Predicting Course Engagement using Student Data

Mahdi Islam
mahdi@uni-bremen.de
University of Bremen
Bremen, Germany

Kamela Chowhdury
kamela@uni-bremen.de
University of Bremen
Bremen, Germany

G M Nazmul Hossain
gmnaz@uni-bremen.de
University of Bremen
Bremen, Germany

ABSTRACT

Online learning platforms collect an exceptionally large amount of student data including their interaction with course material, performances, and other relevant information. A proper study of this data can reveal multidimensional information about the learning environment and help both teachers and students to design better course modules. In this project, we have prepared a data predictive model to predict course engagement rate to identify courses that have higher or lower student engagement using multiple machine learning models. The data of a Virtual Learning Environment, Students and their Assessment are used as input to obtain an output which reveals student's engagement rate with a given course. From the outcomes, we derived that machine learning algorithms can be used to predict course engagement rate and help education facilitators to evaluate the student-course relationship, and tailor their courses to increase the involvement which in turn can generate better student performance.

KEYWORDS

Course Engagement, Machine Learning Model, Online Learning, Binary Classification

ACM Reference Format:
Mahdi Islam, Kamela Chowhdury, and G M Nazmul Hossain. 2019. Predicting Course Engagement using Student Data. In *Proceedings of ACM Conference (University Of Bremen).* ACM, New York, NY, USA, 7 pages. 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Online learning has become even more popular, thanks to more students with connected smart devices, and advanced online learning environment. However, retaining students has also become tougher for online education facilitators due to absence of face-to-face communication and in effect lack of understanding of student relationship with the course material. In such scenario, machine learning can help to portray a better picture and allow teachers to identify low engagement courses to design their material better. Advanced Educational Data Mining (EDM) techniques can be used get an insight of the data. Various indicators such as student's engagement with course material, interaction with teachers, performance analytic all can be a good resource for a bigger picture of an education system and its evaluation.

Educational Data Mining(EDM)

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data obtained from educational settings, and uses those methods to better understand students and the settings in which they learn [16].

EDM techniques can be referred to as formative evaluation [9] technique which involves the evaluation of an underdevelopment educational program to continually improve the program. These techniques can help education program designers to better understand student's interaction with online learning materials and assist in enhancing program design [14].

In our project we study the role EDM can play and how data predictive models can be used to mine student data. we attempted to answer the following learning questions:

(1) Can we use machine learning to predict student engagement with a given course?
(2) Which model is best suited to our project goal and chosen data set?
(3) Does such prediction offer any valuable insight?

2 BACKGROUND

In this project, we used Python as a data science programming language and multiple machine learning (ML) techniques to develop a predictive model of course engagement. The ML techniques, tools and model evaluation metrics used in this project are described as below:

Machine Learning

Arthur Samuel, one of the key pioneers of machine learning, in 1959 defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed". Complementary to this rather obscure definition, in 1998, Tom Mitchell provided a more precise definition of machine learning: "Well posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E" [12]. Machine learning, in general, is a sub-area of

Predicting Course Engagement
University Of Bremen, July 2019, Germany

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
10.1145/nnnnnnn.nnnnnnn

artificial intelligence. The central goal of machine learning is to develop learning models that can learn automatically without any human intervention or support .

Supervised Learning: Classification

Supervised learning is a method of training a model using known input and output data. In Supervised learning, the already known final result is obtained by feeding a model "labeled" input/output data. Supervised learning is subdivided in two machine learning problems: Regression and Classification. We chose supervised learning as it is more fitting to our properly labeled input data and our goal to compare the final test result with training data. Classification is one of the most common supervised machine learning problems. It is used to categorize data into a distinct number of classes which can be used to predict a labeled class.

Tools

Jupyter Notebook: Jupyter Notebook is a popular data science tool that contain live code, visualizatio and narrative text. We use this for data cleaning, simulation, transformation etc.

Libraries

Pandas: This is an open-source, robust popular Python package that makes data manipulation and analysis very easy.

Numpy: This is a Python scientific computing library that provides support for the large, multi-dimensional array.

Evaluation Metrics

In order to measure and evaluate the performance of the ML algorithms, fours evaluation matrices have been used: *Confusion Matrix , Recall, Precision, F1 Score, AUC - ROC Curve.* The *Confusion Matrix* holds the value of about predicted and actual class labels information which is generated by a predictive model. The data in the matrix are used to evaluate. *Accuracy score* is very common evaluation metric for the classification models. It can be calculated as the number of correct predictions divided by the total number of predictions. *Recall* is the number of correct positive predictions and the ratio of the total number of positives. This is also known by rate of true positive. *Precision* is the number of correct positive predictions as a ratio of the total number of positive predictions. *F1 score* is the harmonic mean between recall and precision. Lastly, *The Area Under Curve (AUC) and Receiver Operating Characteristic (ROC)* is used to measure the performance of a classification model.

3 METHOD

Data and Collection

In this project, we used data from Open University Learning Analytics data set by The Open University, UK. As there had a lot of student data we only used one season of 2013j as the winter semester and the subject name is AAA(Social science). In this course we had total 383 students as registered according to data. Base on all assignment score and attendance of assignments,we try to find out the engagement of the students. We took all the assignment score and attendance because our findings is to course engage, and we here try to show depends on student engagement, we can evaluate the course engagement.

One problem is that the selected attributes are stored in different tables (student info, student assessment, assessments, student VLE, courses, and VLE) in the OU data, as shown in Figure. Student info table contains the student's demographic information and the results of each course. Course table contains information about the courses in which students are enrolled. Registration table contains student record timestamps and course enrollment dates. Assessment information is recorded in the assessment table . Student-assessment table contains the assessment results for different students. Interaction information of different students regarding different materials and activities is stored in the student-VLE and VLE tables. VLE interaction data consist of the numbers of clicks students made while studying the course material in the VLE. Each course activity is identified with a label, for example, data plus, forumng, ou content, etc.

The Open University (OU) data set could not be used directly as inputs in the Machine Learning Model. They have 7 different table and target to make it one single table with all important attributes [8]. We have done many prepossessing steps on the data using Python to format the raw data converted into a machine acceptable form and build our feature table.The input matrix we use 15 columns from our feature table without EwC attribute and convert it as array. The value of Engagement with Course (EwC) is our output matrix value, we convert it to array. The prepossessing steps are given below,

Attribute Selection

Student Info. Firstly, we take all the students who had taken the course AAA and the session is 2013J from the studentInfo table and we select only 3 attributes idStudent, highestEducation and final result. In that case, we found only 383 students.

Assessments. We take all the assessments for only the course AAA and 2013J session. Here we found 6 types of assessments. From this table we take only one attribute idAssessment

Scores. From the studentAssessments table we get the score for each assessment for the students.

Feature Selection. We also add some feature in our feature table. Every attribute of our table we calculate for each student. The features are -

- Total Sub Click (TSC) = studentVle.sumClick.sum()
- Total Day of Click (TDC) = studentVle.date.count()
- Excellent = We put value 0 or 1. If the average score for all assessments > 60 percent, then we take this positive and put value 1, otherwise we put value 0. The value 1 means the student has a highest engagement on the course.
- Qualified = If the final result equals to pass then we put 1 otherwise 0.
- Active Student (AS) = If the total TSC mean is less than individual student TSC, then we put 1, else we put 0.
- Engagement with Course (EwC) = Finally, we calculate the EwC attribute value depending on Excellent, Qualified, AS. The highest engagement is 1 and lowest engagement is 0. The value we find by calculating this formula , if(excellent | (qualified and AS)) EwC=1, Else EwC = 0

When we found the value for EwC then we remove this 3 columns excellent, qualified and AS. Because we don't need this column for our model. The value of EwC this is our output value. Removing this 3 columns we made our final feature table and it has total 16 columns. The output columns is EwC and other 15 columns we use as our input attribute.

Label (High and Low Engagement) Extraction. Engagement is very important in a web based learning system, because it affects the student performance. The most highly engaged students make more high scores. First we define a label of engagement before developing the Machine Learning model. The total number of times a student accessed the vle activities is an indicator of engagement with course. However, the prediction of student engagement in VLE course [5] only counting clicks is very difficult because sometimes they are going to click on other sites like facebook, twitter or unimportant activities. Additionally, in some cases, students spend a little bit of time on VLE course but made a good score on assessment. So not only clicks but also we choose other 4 main activities as a criteria for measuring the student activities. They are assessment score (score on the assessment), final exam results after completing the course, student education degree before registering the course (highest education level) and total number of clicks on VLE activities. But it's not always correct. Therefore, we measured the student engagement based on Excellent, Qualified and Active during course. Finally, the results conclude that, high scores is achieved by high engagement students on assessments (excellent) and pass (qualified) in the final exams are more active during the VLE course (active).

Feature Selection. Here we assume that, when students interact with course materials, attending assessments and receive information that means the students are more active and high engagement with course. Depends on the number of clicks on VLE activities we predicted the student engagement. Here we considered only activities related feature and this feature represents the student participation while taking the course.

Missing Values. We found some missing values like nan values, empty values in OU dataset and we fill this values with zero. The zero value means the students have no activity on this section.

Predictors that Affect Student Engagement in Web-Based Systems

In e-learning systems, student activities and materials are the most important predictors to find out the student engagement on the course.

Student ID. This is the unique variable for a student on our whole dataset.

Highest Education Level. From this variable we get the information for highest education level for a student before getting the registration online course. This also have some effects of student online education systems.

Total Number of Clicks on VLE Activities. The variable means how many times student accessed VLE systems. As much as the value is more that means students are very engaged with this course, and decrease the value also sign of low-engagement with course. So, this is a very important predictor for student participation into the course [11].

Score on the Assessment. There are some assessment for testing the student. After first assessment of course this score on the assessment obtained by the students.

Final Results. After finishing the course, this variable represent the student final exam result and possible values are fail or pass. If the student passes the exam that means he/she is more active on this course. Some students also withdraw their course subscription at the middle of the course. So this is also one of the most important predictors for student engagement on the course [3].

Building and Testing the Predictive Model

After preprocessing data, now we have our feature Table. We trained our model using the student training data. We create an excel file from our training data (feature table).

Train and Test Data Splitted. Here we use our 75 percent data for our training and other 25 percent data for testing. Here we are using three machine learning model. First we are

trying using knn model, then svm model and decision tree model.

In the training process, we put the input matrix as an input data to model and the corresponding data classes to the Machine Learning classifier to find out patterns between the input and output. Finally the trained model find a pattern and used this pattern to classify others data. The average performance gain this method provides a good estimation of model performance.

Performance Metrics. After training the model, we evaluate the performance of the learning model using our other 25 percent unseen data. Then we get the prediction results for the models with test data. Here we calculate also True Positive(TR), True Negative (TN), False Positive (FP) and False Negative (FN) and used all this attributes value for evaluation the performance.

The main goal of this study to decrease the False Positive Rate (i.e the number of low-engagement students identified incorrectly as a high-engagement students) [15]. We used some performance metrics to calculate the quality of Machine Learning model predictions. And here are the following techniques,

Confusion Matrix. The confusion matrix holds the value of about predicted and actual class labels information which is generated by a predictive model. The data in the matrix are used to evaluate. In the confusion matrix data have the following meanings

Table 1: Confusion Matrix

		Predicted	
		Positive	*Negative*
Actual	*Positive*	TP	FN
	Negative	FP	TN

- True positive (TP): Number of correct positive predictions.
- False positive (FP): Number of incorrect positive predictions.
- True negative (TN): Number of correct negative predictions.
- False negative (FN): Number of incorrect negative predictions.

Accuracy Score. Accuracy score is very common evaluation metric for the classification models. It can be calculated as the number of correct predictions divided by the total number of predictions. If the accuracy is good that means it's going to predict most high-engagement of students with this course. The best accuracy level is 1 and wors value is 0.0 [6, 15] . The accuracy is measured by given below formula.

$$Accuracy = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN}$$

Recall. This is the number of correct positive predictions and the ratio of the total number of positives. This is also known by rate of true positive. The worst value is 0.0 and the best value is 1 [6, 15]. It can be formulated by

$$Recall = \frac{\sum TP}{\sum TP + \sum FN}$$

Precision. Precision is the number of correct positive predictions as a ratio of the total number of positive predictions. The best value of precision is 1.0 and the worst value is 0.0. It can be formulated by,

$$Precision = \frac{\sum TP}{\sum TP + \sum FP}$$

F1-Score. F1 score is the harmonic mean between recall and precision. The best F1 score value is 1.0 and the worst is 0.0. It can be formulated by

$$F1 - Score = \frac{Precision.Recall}{Precision + Recall}$$

AUC - ROC Curve. The Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) curve, we used both for showing the multi class classification model. This is very popular for classifying when a dataset is not balanced [2]. By using this we can measure the performance of classification model.

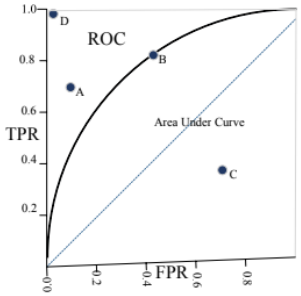

Figure 1: AUC-ROC Curve

- ROC Curve => Probability Curve.
- AUC Curve => Showing the measure of separability.

On this type of curve, the X-axis represents the False Positive Rate (FPR) and Y-axis represents the True Positive Rate (TPR). The AUC range is from 0 to 1. If the value is greater than 0.5, then the model considered as a good model [4].

AUC curve calculated by this formula ,

$$AUC = \frac{1}{2}(TPR + TNR)$$

The terms are used in AUC and ROC curve as follows.

$$TPR = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{TN + FP}$$

4 RESULT

To comprehend our data better, we analyzed the sum of clicks and student's engagement level. From this analysis, we summarized that the summation of click, an average of student assessment score and pass/fail has a connection to the level of student's engagement. We used Spearman correlation statistical analysis to find the relationship between our independent and dependent variables. Next, we used the final result, assessment pass/fail status and score we defined excellent, qualified and active student engagement. Based on this, we trained our training model to predict the engagement rate for a course as EN (Engagement) and NE(Not engagement). To get the engagement level, we defined that collection of all students engagement rate can be used as the course engagement rate. In order to train the model, we used three machine learning algorithms: K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree Model (Table 2). To use these algorithms, the entire dataset was divided in two parts: the train data (.75) and test data (.25). Then, to evaluate the model we used confusion Matrix , Recall, Precision, F1 Score, AUC - ROC Curve. Binary classification was used to differentiate between two engagement levels(Engagement = 1, Not Engagement = 0). We also calculated all assessment attendance and score of all registered students within the course.

Through our machine learning algorithms, we observed that the Support Vector Machine with a Kernel Linear performed the best among all three algorithms. The SVM gave us an accuracy of 0.94 and F1 score of 0.95. It needs to be noted that, another algorithm KNN also came very close to the SVM with an accuracy of 0.92 and F1 score of 0.93. Additionally, a measurement of the performance of the model using the AUC-ROC curve revealed how much the model is capable of distinguishing between classes. For our SVM model, the AUC is 0.9187 (Figure 7) where Decision also came close with 0.8969 (Figure 5). From this, we can summarize the SVM covers more area than other models.

Furthermore, we also measured the performance of the classifier model, where from the confusion matrix we average precision(Precision is calculated as the number of correct positive predictions divided by the total number of positive predictions) of the SVM is 0.95 (Figure 4) and Decision Tree is 0.93 (Figure 5).

Table 2: Result for machine learning task for predicting course engagement depending on student engagement

		KNN		SVM		Decision Tree	
		NE	*E*	*NE*	*E*	*NE*	*E*
Metric	*Precision*	0.71	0.84	0.96	0.94	0.86	0.96
	Recall	0.56	0.91	0.85	0.99	0.89	0.94
	F1-Score	0.63	0.87	0.90	0.96	0.87	0.95
	Support	27	69	27	69	27	69
Accuracy		0.81		0.94		0.92	
Evaluate Course Engagement		78.13		75.00		71.88	

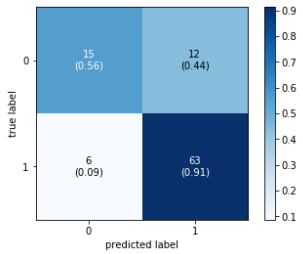

Figure 2: Knn Model Confusion Matrix

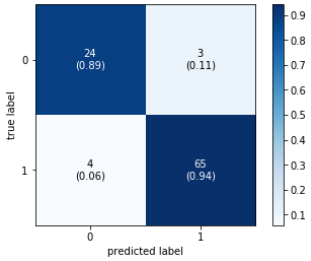

Figure 3: Decision Tree Model Confusion Matrix

5 DISCUSSION

One of the key goal of this project is to answer the learning questions which we defined before this project. The first two questions were:

- Can we use machine learning to predict student engagement with a given course?
- Which model is best suited to our project goal and chosen data set?

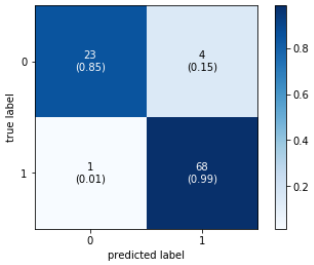

Figure 4: SVM Model Confusion Matrix

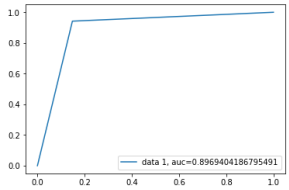

Figure 5: Decision tree AUC ROC Curve

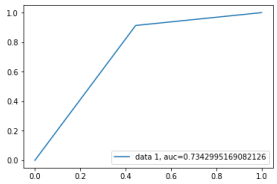

Figure 6: Knn AUC ROC Curve

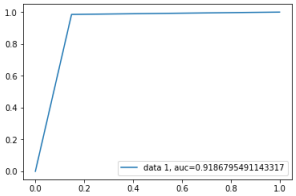

Figure 7: SVM AUC RoC Curve

In order to determine answer to these questions, we used three machine learning algorithms to develop the predictive model. We then evaluated the three models which includes KNN, SVM and Decision tree using multiple evaluation metrics. Form this evaluation, we derived that machine learning can indeed be used to predict a course engagement rate and the most suitable algorithm in this case would be the SVM.

Another key questions was: Does such prediction offer any valuable insight? To answer this question, we will discuss what are the some key outcomes of our project. First, the data-preprocessing and selection step illustrated a connection between the sum of clicks and student's performance such their assessment scores and final result. Here, we also have to consider that using data scientific approach towards education is not limited to student and their assessment but also to reveal insight about the course material as well. Our final course engagement level ranged from 78.13 to 71.88 for all three algorithms. This can be used by teachers to understand how is a course performing and what kind of intervention can be introduced to increase the engagement rate. The correlation between the input data can be used as indicators in this case.

However, there are some limitations and challenges in these insight. A student's engagement with learning materials alone can not be used to conclude a course's performance. It is often observed that, despite having low interaction with course material students perform better or the vice versa scenario. Hence, while a model such as ours can be a good tool to find insight, it can not be used as the only tool and further study needs to be conducted.

One of the hurdle we had to face in this project is selection and evaluation of data that we will use for our ML model. Since this data set has data in six different tables, we had to closely analyze to determine the feature inputs for our ML model. Next, during pre-processing we also realized that there were some missing data and data that needs to be reevaluated. If a student interact with the VLE multiple times a day the Sum of Click data for the same student are listed in different rows. These issues were resolved using variant data prepossessing techniques. Since our featured table data was extracted and merged from different tables, we also had to label the input features properly.

Another difficulties we had to face was to determine the correct method or formula that will yield the course engagement rate. In order to attain the result, we defined the engagement using certain indicators: the total number of times student interacted with the VLE which means the *Sum of Clicks*, student's assessment scores for two assessments, student's previous education degree, and the final exam result. To shape our data to fit into this definition we used Spearman's Correlation method to determine the relationship between the indicator as specified.

With this particular data set we observed that the same trained data can be used to define different scenarios in different context. The course engagement rate can be used in the premise of identifying a connection between student's engagement with course and the impact on their final result. On the other hand, we defined the course engagement rate as an insight to help with evaluation of a course and its material.

One of the key approach that failed for us was the attempt to use too much data as our feature input. This eventually lead us to a point where there were too much data to process and comprehend with no practical relevance to our

desired outcome. Moreover, lack of proper visualization of our findings restricted us from showcasing the potential of this predictive model to our audience.

The key takeaway from these problems and the entire journey of developing this predictive model is the realization of the power and complexity of big data. The most highlighted point would definitely be the point where we were able to process our data and establish a relationship between the indicators.

6 CONCLUSION

Predictive models used to forecast important aspects of an educational system. In case of online learning institutes like the Open University UK, it is rather crucial to balance the lack of direct communication or opportunity to observe and learn. There, a predictive model that produces significant insight can be used by educational instructor and administration for many use cases to improve their teaching material and student performances. For our project, we used data from the Open University Learning Analytic Dataset to collect our input feature and train a model using three machine learning algorithms: K-nearest neighbour, Support Vector Machine and Decision Tree. Form the result of these models and their evaluation metrics, we concluded that predicting a courses student engagement rate can play vital role in assessment of course material and study relationship between student performance and course assessment. Further study to overcome the limitations of this project can help to identify the intervention required improve a course engagement rate, and in return improve student academic performance.

REFERENCES

[1] decision-trees. https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb. Accessed: 16-07-2019.

[2] Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A., and Goodrich, V. Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (2014), ACM, pp. 103–112.

[3] Beer, C. Online student engagement: New measures for new methods. *Unpublished Masters dissertation, CQUniversity, Rockhampton, Qld, Australia* (2010).

[4] Corrigan, O., Smeaton, A. F., Glynn, M., and Smyth, S. Using educational analytics to improve test performance. In *Design for Teaching and Learning in a Networked World*. Springer, 2015, pp. 42–55.

[5] Holmes, N. Engaging with assessment: Increasing student engagement through continuous assessment. *Active Learning in Higher Education 19*, 1 (2018), 23–34.

[6] Hussain, M., Zhu, W., Zhang, W., Abidi, S. M. R., and Ali, S. Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review 52*, 1 (2019), 381–407.

[7] Kumari, P. M., Nabi, S. A., and Priyanka, P. Educational data mining and its role in educational field. *International Journal of Computer Science and Information Technologies (IJCSIT) 5*, 2 (2014), 2458–2461.

[8] Kuzilek, J., Hlosta, M., and Zdrahal, Z. Open university learning analytics dataset. *Scientific data 4* (2017), 170171.

[9] LÃşpez-Cuadrado, J., PÃ‰rez, T., Vadillo, J., and Arruabarrena, R. Integrating adaptive testing in an educational system. pp. 133–149.

[10] Narkhede, S. Understanding AUC - ROC Curve. https://towardsdatascience.com/understanding-auc-roc-curve68b2303cc9c5. Accessed: 2019-06-23.

[11] Prince, M. Does active learning work? a review of the research. *Journal of engineering education 93*, 3 (2004), 223–231.

[12] Puget, J. F. What is machine learning?

[13] Romero, C., and Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications 33*, 1 (2007), 135–146.

[14] Romero, C., Ventura, S., Pechenizkiy, M., and Baker, R. S. *Handbook of educational data mining*. CRC press, 2010.

[15] Rovira, S., Puertas, E., and Igual, L. Data-driven system to predict academic grades and dropout. *PLoS one 12*, 2 (2017), e0171207.

[16] Tetsuya, J. Why is Educational Data Mining important in the research? https://towardsdatascience.com/why-is-educational-data-mining-important-in-the-research-e78ed1a17908. Accessed: 2019-05-29.