Neural Networks; SVM, Support Vector Machine

* Corresponding author.
E-mail address: emma.howard@ucdconnect.ie (E. Howard).

[1] Abbreviations: Bayesian Additive Regressive Trees (BART); Random Forests (RF); Principal Components Regression (PCR); Multivariate Adaptive Regression Splines (Splines); K-Nearest Neighbours (KNN); Neural Networks (NN) and; Support Vector Machine (SVM)

https://doi.org/10.1016/j.iheduc.2018.02.001
Received 21 August 2017; Received in revised form 25 January 2018; Accepted 6 February 2018
Available online 09 February 2018
1096-7516/

systems. In Section 3 we outline our approach to developing an accurate prediction method for an early warning system. We extend current research on the development of early warning systems through: using 'new' prediction methods including BART; identifying an 'optimal time'; and including cluster membership. In Section 4, we discuss the data analytics decisions made and present the results for our course *Practical Statistics*. Finally we progress to the discussion and conclusion of these results in Section 5.

2. Previous work on early warning systems

2.1. Prediction modelling for early warning systems

In this section, we examine the stages in creating a prediction model for an early warning system (detailed data collection; variable selection; prediction modelling; and clustering). Advancements in learning interfaces allow for fine-grained collection of data. Azcona and Casey (2015) highlight that "fine-grained (microscopic) analytics data should yield better results than coarse-grained (macroscopic)" (p. 223). An example of a coarse-grained variable is total count of resources accessed online. In comparison fine-grained data analytics refers to extracting each log entry of a student, and all the information it contains for example: the number of slides visited; number of successful compilations; and time spent on platform (Azcona & Casey, 2015). Their argument is that through using more detailed variables, more powerful prediction models can be created. However this trades off against simplicity; simple models with a small number of variables are easier to interpret and understand.

Variables based on students' demographic/historic data, continuous assessment results and LMS usage have been collected for early warning systems (Pistilli & Arnold, 2012; You, 2016). LMS data can include length of time on a LMS system, number of visits to a module page, contributions to a module discussion thread et cetera. Depending on the prediction models selected, the dataset is reduced to a small number of 'important' variables.

There are numerous types of prediction models used for learning analytics. Gašević, Dawson, Rogers, and Gašević (2016) note that researchers have produced prediction models by using classification algorithms such as EM, C4.5, Naive Bayes Classifier, and Support Vector Machines. Logistic regression and multiple regression modelling are often used as prediction models (Macfadyen & Dawson, 2010; Waddington, Nam, Lonn, & Teasley, 2016), with logistic regression being considered the most popular prediction method for educational settings (Marbouti et al., 2016). Hierarchical mixed models (Joksimović, Gašević, Loughin, Kovanović, & Hatala, 2015; You, 2016), K-nearest neighbour (Marbouti et al., 2016), neural network models (Calvo-Flores, Galindo, Jiménez, & Pérez, 2006), and decision tree methods (Azcona & Casey, 2015) are also methods employed. A common use of prediction models in learning analytics is to identify whether a student will pass or fail the course based on the binary response variable 'pass/fail'. The use of a binary response variable dichotomises students' performance percentage marks. Studies using binary response variable include Azcona and Casey (2015); Marbouti et al. (2016) and Calvo-Flores et al. (2006). However, there are studies that use a continuous response variable (Huang & Fang, 2013; You, 2016) for example students' final module grades.

A key point to note is that predictive models are usually applied to a single course rather than used for several courses. Wolff, Zdrahal, Nikolov, and Pantucek (2013) propose that this may be because each course is structured differently, and therefore dictates what learners are doing. Gašević et al. (2016) investigate generalised predictive models that can be applied to multiple courses, however they note that the inherent differences in disciplines cause specific variables to be strong for some courses, and weak for other courses. Hence, the nature of the course should be considered before selecting variables for an early warning system. Gašević et al. (2016) believe "the understanding of practical needs in specific instructional and learning contexts is the primary driver for the development and deployment of learning analytics methods" (p. 83).

Clustering also plays a significant role in learning analytics through its ability to identify students' engagement levels or learning strategies statistically. When investigating a blended course Lust, Vandewaetere, Ceulemans, Elen, and Clarebout (2011) identify three patterns of tool-use using k-means clustering: the no-users; the intensive users; and the incoherent users. White and Carroll (2017) use Latent Class Analysis to identify four clusters of engagement in a large blended business course. In their discussion they identify what resources each cluster engaged with, and when during the semester these resources were engaged with.

2.2. Early warning systems in practice

One of the best known examples of an early warning system is in Purdue University (Ferguson, 2012; Pistilli & Arnold, 2012; Sclater, Peasgood, & Mullan, 2016) who introduced 'Course Signals' (CS) or a 'traffic light system' whereby students can see whether they are likely to succeed in their course based on a traffic light colour on their learner interface. For example a green colour indicates a high likelihood of succeeding. This prediction of success is based on prediction models using all available student background information and LMS interactions. If a student is identified as at-risk, the lecturer has the option of providing corrective measures including: posting of a traffic signal indicator on the student's CMS home page; sending e-mail messages or reminders; sending text messages; referring the student to an academic advisor or academic resource centre; or organising a face-to-face meeting. Pistilli and Arnold (2010) found that the results of their interventions (based on a control group versus an experimental group) were: students seeking help earlier; lower D's and F's recorded; more B's and C's; and students felt more than a 'number', that is less isolated. Other benefits of Course Signals discussed by Sclater et al. (2016) are students using the subject help desks more, and greater attendance at additional tutorials.

One prime reason for the implementation of an early warning system is to detect students at-risk of dropping out of courses. Pistilli and Arnold (2010) state that most early warning systems rely on midterm grades reported by lecturers. By the time midterms have been corrected it is often far into the semester, and students may have already dropped out. It is crucial that early warning systems operate in the early stages of the semester. However, a balance has to be achieved with the accuracy of the model. As the methods, models, variables and response variable used in identifying at-risk students vary from study to study, it remains difficult to contrast the studies and identify which study has obtained the most accurate results. Results are impacted by the truncating of students' performances to the binary pass/fail variable. Dichotomizing is usually performed for simplicity however this can lead to: lower accuracy through loss of valuable information; a decrease in the predictive power; and in general there is a risk of getting results that may not make sense (Fedorov, Mannino, & Zhang, 2009; Royston, Altman, & Sauerbrei, 2006). Many studies have reported results of identifying at-risk students at the end of the course/semester however for early warning systems this is impractical. Ideally we wish to support all students from the beginning of the semester. For a prediction model, the beginning of the semester is too early to identify at-risk students. For early warning systems, a balance needs to be obtained between the increasing accuracy of the system and the diminishing impact of intervening as we move through the semester. In this paper we refer to the balance between the two as the 'optimal time'.

2.3. Research questions

Our study aims to explore developing a prediction model for an early warning system taking into account the benefits of cluster analysis. Furthermore our study aims to identify an 'optimal time' in the

semester when an early warning system could be implemented. Hence our research questions, in context of *Practical Statistics*, are:

- Which prediction methods work best for predicting students' final grades?
- How do we identify a stage in the semester that can adequately balance the required timing of intervention with the quality of the prediction?
- What effect do cluster memberships based on student engagement have on prediction error?

3. Material and methods

In this section we discuss the course background of *Practical Statistics*, as well as the data collection process and analysis used in this study.

3.1. Course background information

This study took place in University College Dublin (UCD). Many of the large first year courses in UCD start in week 1 with material which links to the country's main State Examination and builds from there. Owing to the large class sizes with mixed-ability and the progression of material beyond prior knowledge, it may be several weeks before we can identify students who are struggling with the course. *Practical Statistics*, a large online undergraduate course aimed at first years, was selected as an example of a STEM course with weekly continuous assessment. It is designed as an introductory course in statistics for a class of mixed ability students. The lecturer allocates 40% of the final mark to continuous assessment and distributes the continuous assessment throughout the course semester to encourage students to continuously engage with the course. *Practical Statistics'* lectures are completely online but the students have 24 h of software labs. The continuous assessment is achieved through: lecture questions based on the course material (weeks 1–12; 0.5% per week; included in model from week 3); watching all of the online videos (2%); Minitab lab questions (weeks 3–5; 1% per week; included in model at week 5); R lab questions (weeks 7–11 excluding week 8; 1% per week; included in model at week 11); Minitab lab examination (week 6; 10%; included in model at week 6); and R lab examination (week 12; 15%; included in model at week 12). Answers to lecture questions and lab sheets are submitted to the LMS and automatically marked by the system, with the marks being returned instantaneously. Students have until midnight of the following Sunday to submit answers. In *Practical Statistics*, students have three attempts at their weekly lecture questions. Upon submitting their answers, students immediately receive "try again" feedback as defined by Shute (2008) i.e. whether specific answers are right/wrong, the percentage correct for that assessment and students have the ability to try the assessment again. For the Minitab lab examination, students receive their overall mark for the assessment. For the R lab examination (worth 15%), students receive no feedback as the examination takes place in the final teaching week of the semester. On the LMS, students are able to review their continuous assessment for each assessment and overall at any stage of the semester. We believe continuous assessment combined with summative feedback encourages student engagement in online courses.

3.2. Participants

In the first semester of 2015/16 there were a total of 144 students registered for *Practical Statistics*. Students' data was removed from the study if: students opted out of the research study; students did not take the end of semester examination; or students had personal circumstances which affected how they were officially graded for the course. Students with extenuating circumstances were excluded as these circumstances could impact students' continuous assessment and LMS use. This could impact predictions. In accordance with our ethical permissions from UCD, we removed these students rather than investigating individual student's circumstances. Subsequently our analysis sample included 136 participants from *Practical Statistics*.

3.3. Data collection and measurements

Data were recorded for students in regards to three categories: students' background information; continuous assessment; and LMS usage on a fine-grained scale. Background information of students (gender, course type (elective, option or core), registration of students (repeating course, etc.), students' year of study, students' programme and Irish/non-Irish) were included as variables to account for differences in educational background and prior experience of students. Online resources (for example videos, lectures slides, pdfs) were grouped into folders based on the material content. In total, there were 15 folders (week 1 course material, ..., week 12 course material, lecture questions solutions, course information, and past examination solutions). We refer to the number of times the resources in a folder were accessed and the folder itself was accessed as the 'activity level' of the folder. For each folder, we included the activity level for the folder for a given week as a variable, for example, in week one of the semester student '8979' had an activity level of 12 for the 'week 1 course material folder' (see Table 1). In week 2 of the semester, student '8979' had an activity level of 9 for 'week 1 course material folder'. These counts are not cumulative. In summary, there are 15 folders which could be accessed over 14 weeks (12 teaching weeks, 1 revision week and 1 exam week). Generally, only some of the folders would be accessed in a given week, for example the first time week 5 folder was accessed was in week 4 of the teaching semester. For a given folder, the activity level ranged from 0 to 55 for a given week. On average, the main resource folder for the week (for example folder 4 for week 4) had an activity level of 11 per student. The dataset was designed to be flexible whereby statistical analysis could be performed to incorporate data up to any stage/point in a semester. We performed statistical analysis for the end of each week in the semester (12 teaching weeks) as well as initially (when only background information was available), the end of revision week, and for the end of semester when the written examination was completed. In total this forms fifteen stages.

3.3.1. Prediction methods

A fundamental problem which occurs in prediction analysis is 'overfitting' whereby a prediction model fits training data very well but predicts poorly on new test data (Baumer, Kaplan, & Horton, 2017). This often occurs when the entire dataset is used in training the prediction model. K-fold cross-validation is a popular method for handling this issue and has been used in multiple prediction model studies (Azcona & Casey, 2015; Wolff et al., 2013). In K-fold cross validation, the dataset is divided into k equal-sized subsets or folds. K-1 folds are used to train the model and subsequently prediction analysis occurs on the kth fold. This is repeated k times such that all cases have a predicted

Table 1
Example dataset to be used to predict students' final module mark.

Student code	Gender	...	Major	Lecture Q results Percentage	Week 1 folder Count for week 1 only	Week 1 folder Count for week 2 only	...
8979	F	...	Science	70.6	12	9	...
9079	M	...	Science	95.1	8	15	...
4567	M	...	Arts	56.8	3	2	...
4547	M	...	Arts	64.7	7	12	...

value associated with them. Our prediction models (Random Forest; BART; XGBoost; Principal Components Regression; Support Vector Machine; Neural Network; Multivariate Adaptive Regression Splines; and K-Nearest Neighbours) were run using 10-fold cross-validation for the same folds. The final percentage grade was used as the response variable.

- Random Forest (RF) is an ensemble learning method. Breiman (2001) states "random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" (p. 1). For RF regression prediction, the mean prediction of the individual trees is returned.
- Kapelner and Bleich (2016) explain that BART is a Bayesian approach to nonparametric function estimation using sums of regression trees which allows for flexibility between non-linear interactions. BART differs from other tree ensemble methods (for example RF) owing to the underlying probability model and use of priors. A benefit of this is that we can create confidence intervals for our predicted values.
- XGBoost is a popular scalable machine learning system for tree boosting (Chen & Guestrin, 2016). It can handle large datasets as well as sparse matrices. As XGBoost cannot be applied to categorical data, any categorical variables were recoded as binary variables. For XGBoost modelling $\sqrt{n}$ iterations were run where n is the number of variables. XGBoost was applied to 15 stages in the semester (Initially, week 1, ...). Initially XGBoost was used on 18 variables (where categorical variables were transformed to multiple binary variables). The number of variables and iterations increased on a week by week basis as additional Blackboard data became available.
- Principal Components Regression (PCR) is a technique that reduces a high dimensional dataset to a lower dimension dataset and then performing regression. It does this by finding linear transformations of the data whereby the maximal amount of variance is retained (Ilin & Raiko, 2010).
- "Kernel-based learning methods (including Support Vector Machines (SVM)) use an implicit mapping of the input data into a high dimensional feature space defined by a kernel function" (Karatzoglou, Smola, Hornik, & Zeileis, 2004). The training of the model is then performed in the feature space.
- Feedforward Neural Network (NN) is a system of nodes which is an imitation of the human brain. A feedforward neural network consists of nodes in layers providing information forward through the layers using the equation $y_i = wx_i + b$. Training neural networks is considered to be difficult (Larochelle, Bengio, Louradour, & Lamblin, 2009).
- Multivariate Adaptive Regression Splines (Splines) is a non-parametric stepwise regression procedure (Friedman, 1991) When including variables, the range of the variable is partitioned into subsets and a constant is applied to each subset for regression. In the backward pass, the model is pruned to limit overfitting.
- K-Nearest Neighbours (KNN) is a nonparametric method whereby the 'k' nearest neighbours or 'k' most similar cases impact the prediction/classification of the case of interest (Hechenbichler & Schliep, 2004). In the case of regression, the 'k' nearest neighbours response values are averaged with importance weightings being considered.

We used mean absolute error (MAE) between the predicted grade and actual grade as a comparison basis to observe the improvement in the accuracy of the prediction model on a week-to-week basis. This allowed us to identify an 'optimal time' for an early warning system to be employed. To improve the accuracy of the initial models, our prediction models were applied to different feature sets which included a combination of: continuous assessment data; background information; as well as varying the levels of LMS data. To further reduce the prediction error we considered: Sunday count variables[2]; cumulative count variables; and resultant cluster analysis. The feature sets discussed in Section 4 are:

- Initial Model - Variables include background information, continuous assessment, and LMS activity level per folder.
- No LMS Variables - Variables include background information and continuous assessment.
- Cumulative Variables - Variables include background information, continuous assessment, and cumulative activity level for each individual folder (for Sundays and for weekdays).
- Cluster Variables - Variables include background information, continuous assessment, cumulative counter of views for each individual folder (for Sundays and for weekdays), and cluster membership variables.

3.3.2. Clustering methods

Cluster analysis is not an essential part of predicting students' final examination marks, however, it can provide meaningful variables which reduce the prediction error. Studies have shown that students can be grouped according to their engagement patterns or module resource use (Lust et al., 2011; White & Carroll, 2017). Since these groups are generally distinct, with a strong relationship with modules marks, variables(s) which classify students according to these groups are potentially good predictors. There are two formats for cluster membership variables. In the first, there is a single nominal variable where students are assigned a number which specifies the cluster that they are associated with. In the second format, there is a variable for each cluster and a student has a probability value of membership to each cluster.

The dataset used for clustering contained fine-grained LMS data (the activity level for each individual folder per week and per Sunday). We use the model-based clustering package *mclust* (Scrucca, Fop, Murphy, & Raftery, 2016) to create an additional clustering of our variables. We use this package because of its repeated superior performance compared to other clustering algorithms (Scrucca et al., 2016), and its ability to model a wide variety of cluster sizes and shapes. Owing to its model-based nature, an advantage of using *mclust* is its ability to calculate probability memberships for each individual to each cluster. Only a limited number of cluster methods has this advantage. Clustering was performed for each stage in the semester. The estimated Bayesian Information Criterion (BIC) was compared for the different combinations, and the combination which maximised the BIC was selected. The resultant cluster membership was considered as a variable for prediction modelling.

4. Results

We now describe the development of prediction methods for an early warning system. Continuous assessment played an important role in our modelling. When developing an early warning system, we need to account for any delays in the correction of continuous assessment or collection of data for example if a midterm in week 5 takes 2 weeks to correct, we should include it in week 7. *Practical Statistics* benefits from the instantaneous nature of online LMS assignments. Through the development of our early warning system, we are able to identify an optimal time (week 5–6) in the *Practical Statistics'* semester to apply an early warning system.

4.1. Student engagement and continuous assessment

Holmes (2015) and Cole and Spence (2012) have suggested that continuous assessment encourages student engagement. As previously mentioned, *Practical Statistics* was designed to ensure consistent student

[2] This is discussed further in Section 4.3.

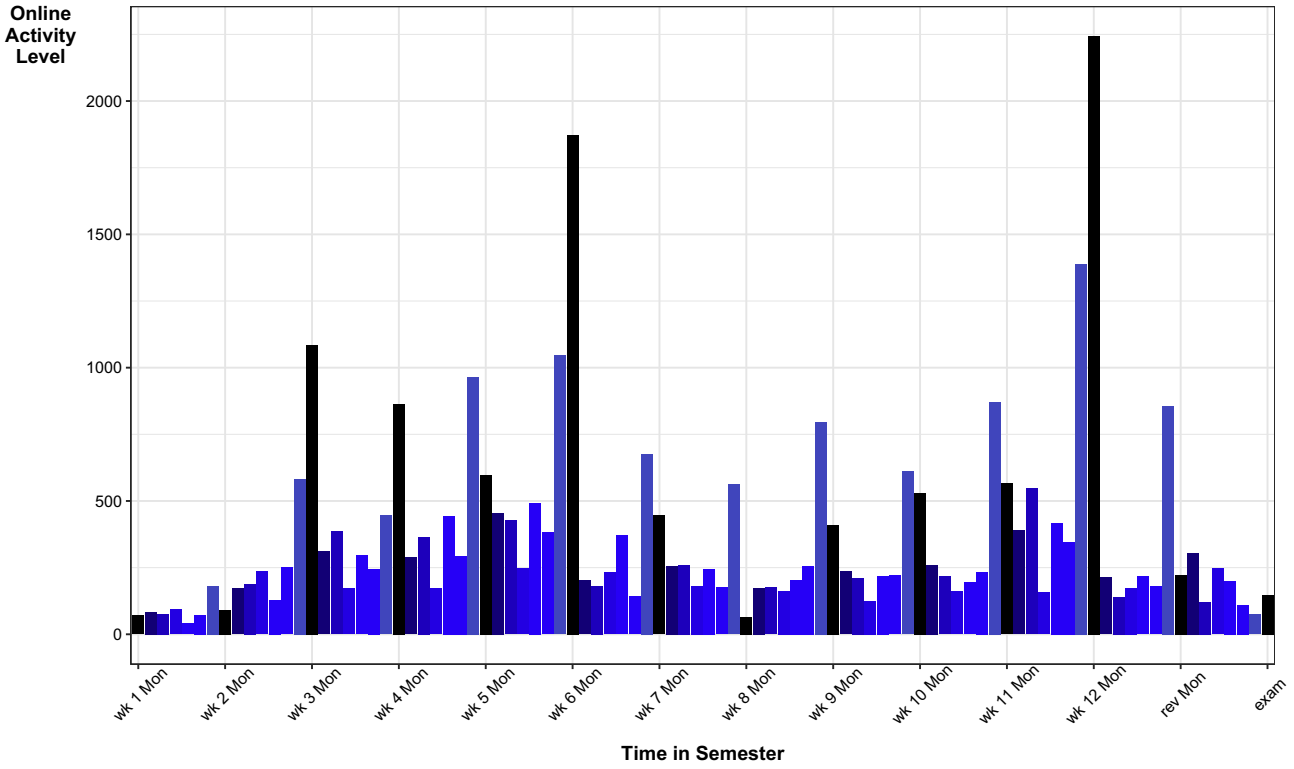

Fig. 1. Activity level of online resources per day over the course of *Practical Statistics'* semester.

engagement through having continuous assessment on a weekly basis throughout the course. Fig. 1 shows that online materials were accessed throughout the semester, however the level of activity, not surprisingly, peaked prior to assessments. The deadline for weekly online lecture questions for credit was on Sunday nights, and this corresponds with the weekly peak in online resource activity. These peaks might suggest two types of students: students who study immediately prior to assessments; and students who study in advance of assessment. Similarly, as expected, the time when the greatest number of resources was accessed corresponds to the day of the R lab examination (the Monday of week 12). A similar peak occurs on the Monday of the Minitab lab examination in week 6. This connection between online views and continuous assessment suggests that a key driver of students' interaction with online resources is continuous assessment.

4.2. Clustering analysis

mclust was applied to several variations of the dataset. Considering the high number of view counts on Sunday, this included investigating the potential of Sunday online activity as separate to weekday[3] online activity. After investigating resultant clusters, *mclust* was only applied to fine-grained LMS data (the activity level for each individual folder per week and per Sunday). Continuous assessment variables and background information of students, were not included as cluster variables. The resultant clusters identified differences in students' frequency levels of using online resources. In comparison to Lust et al. (2011) who divide online resources into tool types, this method is cruder as the clustering is unlikely to pick up subtle differences in students' learning strategies.

For example, for week 5 (identified optimal time) the variables used were fine-grained LMS data (the activity level for each folder per week and per Sunday) for weeks 1–5. *mclust* identified 3 clusters ($n_1 = 61$ (44.9%), $n_2 = 69$ (50.7%) and $n_3 = 6$ (4.4%)). The distinct clusters are

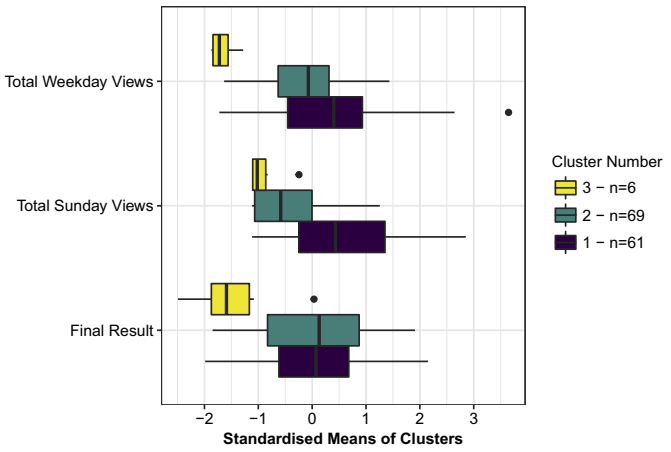

Fig. 2. Identifying engagement patterns of Practical Statistics through boxplots of selected standardised variables for week 5. for example cluster 3 contains six students who have below average resource usage.

best represented in 2D format by boxplots (see Fig. 2) showing the standardised means and spread of the selected variables for each cluster. Three variables (Total Weekday Views (up to week 5), Total Sunday Views (up to week 5), and Final Grade) were selected to show the distinct clusters (see Fig. 2). Cluster 3 students are students who display below average engagement with resources and have the widest final grade range. Lust et al. (2011) categorize these as no-users or low frequency users. Cluster 2 represents the students who have below average resource use on Sunday, and average resource use during the week. In comparison, Cluster 1 represents students who engage (above) average with resources overall; both on Sunday and during the week. Despite this higher average engagement than the other clusters, they have the median final examination grade. Subsequently, as the cluster analysis displayed distinct clusters with various engagement patterns, students' cluster group membership was used as variables in the prediction analysis.

[3] In this study weekday view counts includes Saturday view counts.

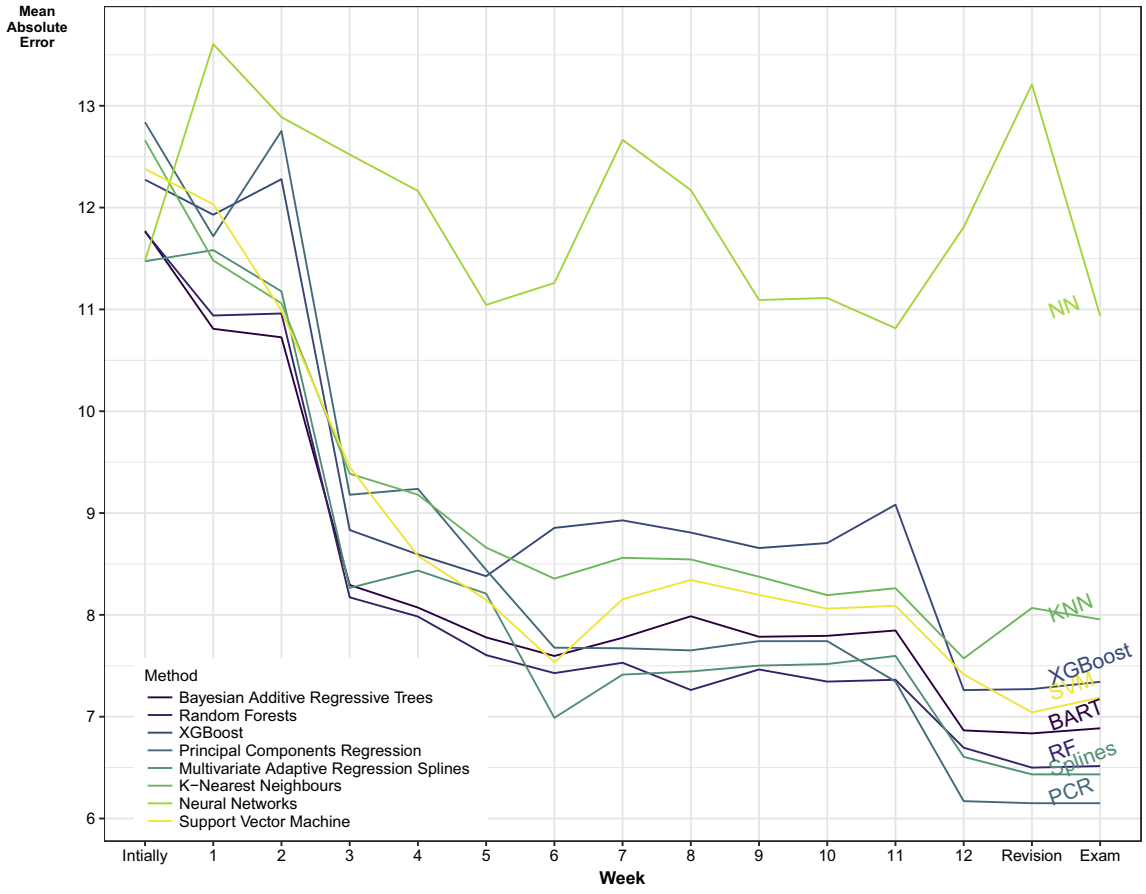

Fig. 3. Average MAE per student on a week-by-week basis from multiple out of sample prediction methods.

4.3. Prediction modelling

Initial prediction modelling was performed on the dataset for each week (all variables available up to that date were included - see Initial Model Section 3.3.1) to determine an optimal time for corrective measures. The initial stage (before teaching semester began) and final stage of the semester acted as a baseline for comparison for the power of the prediction model (see Fig. 3).

In Fig. 3, PCR achieves the lowest MAE value (approximately 6 points at week 12). PCR reduces the number of variables before performing regression. Out of the methods investigated, Neural Networks is clearly the inferior method. An interesting feature of Fig. 3 is the substantial decrease in error from week 2 to week 3. This decrease in error coincides with the inclusion of continuous assessment in the prediction model (the deadline for week 1 lecture questions was in week 3). This emphasises the role continuous assessment plays as a predictor in online STEM courses. To confirm the importance of continuous assessment, we investigated the variable importance of the models. Every model selected continuous assessment variables as the main variables in the model. Fig. 3 shows us the MAE for each prediction model over the entire semester for *Practical Statistics*, however, between weeks 6 and 11 there is relatively little change in the predictive power of the models. From the lack of decrease in predictive error between weeks 6 and 11, we believe that there is little benefit in waiting beyond week 6 to implement an early warning system for *Practical Statistics*. For early warning systems, a balance is required between the accuracy of the prediction models and the stage in the semester in which they are implemented. The stage in the semester needs to reflect where corrective measures could most effectively be given to students. We believe that any small prediction decreases in waiting beyond week 6 would not overcome the additional effort of collecting data after week 6 for prediction modelling and the delay in implementing the early warning system. While there is a decrease in prediction error from week 11 to week 12, after this time is too late in the semester to effectively implement an early warning system in *Practical Statistics*. Subsequently, the stages up to week 6 were identified as important for further data analysis. For *Practical Statistics*, week 5 is potentially the optimal time for implementing an early warning system. We have included our R code for this in the Appendix A with more detailed R code and fictitious datasets available on GitHub at https://github.com/ehoward1/Early-Warning-System.

To reduce our prediction error, we considered alternative feature sets. Investigating alternative variables or alternative formats of variables can be a straightforward way to improve prediction accuracy. We considered alternative feature sets including removal of the LMS data (which provided slightly less accurate predictions), including cumulative activity level for each folder (Cumulative Variables dataset), and including cluster membership variables. These datasets did not involve collection of additional data and are a sample of all the datasets examined. Progressing, we will look at the Cluster Variable dataset in further detail as cluster membership is not a common variable in prediction modelling for early warning systems, and as the prediction models had lower error rates on the Cluster Variable dataset than on the initial dataset. The Cluster Variables dataset for each student consists of: background information; continuous assessment; and cumulative counter for the activity level of each individual folder (for Sundays and for weekdays) as well as cluster membership variables. While including the cluster variable (in most cases) does not alter the MAE significantly, clustering can provide us with information about student engagement in general which may be of value (see Section 4.2 Clustering Analysis). Fig. 4 gives the average MAE per student for the Cluster Variables dataset up to the optimal time of week 6. The second substantial decrease

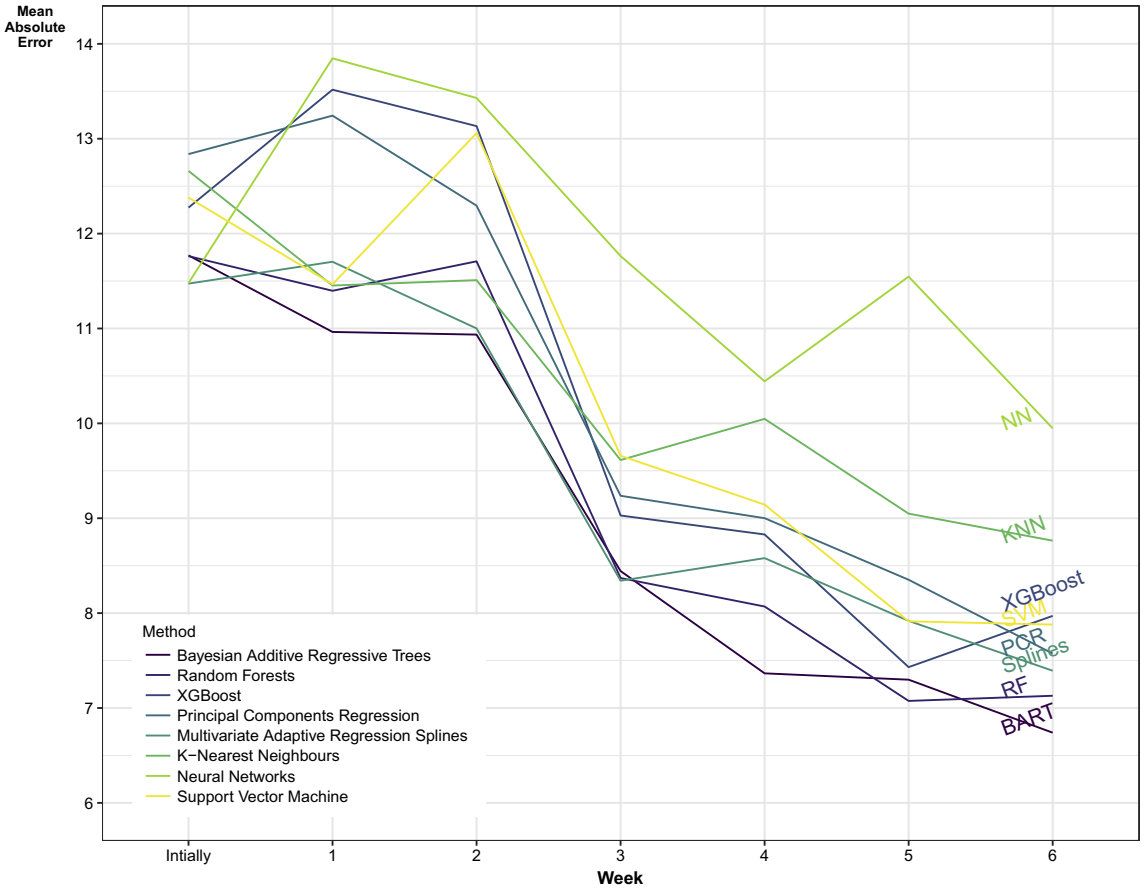

Fig. 4. Improving the Prediction Model by changing the feature dataset to include cumulative variables for LMS interactions and cluster membership variables.

in MAE between weeks 4 and 5 corresponds to the second inclusion of continuous assessment (Minitab lab results). Using our BART predictive model we can identify the final mark the student will obtain to approximately a MAE of 6.5 at week 6.

We will proceed by discussing in further detail the BART prediction model at week 5 using the Cluster Variable dataset. This dataset consists of 29 explanatory variables. We can visually determine the performance of our predictive model by plotting the predicted final grade against the true final grade for each student. Fig. 5 shows the predicted grade plotted against the actual grade of each student, both initially and at the end of week 5. An identity line, showing when the predicted grade equals the actual grade (i.e. a perfect prediction), has been included in Fig. 5. The initial plot acts as a baseline, displaying how the initial prediction of final grade has very low correlation with the actual grade

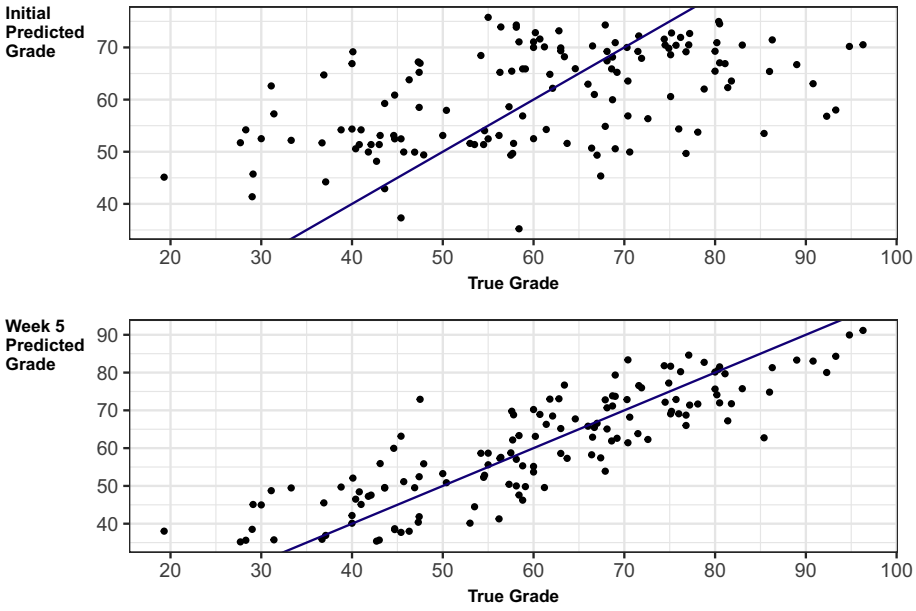

Fig. 5. Scatter plots showing predicted grade versus actual grade initially and at week 5 via an out of sample 10-fold cross-validation.

of students i.e. a poor predictive performance. The initial model relies on a limited number of background/demographic variables. As several students have the same background information, this has resulted in multiple students receiving the same predicted grade. This has resulted in 'bands' of predicted grades. In comparison, the second plot's data is quite linear ($R^2 = 0.74$) and tighter to the identity line, with some outliers. It suggests that by week 5 we can make reasonable grade predictions as the grade predictions are strongly correlated to the actual grade. This supports the belief that week 5 is an optimal time to implement an early warning system, and that the selected BART model (Method - Cluster Variables) performs competently.

5. Discussion and conclusion

5.1. Continuous assessment

The variables used in this study were divided into three categories: students' background information; students' engagement with LMS; and continuous assessment results. Continuous assessment proved unsurprisingly the most important category. Continuous assessment variables were repeatedly chosen as the most important variables by all of the prediction models. Continuous assessment encourages students to engage with a course (Holmes, 2015) and partially accounts for the different levels of LMS interaction throughout the semester. This is observable from the spikes in LMS resource use prior to continuous assessment tests and deadlines (Fig. 1). We suggest the inclusion of consistent continuous assessment in online courses encourages students' engagement over the entire semester (as stated by Cole & Spence, 2012), and limits the number of students studying only in the weeks prior to the final examination. The addition of continuous assessment also contributes to minimising prediction error when building early warning systems however this should not be the main reason for its inclusion. We hypothesize that a low percentage for continuous assessment would also achieve the same effect provided that the continuous assessment is throughout the semester.

This study investigates how to approach developing an accurate prediction model for an early warning systems. The dataset which only had continuous assessment and background information variables, performed comparatively well to the other feature sets, and enjoyed the benefit of being the simplest model. However, by using this model we fail to identify areas of the curriculum where students struggled. A key element in learning analytics is using the resultant analysis for the benefit of the student and teacher. By including the extra LMS variables, we are able to investigate for individual students aspects of the curriculum that they failed to engage with or had overly high engagement with (potentially a sign of a harder concept or an area with which the student struggled). This advantage for the inclusion of LMS variables is considerable, and should be weighed against the simplicity of the 'No LMS Variables' data set.

5.2. Advancements in developing early warning systems

This study summarises the methods employed in developing prediction models for early warning systems, and builds upon the current work. Unlike some of the other studies, we do not dichotomise students' final grades to pass/fail. Instead, we predict students' final percentage mark similar to Huang and Fang (2013) and You (2016). We discuss how one may reduce the prediction error through: use of fine-grained variables; manipulation of variables; and the inclusion of cluster membership in prediction modelling. The detail provided by fine-grained variables gives more information on students' engagement patterns. Subsequently, we hypothesize that analysis of fine-grained variables will allow for more personalised corrective measures. We have used predictive methods (BART and XGBoost) which are uncommon in the data analytics literature as well as common predictive methods (Neural Networks, K-Nearest Neighbours and Random Forest). We found that decision tree methods perform particularly well (BART and Random Forest). Decision tree methods are suitable when using a large number of variables. Hence BART, a decision tree method, is appropriate when using fine-grained variables. BART may be preferable over other decision tree methods, for example Random Forests, owing to its Bayesian nature which allows for the inclusion of error variance which is independent of tree structure and leaf parameters (Kapelner & Bleich, 2016). In our study BART outperformed the other prediction models tested at the optimal time of weeks 5–6.

Clustering is not a necessary step in developing prediction models. However, we have shown that clustering can be used to identify distinct student patterns of engagement which can be used to further reduce the prediction error. Also, clustering may help to identify how students approach learning and subsequently be used to provide corrective measures. The method outlined in this study is appropriate for both online courses and large classes with a significant amount of online material. Through combining these methods, we obtain an average prediction error (based on out of sample 10-fold cross validation and MAE) of 6.5 percentage points by week 6.

A key part of this study was identifying an 'optimal time' to implement an early warning system. Implementing an early warning system too early would result in inaccurate identification of (at-risk) students. In contrast, implementing it too late would diminish the effect of supporting and helping students. Data analysis of prediction models identify week 5/6 as the critical time in the semester for *Practical Statistics* whereby prediction models have reasonably accurate forecasts balanced with sufficient time to intervene and support at-risk students. Identifying at-risk students is only one stage in an early warning system, another stage is understanding what effective supports should be provided to students. Consequently, our current research involves identifying at-risk students during the 'optimal time' in *Practical Statistics* and examining which feedback/intervention measures are effective for large STEM courses.

5.3. Limitations

The method outlined in this study discusses how to develop an accurate predication model for an early warning system for a course, and how to recognise an optimal time to provide students with corrective measures during a course. *Practical Statistics* is an example of a STEM course which has continuous assessment distributed weekly throughout the semester. The method discussed in this study may not be an optimal method for other online courses, particularly if the course is from a significantly different academic field. Each course is unique and will have its own unique feature set. STEM based courses, particularly early undergraduate courses, tend to have continuous assessment which ties to the final examination. We believe BART is applicable for these STEM courses.

For the purpose of reproducibility, the R code for comparison of the prediction models has been included in Appendix A. Further code for this study is available on GitHub at https://github.com/ehoward1/EarlyWarning-System with fictitious datasets (owing to ethical constraints).

Ethics

This study was conducted in accordance with UCD ethics guidelines, and approved by the UCD Ethics Committee under application number LS-15-53-Meehan.

Acknowledgements

We would like to thank UCD IT services for providing us with Blackboard data. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. R code

Function to run and compare 10-cross fold validation for all prediction methods used in the paper. Fictitious datasets for this function and further R code for this paper are available on GitHub at https://github.com/ehoward1/Early-Warning-System.

```
require(xgboost)
require(randomForest)
require(bartMachine)
require(pls)
require(caret)
require(magrittr)
require(earth)
require(nnet)
require(car)
require(kknn)
require(kernlab)
prediction_function<- function(dataset, dataset_boost){ # dataset boost is for xgboost
  set.seed(123)
  folds = createFolds(1:nrow(dataset), k = 10, list = FALSE)
  dataset_boost = apply(dataset_boost, 2, as.numeric)
  # XGBoost runs for numeric data, not integers
  # Vectors to store error for each prediction methods
  pred_bm = vector("numeric")
  pred_rf = vector("numeric")
  pred_pcr = vector("numeric")
  pred_xg = vector("numeric")
  pred_kknn = vector("numeric")
  pred_svm = vector("numeric")
  pred_nnet = vector("numeric")
  pred_earth = vector("numeric")
  grades = vector("numeric")
  # Loop through the folds for each prediction method
  for(i in 1:10){
    # Setting up the data
    train = dataset[folds!=i,] %>% data.frame %>% na.omit
    train_b = dataset_boost[folds!=i,] %>% data.frame %>% na.omit
    test = dataset[folds==i,] %>% data.frame
    test_b = dataset_boost[folds==i,] %>% data.frame
    # BART
    bm = bartMachine(train[,-1], train[,1], seed = 123, alpha = 0.95, num_burn_in = 400,
    num_tree = 100, num_rand_samps_in_library = 20000, k = 2, q = 0.9, nu = 3)
    pred_bm = c(pred_bm, predict(bm, test[,-1]))
    # Random Forest (RF)
    rf = randomForest(train[,-1], train[,1], ntree = 100)
    pred_rf = c(pred_rf, predict(rf, test[,-1]))
    # Principle Components Regression (PCR)
    pcr = pcr(FINAL~., data = train)
    var_exp = compnames(pcr, explvar = TRUE)
    var_e = unlist(strsplit(var_exp, "[ ()")) %>% as.numeric() %>% setdiff(c(1:150, NA))
    var_total = 0

# Calculating number of variables to include based on variation explained
for(j in 1:length(var_e))
{
var_total = var_total + var_e[j]
if(var_e[j] < 1 || var_total > 90)
{
n_comp = j
break
}
}
pred_pcr = c(pred_pcr, predict(pcr, test[,-1], ncomp = n_comp))
# Xgboost
iter = train_b %>% ncol %>% sqrt %>% ceiling
xg = xgboost(data = as.matrix(train_b[,-1]), label = train_b[,1], eta = 0.5,
nround = iter, max.depth = 4, objective = "reg:linear")
pred_xg = c(pred_xg, predict(xg, as.matrix(test_b[,-1])))
# K-Nearest Neighbours (KNN)
kknn = train.kknn(FINAL ~., kmax = 15, distance = 1, data = train)
pred_kknn = c(pred_kknn, predict(kknn, test[,-1]))
# Neural Network (NN)
my.grid = expand.grid(.decay = c(0.05, 0.5, 0.75), .size = c(4, 9))
nnet = train(FINAL~., data = train, linout = 1, method = "nnet", maxit = 500, tuneGrid = my.grid, trace = FALSE)
pred_nnet = c(pred_nnet, predict(nnet, test[,-1]))
# Support Vector Machine (SVM)
svm = ksvm(FINAL ~., data = train, C = 5)
pred_svm = c(pred_svm, predict(svm, test[,-1]))
# Multivariate Adaptive Regression Splines (Splines)
earth = train(FINAL~., data = train, method = "earth", tuneGrid = data.frame(degree = c(1,2), nprune = 5))
pred_earth = c(pred_earth, predict(earth, test[,-1]))
grades = c(grades, test$FINAL)
}
# Calculating the error for each method
error_rf = sum(abs(pred_rf - grades))/nrow(dataset)
error_pcr = sum(abs(pred_pcr - grades))/nrow(dataset)
error_xg = sum(abs(pred_xg - grades))/nrow(dataset_boost)
error_bm = sum(abs(pred_bm - grades))/nrow(dataset)
error_earth = sum(abs(pred_earth - grades))/nrow(dataset)
error_kknn = sum(abs(pred_kknn - grades))/nrow(dataset)
error_nnet = sum(abs(pred_nnet - grades))/nrow(dataset)
error_svm = sum(abs(pred_svm - grades))/nrow(dataset)
# Returning Values
my_list = list("MAE_bm" = error_bm, "MAE_rf" = error_rf, "MAE_pcr" = error_pcr,
"MAE_xg" = error_xg, "MAE_kknn" = error_kknn, "MAE_nnet" = error_nnet,
"MAE_svm" = error_svm, "MAE_earth" = error_earth)
return(my_list)
}
```

References

Azcona, D., & Casey, K. (2015). Micro-analytics for student performance prediction leveraging fine-grained learning analytics to predict performance. *International*

Journal of Computer Science and Software Engineering, 4(8), 218–223.
Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2017). *Modern Data Science in R*.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5— 32.
Calvo-Flores, M. D., Galindo, E. G., Jiménez, M. C. P., & Pérez, O. (2006). Predicting students' marks from Moodle logs using neural network models. *Current Developments in Technology-Assisted Education, 1*, 586—59 0.
Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *22nd SIGKDD Conference on Knowledge Discovery and Data Mining* San Francisco, CA, USA: KDD '16.
Cole, J. S., & Spence, S. W. T. (2012). Using continuous assessment to promote student engagement in a large class. *European Journal of Engineering Education, 37*, 508–525.
Corrigan, O., Smeaton, A. F., Glynn, M., & Smyth, S. (2015). Using educational analytics to improve test performance. *Proceedings of Design for Teaching and Learning in a Networked World, 10th European Conference on Technology Enhanced Learning* (pp. 42–55). Toledo, Spain: Springer.
Fedorov, V., Mannino, F., & Zhang, R. (2009). Consequences of dichotomization. *Pharmaceutical Statistics, 8*, 50–61.
Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5/6), 304–317.
Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*(1), 1–67.
Gašević, D., Dawson, S., Rogers, T., & Gašević, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education, 28*, 68–84.
Hechenbichler, K., & Schliep, K. (2004). Weighted k-nearest-neighbor techniques and ordinal classification. *Discussion Paper 399, SFB, Ludwig-Maximilians University Munich* (pp. 1–16). .
Holmes, N. (2015). Student perceptions of their learning and engagement in response to the use of a continuous e-assessment in an undergraduate module. *Assessment & Evaluation in Higher Education, 40*, 1–14.
Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education, 61*, 133–145.
Ilin, A., & Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research, 11*, 1957–2000.
Joksimović, S., Gašević, D., Loughin, T. M., Kovanović, V., & Hatala, M. (2015). Learning at distance: Effects of interaction traces on academic achievement. *Computers & Education, 87*, 204–217.
Kapelner, A., & Bleich, J. (2016). BartMachine: Machine learning with Bayesian Additive Regression Trees. *Statistical Software, 70*(4), 1–40.
Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab – An S4 package for kernel methods in R. *Journal of Statistical Software, 11*, 1–20.
Larochelle, H., Bengio, Y., Louradour, J., & Lamblin, P. (2009). Exploring strategies for training deep neural networks. *Journal of Machine Learning Research, 1*, 1–40.
Lust, G., Vandewaetere, M., Ceulemans, E., Elen, J., & Clarebout, G. (2011). Tool-use in a blended undergraduate course: In search of user profiles. *Computers & Education, 57*, 2135–2144.
Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*(2), 58 8—599.
Marbouti, F., Heidi, D.-D., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education, 103*, 1–15.
Pistilli, M. D., & Arnold, K. E. (2010). In practice: Purdue signals: Mining real-time academic data to enhance student success. *About Campus, 15*(3), 22—2 4.
Pistilli, M. D., & Arnold, K. E. (2012). Course signals at Purdue: Using learning analytics to increase student success. *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 2–5). New York, NY, USA: ACM LAK' 12.
Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine, 25*(1), 127 —141.
Sclater, N., Peasgood, A., & Mullan, J. (2016). Learning analytics in higher education: A review of UK and international practice. *Technical report*London: Jischttps://www.jisc.ac.uk/sites/default/files/learning-analytics-in-he-v2_0.pdf April.
Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Foundation for Statistical Computing, 8*(1), 289–317.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.
Waddington, R. J., Nam, S., Lonn, S., & Teasley, S. D. (2016). Improving early warning systems with categorized course resource usage. *Journal of Learning Analytics, 3*(3), 263–290.
White, A., & Carroll, P. (2017). Identifying patterns of learner behaviour: What business statistics students do with learning resources. *INFORMS Transactions on Education, 18*(1), 1–13.
Wolff, A., Zdrahal, Z., Herrmannova, D., Kuzilek, J., & Hlosta, M. (2014). Developing predictive models for early detection of at-risk students on distance learning modules. *Proceedings of the fourth international conference on learning analytics and knowledge*New York, NY, USA: ACM LAK '14.
Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). The Open University's repository of research publications improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment conference item analysing clicking behaviour in a virtual learning. *Third conference on learning analytics and knowledge.*New York, NY, USA: ACM LAK '13.
You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet and Higher Education, 29*, 23–30.