DATASET DESCRIPTION

The data set analyzed in the project is composed of 29,305 entries describing the video-watching behaviors of 3876 students over a total of 93 videos. The entries also contain the results the students obtained on their first attempt on the quizzes corresponding to the videos. Each video has a unique ID, which is a number between 0 and 92. Each student is characterized by a unique user ID, and each entry is unique for the behaviors/score of a particular student on a particular video/quiz. There was no specification on the number of videos a student had to watch. Therefore, if a student only watched one video and completed its corresponding quiz, the user ID corresponding to the student and the corresponding behaviors/score appear only once in the data set. Similarly, if a student watched multiple videos and completed the corresponding quizzes, the user ID corresponding to the student and the corresponding behaviors/score appear multiple times in the data set.

There are 9 video-watching metrics in the data set. The metrics are defined as follows:

1. fracSpent - the fraction of time the student spent watching the video relative to the length of the video.
2. fracComp - the fraction of the video the student watched. This number ranges between 0 and 1 where 0 corresponds to watching none of the video and 1 corresponds to watching the video completely.
3. fracPlayed – no definition provided.*
4. fracPaused - the fraction of time the student spent paused on the video, relative to the length of the video.
5. numPauses - the number of times the student paused the video.
6. avgPBR - the average playback rate that the student used while watching the video, ranging between 0.5x and 2.0x.
7. stdBPR – no definition provided.*
8. numRWs - the number of times the student skipped backwards (rewind) in the video.
9. numFFs - the number of times the student skipped forward (fast forward) in the video.

* Even though one might think the labels of these metrics are self-descriptive, since no concrete descriptions were provided for these metrics, we did not include them in our analysis to avoid any corruption in the data.

Finally, the last column of the data set corresponds to scores of the students on their first attempt at answering the question given directly after the video. A score of 1 corresponds to the student answering the question correctly, and a score of 0 corresponds to the student answering incorrectly.

ANALYSIS METHODS

For this project we considered three main analyses questions:

1. How well can the students be grouped based on their video-watching behaviors?
2. Can the video-watching behaviors of a student be used to predict the student's performance on video quizzes (average score across n videos)?
3. Can the behaviors of a student on a particular video be used to predict the student's score on the corresponding video quiz question?

1.

For this first analysis question we decided to implement a clustering algorithm on a subset of the data set. Specifically, we implemented the k-means algorithm. This is an iterative centroid-based algorithm which finds a simple structure in the data. For this analysis we only considered the students that completed at least five of the videos. We determined that a student completed at least five videos if the user ID corresponding to the student appeared in at least five entries and those entries had a *fracComp* score above 0.98. For this analysis we considered all seven metrics defined above. The metrics were averaged for each student to perform the analysis. Finally, we implemented the algorithm for a range of clusters between 2 and 15 to find the optimal number of clusters into which to divide the data.

We chose to implement the k-means algorithm for this analysis because it allowed us to find a simple structure given a multi-dimensional data set. We also chose this technique because it allowed us to call the algorithm with a varying number of centroids to find the optimal number of centroids. Additionally, this model does not require an assumption about any preexisting model in the data. We evaluated the clustering models by computing the distortions in each of the models for the range of centroids we defined. Additionally, we computed the Calinski-Harabasz score for each model given a number of centroids to further identify the optimal number of centroids. We expected this analysis to tell us if the students can be naturally grouped into some number of clusters, and what that number would be.

2.

For this analysis question we decided to implement Ridge regression with cross validation on a subset of the data set. We considered only the students that completed at least half of the quizzes in our analysis. We determined that a student completed at least half of the quizzes if the user ID corresponding to the student appeared more than 46 times in the data set. All the seven metrics defined above were considered in our analysis. The seven metrics were averaged for each student to create a feature matrix of all students. The score entries for each student were also averaged to create a target vector of all students. For the Ridge regression analysis, we chose a range of regularization parameters in a logarithmic space between 1 and 2. We considered a total of 100 regularization parameters in this range. After performing the Ridge regression analysis with cross validation, we obtained the most optimal model for the data. We evaluated this model by computing the coefficient of determination for the model obtained.

We decided to use Ridge regression with cross validation because this method provides a way to obtain an optimal linear model with a regularization parameter. With the regularization parameter, we can compute a model that best fits the data but avoids overfitting. Ridge regression also provides a set of coefficients that allowed us to infer the effects of each behavior on the average performance of each student. From the analysis we expected to identify whether there exists a reliable model for predicting the performance of students of the video quizzes given their behaviors.

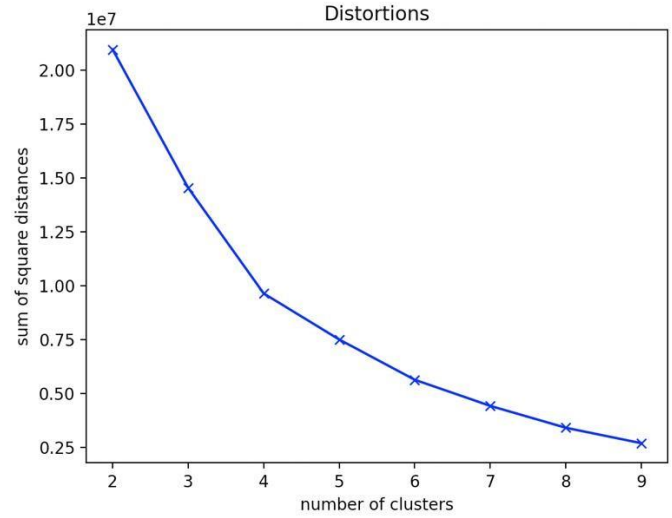
3.

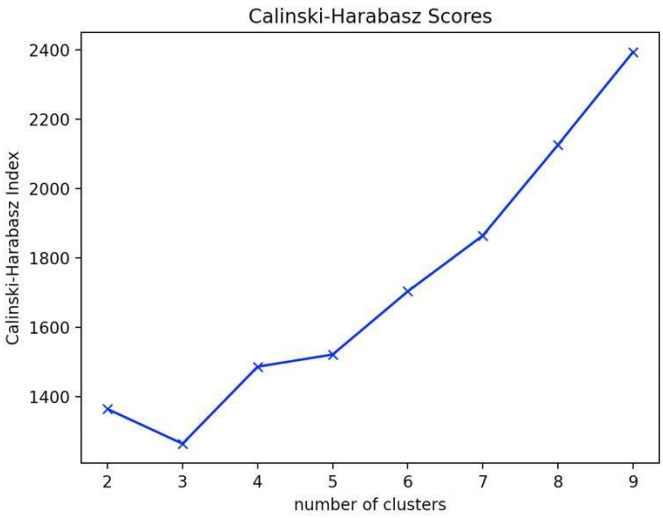
For this analysis question, we decided to implement a logistic regression analysis with cross validation on a subset the data set. We chose a logistic regression analysis since the score a student received on a particular video quiz could only take two values, zero or one. Due to the nature of the scores, we identified that the problem of predicting the scores of students could be shaped into a classification problem. Logistic regression was the ideal tool for this analysis because we could classify each entry in one of two categories. For this analysis we considered the same metrics as in the previous two analyses. However, in this analysis we did not perform any averaging, rather we analyzed the behaviors in each entry and the respective score on the quiz. From the analysis we expected to identify whether there exists a reliable model for predicting the score of a student on a particular video quiz given his behaviors on that video.

RESULTS

1.

Based on our analysis, we concluded that there is not a natural way in which the students can be grouped based on their video-watching behaviors. Since we did not have a ground truth table for comparing our clustering results, we used two other internal techniques for determining our answer. For determining this answer, we first plotted the Calinski-Harabasz score for each number of clusters we tested (see respective graph below). We did this to get an initial idea of what number of clusters performed better when compared to the other numbers in the range. After this initial evaluation, we plotted the sum of squared distances for each number of clusters we tested. We visually analyzed our plot to determine the number of clusters at which the graph decreased most abruptly (the “elbow” of the graph). From the first plot, one can see that the sum of squared distances drops most abruptly before stabilizing at 4 clusters. Additionally, we see that in the Calinski-Harabasz score graph, there is a “spike” at k=4. This gives evidence that if the students were to be grouped, 4 clusters would be the most ideal number into which to divide the students.





2.

Based on our analysis, we concluded that the video-watching behavior of students cannot be used to predict their performance. After performing a Ridge regression analysis with cross validation, we obtained a model with the following coefficients:

[-2.76988980e-05 1.25058124e-03 1.67409202e-05 2.61141678e-03 2.37627953e-03 2.84053174e-04 -6.53244686e-04]

The coefficients above correspond to the seven metrics defined in previous sections in their respective order. One can observe that these coefficients have a small order of magnitude, which might be an indication that the behaviors are not very useful for predicting the performance of students. Additionally, the first and last coefficients have a negative sign. The first coefficient corresponds to the fraction of time student spent watching a video relative to the length of the video. In the raw data, the entries for fracSpent had a very large variation (standard deviation of 105.41). Some of the large coefficients might indicate that the students left the video paused for a long time, which might indicate

the students not watching the video at all. The last coefficient corresponds to the number of fast forwards per video. If this number is high, it would mean that the student fast forwarded many times. This behavor might indicate that a student was not paying very close attention to the video.

The intercept of the model is:

[0.63715928]

One can observe that the intercept of the model has a large order of magnitude relative to the orders of magnitudes of the coefficients. From the intercept one can also observe that if a student did not watch a video at all, the expected performance of that student is 63.7%. This expected performance is high in our opinion, but it is not as high as it may seem. Since the score of a student can only be one or zero, if a student selected an answer at random for all the quizzes the student took, the student would have an expected performance of 50%. Therefore, the expected performance of a student that did not watch any videos before the quiz is 13% points higher than that of a student randomly selecting answers.

For evaluating this model, we computed its coefficient of determination r^2. We obtained a very low coefficient of determination (0.022099057636081887). Based on this value, we concluded that the model was not reliable for predicting the students' performance on the video quizzes.

3.

Based on our analysis, we concluded that a student's performance on a particular video quiz question cannot be very well predicted. After performing a logistic regression analysis with cross validation, we found that the highest prediction accuracy over 100 different folds was 66.3%. This means that in the best case, our logistic regression model can predict the score of a student on a quiz with 66.3% accuracy.

The coefficients of this model are:

[-0.00764271 0.0121332 0.00307826 0.00605778 0.02910122 0.00988976 -0.00278978]

The coefficients above correspond to the seven metrics defined in previous sections in their respective order. One can observe that avgPBR has the highest impact on the odds for this model. Additionally, as in the Ridge regression analysis, fracSpent and numFF have a negative effect on the odds for this model.

Since the accuracy, our logistic regression model was about 66.3% of the time, we do not expect this model to be able to correctly predict the scores based on the data provided.

For the last two analyses we did not get positive results. One reason for this might be the large amount in the variations of the features in the data. The first large variation in the data is the variation in the number of videos students watched. Some students watched only one video and some students watched many more times. Additionally, most of the features had a relatively large standard deviation. Some had standard deviations larger than 100. The large variations in the number of videos students watched and in the video0watching behaviors of the students makes finding an accurate model for predicting scores harder.

Another reason for why finding an accurate model might be the binary nature of the scores. Since scores can only be 1 or 0, if a student selected an answer at random, his expected performance would be 50%.

This makes the behaviors of the students have a smaller impact on their scores. This also makes finding an accurate model harder.