Path 1

Final Project: Student performance related to video-watching behavior

Section Number: 003, Shounak Mukherjee

Purdue Username: mukher53

Github Username: @Shounak007

Dataset Description:

In this project we are given a file which contains data about a online course with videos. Students are meant to watch these videos and answer a quiz upon the video after completion. In this dataset we have multiple students and to identify each student, they are each provided with a unique ID. The ID's provide information about how the student answered and identified the question. The other fields that were provided is the video ID. Each video has a video ID to identify where it came from. This is a numerical value from 0 to 92 due to there being 93 videos in the course. We are also provided with the ratio of time the student spent watching the video versus the length of the video. A fracComp variable is provided which measures the fraction of the video we watched, ranging from zero to one (or a hundred percent). Accompanying this, fracPaused is given which shows the amount of time the student spent paused on the video relative to its length. For example, if we paused a three-minute video for 30 seconds, fracPaused would be one-sixth. We also have numPauses, which is the total number of times the student paused the video. Additionally, there's avgPBR, which stands for average playback rate, measuring the speed at which the student watched the video. The playback rate can range from 0.5x to 2.0x, meaning the student can watch the video at half speed or twice the regular speed. Next, we have numRWs, which is the number of times we rewound or skipped backwards in the video. Conversely, numFFs is the number of times we skipped forward or fast-forwarded. Finally, there's the variable s which indicates whether we got the answer to the question given after the video correct (s=1) or incorrect (s=0). The s variable is binary so this means that the student needed to get every single question right in order to display s = 1 for a video or we can assume that there was a single true/false question after a video which the student needed to answer. Lastly, we are given the variable fracPlayed which represents the fraction of the video watched. Combining one or more of these metrics allows us to come to conclusions about the students and how well they did.

This data analysis project aims to answer three questions:

1) How well can the students be naturally grouped or clustered by their video-watching behavior (fracSpent, fracComp, fracPaused, numPauses, avgPBR, numRWs, and numFFs)? You should use all students that complete at least five of the videos in your analysis.

2) Can student's video-watching behavior be used to predict a student's performance(i.e., average score s across all quizzes)?

3) Taking this a step further, how well can you predict a student's performance on a particular in-video quiz question (i.e., whether they will be correct or incorrect) based on their video-watching behaviors while watching the corresponding video?

Method:

Method for Problem #1:

For this problem we needed to consider the students that only watched 5 videos or more and cluster them based on video watching behavior. This was done using K means clustering. First, we created a list "l" which makes a list of the various userIDs that have watched more than 5 videos. After we have our new userIDs, this list needs to be used in our dataframe , so we create a new data frame with the valid userIDs. From that we redefine and use this in our new data set. Now, since we have all our data that we will be using, we can make a K means model with 6 clusters and fit the k means object to our data to receive our score. To receive our score, we use a silhouette coefficient which is a metric that calculates the goodness of fit for clustering. In order to figure out which number of clusters would give us the most accurate result; we plotted the clustering of k means from 1 to 10 and calculated a average score for each cluster. This is called the "elbow method" and it was seen that there were two elbows, one at 3 and one at 6. This is the reason we chose 6 clusters to represent our model. From this we received a value which told us how well the students were naturally grouped.

Method for Problem #2:

For this problem we first counted every time a student got a video quiz correct which was denoted with s =1. Afterwards, these rows were grouped via user ID and the sum of the scores were calculated for each group. To technically present this new variable, a new column with the sum of scores for each user ID was created so the data could be grouped better called finalScore. After this was found, we needed to compare with certain variables to see if there was any correlation or if we could find any sort of similarities between variables. For this problem we found the score for each userID from checking the relation with fracSpent and fracPaused using a linear regression model. We then fit the model with the average of fracSpent with the final score to see if we could come to a conclusion about the relationship of the two. This was done with the aid of linear regression, so our score was the r squared value which means the higher the r squared value is, the more chance the two variables are related. Lastly for this problem we found the mean squared error to be certain about our findings.

Method for Problem #3:

For this problem we needed to find how well can we predict a student's performance on one in video quiz question based on video watching behaviors. We belived the best way to see this was through visualization and using scatterplots. For this we visualized the relationship between fracPasued and fracSpent and compared that relationship with the coloum we created in the method for problem two "final_score". From this we could make conclusions about how well a student could do on one particular quiz question based on the time they spent watching the video and the time the video was paused.

Results:

Results for Problem #1:

In this problem after a k means was used with 6 centroids we received a score of 0.957. This score ranges from 0 to 1 and the closer it is to 1 means it is more accurate. This means that our model could group the students extremely well regarding various groups.

Results for Problem #2:

In this problem we concluded that a student's video watching behavior cannot be used to predict a students score. Our regression model gave us an r squared score of 0.0008. The reasoning for this value will be explained in the analysis section. We also calculated a mean squared score of 0.0436. This means that the average score does not provide enough information as to whether the student watched the video or not.

Results for Problem #3:

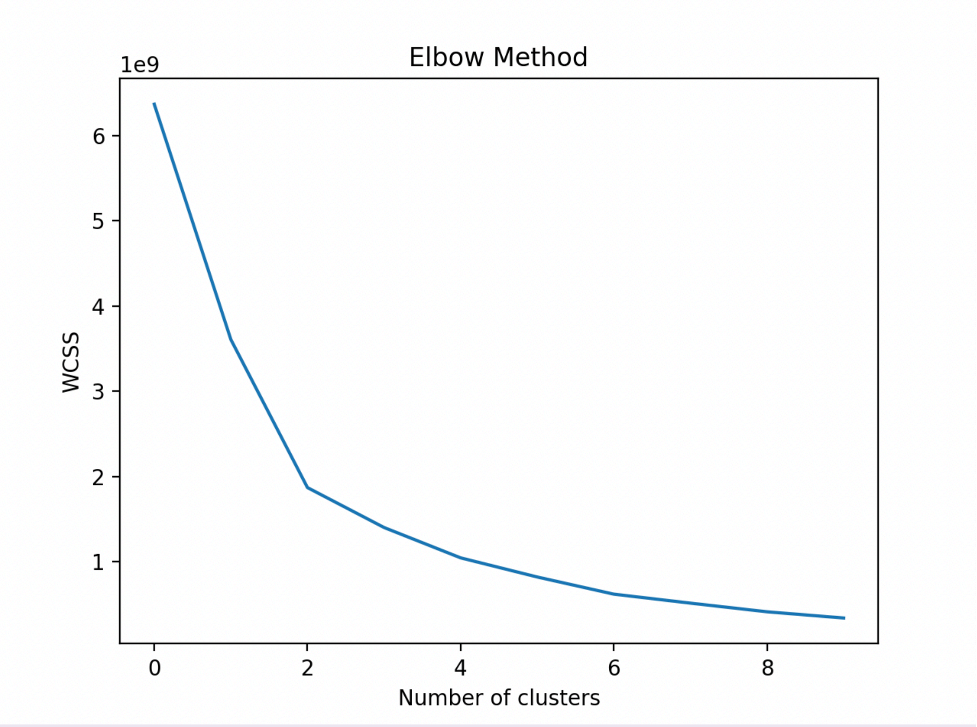
In this problem we saw that we could somewhat predict a student's performance based on their video watching behaviors. The reasoning behind this is because there were too many exceptions. An example is that, multiple students did not watch video number 5 but they still received points for the problem quiz. This skewed the data and did not allow us to come to any concrete conclusions. Like we did in the second problem we checked the variables fracSpent versus final score and fracPaused versus final score and got values that supported the conclusion that video watching behavior cannot be determined.

	FracSpent Vs Final Score	**FracPaused Vs Final Score**
R squared	0.000828	0.001331
MSE	0.043644	0.043622

Analysis:

Analysis for Problem #1:

This part of the problem showed a valid score. We saw a very high accuracy which means that the students can be grouped very accurately based off their video watching behavior. Originally it was seen that even without considering the students that watched less than 5 videos we still got a score of 0.956 but filtering out those students allowed us to gain a more accurate score. It was seen that this was done using the elbow method. The elbow method is a way to find the correct number of centroids that should be used in a k means cluster. We graphed it and saw that there were two elbows, one at 3 and one at 6. When both were tried, it was seen that 6 centroids gave us the best value. Here is the screenshot displaying the elbow method:

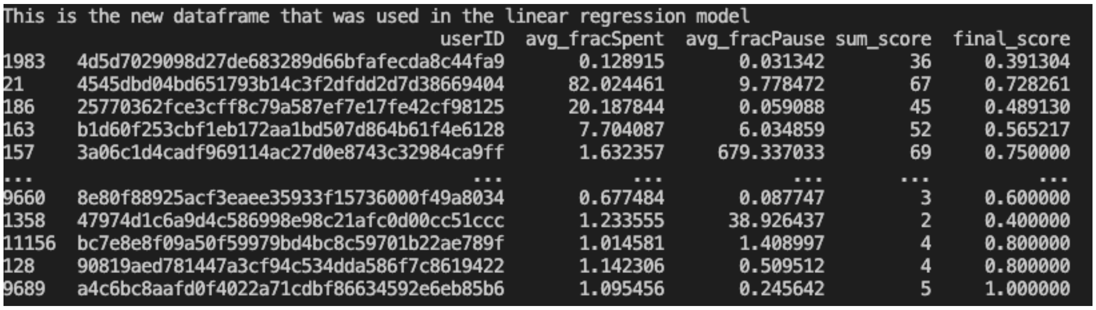


After this was decided we finished making the unsupervised k means model. We had to train the data with given label data used the model to predict on the test data. This elbow method showed us that as we increase the amount of clusters, the less accurate the data is. When we used 7 clusters our accuracy fell to 0.953 and when 8 clusters were used it fell to 0.917. Therefore, from both the elbow method and the training model we can be 95.6% certain we are grouping the students correctly.

6 cluster Accuracy:	0.95709
7 cluster Accuracy:	0.95348
8 cluster Accuracy:	0.91788

Analysis for Problem #2:

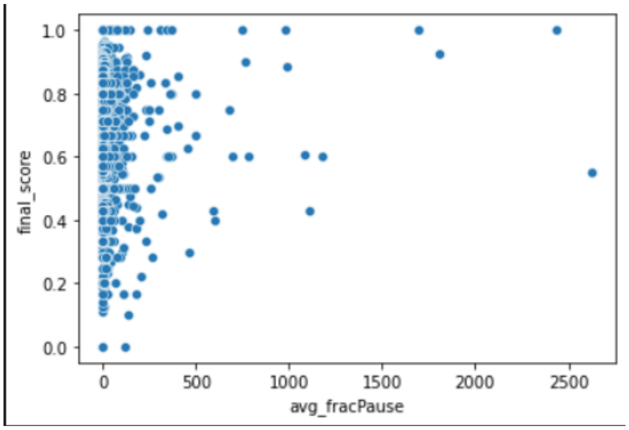
For this problem it was concluded that we are not able to determine whether a student's video watching behavior can predict their score. This is because it requires more factors than simply if they watched the video or not. In the data it was seen that a lot of the times the students were able to score points on the problem quizzes even if they did not watch the video. This causes a skew in the data and does not allow us to conclude valid results. It does allow us to conclude that there are other factors at play. This happened numerous amounts of times which caused us to determine that a valid prediction cannot be made solely of this. Our MSEE value also outputted a similar answer which further supports this conclusion. The updated data frame used to come to this conclusion is shown here below:



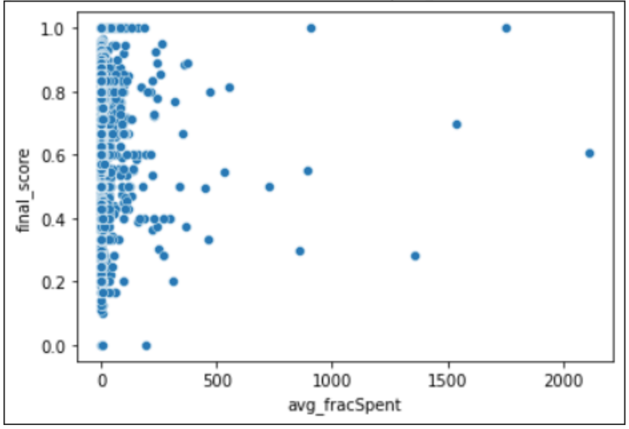
```
This is the new dataframe that was used in the linear regression model
                                         userID  avg_fracSpent  avg_fracPause sum_score  final_score
1983   4d5d7029098d27de683289d66bfafecda8c44fa9       0.128915       0.031342        36     0.391304
21     4545dbd04bd651793b14c3f2dfdd2d7d38669404      82.024461       9.778472        67     0.728261
186    25770362fce3cff8c79a587ef7e17fe42cf98125      20.187844       0.059088        45     0.489130
163    b1d60f253cbf1eb172aa1bd507d864b61f4e6128       7.704087       6.034859        52     0.565217
157    3a06c1d4cadf969114ac27d0e8743c32984ca9ff       1.632357     679.337033        69     0.750000
...                                         ...            ...            ...       ...          ...
9660   8e80f88925acf3eaee35933f15736000f49a8034       0.677484       0.087747         3     0.600000
1358   47974d1c6a9d4c586998e98c21afc0d00cc51ccc       1.233555      38.926437         2     0.400000
11156  bc7e8e8f09a50f59979bd4bc8c59701b22ae789f       1.014581       1.408997         4     0.800000
128    90819aed781447a3cf94c534dda586f7c8619422       1.142306       0.509512         4     0.800000
9689   a4c6bc8aafd0f4022a71cdbf86634592e6eb85b6       1.095456       0.245642         5     1.000000
```

Analysis for Problem #3:

In this part of the problem, we used the final score column that we made which has the valid user IDs and plotted them against the fraction of time spend watching the videos. This gave us a fairly linear scatterplot but once again there are a lot of extraneous data points which skewed our score a lot which is why we cannot be certain that the fraction of time spent leads to a true score, but the trend below seems to suggest so. Therefore, seeing all the datapoints be somewhat in a line we could be somewhat confident in predicting whether a student receives a "1" value based on the graph below based purely off of the variable "fracSpent"



The same trend was done with fraction paused as well. We see a slightly more compact graph but the result is the same as fracSpent so we arrive at the same conclusion as before. Therefore, here we can once again be fairly sure about a student's result based on the amount of time it was paused. (see below:) (*These graphs can be found on the .ipynb file*)


final_score
avg_fracSpent