Omar Tounsi (otounsi)
Logan Stapleton (stapletl)
ECE 20875
Miniproject report
4/21/2021

Miniproject report

Path 2: Student performance related to video-watching behavior

Introduction

The purpose of this study is to analyze and draw conclusions from a dataset of student video-watching behavior and quiz scores for an online course. This dataset contains information such as fraction of time spent, number of pauses and playback speed for each student-video pair as well as the corresponding quiz score. Our goal is to identify how these parameters play a role in determining the student's performance in the course.

Objectives:

- To determine how well the students can be clustered by their video-watching behavior.
- To determine whether video-watching behavior can be used to predict a student's general performance.
- To determine whether video-watching behavior for a particular quiz can be used to predict a student's specific performance on that quiz.

Description of the dataset

This dataset contains 29304 datapoints. Apart from the student and videos IDs, each data point has 8 features, namely:

- fracSpent: the fraction of time the student spent watching the video relative to the length of the video.
- fracComp: the fraction of the video the student watched.
- fracPaused: the fraction of time the student spent paused on the video, relative to the length of the video.
- numPauses: the number of times the student paused the video.
- avgPBR: the average playback rate that the student used while watching the video.
- numRWs: the number of times the student skipped backwards in the video.
- numFFs: the number of times the student skipped forward in the video.
- S: the quiz score of the student.

There are two additional features in the dataset, fracPlayed and stdPBR, which we will not consider in this study.

Omar Tounsi (otounsi)
Logan Stapleton (stapletl)
ECE 20875
Miniproject report
4/21/2021

We start by computing the basic descriptive statistics of each feature:

	fracSpent	fracComp	fracPaused	numPauses	avgPBR	numRWs	numFFs
Mean	24.07652	0.7678704	36.3401342	2.8255187	1.1044	2.238022	1.567602
Std dev	308.2705	0.3405872	375.75223	59.1009056	0.3156	15.56432	6.369905
Max	18215.98	9.2887314	15957.392	10083	2	2237	309
Min	0	0	0	0	0	0	0

Figure 1: Descriptive statistics

We observe that certain features have very standard deviations. This is due to the presence of certain disproportionately large outliers in the data. These outliers can potentially harm the data analysis since some of them clearly indicate an abnormal video watching behavior. For example, a fracSpent value that is in the order of the hundreds is likely the sign that the student has physically left the watching station for an extended period of time.

As such, we will determine and apply value thresholds for each feature in order to filter out possible outliers and discard them from the dataset. After considering the statistics and reviewing each feature's data carefully, our team has determined the following ranges as acceptable data:

	fracSpent	fracComp	fracPaused	numPauses	avgPBR	numRWs	numFFs
Range	[0 , 5]	[0 , 1.1]	[0 , 60]	[0 , 60]	All data	[0 , 30]	[0 , 25]

Figure 2: Data thresholds

In addition to these thresholds, we will limit our analyses in the upcoming sections to only certain datapoints of the dataset, depending on the number of video quizzes completed by the student.

Omar Tounsi (otounsi)
Logan Stapleton (stapletl)
ECE 20875
Miniproject report
4/21/2021

Student clustering based on video-watching behavior

Our objective is to determine how well the students can be clustered by their video-watching behavior. We will only use students who have completed at least five of the videos in this analysis.

Analysis

We will use k-means clustering as our analysis. We choose k-means because we do not start with an assumed model or distribution of the data. Furthermore, we consider that a simple, centroid approach method is sufficient for this study, particularly since the quiz scores of the students are either 0 or 1.

We define the four following initial cluster centers based on the level of interaction the student has with the video:

	fracSpent	fracComp	fracPaused	numPauses	avgPBR	numRWs	numFFs
Zero	0	0	0	0	0	0	0
Low	0.3	0.25	1	1	0.5	1	1
Med	0.7	0.5	5	2	1	2	2
High	1.2	1	30	5	1.5	5	5

Figure 3: Initial cluster centers

The zero level denotes a non-existent student interaction with the video. It groups students who do not watch the video.

The low level denotes a minimal student interaction with the video. It groups students who watch a small portion of the video at a half playback rate, with short and rare pauses, rewinds and fast-forwards.

The medium level denotes an average student interaction with the video. It groups students who watch most but not all of the video at a normal playback rate, with a moderate number of medium-length pauses, rewinds and fast-forwards.

The high level denotes a high student interaction with the video. It groups students who watch the entirety of the video at a high playback rate, with long and frequent pauses, rewinds and fast-forwards

We expect the coordinates of the centers to be updated during the clustering process, but to preserve all or most of their above respective characteristics.

Omar Tounsi (otounsi)
Logan Stapleton (stapletl)
ECE 20875
Miniproject report
4/21/2021

Results

After running clustering.py, we obtain the following output. The decimal values were rounded, and the coordinates of the non-center points have been omitted for display convenience:

Cluster: 18529 points and center = [0.8226, 0.7947, 0.3865, 1.4188, 1.1138, 0.5303, 0.6123]

Cluster: 2299 points and center = [1.1188, 0.6200, 1.3030, 6.7777, 1.1182, 9.2292, 3.9830]

Cluster: 548 points and center = [1.0424, 0.7340, 17.0264, 4.0164, 1.1232, 2.1405, 1.5018]

Cluster: 325 points and center = [1.0829, 0.7291, 44.1946, 5.2338, 1.1039, 2.7815, 1.8953]

Figure 4: Final cluster centers

The coordinates of the third and fourth centers, the medium and high interaction levels respectively, have not considerably shifted from their initial values. However, the coordinates of the first and second centers, the zero and low interaction levels respectively, have changed so significantly that they no longer describe the video-watching characteristics they were designed to represent.

At this point, we will stop considering the group of students who do not watch the video as a separate cluster. Instead, we will change the definition of the first center to encompass both zero and low levels, bringing the number of clusters to three:

	fracSpent	fracComp	fracPaused	numPauses	avgPBR	numRWs	numFFs
Zero-Low	0.9	0.6	1	2	1	0.5	0.5
Med	1	0.7	15	3	1.05	2	2
High	1.1	0.8	35	5	1.1	6	4

Figure 5: Updated initial cluster centers

After running clustering.py, we obtain the following output. The decimal values were rounded, and the coordinates of the non-center points have been omitted for display convenience:

Cluster: 18978 points and center = [0.8250, 0.7919, 0.5577, 1.4471, 1.1146, 0.5687, 0.6744]

Cluster: 2219 points and center = [1.1439, 0.6286, 1.9755, 7.1969, 1.1145, 9.4790, 3.7160]

Cluster: 504 points and center = [1.07370, 0.7278, 37.46649, 4.642, 1.1057, 2.5674, 1.7777]

Figure 6: Updated final cluster centers

We observe that:

- The datapoints are overwhelmingly centered around the zero-low level center. It is followed in importance by the medium level cluster, and by the high level cluster.

- The zero-low level cluster groups datapoints with the lowest feature values, except for fracComp for which it has the highest average value.

- The medium level cluster groups datapoints with the highest fracSpent, numPauses, numRWs and numFFs as well as the lowest fracComp values.

- The high level cluster groups datapoints with the highest fracPaused values. This feature is the one for which we observe the highest increase from cluster to cluster.

Conclusions

The students from the dataset can be split into three different groups:

- The 87% majority of the students demonstrate a level of interaction with the videos that can be described as zero to low when compared with the rest. Although these students are the ones who generally watch the highest portion of a video, they fall short in all of the other behavior categories, particularly the total time they spent on a video as well as the length and frequency of their pauses.

- A 10% minority of students demonstrate a medium interaction level with the videos. This group of students spend the most time on the videos, pause the most frequently and use the rewind and fast forward functionalities the most. However, this group of students generally watch the lowest portion of a video.

- A 2% minority of students demonstrate a high interaction level with the videos. These students have average scores in most behavior categories which generally place them between the previous two groups. However, they are set apart from the rest by their superior scores in the pause length category, which is the feature with the highest variance in the dataset.

Omar Tounsi (otounsi)
Logan Stapleton (stapletl)
ECE 20875
Miniproject report
4/21/2021

Video-watching behavior and general quiz performance

Our objective is to determine whether video-watching behavior can be used to predict a student's general performance. We will only use students who have completed at least half of the quizzes in this analysis.

Analysis

We will use linear regression to find a model that describes the relationship between general quiz performance and video watching behavior. We will consider several different models intended for predicting the average score for each student, each model using either one or a combination features as input.

The training and testing data were normalized prior to the regression, a 90/10 split was used. We applied ridge regression with a regularization constant of 10.

We expect to generate at least one model with a high enough r squared value to consider its feature inputs as correlated to average score.

Results

Our first attempt is to produce a single model which uses all of the seven features as inputs. We obtain the following r squared value for the model:

Input features	All features
r squared	-0.151

Figure 7: r squared value for the seven-features model

Since this model yields a negative r squared value, we discard it as non-valid.

Our second attempt is to build models using one feature each as single input. We obtain the following r squared values:

Input feature	fracSpent	fracComp	fracPaused	numPauses	avgPBR	numRWs	numFFs
r squared	-0.191	-0.211	0.006	0.148	-0.178	-0.078	-0.220

Figure 8: r squared values for each one-feature model

Only two out of the seven models yield a non-negative r squared value. We select the model using numPauses for testing because it has the highest r squared value.

Omar Tounsi (otounsi)
Logan Stapleton (stapletl)
ECE 20875
Miniproject report
4/21/2021

The coefficients of the "numPauses" model are the following. Decimal values were rounded:

Slope: 0.04788
Intercept: 0.63962
Figure 9: "numPauses" model coefficients

We observe that the slope is very small, and the intercept is large considering the scale is only from one to zero for the average scores. This means that, despite its relatively higher r squared value, the linear regression model did not find that the numPauses feature had much impact on the scores.

Using the model, we calculate the following score values for seven random students:

Predicted Value	0.646443	0.627431	0.659785	0.63436	0.62291	0.59545	0.636	0.64897
True Value	0.83098	0.80487	0.55555	0.80487	0.64814	0.5375	0.78571	0.41818

Figure 10: Predicted vs True average score values

These seven predicted-true value pairs have an average mean squared error of 21.60%, which is a figure way too high for us to describe the model as accurate.

Conclusions

We started our linear regression analysis by building a seven features model. This model yielded a negative r squared value, which showed that the seven features grouped together in a single model were not a good predictor of average score.

We then built seven different models, each using a single feature as input. Although two of the produced models had non-negative r squared values, none had an r squared value greater than 0.5, which we considered as the minimum acceptable value to consider any correlation with average score.

As such, we conclude that no feature or group of features present in this dataset is a good enough predictor of student performance.

Omar Tounsi (otounsi)
Logan Stapleton (stapletl)
ECE 20875
Miniproject report
4/21/2021

Video watching behavior and specific quiz performance

Our objective is to determine whether video-watching behavior for a particular quiz can be used to predict a student's specific performance on that quiz. We will use all student-video pairs in this analysis.

Analysis

We will use logistic regression to create a model for predicting individual quiz scores based on a particular video's watching behavior features. The choice of logistic regression is optimal since it is the most appropriate data analysis technique for predicting zero or one values. We will consider several different models intended for predicting a student's score, each model using either one or a combination features as input.

The process for data was similar to that of part two. The data was normalized prior to the regression and there was a 90/10 split for the training and testing data.

We expect to generate at least one model with a high enough r squared value to consider its feature inputs as correlated to score.

Results

Our first attempt is to build models using one feature each as single input. We obtain the following pseudo r squared values:

Input feature	fracSpent	fracComp	fracPaused	numPauses	avgPBR	numRWs	numFFs
Pseudo r squared	-0.067	-0.067	-0.066	-0.067	-0.067	-0.067	-0.066

Figure 11: Pseudo r squared values for each one-feature model

Since these models all have negative pseudo r squared values, we discard them as non-valid.

Our second attempt is to build models using combinations of features as inputs. We choose to group certain features based on the r squared values of their linear regression models obtained in figure 8 as well as the pseudo r squared values of their logistic regression models obtained in figure 11. We obtain the following pseudo r squared values:

Input features	All features	fracPaused, numPauses, numFFs	fracPaused, numPauses
Pseudo r squared	-0.058	-0.062	-0.064

Figure 12: Pseudo r squared values for each multi-feature model

These models also all yield negative pseudo r squared values. As such, we will select the model using all seven features for testing because it is the one with the highest r squared value.

Input feature	fracSpent	fracComp	fracPaused	numPauses	avgPBR	numRWs	numFFs
Coefficient	-0.0412	-0.0555	0.1126	-0.1118	-0.0212	0.1109	-0.1383

Figure 13: Logistic regression model coefficients

Similarly to the linear regression model, the coefficients of the logistic regression model do not indicate a strong correlation between the features and the quiz score.

Using the model, we calculate the following score values for ten random students:

Predicted	0	1	0	0	0	1	0	0	0
True	1	0	1	0	0	1	1	1	1

Figure 14: Predicted vs True score values

We observe that the model predictions have virtually no bearing on what the actual values are.

Conclusions

We used a similar procedure for this logistic regression analysis as we did for linear regression. We started by building models using one feature each as single input. The negative pseudo r squared values that were obtained showed that none of the features were individually good predictors of quiz score.

We then decided to look at multi-feature models, building three models which used various combinations of features as inputs. After calculating their pseudo r squared values, it appeared that the model with all seven features as inputs had the lowest pseudo r squared value. That being said, that value was negative. This meant that the model was unable to explain the variance of the data and as such did not lend itself well to predicting scores

That the data logistic regression model falls short of predicting individual score performance is not surprising, considering that the linear regression model used in the previous section was also unable to accurately predict average student score.

The failure of the model may also be explained by the size of the dataset that the model is exposed to. Compared with the other two analyses of this study, which filtered out certain student-video data pairs from the dataset based on the number of quizzes taken, this analysis used all student-video pairs present in the dataset. This increased the probability of outliers being present in the data.

In conclusion, video-watching behavior is not a good predictor of a student's performance on a particular in-video quiz.