Section. __2______ Instructor ___Yi Ding_____ Date __04/26/24_____

ECE 20875 : Introduction to Python for Data Science

Final Project

Path 1 : Analysis of Video Watching Behaviors

Manav Jhaveri, Michael Lollino

Jhaveri0, mlollino

Github Link : https://github.com/ECEDataScience/miniproject-s24-manavjhaveri5

Backup Link : https://github.com/manavjhaveri5/ECE20875_MiniProject

Instructor's Comments:

Data Set Description

The file behavior_performance.txt contains data for an online course on how students watched videos and how they performed on in-video quizzes. It tracks a variety of video-watching behaviors, such as the fraction of video watched (fracSpent), completion rate (fracComp), frequency of pauses (fracPaused), number of pauses (numPauses), average playback rate (avgPBR), number of rewinds (numRWs), and number of fast-forwards (numFFs). These indicators collectively provide a nuanced portrait of student interaction with the educational material. Additionally, the dataset records performance outcomes via students' average scores on in-video quizzes, offering a quantitative basis for correlating engagement patterns with academic success. There are a total of 29305 data points in this dataset, and 92 unique videos. Our goal was to analyze the dataset by focusing on the unique users (grouped by their userID) and gauge their performance on the quiz through their video watching behaviors.

This data will be leveraged to explore natural groupings within the student population based on viewing behaviors, employing methods like KMeans clustering or Gaussian distribution analysis. Furthermore, it will facilitate predictive modeling to ascertain the extent to which these behaviors can forecast overall quiz performance and individual question-by-question outcomes, aiming to uncover any potent predictors of student success within the digital learning environment.

Method

Our general methodology in this project was to separate the data into clusters that make it easy to generalize predictions and analyze video watching behaviors on a macro scale. We then focused on implementing prediction algorithms to predict the performance of the users based on their video watching behaviors. For overall accuracy, we utilized multiple checks for cross-validation, such as using silhouette scores to benchmark our clustering, using multiple different prediction algorithms.

To methodically analyze the behavior_performance.txt dataset, we will adopt a multi-phased approach:

Data Cleaning: We began by cleaning the data to handle missing values, ensuring all users have complete data for at least five videos.
Feature Selection: We then extracted the relevant features for analysis: fracSpent, fracComp, fracPaused, numPauses, avgPBR, numRWs, and numFFs.

Normalization: Lastly, we applied StandardScaler to normalize the features, preparing them for clustering and predictive modeling.

Question 1

We then applied the KMeans clustering algorithm to the normalized dataset, where the silhouette score—a measure of how similar an object is to its own cluster compared to other clusters—guided our selection of three as the optimal number of clusters. This choice was later validated by the Gaussian Mixture Model, which also favored three clusters as indicated by the lowest Bayesian Information Criterion (BIC).

Question 2

The analysis we used for question 2, "Can students' video-watching behavior be used to predict a student's performance?" is ridge regression. This is because ridge regression reduces overfitting and it performs automatic regularization. We also chose this analysis because ridge regression can be used when the linearity of the data is uncertain, which is the case for the data in this path. By using the video watching behaviors as the feature matrix and the score, 's' as the predictive feature, we can perform ridge regression to create a model that analyzes the two. Once the model is constructed, for each video, we will analyze the MSE and coefficient of determination in order to determine the accuracy and predictability of the model. We hope to find that the MSE is a relatively low value while the coefficient of determination is relatively high.

Question 3

The analysis we used for question 3. "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?" is logistic regression. Logistic regression was used because it is very effective when predicting binary outcomes, such as the one specified in this question where the student is either correct (s=1) or incorrect (s=0). The 'accuracy' parameter is a good indicator for whether the logistic regression model performs well. Therefore, we will use the average accuracy value, using all accuracy values for each video, in order to determine if a student's performance on a particular in-video quiz question can be predicted.

Results

Question 1

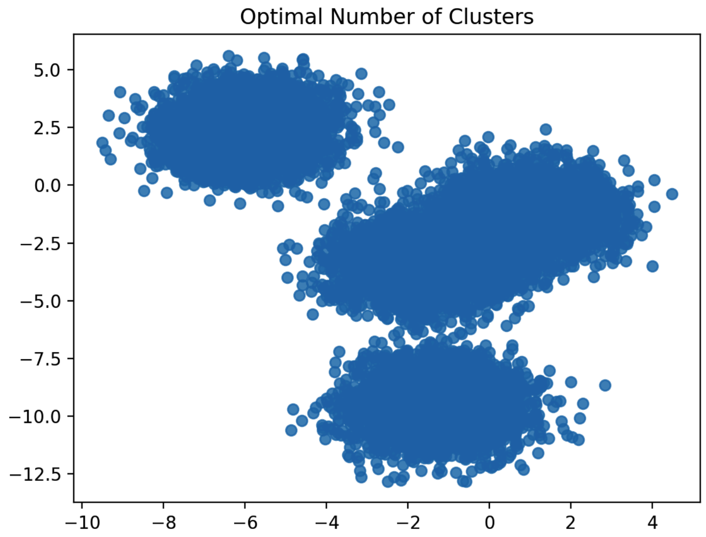

Figure 1: Optimal Number of Clusters

Cluster	fracSpent	fracComp	fracPaused	numPauses	avgPBR	numRWs	numFFs
0	11.26%	99.10%	15.18%	2.69	1.16	0.53	0.19
1	6.09%	98.02%	7379%	3.61	1.05	1.16	0.01
2	4110.86%	97.95%	38.41%	1.01	1.23	3.60	2.01

Table 1: Average Values to Analyze Video Watching Behavior for each Cluster

Question 2

The following two plots show the MSE values vs Video IDs for all videos, then for the filtered videos. The videos removed for the filtered videos are those that had an MSE value of greater than two. It can be assumed that these values are outliers. This was done to get a better visualization for every video that possessed a low MSE value.

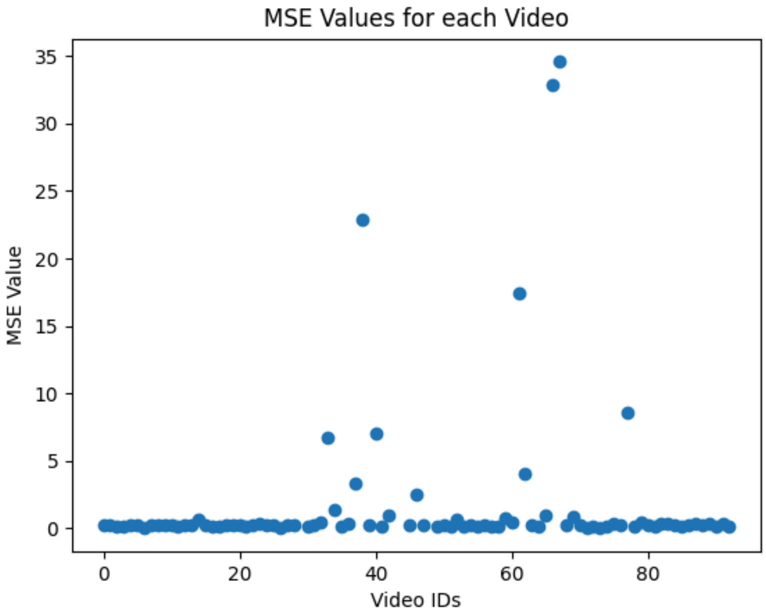

Figure 2: MSE Values vs Video IDs for all Videos

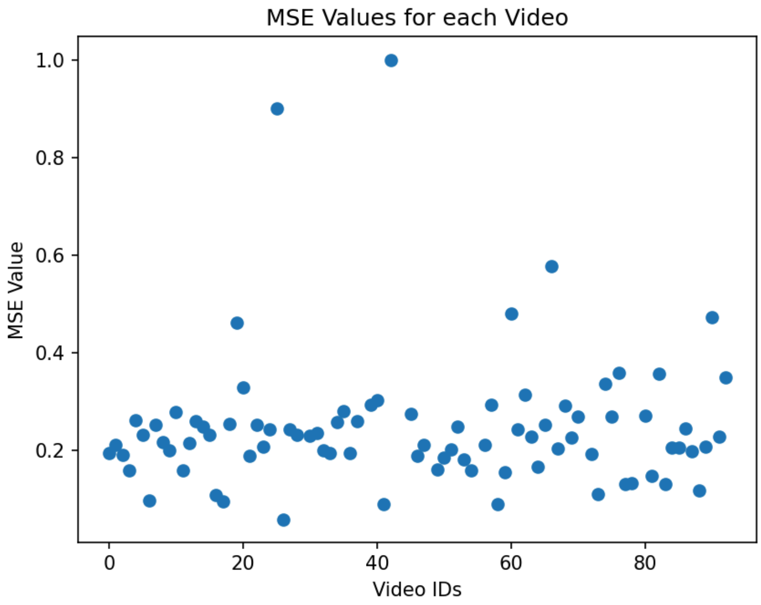

Figure 3: MSE Values vs Video IDs for filtered Videos

The following plot shows MSE values vs coefficient of determination for the filtered videos. The filtered videos are the same as previously explained.

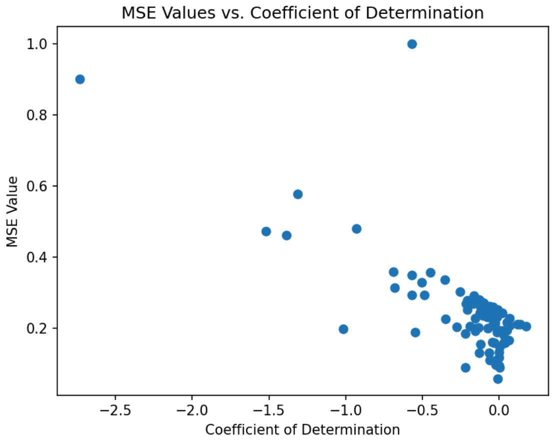

Figure 4: MSE Values vs Coefficient of Determination for Filtered Videos

The following values are outputs of our code. It shows the lowest MSE value, average MSE value, and Average Coefficient of Determination.

Lowest MSE value: 0.05924534620669295

Average MSE Value: 0.2474645671882882

Average Coefficient of Determination: -0.2188084451942339

Question 3

The following figure shows the MSE Value splotted against the accuracy value for each filtered video. This plot is a result of the logistic regression performed.

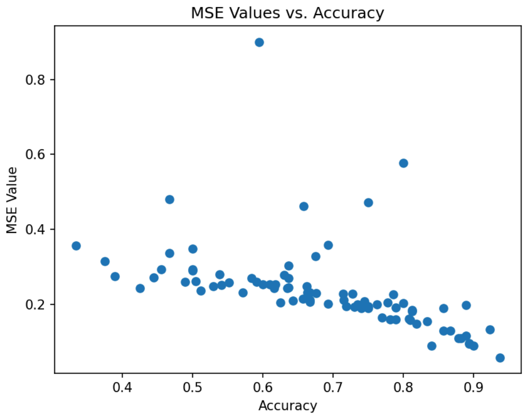

Figure 5: MSE Values vs Accuracy Values for Filtered Videos

The following values are direct results of our code and provide insight into the accuracy parameters when logistic regression is performed on the dataset.

```
Highest Accuracy Values:  1.0
Average Accuracy Value:  0.6924339887166118
```

Analysis

Question 1

Students in cluster 0 exhibit moderate engagement with videos, having a reasonably high completion rate and a slightly faster than normal playback rate. The number of pauses and fast-forwards is low, suggesting they follow the video content without much need to skip or rewatch sections. Cluster 0 could represent students who are keeping pace well with the video content. They may benefit from regular-paced coursework and a consistent structure.

Students in cluster 1 complete almost the entire video but have an anomalously high fracPaused value, likely due to leaving the video on pause. The average playback rate is normal, and the number of fast-forwards is negligible, which implies they prefer to watch content in real-time without skipping. Cluster 1, with its high fracPaused, might indicate students who are either multi-tasking or struggling to maintain focus. Interventions for these students could include techniques to improve concentration or a review of the video content's difficulty level.

Cluster 2 has extremely high fracSpent, indicating students leave the video open much longer than its length, which may imply heavy pausing or replaying. The pause time (fracPaused) is quite high as well. Students in this cluster are the most likely to rewind and review content, as seen by the highest numRWs. They also use fast-forward but less frequently than rewinds, suggesting selective attention to the content. Cluster 2 likely includes students who may be struggling with the content, given their behavior of leaving videos open long past their length and frequent rewinding. These students might benefit from additional resources, such as summaries, to help them understand the content without excessive replaying.

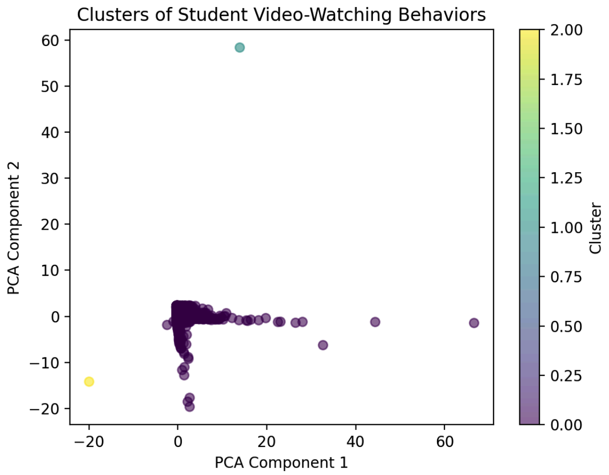

Figure 6: Clusters of Student Video-Watching Behaviors

The selection of three clusters for analyzing student video-watching behaviors was validated through a Gaussian Mixture Model (GMM), which indicated that the lowest Bayesian Information Criterion (BIC) was achieved with three clusters. This finding was corroborated by a high silhouette score of 0.9850, suggesting a strong separation and cohesion within the clusters. Additionally, visual inspection of the PCA plot and the make_blobs visualization further reinforced the distinction and clear segregation of the data into three well-defined groups.

Question 2

Figure 2 shows MSE values when comparing testing data of the feature matrix and predicted score (‘s’) values using Ridge regression. Each data point represents a video and how well the model performed for that respective video. Evidently, there are a few videos that possessed very large MSE values and, therefore, the video watching behaviors for that video are not good indicators for how well the students performed on the in-video quiz. This is likely due to a data entry error as some of these MSE values are extremely high. The average MSE value is also relatively high (Avg. MSE = 1.8054506625128603) indicating that a students’ video watching behavior may not be indicative of their average score on a given video. However, if we remove outliers (MSE > 2), we get an MSE value that is much smaller. These filtered MSE values can be seen in Figure 3.

Lowest MSE value: 0.05924534620669295

Average MSE Value: 0.2474645671882882

These MSE values suggest a better model and that the error between the predicted ‘s’ value and the actual ‘s’ value is not that high. Yet, there is a significant difference between the average MSE value and the minimum MSE value. This means that the error of the predicted score varies with the specific video. This makes sense as some videos may be more in-depth or more detail oriented than others, allowing for less variability.

Figure 4 shows that as MSE decreases, the coefficient of determination increases. This makes sense because as the error decreases, the more of the variance can be explained by the independent variable (larger R^2 value). However, the average coefficient of determination is negative. This means that for a majority of the videos, the video watching behaviors could not explain the variance in the students' score, ‘s’. I believe this may be because of a few things. First, the predictive variable, ‘s’ is not continuous and therefore it is difficult to assess the linearity of the model using ridge regression. Next, the amount of data for a respective video may be too small. Finally, the model is extremely under-fitted. This makes the most sense since the MSE is relatively low while the coefficient of determination is essentially 0 meaning that the model generalizes well (low MSE) but it cannot explain the variance in the data at all. Using ridge regression may not have been a great model for this dataset for reasons previously stated, and because it penalizes large coefficients. A certain video watching behavior may dominate the model and that would be for a good reason, because this behavior has a great influence on the students score. Using ridge regression, this influence would be diminished, which may be contradictory and can explain the poor fit of the model.

Question 3

Figure 5 shows the MSE values plotted against the accuracy values for each filtered video. There is a clear negative relationship between the two variables as MSE values decrease, accuracy values increase. This makes sense since as error is decreasing (from the previously predicted model), the accuracy of the predictions will increase. The average accuracy value was around 0.7, meaning that the model correctly predicts whether a student will get the quiz question correct or incorrect 70% of the time. This means that the logistic regression method is a fairly good model for predicting a student's score. This model seemed to perform much better than the ridge regression, this is likely because ‘s’ is binary, and that is exactly what logistic regression is made for. It is important to note that the maximum accuracy value for a specific video was 1. This means that the model correctly predicted students' scores 100% of the time. The variance in accuracy scores is likely explained by the quality of the respective video and how well it taught students.