Master’s Project Prospectus

**Predicting Online Learning Outcomes With Clickstreams**

**by Applying Exploratory Factor Analysis (EFA) Model**

Department of Educational Studies

Workforce Development And Education

Qiwei Men

March 20, 2020

**Problem Statement**

As the development of the internet and information technologies, online education has become not only a useful complement, but also a powerful competitor to the traditional education. In the recent 10 years, the model of massive open online courses (MOOCs) brought a revolution to the educational business. MOOCs platforms like Coursera, Edx and Udemy have changed people’s cognition and expectations of online learning. As what distance education used to promised, MOOCs is making education more accessible, flexible and affordable for potential learners. Moreover, it is also changing people’s choice and habit of learning behavior. For example, as of February 2019, Coursera had 45 million registered users signed up for its programs, and offered more than 2,000 online courses. According to a [survey](http://www.onlinelearningsurvey.com/reports/gradechange.pdf) by the Babson Survey Research Group, about 33 percent of college students are taking at least one course online.

Before the age of online learning, getting feedback from students and mastering the overall situation of massive students in distance is nearly impossible. However, with the help of new technology and big data, we can record, store and analyze the actions of learners and find out how this information related to their learning performance. In this project, I will use the OULAD dataset from the Open University and the method of Exploratory Factor Analysis (EFA) to detect the relationship between learner’s click actions and their learning performances. Then based on the results, provide some suggestions to learners, teachers and course designers, for elevating the quality of education and the performance of learners.

Though online courses has evolved into a more advanced version as the technology development, there are still some unsolved problems that are critical to the future development of online education. In this project, I will focus on two questions:

1.     Online courses usually have much more enrolled students than face-to-face courses, tracing the learning state and performance of each student is more difficult than that of the physical classroom. Can we use learners' clickstream to predict their final outcome (fail, pass or distinction)?

2.     Without the limit of time and space, there’s massive contents can be posted online, understanding how different kinds of contents influence the learner's performance will be helpful for optimizing the course design.

The purpose of this project is establishing an effective model for learning state tracing and performance prediction base on the clickstream made by learners. This subject is significant for the development of online education, since for a long time, the quality of online courses was thought to be questionable, the diplomas and certificates obtained online are not highly recognized by the public. Generally, learners do not treat virtual courses as seriously as physical ones, if we can address the properties of clickstream patterns leads to different outcomes and identify students at risk, we can make necessary interventions for avoiding potential fail. In addition, if we know some click actions have significant influence to the final outcome, we may find out what components of the course are more important and determine how to organize the contents and assignments of the course.

This research also provides an opportunity to verify the classical theory of distance education like the theory of transactional distance ( Michael Grahame Moore,1993), the interaction equivalency theorem (Terry Anderson,2003), and the community of inquiry models(Garrison, Anderson & Archer, 2000). These theories all suggested some models for learning and interaction pattern in the virtual environment, we will check whether the results implied by the data are consistent or divergent with the theoretical model.

**Literature Review**

The clickstream data analysis has been widely used in industry and business areas, companies like amazon use clickstream to learn customers’ preference, and use this information to predict their purchase behavior, clickstream analysis is useful in adjusting the promotion strategy and improving the customers’ experience. Data analysts believe that clickstream data reflect the user’s behavior pattern, which can also be used to predict their decision and outcome. As the development of MOOCs, the approach of clickstream data analysis has also been introduced to the educational studies, developed into a discipline named educational data mining (EDM), which concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using these methods to better understand students, and the settings which they learn in.

student clickstream data has been the subject of a number of prior studies, such as the investigation of potential predictive relationships between online student activity and student outcomes (Jihyun Park et al., 2017)[1]. Some of these analyses have focused on using the clickstream data to predict MOOC completion, and drop rates. Since one major problem with MOOCs is that they have extremely high rates of dropouts. Gutl et al. (2014)[2] ompared dropout rates between traditional face-to-face courses and online courses and found that the drop rate of MOOCs can be 10 or 20 percent higher than traditional classes. Besides, another popular topic is making prediction about the outcomes of students who have completed the course, this is also the topic we want to investigate in this project. The main purpose of performance prediction is “identifying students at risk of failing a course with a view to performing interventions with the aim of encouraging students to work harder, and have a greater chance of passing the course.”( Pedro J et al., 2019) [3].

Learning analysts also studies different components of the clickstream data, for example, Christopher G. Brinton[4] conduct a research focus on the video-watching behavior of MOOC students, where users spend the majority of their time learning. Gábor Kőrösi[5] use data of mouse behaviours (move, scroll, click) to make prediction. Niki Gitinabard et l.[6] focus on data related to social activity in the forum. Robinson et al. [7] and Chen et al. [8] also included the number of posts as a feature in their studies. In this review, I will compare three different prediction models and discuss their advantages and limitations.

Naif et al. (2019)[9] deployed a long short-term memory(LSTM ) model using Open University Learning Analytics dataset to predict the performance (pass or fail) of students with around 90% accuracy by analyzing the clickstream data generated in the first 10 weeks of student interaction in a virtual learning environment (VLE). This study makes the prediction by converting the problem into a sequential weekly format and measuring the effectiveness of the deep sequential model versus the conventional machine learning baseline models.

The method of LSTM tended to monitor the sequential week-wise pattern of students and their activities, other than deal with students’ interactions in a collective and aggregated manner, which is effective for sequential data; the advantage of this model is it captured the learning behavior of students and efficiently predicted the at-risk students on the basis of their interactions and engagement patterns. However, there are also some limitations. Firstly, this model only used interaction data, but there are also other behaviors hidden in the clickstream, for example, how many clicks the learner made on certain contents (lectures, quiz or forum), what their overall activity and engagement level, this information may be helpful for making prediction and understanding how different factors are correlated. Secondly, this model only predict the outcome fail or pass, however, there are four outcomes : withdraw, fail, pass and distinct, the authors don’t interest in the dropout rate in this article, so omitting "withdraw" is reasonable, but the difference between pass and distinct students maybe matter, since our interest not only limited to saving potential fail students to pass, but also promote mediate students to a higher level, in another word, the further goal should be more than just reducing the fail rate but also improving the excellence rate. However, no corresponding contribution was made in this work.

In summary, the method of LSTM is effective in predicting the at-risk students by analyzing the interaction data, however, the model does not include other information in the dataset and cannot reflect the final outcome as a result of the combination of multiple factors, the accuracy of the prediction can be improved by including more variables. In addition, this work cannot discriminate pass and distinct students, it also did not provide any information about how to increase the excellence rate, which is critical for improving the quality of the course. Hence, more jobs can be done in the further study.

Ren, Rangwala, and Johri (2016) [10] used the personalized multiple linear regression (PLMR) model to analyze the data from Edx and predict the scores that a student may achieve on a given grade-related assessment based on information, considered as prior performance or prior activity in the course. They also track the participation of a student within MOOC (via click-stream server logs) . The study also found that features associated with engagement (logging multiple times), studying materials (viewing videos and attempting quizzes) were important along with prior homework scores for this prediction problem.

In this study, Ren et al. use features related to session, quiz, homework, video and time to capture a student’s studying behavior and learning habits, the evaluation shows improved performance in terms of prediction of real time homework scores compared to baseline methods. The advantage of this study is it provides an effective method for tracking and predicting the performance, this is helpful for the teaching team to learn the state of students in every short period and offer necessary association and help. The limitation of this research is that it only focuses on predicting the short term outcomes, more specifically, the homework grade, while in most MOOCs, homework grades only make up a small percentage of the final outcome, the author didn’t give a further prediction for the final result, which is a big concern in this subject.

In summary, Ren et al.’s work identify the most important factors which may influence the student’s performance in homework, the model they built offer an effective method for tracking and predicting student’s short-term performance, but didn’t give a further investigation on the long-term outcome. That limited the application scope of the model.

In another study, Yang et al. (2017)[11] used the method of time series neural network to analyze the lecture video-watching clickstreams and predict students’ average CFA grade in two MOOCs based on only previous assessment. The accuracy of the prediction outperforms a baseline of average past performance by more than 60% on average. The study also shows a result that when taken alone, none of the behavioral features are particularly correlated with performance, which suggests the importance of taking the learning performance as a result of the combination of multiple factors. They also gave some suggestions to course instructors about how to use this predictive model to stage student interventions.

In this study, the authors used neural network prediction models for personalized prediction. They firstly only used the past quiz data to make the prediction, then move to a developed model by adding the clickstream data. The result show that the involvement of clickstream data improve the prediction accuracy significantly. The limitations of this study is that in MOOCs, plenty of students never do quiz, which makes the model largely depend on the previous quiz performance inapplicable or questionable. Moreover, neural network is a complicated model, for implementing the prediction, a sophisticated pre-processing needs to be employed to handle the sparsity of available data, the process and the result are also difficult for understanding and interpretation.

In summary, Yang et al.’s work is effective in making prediction based on the existing quiz performance and clickstream data but not applicable for students who do not take quiz a lot. Besides, the method of neural network is also more complicate than previous two.

After reviewing and comparing all these prediction models for MOOCs performance, we can see that they all have satisfied accuracy and effectiveness for the scheduled purpose. In fact, there’s no perfect model but only proper model for prediction, the choice of the model should depend on the research interest and the situation of data, not all types of variables are useful in all contexts. Therefore, the listed limitation does not mean there’s any default in these studies, but only further study suggestions. The purpose of this review is offering a general understanding about questions like which indicators are more frequent in the literature, what problems have been addressed and which ones offer new possibilities for research, and what predictive models achieve accurate results for each context. This information will help us to find a proper model for our data and interested questions. In 2019, Pedro J et al.[3] published the article “Prediction in MOOCs: A Review and Future Research Directions”, which is the most comprehensive review in this area.

**Dataset**

The Open University Learning Analytics dataset (OULAD) dataset contains data from courses presented at the Open University (OU). It contains student demographics, clickstream history, and assessment submission information about 32,593 students over a course duration of 9 months, from 2014 to 2015 [12]. The data were composed of several courses, with each course being taught at different intervals in a year. Four distinct performance classes were defined: distinction, pass, fail, and withdrawal. The OULAD comprised students’ information regarding their interaction with the VLE—their assessments, quizzes, and course performances. The interaction with the VLE was categorized into different activity types with each activity referring to a specific action, such as downloading or viewing lectures, course content, or quizzes. The names of each of these activity types are as follows: dataplus, forumng, homepage, subpage, oucontent, resource, ouwiki, quiz, and url. The aggregated average clicks per student were processed weekly to visualize the students’ weekly interactions.

**Data Cleaning**

In this project, we only use the data of modules F which is a STEM course with 4 presentations in two years, we select the presentation in spring 2013. In order to make the result interpretable, we only select click data that can represent specific behaviors and intentions as variable, in addition, we calculate variables like active days and viewed contents from the original data. Since we just interest in the performance of students who had completed the course, we also ignore the data of students who had dropouts. After the pre-processing, the dataset includes 8 variables and 1538 observations, which are:

·       active\_days - How many days the students make at least 1 clicks.

·       viewed\_content - How many contents the student has viewed, this course has 370 contents online.

·       oucontent - videos and lectures produced by the Open University teaching team.

·       resource - other contents

·       forumng - forum for students to asking questions, making discussions and other interactions.

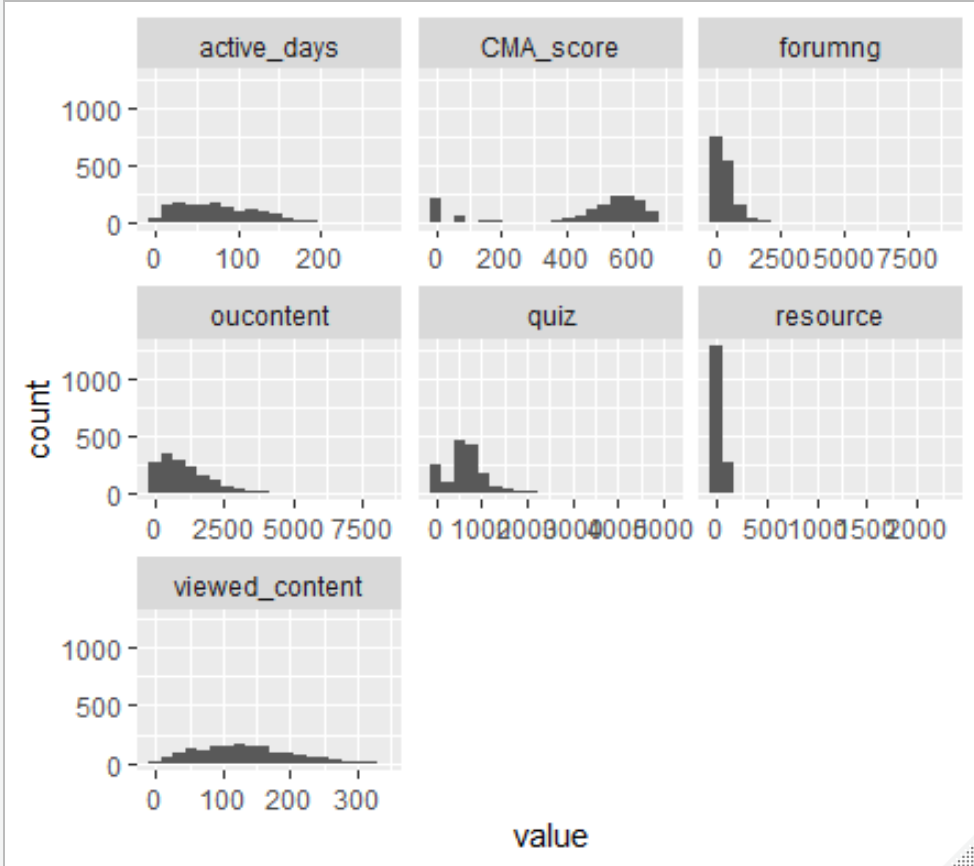
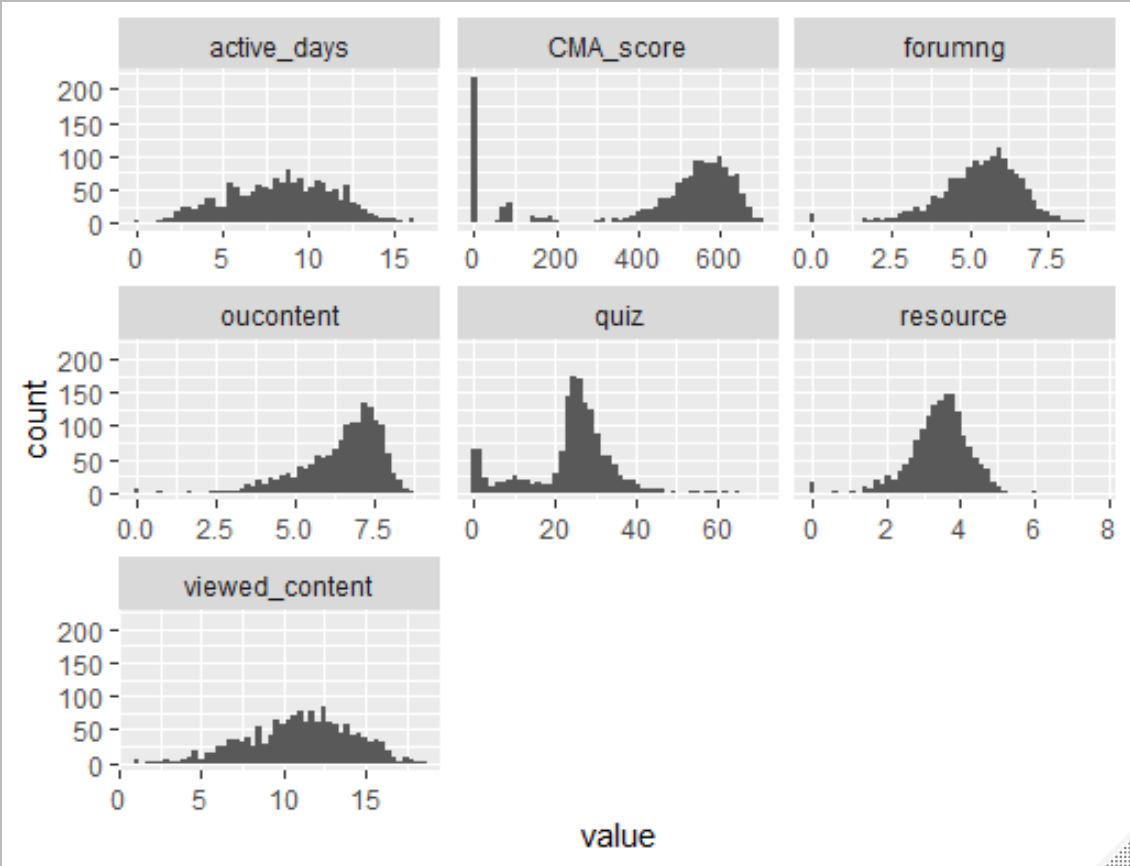
·       quiz - quiz related to the content. • Resource - other learning resources like articles and videos not created by OU.

·       CMA\_score - score earned in the quiz, CMA\_score weights 5-10% in the final result. The score was given by a computer.

·       final\_result- 4 levels withdraw, fail, pass and distinction.

**Data transformation**

The plots show that some variables are not normally distributed, in order to satisfy the assumptions of multivariate analysis, variables have to be transformed, From the suggestion of BoxCOX transformation , we decide to conduct square root transformation to active\_days,viewed\_content and quiz. conduct log(x+1) transformation to forumng, oucontent and resource.

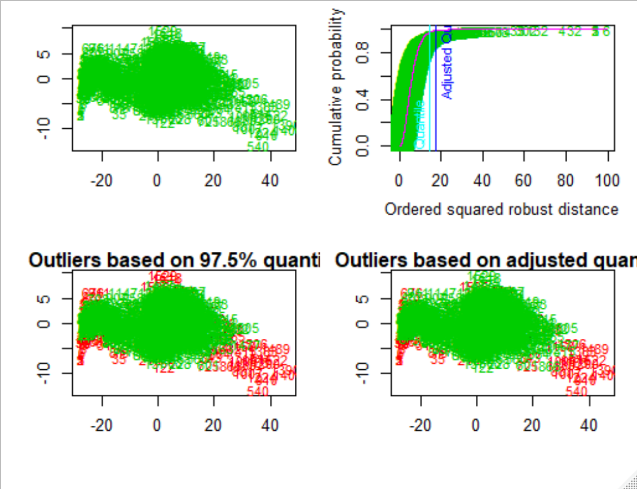
 

**Outlier detecting**

The method of multivariate analysis need variables be normal or approximately normal distributed, after the transformation, we need to detect and remove outliers to improve the multivariate normality of the data. Firstly, we use package mvoutlier to detect outliers, then remove the outliers from the data,

|  |  |  |
| --- | --- | --- |
| distinction | failed | pass |
| 181 | 397 | 874 |

|  |  |  |
| --- | --- | --- |
| distinction | failed | pass |
| 187 | 443 | 908 |



**Before**

**After**

In the outlier detecting procedure, we use quan = 0.8 and alpha = 0.05, which give a high tolerance for outliers. The purpose is keeping as much information as possible in the data. After we remove outliers, There’s still 1412 observations, the percent of outliers that been removed from the data is 0.057, we believe that such a minor reduction will not influence the result of the analysis.

**Methodology**

Exploratory factor analysis (EFA) is a technique that is used to reduce a large number of variables into fewer numbers of factors. The theory behind factor analytic methods is that the information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. In matrix notion.

**X - µ = L F +** **ε**

(p \*1) (p×m) (m×1）(p×1)

Where X is the observable random vector, µ is the mean, L is the matrix of factor loadings, F is common factors, ε is the error term. This technique extract maximum common variance from all variables and puts them into a common score. As an index of all variables, we can use this score for further analysis.

Factor analysis is part of general linear model (GLM) and this method also assumes several assumptions: variables should be continuous, there is linear relationship, there is no multicollinearity, and there is true correlation between variables and factors. In general, linear FA does not require normality of the input data. Moderately skewed distributions are acceptable. Several methods are available, in this project , we use orthogonal factor model (PCA) analysis and maximum likelihood method (ML).

**Orthogonal Factor Model**

The orthogonal factor model is based on the method of the principal component (PCA). It starts extracting the maximum variance and puts them into the first factor. After that, it removes that variance explained by the first factors and then starts extracting maximum variance for the second factor. This process goes to the last factor.

One critical question in EFA is selecting appropriate number of factors, by conducting PCA, we can see that the first three principal components can explain over 90% variance of the data, which means that these three factors can capture most of information about students' behavior and the difference in performance between different students. In this project, we will use both of these two methods to investigate the relationship between different factors and how they influence the outcome.

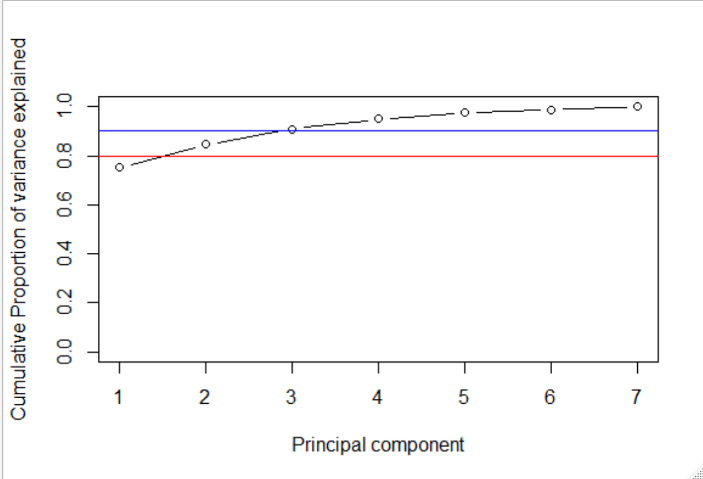
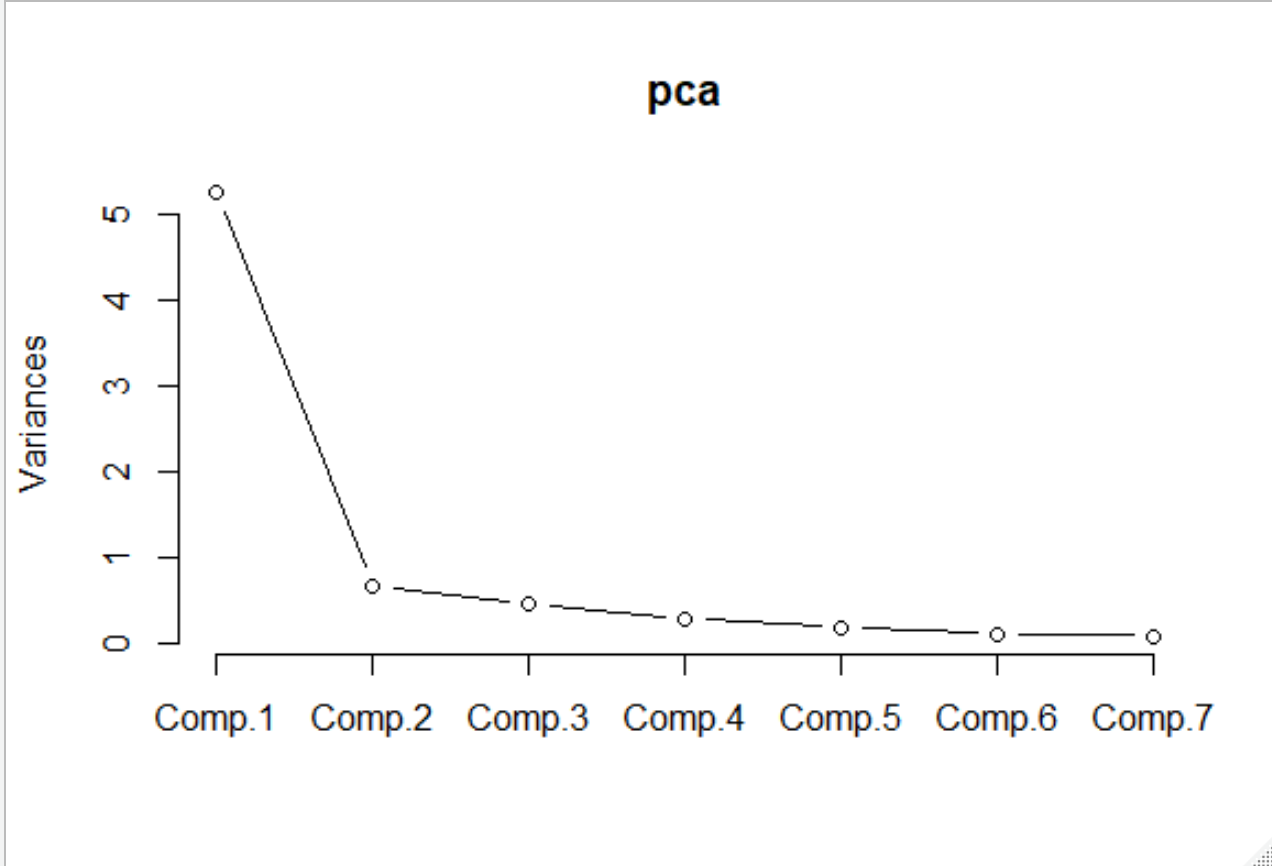


Figure 1

The factor map shows the relationship between variables and how individuals loading scores projected on different factors dimensions.

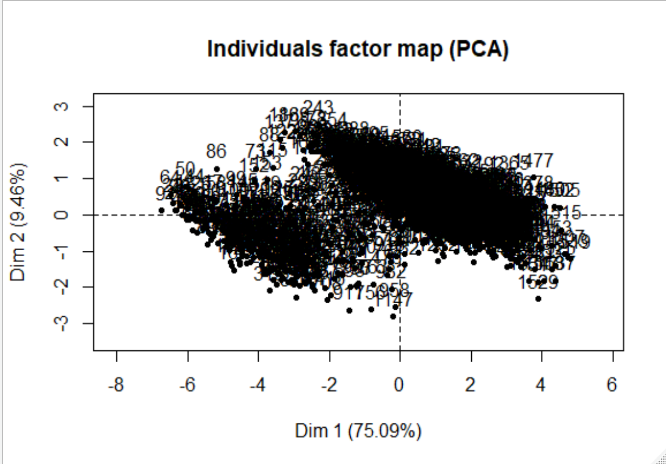
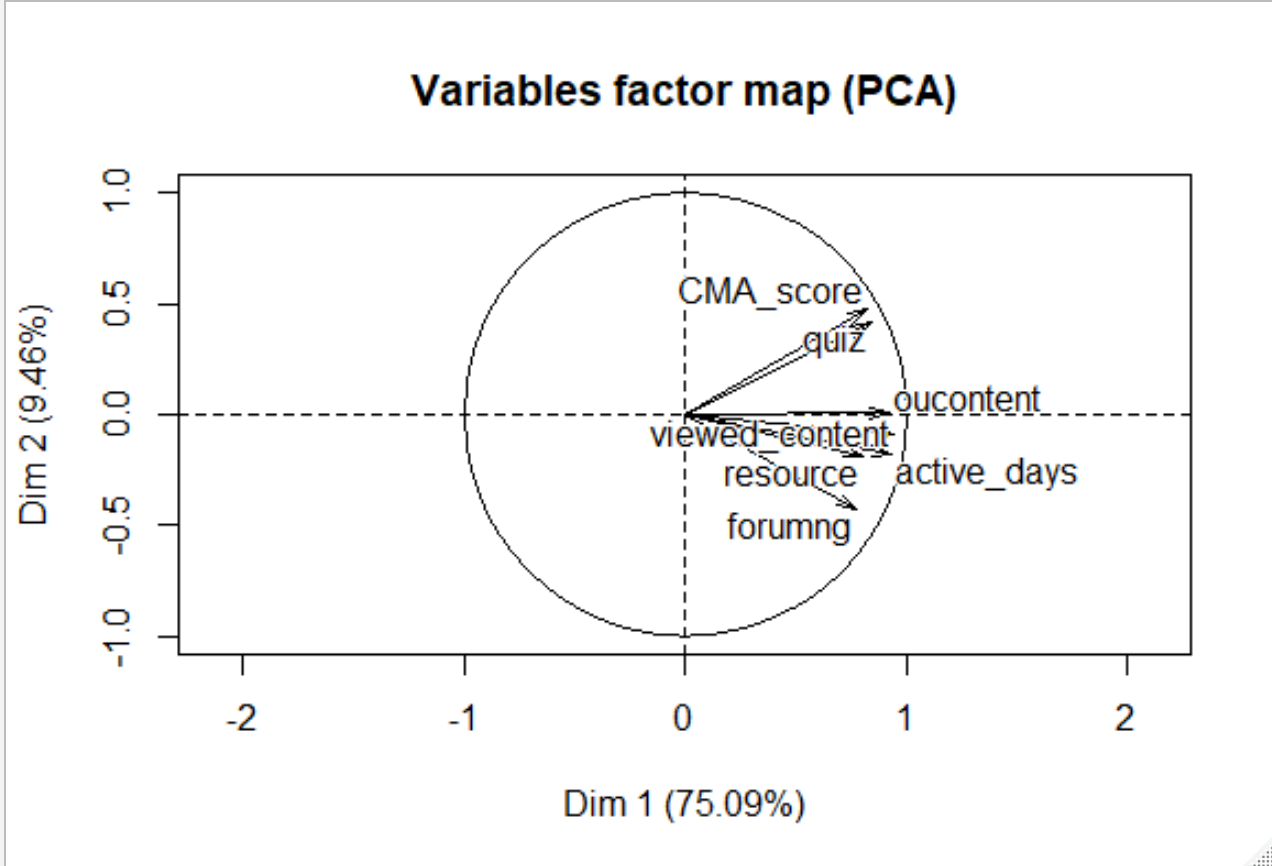


Figure 2

**Maximum likelihood method**

Maximum likelihood, also called the maximum likelihood method, is the procedure of finding the value of one or more parameters for a given statistic which makes the known likelihood distribution a maximum. The Maximum likelihood method also works on correlation metrics but it uses maximum likelihood method to factor. The differences between the maximum likelihood estimates and the “principal factors” approach can be substantial. If the data appear to be normally distributed (as shown in the transformed data), then the additional efficiency of maximum likelihood estimation is highly worthwhile.

**Fisher’s Linear Discriminant Analysis(LDA)**

Discrimination and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups. Fisher's idea was to transform the multivariate observations x to univariate observations y such that the y's derived from population π1 and π2 were separated as much as possible. Fisher suggested taking linear combinations of x to create y's because they are simple enough functions of the x to be handled easily (Richard A. Johnson and Dean W. Wichern, Applied Multivariate Statistical Analysis, 6th Edition). The discriminant scores are calculated for each observation for each class based on these linear combinations. The class with the largest score will be the classification prediction for that observation. Figure 3 shows the mechanism of LDA.

A close up of a map

Description automatically generated

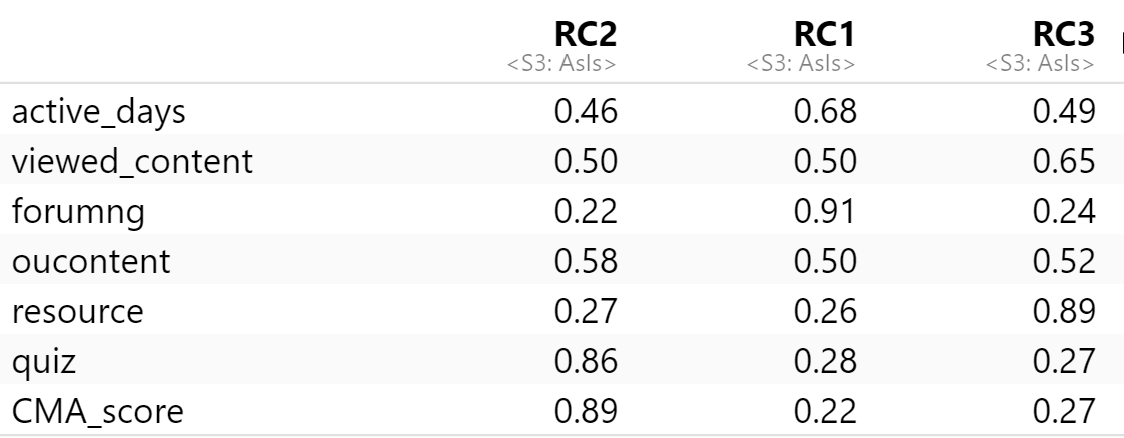
Figure 3

The assumptions of discriminant analysis are the same as those for factor analysis. The analysis is quite sensitive to outliers and the size of the smallest group must be larger than the number of predictor variables. Independent variables are normal for each level of the grouping variable. It has also been shown that discriminant analysis may still be reliable when multivariate normality is often violated. In our project we had already removed outliers and did the transformation. By checking the distribution of the data, we believe that the data is acceptable for these multivariate analysis methods.

**Results**

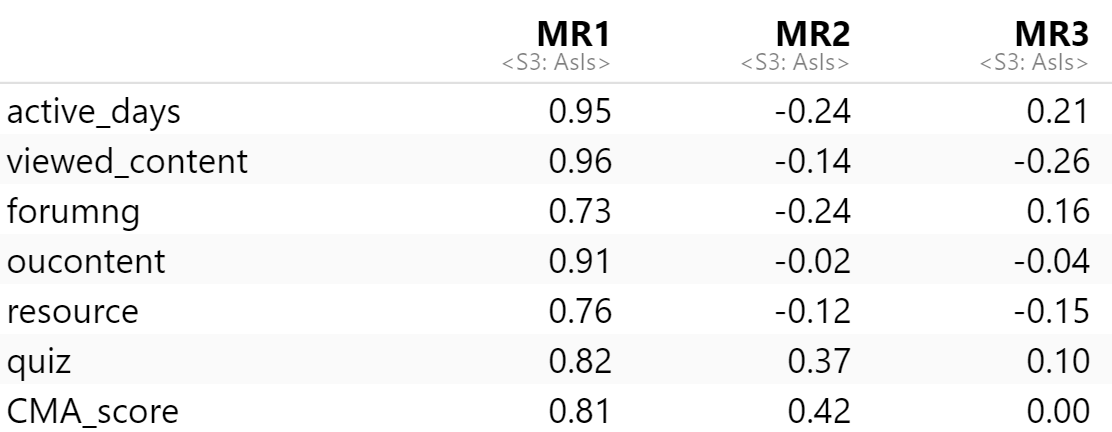
**Factor Analysis**

We carried out factor analysis using both principal component (PCA) method and maximum likelihood method to inspect the most influential factors of the learning performance, as shown previously, the first three factors can explain over 90% of the variance, it’s reasonable to make the interpretation based on the result of three factor analysis.



In the output of orthogonal factor model, columns show the estimated factor loadings in different variables, high loading values means that the corresponding variable plays a dominant role in this factor. In the first row, variable “quiz” and “CMA\_score” have the highest-loading score, these two variables are all related to the practice behavior like doing quiz and earn good points. The result indicates that doing practice or not is an important factor which may influence students’ performance, and is also a good indicator for predicting the outcome. Hence, the first factor can be interpreted as “practice factor”. Similarly, the second factor can be explained as “interaction factor,” since “forumng” has the highest-loading score, which means making interactions with other students or not is a critical discrepancy between different groups. In the third factor “resource” and “viewed\_content” have higher loading scores, which indicates that viewing contents other than “oucontent” (which most students will do) can make difference in the final outcomes. So factor can be interpreted as “content factor”.

Next, we carried out factor analysis using maximum likelihood method, in order to make the result easier for interpreting, we also apply oblique rotation. Different from the orthogonal method, loading scores here can be positive or negative, so not just the value but also the contrast between variables can be used to interpret factors.



In the first column, all the variables have pretty high-loading score, which indicate the overall active level of students online. The second column shows the contrast between “quiz,” “CAM\_score” and other variables, same as above, this factor can be interpreted as a practice factor, which reflects the difference in the behaviors related to practice. The third column shows the contrast between variables related to content (oucontent, viewd\_content and resources) and all other variables, reflect the difference in how students check learning materials, so this factor can be interpreted as “content factor.”

In summary, the result of orthogonal factor analysis shows that practice, interaction and content are the three most important factors that influence student performance. While the result of maximum likelihood method implies that the overall active level is also critical, since this “activity factor” can be taken as composite index, the results are fairly consistent.

**Discrimination and Classification**

In this project, we believe that students with different performances have different behavior patterns, these patterns can be reflected by their click actions, and different click actions can be identified and classified by using measured variables to put the observations into different classes.

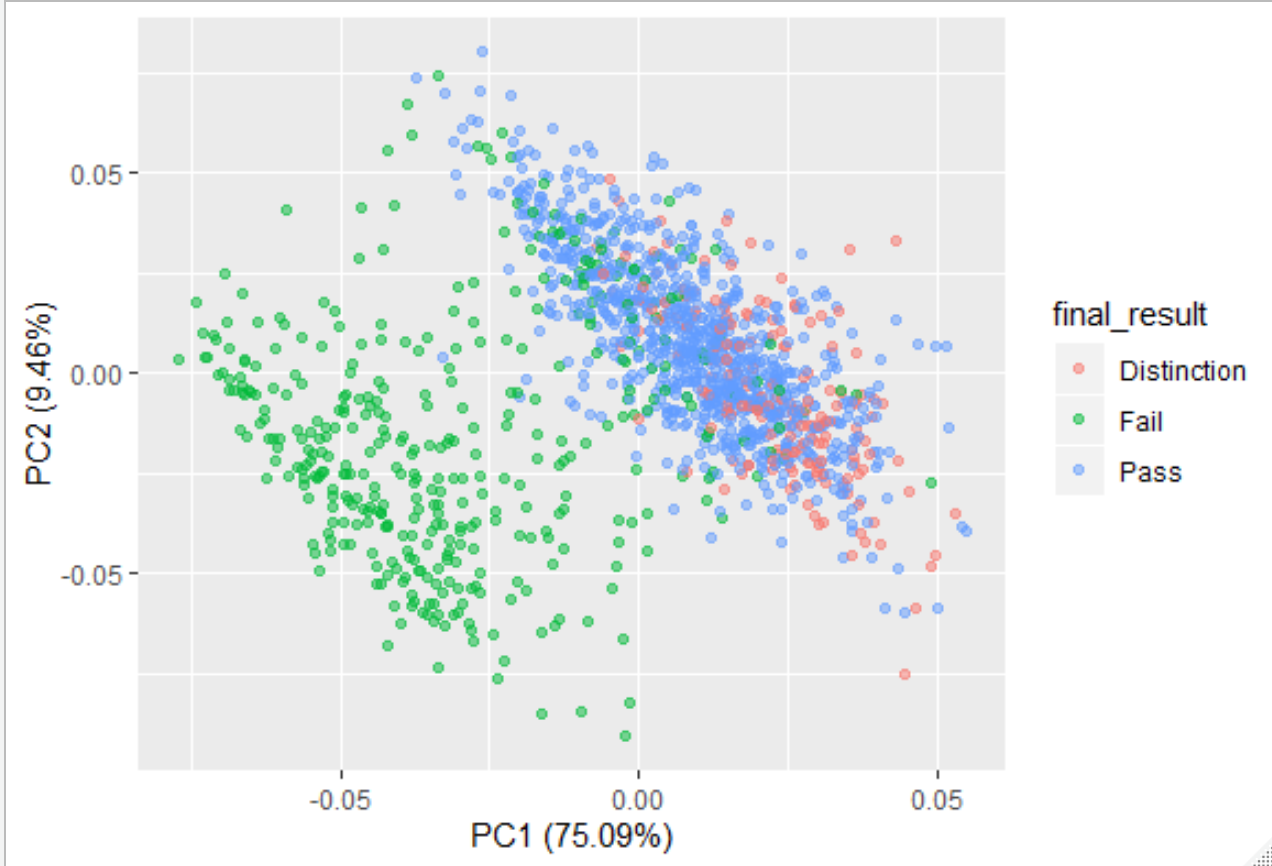


Figure 4

The loadings for the first two factors indicate a clear difference between the failed students and the other two groups. The points of pass students and distinction students are quite overlapped in the projection. Points of fail students spread more widely, which reflect that there’s more diversity in their learning behaviors. While points of fail and pass students are pretty concentrated. Next, we compare two pairs of groups and address what’s the difference in distinction vs. fail and distinction vs pass.

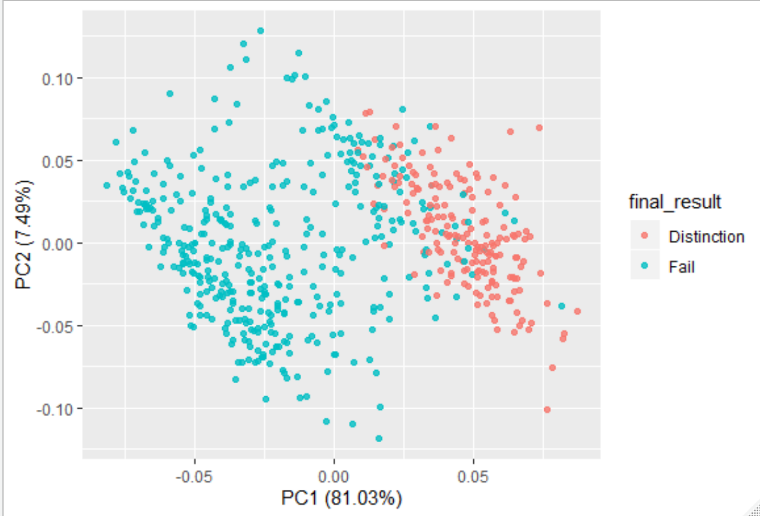
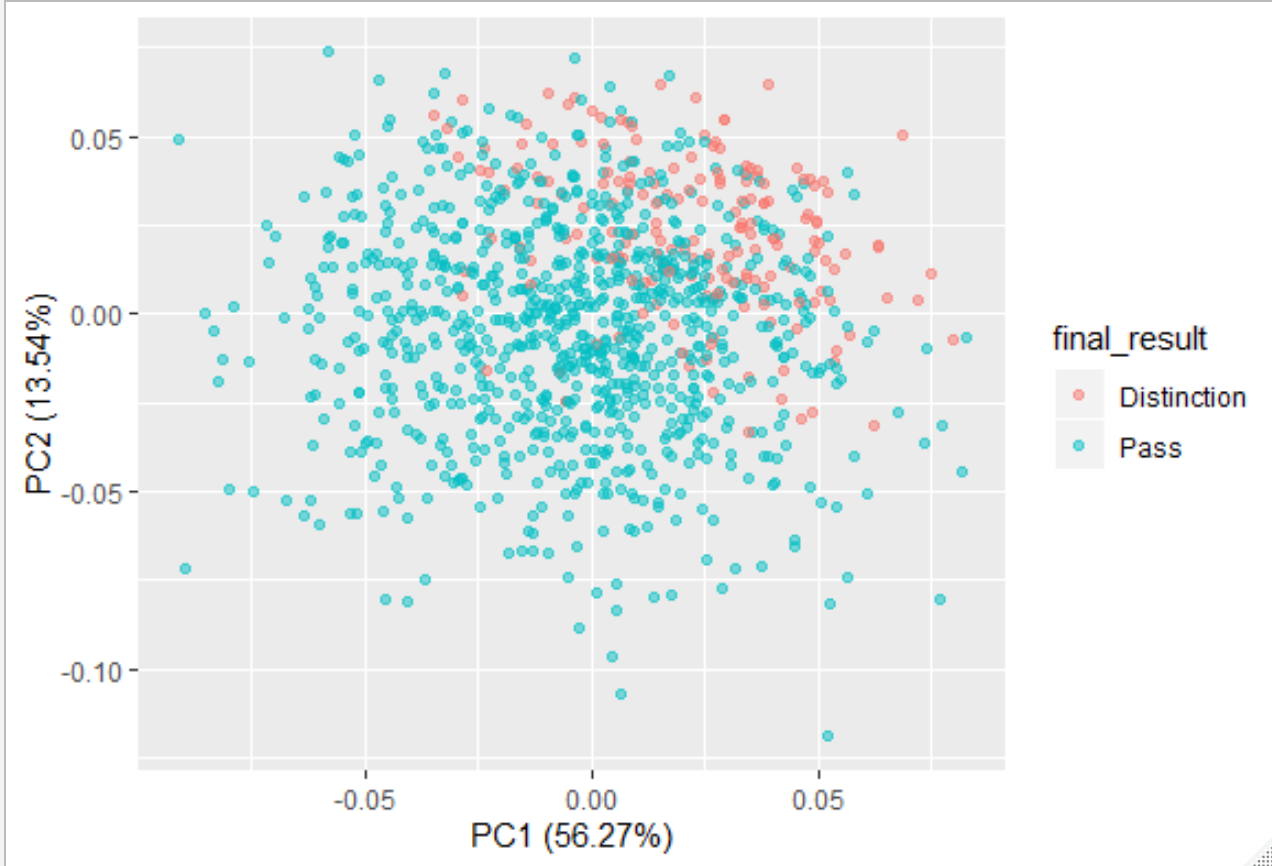


Figure 6

Figure 5

Figure 5 shows that distinction students have higher value and smaller variance in PC1 than fail students. This also happens between pass students and fail students. In figure 3, the points of distinction students are highly overlapped with pass students, but the overall value in PC2 is much higher than pass students. This difference is not easy to detect in figure1, but when shifting to another project dimension, the pattern became obvious. Then we will use LDA model to make the discrimination.

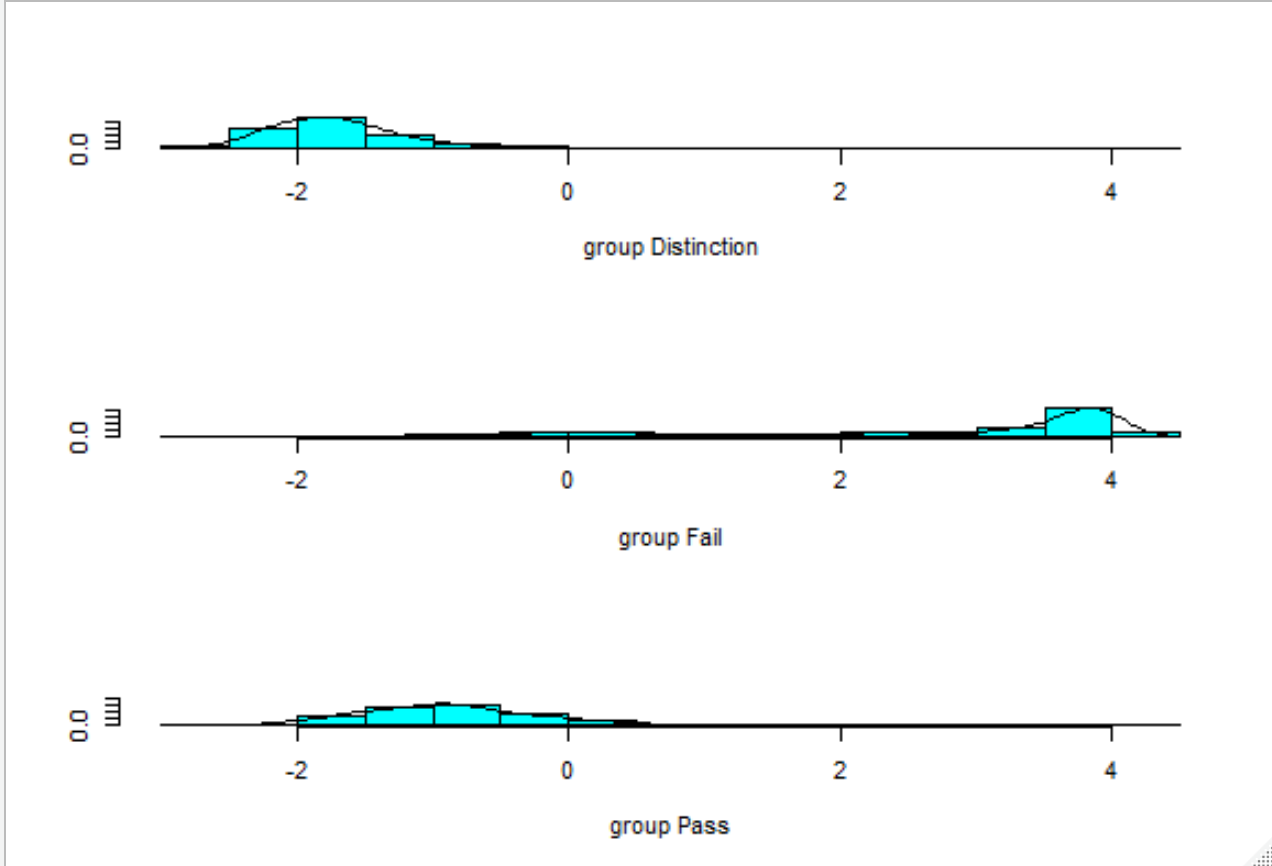
.

Figure 7

The LDA projection shows how students with different outcomes have been discriminated. Most fail students are located on the right of the axis, while there’s a long tail on the left, which implies higher diversity in the behavior of these students. On the other hand, pass group and distinction group are close to each other and partly overlapped. The plot also implies that applying EFA for prediction is reasonable and effective, especially for identifying fail students which is clearly separated from the other group.

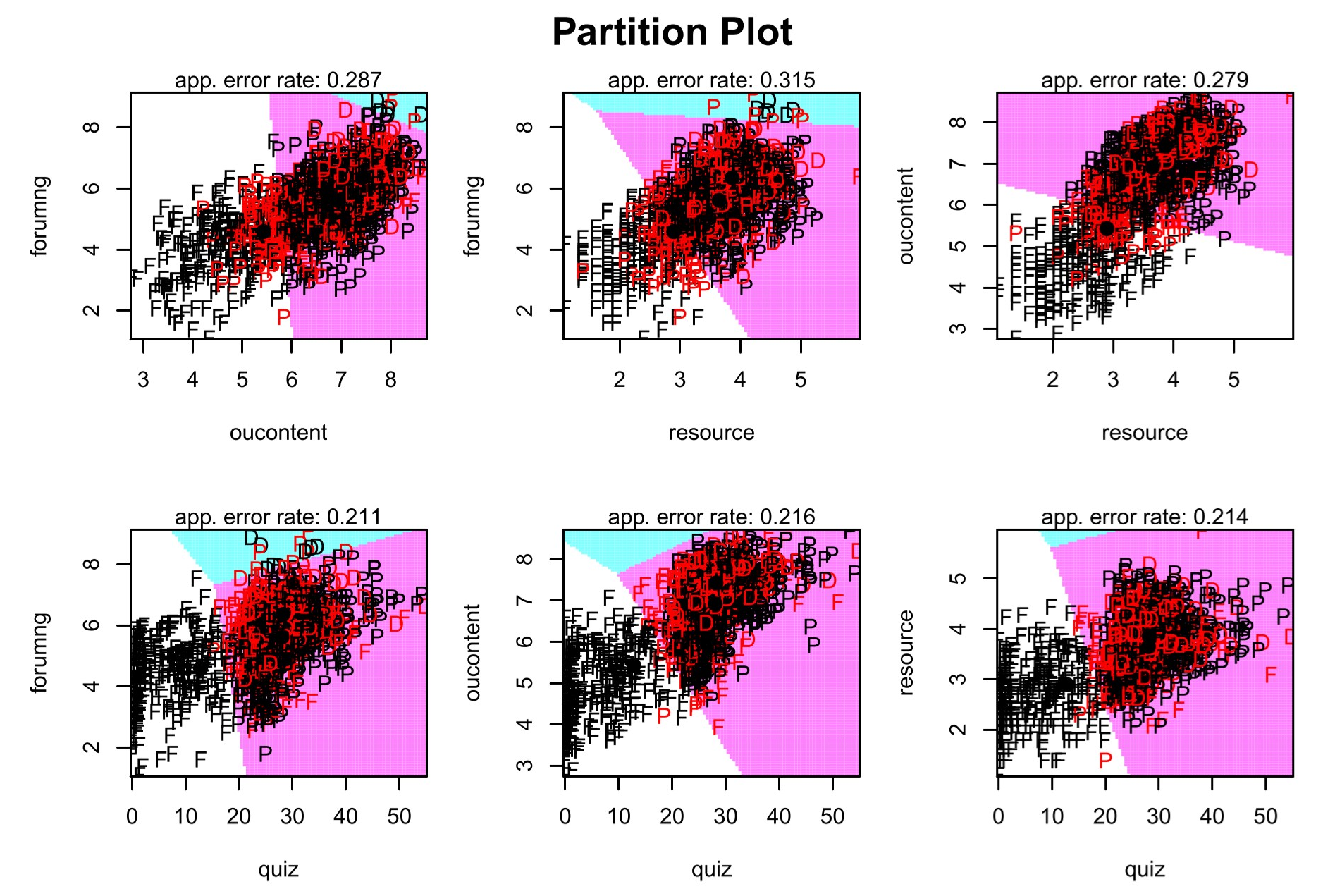


Figure 8

The partition plot (Figure 8) shows how students were discriminated by the combination of two factors, in most cases, the error rate is below 0.3, the lowest error rate occurs with the combination of quiz and forumng. As the plot shows, pass and distinction students make more clicks on quiz than fail students and distinction students makes more clicks than pass students.

**Model Performance Evaluation**

The last step is evaluating the accuracy of the model, as the table shows, the overall accuracy of the prediction is 0.82, the accuracy of predicting fail is 0.98, the accuracy for predicting distinction is 0.57, the accuracy of pass is 0.79.

**Cross validation evaluation**

**Model performance evaluation**

|  |  |  |  |
| --- | --- | --- | --- |
|  | distinction | fail | pass |
| distinction | 33 | 0 | 27 |
| fail | 0 | 101 | 31 |
| pass | 37 | 4 | 250 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | distinction | fail | pass |
| distinction | 58 | 0 | 44 |
| fail | 0 | 304 | 4 |
| pass | 123 | 93 | 826 |

Next, we use the method of cross validation to test the accuracy. We apply half of the data for training and the other have of the data for testing, the result shows that the overall accuracy is 0.80, the accuracy for predicting fail is 0.79, the accuracy for prediction distinction is 0.55, the accuracy for prediction pass is 0.86.

**Discussions**

Studies of distance education has already shown that when teaching and learning happen in virtual space, the students’ behavior pattern is different from that in face-to-face case. In this study, we use EFA analysis methods to investigate how click actions varies between students with different outcomes. The results verified some theories in online learning studies and also raise some new questions for further studies.

There are three most common types of interaction discussed in the distance education literature, which are the learner-instructor, learner-learner, and learner-content interaction. Terry Anderson’s (2003) notion of equivalency means that deep, meaningful learning can occur as long as one of the three forms of interaction is very high. The other two may be eliminated or offered at minimal levels without degrading the educational experience. Anderson asserts, however, that high levels of more than one type of interaction likely result in a more satisfying educational experience.

Student-teacher interaction has the highest perceived value among students, but in online education, there are so many students and only limited number of faculties are available. Therefore, student-content interaction replaces student-teacher interaction in many mass education systems. In our research, we consider teacher videos(OU content) and other resources as forms of student-content interaction.

Traditional views believe that student-student interaction is more critical for collaborative and cooperative tasks while less critical for learning designs based on cognitive and behaviorist learning theories. The result of factor analysis implies that the student-student interaction can affect students’ learning performance but the extent is less critical than content and practice.

We believe that the individual learner’s self-interaction is also critical. In our research, the practice factor can be explained as a learner-self interaction, in this process, learners interact with themselves and take a reflection on the learning content. In online courses, doing quiz is not a compulsive part, most students can get pass even though they never do quiz. However, our research shows that practice and learner-self interaction is a critical factor in the final performance especially between pass and distinction students.

One limitation of this project is that the data only include the number of clicks but no other information like how long the user stayed on the page, the order of clicks and how they jump between sections. In reality, the number of clicks may not tell the whole story, if we have more detailed data, more accurate and complex approaches can be applied, that can be a direction for further study.

**Conclusions**

As the development of information technology, Big Data is seeping into almost every aspect of modern life, and education is certainly no exception. The rise of MOOCs makes it’s possible to collect, aggregate, analyze, categorize, and learn from the data created by learners. It also offers an opportunity for online educators to learn learners behaviors and make decisions and improvements based on the data analysis.

This project introduced an exploratory factor analysis (EFA) methodology for identifying critical factors and predicting learning performance. Our primary goal was to build an effective model for learning state tracing and performance prediction base on the clickstream data. Our model was designed to predict the outcome (“Fail,” “Pass” or “Distinction”) that the students will get in an online course offered by the Open University, which would be an immense help for MOOCs platforms to identify students at risk of failing and increase their chance of passing the course. We proposed 7 variables to describe clickstream data. We employed two factor analysis methods to investigate the influence of different variables, we also applied LDA techniques to discriminate students with different outcomes. We also use cross validation to evaluate the performance of the model. The result shows that the overall accuracy is 0.80, the accuracy of predicting fail is 0.79.

The project shows that there are obvious discrepancies in click behaviors among students with different outcomes. Failed students show lower activity compared with the other two groups, they also less likely to do quiz or make interactions with others. Pass and distinction students share similar click patterns, they all show fairly activity, no evidence shows that distinction students make more clicks than pass ones, they all did quiz and view contents frequently. The main difference is distinction students are more active in making interactions and exploring extra learning materials.

The project also indicates that fail students have a widespread and long tail distribution, which implies that there are more extreme cases, outliers and diversities within the fail group, though these cases only take a small proportion of the total population. Students with high activity and engagement level may fail because of other factors like education background, while students with low engagement levels, who never do the quiz, rarely check learning materials or participate in discussions nearly impossible to pass the course, that may provide a direction for the intervene aimed to help at-risk students.

**Recommendations**

Based on the result of the project, some suggestions are reasonable for MOOC participants. For learners, being active is very important, most students with higher activity finished the course with good results. Learners should take advantages of online learning, like the lecture videos can be played repeatedly. Doing practice is critical for getting pass, all the homework and quiz should be taken as seriously as in physical classes. Making meaningful interactions is necessary for students who want to improve their performance to a higher level.

For course designers, some components are more important than others, for example, the result suggests that most fail students rarely do quiz, while most pass and distinction students do quiz and earn good points though these points may not count for the final result, they are helpful in Consolidating the knowledge. Hence, designers may apply some mechanisms like the next chapter will be locked until the learner finish the existing quiz. Course designers should create and select content very carefully, for example, each lecture video should not be longer than 30 minutes, avoiding too much hyperlinks or pointless readings which may blur the objectives of the courses. Tagging materials with importance levels, tell learners which contents should have higher priority if they have a limited time.

For MOOCs platforms, a system should be established for tracking the learning states of students based on their click behaviors, the system should create weekly or monthly learning reports for each user, show the progress they made and gaps they need to pass over. If the student shows low engagement, the system should send notifications to remind them check the updates more regularly, if the student shows the tendency of failing the course, the system should send alert and provide suggestions for catching up. Since there are usually massive students in MOOCs. To encouraging students to participate in discussions and help each other, the platform should build a bonus system for praising content producers and volunteer tutors. Realizing this manually is almost impossible, so keep improving the algorithm and building effective model is critical for the further development of MOOCs.

**References**

1. Jihyun Park et al., “Detecting Changes in Student Behavior from Clickstream Data”, LAK ’17, March 13-17, 2017, Vancouver, BC, Canada
2. C. Gutl, R. H. Rizzardini, V. Chang, and M. Morales, “Attrition in MOOC: Lessons learned from drop-out students,” in Proc. Int. Workshop Learn. Technol. Edu. Cloud, Sep. 2014, pp. 37–48, doi: 10.1007/978-3-319-10671-7\_4.
3. Pedro J. Munoz-Merino, Pedro Manuel Moreno-Marcos , Carlos Alario-Hoyos , Carlos Delgado Kloos , “Prediction in MOOCs: A Review and Future Research Directions”, IEEE Transactions On Learning Technologies, Vol. 12, No. 3, July-September 2019.
4. Christopher G. Brinton, Swapna Buccapatnam, Mung Chiang, H. Vincent Poor, “Mining MOOC Clickstreams: On the Relationship Between Learner Behavior and Performance”, IEEE conversion 2015.
5. Gábor Kőrösi, Péter Esztelecki, Richard Farkas, Krisztina Tóth , “Clickstream-based outcome prediction in short video MOOCs”, 978-1-5386-4599-4/18/$31.00 ©2018 IEEE.
6. Niki Gitinabard et al. , “Your Actions or Your Associates? Predicting Certification and Dropout in MOOCs with Behavioral and Social Features”, Proceedings of the 11th International Conference on Educational Data Mining.
7. C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach, “Forecasting student achievement in MOOCs with natural language processing,” in Proc. Int. Conf. Learn. Anal. Knowl., Apr. 25, pp. 383– 387, doi: 10.1145/2883851.2883932.
8. Y. Chen, Q. Chen, M. Zhao, S. Boyer, K. Veeramachaneni, and H. Qu, “DropoutSeer: Visualizing learning patterns in massive open online courses for dropout reasoning and prediction,” in Proc. IEEE Conf.Vis. Anal. Sci. Technol., Oct. 2016, pp. 111–120.
9. Naif Radi Aljohani, Ayman Fayoumi and Saeed-Ul Hassan, “Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment”, Sustainability 2019, 11, 7238; doi:10.3390/su11247238.
10. Z. Ren, H. Rangwala, and A. Johri, “Predicting performance on MOOC assessments using multi-regression models,” in Proc. Int. Conf. Data Mining, Jun./Jul. 2016, pp. 484–489.
11. T.-Y. Yang, C. G. Brinton, C. Joe-Wong, andM. Chiang, “Behavior-based grade prediction forMOOCs via time series neural networks,” IEEE J. Sel.Topics Signal Process., vol. 11, no. 5, pp. 716–728, Aug. 2017.
12. Jakub Kuzilek, Martin Hlosta & Zdenek Zdrahal(2017), Data Descriptor: Open University Learning Analytics dataset.
13. Anderson, T. (2003). Getting the mix right again: An updated and theoretical rational for interaction. The International Review of Research in Open and Distance Learning, 4(2).
14. Stein, David S.and Constance E. Wanstreet (2017). Jump-Start Your Online Classroom: Mastering Five Challenges in Five Days. Chapter 5: Building Spaces and Places for Learning

Agpendix A. Thesis Completion Timeline.

In this appendix, include a proposed schedule and timeline for completion of the thesis. This will aid the committee in determining their availability.

Agpendix B. Proect Faculty Committee.

List the proposed members of the thesis supervisory committee. Include suitable justification for inclusion of each committee member. Two members are required.