CS-411 Digital Education and Learning Analytics

Exploratory Data Analysis on the MOOCs Dataset

Riyadh Alnasser and Deepak Karkala
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE EPFL

1.Introduction

The Internet has revolutionized Learning and teaching in the last decade, large numbers of different course are available to the entire world online. The learning process produces a huge amount of data which can reveal a lot about how people engage and learn in online courses. In this short report we analyze a dataset from two online courses and answer the following questions:

- How can we identify student's engagement in the two online courses?
- Find different group of students based on their activity in the course?
- Which activities correspond to high chance of success?
- What type of learning activity is crucial for student's performance in online courses?

2. Data Description

The dataset consist of 3 weeks worth of MOOCs data for both Java and C++ courses provided by ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE (EPFL). The dataset contains data regarding students' performance in both classes. Each student has anonymized ID as well as some of variables that represent different activities that students may engage in while taking the course. These activities can be for instance, viewing the lectures, assignment submission, viewing the forum, launching a thread as well as posting on a thread. The data also contains information related to students grade, their achievement level as well as the countries they come from.

3. Exploratory data analysis

3.1 Defining an Engagement Index

In this section, we examine two different ways of identifying students engagement index. The First way of defining the engagement index comes from what we have experienced from taking online courses. We believe that in order to be engaged in online class, it is important to watch the lectures video as well as participate in solving the assignments. So we define our first engagement index to be the sum of the activities related to watching lectures video and the activities related to assignment submission. In the Figure1 (a) below, Java and C++ students' engagement index across the three weeks has been represented. Each point represents the engagement index in a week for Java and C++ students. We observe the highest drop in student engagement for C++ students in the second week.

Although the difference in the weekly engagement index between Java and C++ is subtle, using the statistical significance test, t.test() function in R, we find the difference

to be statistically significant, p-value=0.0001892 for the first week, p-value=2.103e-09 for the second week and p-value=3.551e-05 for the third week.

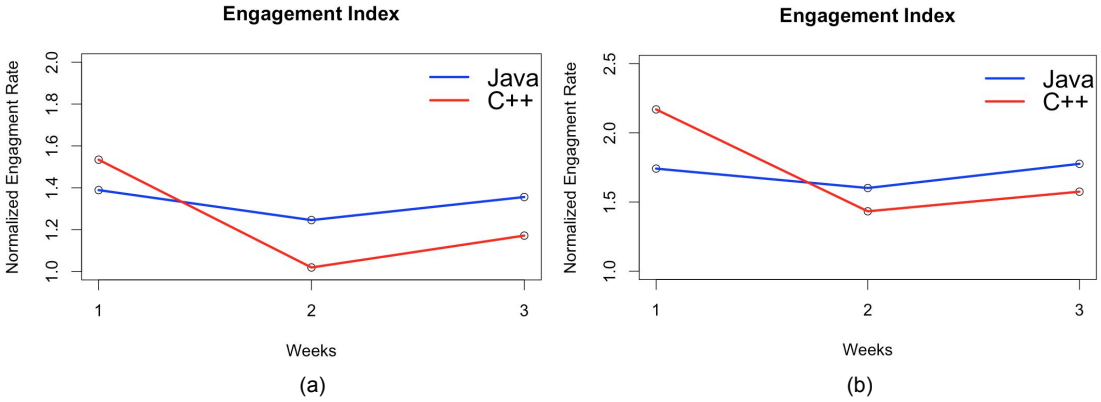


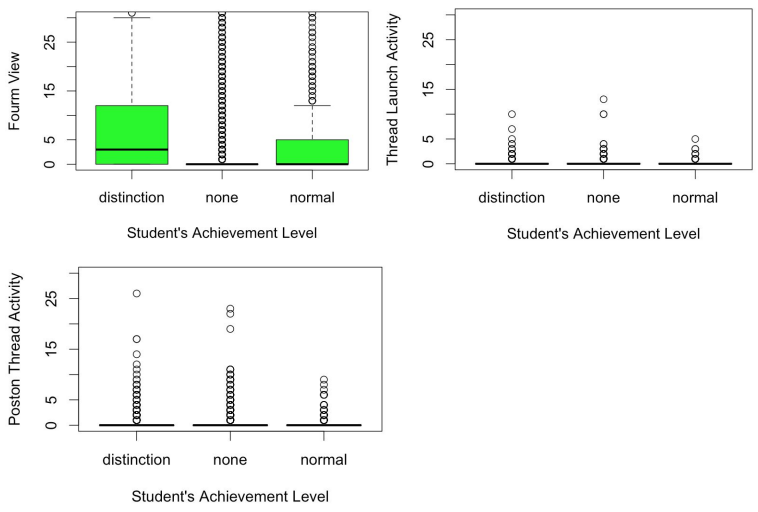
Figure 1: Variation of Engagement index (a) First Index (=Lecture views, reviews + Assignment submissions, resubmissions), (b) Second Index (=First Index + Forum Activities)

The second way of defining the engagement index is to consider the different forum activities in addition to the lecture views and assignment submission. In the second engagement index we assume that all course related activities are equally important and we calculate student engagement index by taking the sum of all the course related activities. The result is represented in Figure1(b). We observe the same pattern of the weekly change in the student engagement for both courses, and similarly we observe the highest drop in student engagement to be for C++ students in the second week.

The difference in the weekly index between Java and C++ is subtle here as well, and using the statistical significance test, t.test() function in R, we find the difference to be statistically significant, p-value=1.576e-09 for the first week, p-value=0.01743 for the second week and p-value=0.01219 for the third week.

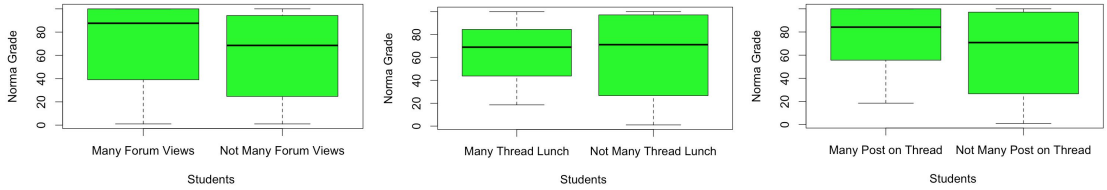
3.2 Forum activities

In order to investigate what type of forums activities in beneficial for students performance in the course, we looked at the achievement level of the students and then we reflected back on what type of forum activities did those students engaged in. Our results show that students who got a distinction in the achievement level have a higher mean of Forum View, as can be seen in Figure2(a). However in the other two forum activities (Thread Lunch & Post on Thread) student with different achievement level have similar behavior in engaging in these forum activities, Figure2(b),(c).



(a) (b) (c)

Figure 2: Effects of forum activities on the student's' achievement level (a) Forum view, (b) Thread Launch, (c) Post on Thread.



(a) (b) (c)

Figure 3: Effects of forum activities on the student's' normal grade. (a)Forum view, (b)Thread Launch, (c)Post on Thread.

To test this result, we use the correlation test function in R *cor.test()* to test the correlation between the normal grade and every type of forum activities. Our results suggests that there is a higher correlation between forum activity and the normal grade cor = 0.219, as opposed to other two forum activities (Thread Lunch & Post on Thread) where we observed a relatively lower correlation with the normal grade, cor = 0.148 and cor = 0.112 respectively.

To further investigate this, we have divided the students into two groups, the first group contains the students with many forum views (>10), and the other group contains students with not many forum views(<10). We have applied the same thing on the other two forums activities (the thread lunch and post on thread). See Figure3.

By comparing the two groups in forum view, we find that students who view the forums many times have a higher normal grade mean than the student who did not view the forums a lot, 70.9 and 59.6 respectively, Figure3(a). And By applying a statistical significance test, t.test() function in R, we find the difference to be statistically significant p-value<2.2e-16.

However, by looking at the thread lunch and post on thread, there is a very small difference between the two groups (the students who lunch and post on a lot of threads and the student who did not participate a lot on launching and posting on threads),

Figure3 (b),(c), and it is not statistically significance, p-value=0.9844 for the thread lunch and p-value=0.1623 for post on thread.

From this result, we can conclude that forum viewing is the most beneficial activity among different forum activities for student performance in these courses.

3.3 Unsupervised learning to analyze MOOC data

In this section, we use unsupervised learning to analyze the data for the two courses. The aim of this analysis is to try to find a hidden structure in the data and hopefully identify a set of activities which contributes to higher chance of success for students in the courses. We use principal component analysis to reduce the dimensions of the data and k-means clustering to obtain clusters. We then analyze the clusters and give appropriate names and also use statistical tests to verify the significance of the results.

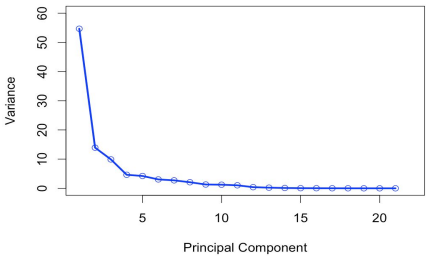
3.3.1 Principal Component Analysis

As described in section 2, the data-set for the two courses, includes a large number of features/dimensions and in-order to form clusters, we would like to use smaller subspace. In order to achieve this, we use Principal Component Analysis or in short PCA. PCA is used to create linear combinations of the original data that capture as much information in the original data as possible.
Using R, we obtained the principal components using prcomp(data) We then computed variance explained by the components as varexp=cumsum(PCA$sdev^2) / sum(PCA$sdev^2).

From the values of the principal components analysis in the summary in Figure 4 (a) and from the plot in Figure 4 (b), we can conclude that retaining 6 components would give us enough information, as we can see that the first 6 principal components account over 90% of the variation in the original MOOC data.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     7.3942 3.7333 3.14757 2.15276 2.05723 1.74679 1.65195
Proportion of Variance 0.5484 0.1398 0.09938 0.04649 0.04245 0.03061 0.02737
Cumulative Proportion  0.5484 0.6882 0.78760 0.83409 0.87654 0.90715 0.93452
                          PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     1.44245 1.14504 1.11419 1.01394 0.63085 0.47715 0.34289
Proportion of Variance 0.02087 0.01315 0.01245 0.01031 0.00399 0.00228 0.00118
Cumulative Proportion  0.95539 0.96854 0.98100 0.99131 0.99530 0.99758 0.99876
                          PC15    PC16   PC17    PC18    PC19    PC20    PC21
Standard deviation     0.2228 0.18872 0.1404 0.08902 0.06160 0.06020 0.05512
Proportion of Variance 0.0005 0.00036 0.0002 0.00008 0.00004 0.00004 0.00003
Cumulative Proportion  0.9993 0.99962 0.9998 0.99990 0.99993 0.99997 1.00000
```

(a)



(b)

Figure 4: Principal components of MOOC data. (a)Summary of the PCA, (b) Plot of Variance.

3.3.2 K-means clustering

Clustering is the process of partitioning a group of data points into a small number of clusters. Using the principal components generated in previous section, we wanted to divide the students into clusters such that students in each cluster share similar features/characteristics. In R we use kmeans() function to form clusters, while k-means

in unsupervised, we need to give to the algorithm how many clusters we expect to find in the data. In our case here we choose four clusters k=4. Figure 5 shows the clusters.

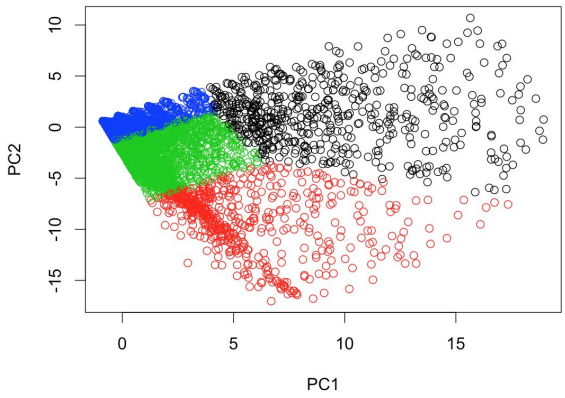
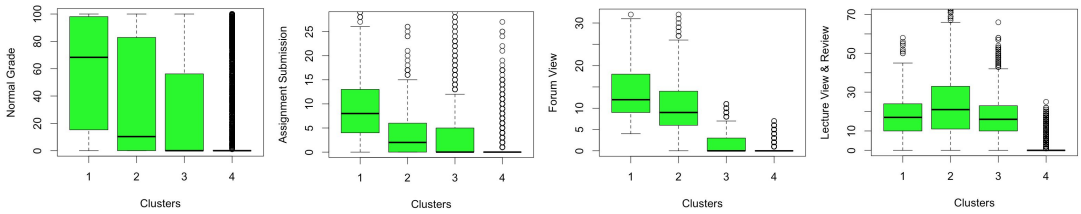


Figure 5: K-means clusters of students in two courses.

3.3.3 Analyzing clusters

By analyzing the behavior of different students belonging to different clusters, we observe that students belonging to a cluster share the same characteristics. In this section, we analyze the data visually using boxplots and in the next section we use statistical tests to verify if the results are significant. The following box-plots shows the distribution of normal grade, assignment submission, forum views, and lecture views for students belonging to different clusters.



(a) (b) (c) (d)

Figure 6: Boxplots of characteristics for students belonging to different clusters (a) Normal grade (b) Assignment submission (c) Forum view (d) Lecture views and reviews

By analyzing the normal grade of students belonging to different clusters, we find that each cluster has different mean of the normal grade, and also by looking at the other activities such as assignment submission, forum view, and lecture views we find that each cluster has different mean of these activities.

3.3.4 Observations from visual inspection

Assigning names to clusters

Based on the above observations, we can assign appropriate names to each cluster. It is evident that the number of assignment submissions is very high among students belonging to cluster 1 whereas the number of lecture views is higher for students in cluster 2. Looking at students belonging to cluster 3, we find a relatively low participation in the course related activities. Finally students belonging to cluster 4 almost didn't participate in the course related activities.

From this observation, we can assign names to these clusters, we give the Assignment Submitter & Forum Viewers to Cluster 1, Lecture Viewers to Cluster 2, Not Very Active Students to Cluster 3, and Non-participants to Cluster 4.

Factors that contribute to higher chance of success

Based on the above box-plots, it can be seen that students belonging to cluster 1 possess higher normal grade. It can be noted that the number of assignment submissions and forum views are also high for students in cluster 1. Thus from visual inspection, it can be concluded that the **assignments submissions** and **forum views** significantly increases the chance of success for students. In the next section we perform statistical tests to confirm if these results are significant.

3.3.5 Statistical analysis to verify if result is significant

The boxplot in Figure6 (a) shows that means of normal grade in the four clusters are different, and the same for the other variables in Figure6 (b),(c),(d). To make sure that these differences in means are not due to the chance, we apply ANOVA test by using aov(Variable ~ Cluster) function in R. We have done the test for all variables: normal grade, Assignment Submission, Forum View, and Lecture View. Our results highlight that the differences of the means of these valuables with the different clusters are statistically significant. (p-value< 2e-16) for all variables.

CS-411 Digital Education and Learning Analytics

Predictive Modeling on the MOOCs Dataset

Riyadh Alnasser and Deepak Karkala
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE EPFL

1. Introduction

This works aims at predicting student's chance of success based on activities in the first few weeks of the course. The methods used include selecting a subset of features which correlates with the final grade, choosing the appropriate regression model and fine-tuning the model parameters in order to achieve higher prediction accuracy.

2. Methodology

This section describes the techniques used to pre-process the data and methodology used in selecting a subset of features which helps in improving prediction accuracy.

2.1 Feature extraction

The given data set consists of sequence of events triggered by the students participating in the class. In order to derive features useful in predicting the success, the following methodologies were used.

1. A starting time was defined and we removed all the events that occurred prior to this time.
2. For each student, we aggregated all the events (occurring across the weeks) belonging to a given event type.

Thus by combining similar events across the weeks, the number of features was greatly reduced and the next goal was to select a subset of these features useful in predicting the students' success.

2.2 Feature selection and adding new features

In order to predict the success of students, it is preferable to have a subset of features which helps in predicting students' success more accurately. The number of features were reduced using two methods.

1. Preliminary visual examination of the data revealed that the students who passed have watched more videos and checked problems regularly.
2. The t-values obtained using stepAIC also showed that the same variables observed in visual inspection had high correlation with the grade.

Based on these observations, the following features were selected: *"Video.Load", "Problem.Check","Forum.Thread.View","Video.Pause","Video.Play","Video.SpeedChange"*

In addition to this, the following new features were added:

1. TotalProblemsCheck = Problem.Check_1 + Problem.Check_2 + Problem.Check_3 + Problem.Check_4

2. TotalVideoEvents = Video.Load + Video.Pause + Video.Play + Video.SpeedChange
3. EngagementIndex = TotalProblemsCheck + TotalVideoEvents + Forum.Thread.View
4. We also used interactions between the following variables:
    a. Problem.ThreadView = TotalProblemsCheck * Forum.Thread.View
    b. Problem.Video = TotalProblemsCheck * TotalVideoEvents
    c. Video.ThreadView = TotalVideoEvents * Forum.Thread.View

2.3 Dimensionality reduction

Since we used only a small subset of features, we observed that using PCA did not improve the prediction accuracy and hence we decided not to use PCA.

2.4 Data Normalization

The variance of different events in the given data varied significantly. In order to avoid the bias to few variables with higher variance, we normalized the data using the *center* and *scale* options in the *preProcess* function in 'R'.

2.5 Outlier removal

The presence of outliers will significantly reduce the flexibility of the model, since the model tries to fit itself to these outlier data points and hence we decided to remove outliers.

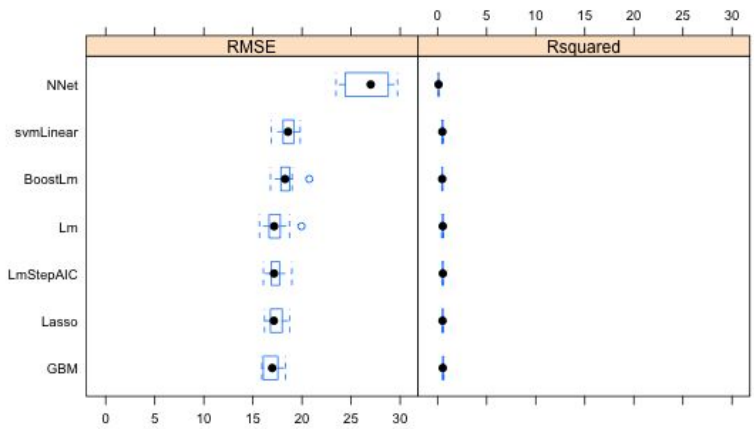
2.6 Imbalanced classes

We observed that the given dataset consisted of very few students with higher grades. This can result in bias during model fitting. In order to have similar number of students who had passed and failed, we used the function *upsample* in 'R'.

3. Model fitting and comparisons of models

3.1 Model selection and Comparison of models

We tried to fit several regression models and the models were evaluated based on the metric *RMSE*. The fitted models were compared using the function *resamps* in 'R'.



It is evident from the figure that the models with least *RMSE* were

1. Linear Regression with Stepwise Selection [*lmStepAIC*]
2. The Lasso [*lasso*]
3. Generalized boosted Models [*gbm*]

Thus we decided to use these three models. The final predictions were obtained by averaging the predictions of the above models. It can be noted that the high RMSE for neural network method could possibly due to suboptimal tuning of the model.

3.2 Parameter tuning

The following options were used in the function *trainControl* for fitting models.

1. Resampling method: Repeated Cross-validation
2. Number of folds in K-folds cross-validation: 10
3. Number of separate K-fold cross-validation: 5

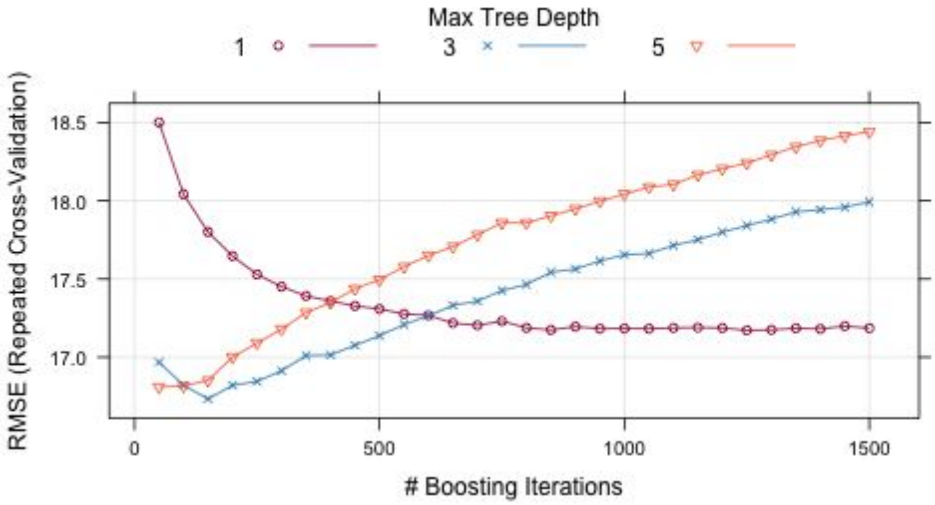
3.2.1 Tuning Lasso

Initially the following range was used-Fraction of full solution: (0.3, 0.5, 0.7, 0.9, 0.95, 0.98)
The least RMSE(17.23) was obtained with fraction = 0.95 and was used in the final model.

3.2.2 Tuning Generalized boosted model

We started by using the following range of values to tune Generalized boosted model.

1. Number of trees: (1:30)*50
2. Complexity of the tree: (1 3 5)
3. Learning rate: 0.1
4. Minimum number of training set samples in a node to commence splitting: 20



The above figure shows the RMSE values for different values of boosting iterations and tree depth. By observing, the RMSE values, we iteratively reduced the range of values and the following values were used in the final model (RMSE:16.76) used for prediction.

1. Number of trees: **150**
2. Complexity of the tree: **3**

4. Conclusion

With the above options used, our final model had a prediction accuracy of 85.42% on the test data. Based on this, we can conclude that the students with higher grades tend to watch more videos, solve more problems and have a higher forum activity. Thus the instructor can monitor these activities in order to predict a student's chances of success. Further we also learned the following two things regarding prediction in data science in general. The first one is the fact that choosing the relevant subset of features improved the prediction accuracy significantly compared to that by selecting different models. Also simpler methods tended to have better prediction accuracy than the complex ones.