Path 2: Student Performance Related to Video-watching Behavior

Brendan McLaughlin
12/11/2022

The dataset given for this path contains 12 total parameters, which are listed in Table 1 along with a description.

Table 1: Dataset features with description

Parameter	Description
userID	Unique alphanumeric tag that assigns the data to a given user
VidID	ID number of the video watched, ranging from 0 to 92
fracSpent	Fraction of time spent by the user on a given video over the video length
fracComp	Fraction of the video that the user watched
fracPlayed	Fraction of the video that the user watched, can be >1 if parts of the video are rewatched
fracPaused	Fraction of time the user spent paused in relation to the length of the video
numPauses	Total number of times the user paused the given video
avgPBR	Average playback speed of the video, between 0.5x and 2.0x
stdPBR	Standard deviation of the playback rate
numRWs	Total numbers of times the user rewound the video
numFFs	Total number of times the user fast-forwarded the video
s	Whether student was correct (s=1) or incorrect (s=0) on their first question attempt

This data set contains roughly 25000 points, averaging 267 data points per video and just under 4000 individual users.

The first analysis of the dataset aims to determine whether the users in the dataset can be effectively grouped into categories based on their video-watching characteristics. The

method used to perform this analysis is to fit a Gaussian mixture model to each video dataset with a range of cluster amounts. In order to determine which number of clusters fits the dataset best, the Bayes Information Criterion (BIC) can be used as a measure of how well the model with a given number of clusters predicts the dataset. This value can be minimized to find the theoretical best number of clusters, although this runs the risk of creating a situation where having more clusters is always more mathematically accurate, up to one cluster for every point in the dataset. However, if the data has some meaningful data about how to group the subsets, this will not be an issue as increasing the number of central points will just split the actual groups that do exist and increase the BIC.

The second step of the analysis is to determine whether an individual user's overall video-watching characteristics can be used to make a prediction about how well they will score on the end-video questions on average. The optimal model to test if this is possible is a linear regression model trained on the average characteristics of the users to output the average user score. Creating this model first requires the total data to be separated by individual user, and then for each user's data to be averaged in its entirety to create a single piece of data. This set of user averages must be then normed before it can be used to train the linear regression model. To ensure a generalizable model, the regularization parameter of the regression model must be varied in order to find the model with the highest fit, which is measured by the mean squared error of the predicted outputs. Finally, the coefficient of determination of the model in relation to the dataset can be used as a numeric basis to determine how well the model can predict the target values within the true data.

The final analysis goal is to determine the possibility of using the dataset for each individual video to predict whether a user will answer the question correctly or incorrectly based on their video-watching behavior for that particular video. Given that there are only two possible outcomes that a model could predict, a classification algorithm will be most effective to predict the outcome of an inputted datapoint, and a logistic regression model offers this type of binary output while remaining relatively calculation-light. To train the model, the dataset will again be separated into one set of data for each video, and each video dataset will be normalized for model training. Once each model has been trained, the accuracy score of each model can be used to show how well the trained logistic regression model fits the data, and the mean of these accuracy scores can show how the model output behaved overall when using a logistic model as a prediction method.

Results:

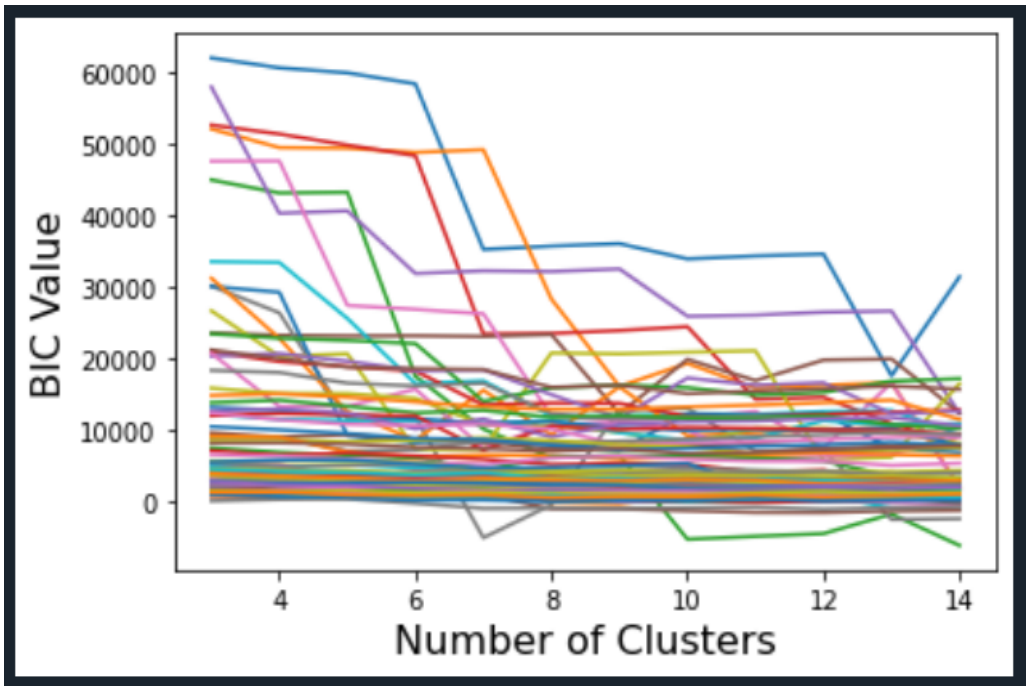

Figure 1: Bayes Information Criterion value plotted against an increasing number of clusters tested for each video dataset

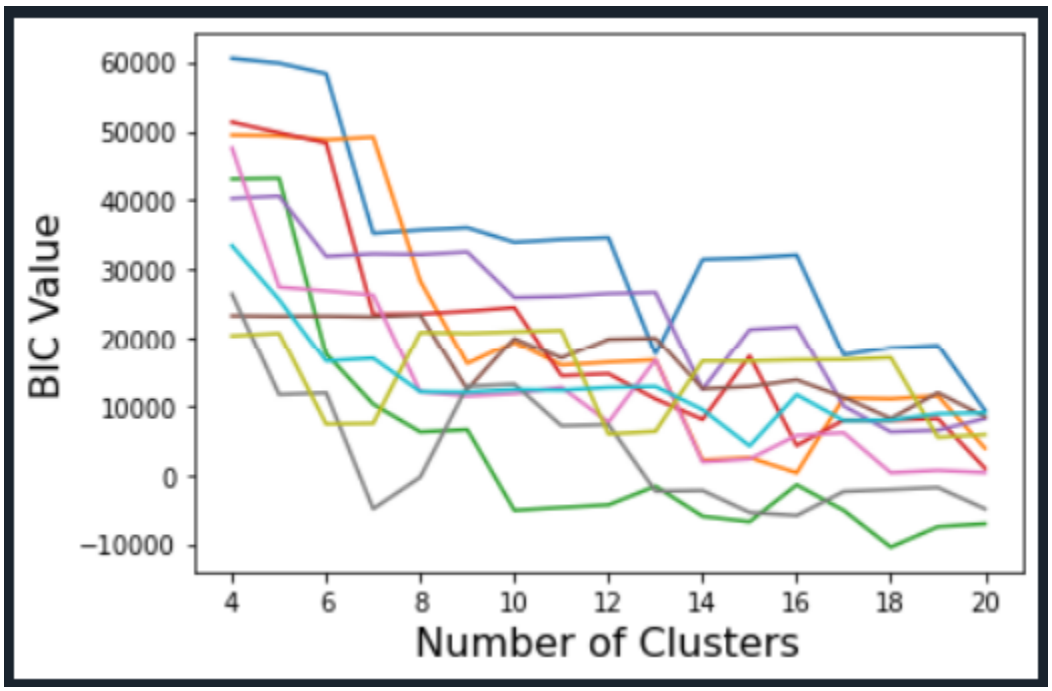

Figure 2: Subset of BIC vs. number of cluster graphs for videos ID 0 to ID 9

Figure 1 shows how for each video dataset the Bayes Information Criterion, the measure of fitment for a given number of means, generally decreases as the number of clusters

increases. This shape is better shown in the graph of the subset shown in Figure 2, but it is clear that not every best fit occurred at the highest quantity of Gaussian means. On average the best number of means was 15.64 separate means, with a standard deviation of 3.98. Based on this data there was a considerable amount of difference in the best quantity of mean, which is to be expected as some questions are likely easier or harder, either grouping the population together or spreading it out across a wider range. This outcome is not that unexpected, as the natural grouping of students or users of this type of service is by letter grades, which following the conventional A+, A, A-, …, F grading scale has a total of 14 groupings. This similarity provides evidence that this data provides meaningful grouping data on a video-by-video basis, and that the Gaussian mixture mean analysis as a whole was a useful avenue of data analysis.

The application of the linear model in order to analyze whether the average score of a user could be predicted based on their video-watching behavior resulted in a model with a coefficient of determination of $r^2 = -0.00162$, which is extremely low. This extremely low prediction score shows that the data cannot be used to predict the average user score by itself. Figure 2 shows the prediction data plotted against the true data, which in an accurate model will be only slightly off of the y =x line. The fitted linear model clearly does not resemble this, and from the plot it is clear that the model predicts nearly all of the average scores to be near 0.5, with impossible outliers of predicted average score above 5. Differing subsets of the data were also fitted to the data using the same linear regression model, with the same result. While it is possible that the data could be used in some form of linear regression analysis to predict average user score, additional transformation of the dataset would be required as the current dataset for each video is not conducive to a linear regression model. Removal of outliers could also increase the accuracy of the linear regression prediction model, although a few outliers are unlikely to skew a model of 25000 data points very much.

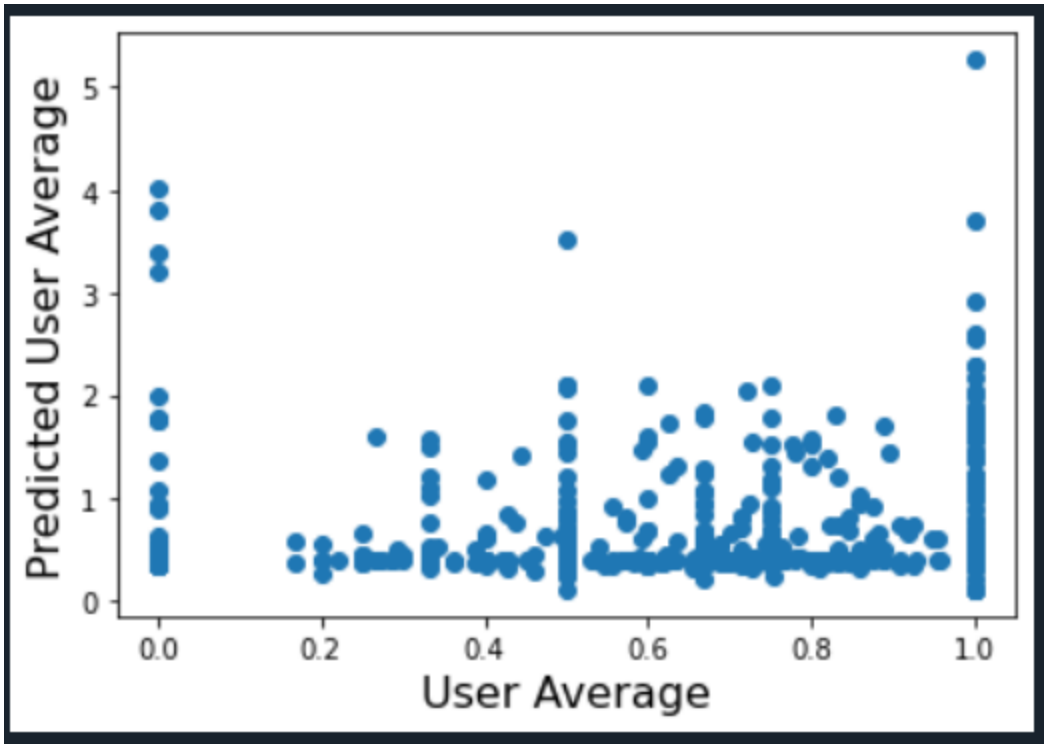

Figure 3: Plot of predicted average score versus actual average score

The logistic regression model designed to predict whether a user answered correctly or incorrectly had similar results to that of the previously discussed linear regression prediction model, however as it applied to videos on an individual basis, there was a large amount of variance in the accuracy of the prediction. For the video with ID 46, the logistic regression model was able to predict a correct or incorrect response with an accuracy of 0.78, but for video ID 80, the specific model only had an accuracy of 0.25. The mean prediction accuracy across all video-specific logistic models equalled only 0.514, much too low for a prediction model to be considered useful. Based on the maximum accuracy value, there may be some qualifiers in a given video's dataset that show it can effectively be predicted using a logistic regression, or some subset and transformation that can make this form of analysis more accurate for the provided data. However, based on this basic form of analysis on the total dataset, the particular outcome of a student's response on a specific question cannot be predicted accurately enough for the data to be useful for this purpose.