Creative EDM Assignment

Jingxuan Liao, Ruoqi (Mark) Wang, Yuchen (Rebecca) Wu

Human Development Department, Teachers College, Columbia University

HUDK 4050: Core Methods in Educational Data Mining

Professor Zhichun(Lukas) Liu

December 21, 2021

Introduction

During the pandemic, many colleges have shifted to remote learning and teaching on online platforms. One of the major problems for instructors is the difficulty of tracking students' learning behavior and offering customized instructions according to students' performance. In this project, we are interested in investigating the factors that affect a student's test scores, predicting students' final test scores based on 'determining factors', and grouping them into different study sessions. This enables teachers to provide more individualized and targeted help toward improving final scores.

In a word, our primary research problem is 'what behavioral factors may affect undergraduate students' test scores in the American history course in one semester for the synchronous online settings such as Zoom & Canvas?'. The study population is undergraduate students in one American history course, around 200 students. The learning activities take place during the period of one semester (around 3-4 months) and in the online learning platforms Zoom and Canvas. Our primary objectives in the process include:

1. Define and operationalize behavioral factors that affect undergraduate students' final scores
2. Collect and analyze data from online platforms and surveys to predict students' final scores
3. Build and test models to predict undergraduate students' final test scores
4. Build clusterings of students according to predicted scores

Our goal for predicting test scores and building clustering is not to make assumptions about students' ability, but to provide more targeted and customized instructions and resources

by dividing students into different review sessions. In this way, students can get more help toward improving final scores.

Before conducting the study, there are a few caveats to take into consideration. First, issues related to privacy and confidentiality need to be addressed. For the personal data and learning data we collected from platforms backstage and surveys, we will make sure such data will only be used for the purpose of this project and we will not disclose or share any information with the school, students, or parents. All records will only be stored with one copy in the researcher's hard drive and one copy in the cloud for security purposes. Second, to prevent bias and ensure the inclusiveness of this project, we will make sure the result (predicted final scores and grouping) will not affect instructors' grading process. This is achieved through substituting students' names with ID so that instructors cannot identify students' predicted final test scores with their actual performance in the final test. Also, to ensure inclusiveness, we will make sure to count each student in the final analysis. Despite having some missing data or invalid data like 0, we will substitute with more statistically meaningful data or exclude the variable itself. Lastly, we will follow the ethics related to conducting research and be transparent to students and instructors about the purpose of this project, but details will not be disclosed to prevent 'biased' data. Students will be given an opted-out option from the project, instead, they will be randomly assigned to review sessions.

Literature review

According to results from previous studies, educational data mining can be used to identify students' behavioral patterns and predict their grades (Romero & Ventura, 2010). Many studies on student performance prediction have focused on drop rate prediction, pass rate

prediction, and grade prediction, little has been utilized for the purpose of classification, like grouping students into study sessions for exam preparation (Zhang et al., 2017). Also, many have used analytic methods such as Bayesian Network, Association rule, PCA, Classification, etc., to build prediction models (Zhang et al., 2017). In our study, we used logistic regression and clustering methods to build our unique models and serve our purposes of prediction and classification. To identify critical factors to improve prediction accuracy, we refer to the standard of accepted range of prediction accuracy from 75% to 95% (Asif et al., 2014; Villagrá-Arnedo et al., 2016).

Data Collection

We would like to retrieve most of our data from the backstage of zoom and canvas. Instructors and teaching assistants of the course usually have access to students' data from the backstage of the canvas. Data they can access includes the students' clicking of each resource, and the time length the students stay on each page. Zoom also provides similar information such that the host of the meetings will have access to meeting attendees' activities during the meeting. Also, we would like to send out online surveys to collect self-reported data from students. Tools such as Google forms will help organize the statistics and do some of the analysis.

We would like to collect a dataset that is around 200 data points, or in reality, it means there are around 200 students in the class. To include all critical factors, we operationalize students' behavioral learning data into four dimensions, which are participation, learning, performance, and course requirements. We would like to collect data on each of the four dimensions and look at them from a holistic perspective. Most of the data are collected on a semester basis, while

some of the self-reported data are collected on a weekly basis, and still, others are collected only once for the semester.

For participation data, we would like to collect data from three different perspectives. The first one is the total number of hand-raising in Zoom live sessions, and this variable is assessed by how many times the student clicks on the "raise hand" button on Zoom throughout the whole semester. The number of times students click on the "raise hand" button in each class meeting will be recorded, and the statistics for each class meeting are going to be added up in order to calculate the total amount for the whole semester. The second one is the total number of discussion posts created by the student in Canvas. This is assessed by how many posts the students posted on the voluntary discussion board of the course. The third one is the number of days the student is absent. This is assessed by the number of times the student did not enter the zoom meeting at all while there is a class scheduled throughout the whole semester. A total number is going to be calculated for the semester.

For data regarding learning, we would like to collect data from one perspective, which is the total number of visits to the class resource page. This is assessed by how many times the student visits the course Canvas page throughout the entire semester. The total number of visits is going to be calculated for the entire semester.

For data in terms of performance, we would like to collect data from three different perspectives. The first one is the rate of completion of online assignments. This data is assessed by the number of assignments completed throughout the semester divided by the total number of assignments assigned by the instructors throughout the whole semester. The second piece of data that we are going to collect is the average essay scores. This is assessed by the total number of essay scores of the entire semester divided by the total number of essays assigned throughout the

semester. The third one is the average previous test scores. This variable is assessed by the total number of points gained in all tests throughout the semester divided by the total number of tests given by the instructors throughout the semester.

The last portion of the data that we are going to collect is self-reported data generated from online surveys. We are planning on collecting three pieces of data in the survey. The first one is whether this course is a mandatory course or an elective course for the student. This variable is going to be categorical, with 0 representing the course being elective for the student, and 1 representing the course being mandatory for the student. The second piece of information that we are going to gather through the self-reporting survey is whether the students have attended the office hours of either the teaching assistant or the instructor throughout the semester. This variable is also a categorical variable, with 0 represents that the student has never attended any office hour sessions while 1 represents that the student has attended the office hour sessions at least once. The last piece of information that we are going to collect through the self-reporting data is the number of hours the students study each week, and then they will calculate the approximate total number of hours they have studied for this class outside of class time.

Data Cleaning

After collecting all the data according to the defined variables, we make a plan to clean data according to the unique characteristics of each variable. For variables that are continuous and have a wide range, we standardize them to improve model performance, such as ‘average essay scores’ and ‘average previous tests’. For variables that are likely to have many zeros, we make it into categorical variables to prevent the skewing of data (left-skewed). For example, ‘office hour attendance’ will be collected from the survey as either ‘have attended’ or ‘have not

attended'. For variables that contain outliers and null values, we will not delete them as we hope to include all students involved in the study. Instead, we will substitute with median, mean, or mode depending on the distribution of data. For example, if the data is not skewed to one side, the mean is safe to use; otherwise, we will use the median. For categorical variables, we use mode to replace the missing data.

Analytical Methods

In data analysis, we would like to run a descriptive analysis to see the overall picture of students' behaviors. Then, we would like to do a multivariate analysis to look at the relationship between every variable and the test score. Before we build a predictive model, we would like to run a clustering analysis to group students based on necessary components such as behavior and performance. Then we calculate the correlation of each variable with the test scores to see which factors have a greater impact on test scores, and finally, we would build a logistic regression model to predict students' next year's history test scores.

Clustering Analysis

We would like to do clustering with k-means. We would like to visualize the clusters by Silhouette Plot. The Silhouette coefficient could help us to evaluate the number of clusters. If it is necessary, we might do Principle Component analysis by scree plot to extract the most critical predictor as an x-axis to visualize the clusters plot.

Prediction Model

The test score would be divided into three categories: Low, Medium, High. The logistic Regression Model would be good to use to predict the range of the next history score. Since one of the goals is to group students and offer effective intervention, the categorical prediction of score would be more reasonable to our goal than the numerical prediction. Thus, the Logistic Regression model fits our analysis.

Expected Result

We expect that the model would be accurate in predicting students' next history scores. We expect that the accuracy score would fall in the range of 65% - 95%. From our data collection and correlation analysis, we expect that the number of absence days should be negatively related to the final exam score. Moreover, we expect to see three clusters from the clustering analysis. We expect that the number of clusters would be aligned with the outcome of the prediction. The students should have common characteristics with others who fall into the same range of test scores. In our prediction model, we expect that the number of hand-raising in Zoom Live Session, the number of office hour attendance, the average homework scores, the average quiz scores, and whether it is a mandatory course or not would be critical predictors to our model.

Implications

This research holds implications in both research and practice fields. First of all, during the period of pandemics, the result from this study might provide educators and students with more insights on what behavioral factors could contribute to the performance of the online synchronous study exam scores. Educators and students can utilize these pieces of information to

make changes to corresponding aspects in order to see possible improvements in exam scores. Also, based on our research, we would suggest future researchers investigate more on the causal effect of the behavioral factors that we identified as critical on undergraduate students' social science final exam scores. If the conclusions from our study are strengthened, more relevant parties could utilize them to make a difference. Lastly, this research sheds light on an instructional method of assigning students to appropriate learning groups to provide more targeted and customized short-term instructions. Based on the clusters that we identified students to be in, customized study materials are going to be designed for them to achieve better academic achievement.

Limitations

The first aspect of the limitations that our study might hold is that, due to the limited number of data we could collect from a class of 200 students, the model may not be accurate enough. The accuracy of the model might be increased if there are larger scales of data points. Secondly, due to different settings of courses at different universities and colleges, some of the data we would like to collect might not apply to similar courses at other universities or colleges. For example, some universities offer recitation sessions, and as a result, the frequency of students going to office hours might greatly decrease compared to universities or colleges that do not offer recitation sessions. Lastly, due to the unique characteristics of the data, which is a combination of undergraduate students, history class, and synchronous online learning, the conclusion might not be able to be generalized to other geographical and cultural settings.

Challenges

Some of the challenges that our team might face while conducting the study include that, while collecting some of the data, it is very hard to make sure the data is actually accurate. Due to the nature of Zoom and Canvas, data we retrieve from the backstage of Zoom and Canvas might not always be able to reflect students' true learning behavior. Also, self-reported data are not always reliable due to the fact that students might have incentives to fake their answers in the surveys. Secondly, we might not be able to include all relevant behavioral factors that have an effect on students' final exam scores, and we hereby call for future researchers to include more factors if appropriate. Lastly, while analyzing the dataset and presenting the study, there might be issues related to the data owner's privacy.

References

Romero, C., & Ventura, S. (2010). Educational data mining: A Review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.

Asif, R., Merceron, A., & Pathan, M. K. (2014). Predicting student academic performance at degree level: A Case study. International Journal of Intelligent Systems and Applications, 7(1), 49-61.

Villagrá-Arnedo, C., Gallego Durán, F.J., Compañ, P., Llorens-Largo, F., & Molina-Carmona, R. (2016). Predicting academic performance from behavioral and learning data. International Journal of Design & Nature and Ecodynamics, 11(3), 239-249.

Zhang, W., Huang, X., Wang, S., Shu, J., Liu, H., & Chen, H. (2017). Student performance prediction via online learning behavior analytics. 2017 International Symposium on Educational Technology (ISET). https://doi.org/10.1109/iset.2017.43