Rushabh Ramesh Ranka

PROJECT REPORT

Description of Dataset (Section 2):

I choose a dataset of the student performance with respect to the video watching behavior. The data contains 11 columns that contain student id, video id, 9 columns of features related to video watching and a 12th column containing the output column. The output is the score the student gained for a one point in-video quiz question. The score has a value of 1 for correct and 0 for incorrect. The description of each feature about the video watching behavior is given in the readme.pdf except for stdPBR (standard deviation of playback rate) and fracPlayed (the total length of the video watched that was on play model divided by the length of the video.

Methodology of Analysis (Section 3):

Question 1:

Question was “How well can the students be naturally grouped or clustered by their video watching behavior (fracSpent, fracComp, fracPaused, numPauses, avgPBR, numRWs, and numFFs)? Using data of students that complete at least five of the videos in your analysis.” To solve this problem, I used the k-means algorithm as this is an unsupervised learning problem.

The question itself also calls for a grouping method. I chose k-means as the data clustered in few areas. In addition, as the score had a value of 0 or 1, I chose two clusters for the visual plots.

Each X feature value was individually compared with the output scores to compute the clusters. This helped in identifying the outliers caused by the specific X feature. If the outliers were too much, the clustering was regarded as a poor match. If the 2 clusters were visually distinguishable then the match was good. A good cluster also implied that there was a noticeable change in the predicted score.

Question 2:

Question 2 was “Can student's video-watching behavior be used to predict a student's performance (i.e., average scores across all quizzes)? Use all students that complete at least half of the quizzes in your analysis”. In order to solve this problem, I chose the logistic regression method. This method seemed quite logical as the score values (y value) were either a 0 or a 1. Since only these two classes exist, it doesn’t seem logical to use a method like linear regression as we use linear regression for interpolating between the points. After computing the prediction model, I analyzed the efficacy of my model using the confusion matrix. The confusion matrix gave the complete picture about the accuracy and precision of the model.

Question 3:

Question 3 was, “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?. Use all student pairs in the analysis” In order to solve this problem, I chose the naïve bayes model as this model allowed to assume different probability distributions for the quiz scores. These models could then be evaluated for their accuracy and precision in predicting the quiz scores. Hence, a confusion matrix was used to evaluate the efficacy of each model.

Results (Section 4):

Question 1:

As stated in the methodology of analysis, I used a k-means clustering algorithm to analyze the problem. The following plots will evaluate the relationship between the specific video watching behavior and the quiz score.

Figure 1

Chart, scatter chart

Description automatically generated

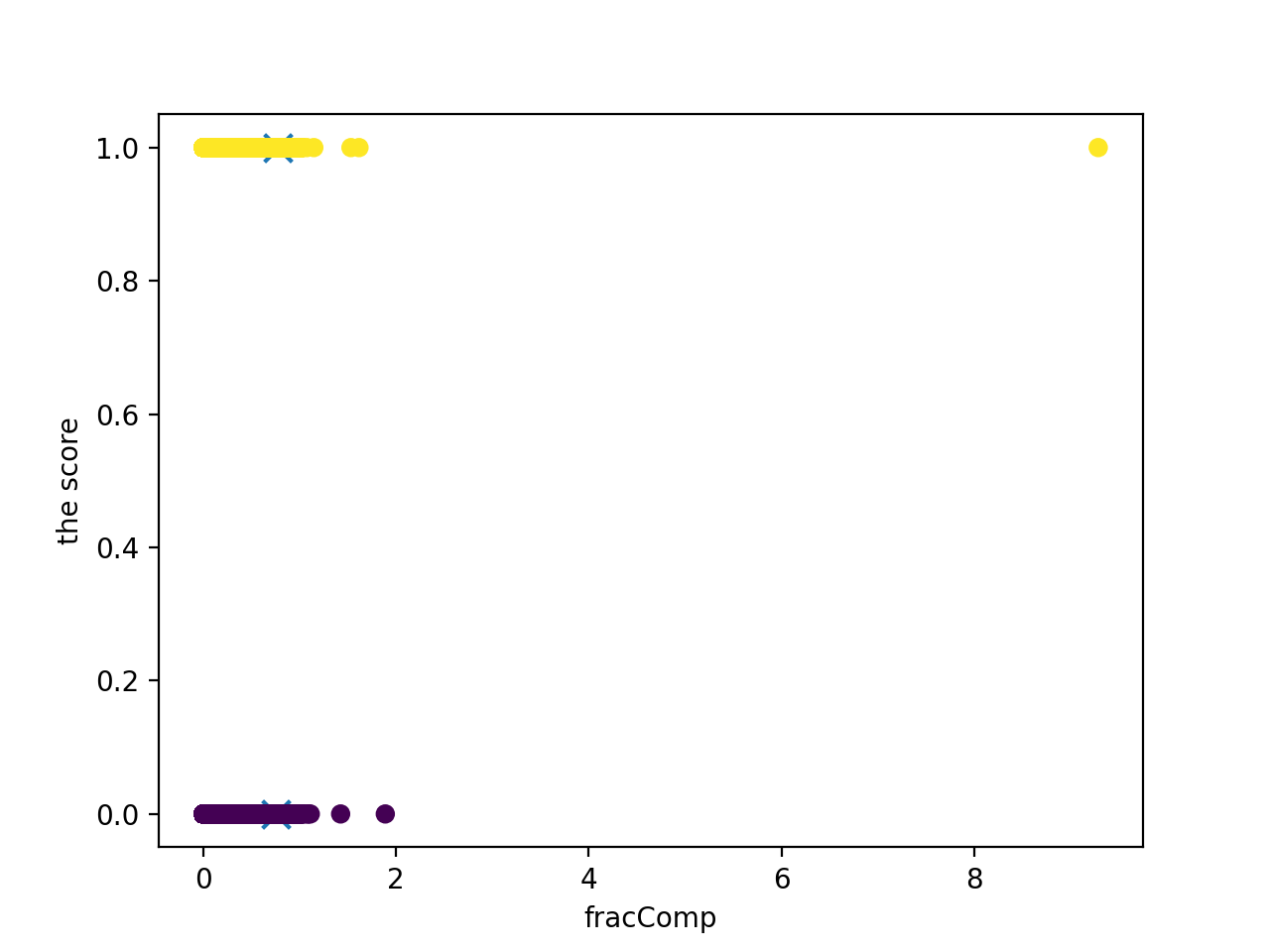
Cluster 1: 1.53293711e+01 , 6.68499735e-01

Cluster 2: 7.89598922e+03 , 4.64285714e-01

In this plot, fracSpent is the length of time the student spends watching the video divided by the total length of the video. There is a chance for lot of anomalies in this situation as a student might spend hours playing, rewinding and pausing the video. Looking at the plot, most of the cluster values are around the 0 value, however, some of the fracSpent go as far as 17,500. Such values seem to be anomalies as most fraWatched hover around 1.

In conclusion, students in the shorter interval (fracSpent = 15) do around 20% better than students in the larger interval(fracSpent = 7890). 20% is quite a low number and hence this clustering can be considered to be poor.

Figure 2



Cluster 1: 7.69662703e-01 1.00000000e+00

Cluster 2: 7.55440068e-01 -8.32667268e-15

Looking at the FracComp clusters, it seems that the clustering looks poor. A 1.5% change in value of fracComp tells whether the score is 0 or 1. Such a drastic jump makes this clustering a poor one for predicting the score.

Figure 3

Text

Description automatically generated

Cluster 1: 2.21600907e+01, 6.68161252e-01

Cluster 2: 7.94977986e+03 ,7.27272727e-01

The clustering for fracPaused is poorly grouped. This is evident in the fact that a video was paused for 15,000 times longer than the video’s length. Such anomalies can only be explained when a student takes a long break. From the fracSpent we saw that a long time spent should yield a low score, however the fracPaused cluster proves the opposite.

Figure 4

Chart

Description automatically generated

Cluster 1: (2.45841471e+00 6.68253581e-01)

Cluster 2: (1.00830000e+04 1.00000000e+00)

By visual inspection itself, it can be deduced that cluster 2 has outliers. This is again poor clustering as one cluster is n – 1 points and the other is only 1 point.

Figure 5

A picture containing chart

Description automatically generated

Cluster 1: (1.12044546e+00 , 1.00000000e+00)

Cluster 2: (1.09654982e+00, -8.32667268e-15)

The above clustering for avgPBR looks like a poor clustering. There are similar average playback rates yet cluster 1 and 2 show a different ore value. Hence, there is no correlation between avgPBR and the score.

Figure 6

Chart

Description automatically generated

Cluster 1: (2.11242676e+00 6.68294271e-01)

Cluster 2: (2.23700000e+03 0.00000000e+00)

It can clearly be seen from the graph that cluster 2 is located at an outlier point. If cluster 2 point didn’t exist, then perhaps the clustering could be good. However, in conclusion, the clustering is poor.

Figure 7

Chart

Description automatically generated

Cluster 1: (0.94737714 0.6689318)

Cluster 2: (34.08126411 0.63205418)

The above clustering does not show any relation between the number of fast forwards and a student’s score. There is a weak trend wherein the more you fast-forward the score decreases. However, in general the correlation is poor.

Question 2:

The question was asking whether it is possible to predict a student’s average score across quizzes using the video watching behavior. One of the criteria for this question was to take into account only students who had watched half of the videos. In addition, it is important to note that the score for a quiz has a value of 1 or 0. As a result this problem becomes a classification problem. Hence, to make a prediction model for this data, a logistic regression approach was used.

In this logistic regression approach, the X values were divided into train and test data with test size being 25%. The model obtained had the coefficients y = -0.01590809(x1) -0.06395957(x2) +0.26940183(x3) -0.00493074(x4 )+ 0.1921098(x5) -0.02365084(x6) + 0.08106618(x7) + 0.09475089(x8) -0.14089043(x9) + 0.60214932.

To evaluate the efficacy of this model, a confusion matrix was computed to deduce the precision and accuracy of the model. The results are depicted below.

|  |  |  |
| --- | --- | --- |
|  | Positive(Predicted class) | Negative(Predicted class) |
| Positive ( Actual) | 11 | 593 |
| Negative (Actual) | 6 | 1053 |

Using these results the accuracy came to be around 63.98% whereas the precision was 63.97%.

The accuracy of this model seems to be quite reasonable and logical. I say this because the parameters in this model are quite prone to anomalies and also some of these parameters might not matter in predicting the score. For example, the amount of time a student spends on the video might not matter as it is how much information the student grasps through the video that matters. In addition, since the quiz is just one question, we need to understand that some students might get lucky as well. In conclusion, I can say that no, we cannot predict a student’s performance through their video watching behavior. This is a subjective matter which needs to take into matter other external factors as mentioned above.

Question 3:

The question was asking “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?”. Using the nayes bayes approach, different models were assumed as the probability distributions for the feature matrix. This helped in evaluating various models and deducing the most accurate model. Hence, a confusion matrix was used to measure the accuracy and summarize the results of the models.

The first model used was a Bernoulli one. Following is the confusion matrix obtained.

|  |  |  |
| --- | --- | --- |
|  | Positive(Predicted class) | Negative(Predicted class) |
| Positive ( Actual) | 96 | 2262 |
| Negative (Actual) | 132 | 4836 |

This model yielded an accuracy of 67.32%.

This model was followed by a Gaussian model. Following is the confusion matrix obtained.

|  |  |  |
| --- | --- | --- |
|  | Positive(Predicted class) | Negative(Predicted class) |
| Positive ( Actual) | 1566 | 792 |
| Negative (Actual) | 3329 | 1639 |

This model yielded an accuracy of 43.75%.

This model was followed by a logistic regression model. Following is the confusion matrix obtained.

|  |  |  |
| --- | --- | --- |
|  | Positive(Predicted class) | Negative(Predicted class) |
| Positive ( Actual) | 1 | 2357 |
| Negative (Actual) | 2 | 4966 |

This model yielded an accuracy of 67.80%.

This model was followed by a multinomial model. Following is the confusion matrix obtained.

|  |  |  |
| --- | --- | --- |
|  | Positive(Predicted class) | Negative(Predicted class) |
| Positive ( Actual) | 113 | 2245 |
| Negative (Actual) | 243 | 4725 |

This model yielded an accuracy of 66.04%.

I would select the logistic regression model to predict the scores as it yields the highest accuracy of 67.80%. This value seems plausible as it ties back to the dataset not being a strong indicator for performance based on the video watching behavior. Since the accuracy is low, it seems that the large portion of the data does not correlate with the score. If there was a good correlation between the score and the other parameters, the accuracy would be higher.

Lastly, it can also be ruled out that underfitting occurred. If we underfitted our model, we would have an accuracy lower than 50% for testing and training. No likely overfitting also occurred as that would imply a very high accuracy.