ECE20875: Python for Data Science Spring 2022
Mini-Project

1. Project team information

Name: Yi-Hsiang Chang

Github: SeanoChang

Email: chang755@purdue.edu

Name: Li Chung Yang

Github: LiChungYang

Email: yang2010@purdue.edu

2. Descriptive Statistics

The dataset we are working with is for the behavior-performance. In this text file, it shows ten different fields of behaviors that may affect students' performance. In question one, we will deal with all the students that complete at least five of the videos. For question two, we will deal with the data with the average score s across all quizzes and the two selected behaviors fracComp, numRWs. For question three, we will further use the data we analyzed in question two but also consider the userID and videoID.

We used Kmeans and Ridge Regression to make predictions for an unknown student. With these techniques, we only could perform relatively close predictions over some of the videos, which are 2 3 8, where they have an accuracy of 70% for prediction. Video 4 has a specifically low performance on predicting the test score, with an accuracy of around 50%.

3. Approach

Problem 1:

For problem one, we need to understand how well can the students be naturally grouped or clustered by their video-watching behavior. To find out the appropriate data that we want, we first got students who fulfilled the criteria that have a video completion rate >= 0.9 with at least 5 videos completed.

Then, we grouped and clustered them by their behavior using the method of k means and Gaussian Mixture Models. In order to search the center point of the different clusters for Kmean, we used the silhouette score and find out the number of centroids from different clusters. The difference between the method of Gaussian Mixture Models and k means is that Kmeans consider the mean to find the centroid of different clusters whereas Gaussian Mixture Models consider both mean and variance.

However, when we try to implement the GMM method, we always found the best number of clusters to be a little bigger than expected (8, 9, 10). Also, since the raw dataset has been grouped by different behavior, we expect the number in each behavior will not be very different. So, we think we will not need to consider the variance. As a result, we decided to implement the Kmeans instead of GMM.

Problem 2:

For problem two, we can use selected data from students' behaviors to perform predictions on the unknown student test score.

After consideration, we think the fields of fracComp and numRWs are the two behaviors that can best predict a student's performance. For fracComp, if the fraction of the video a student completed has a higher value, we assume that the student learns about the content more. For numRWs, we assume that the number of times a student rewind backward shows the student doesn't want to miss out on the video, in other words, the student doesn't want to risk missing out on the crucial points of the video.

The other reason why we chose these two behaviors to analyze is that there are fewer outliers in these two fields. So, the data we got from these two fields will be more useful. As the example of fracSpent, we can find the number 69.5978688827 which means that students spent around 69 times the video time length on that particular video. This indicates that we can not predict the larger the number of fracSpent we get the better the performance of that student on that particular video will be.

After selecting the behaviors, we used ridge regression to predict the relationships between the selected behaviors and s. The reason we use the method of ridge regression instead of polyfit is that we can not predict whether the data is linear or not. By using ridge regression, we can get a more accurate result. We also wrote a function that asks the user to enter the number of random students you want to predict. This function will randomly select the user input numbers of students from the dataset and perform a prediction of how the accuracy of our method.

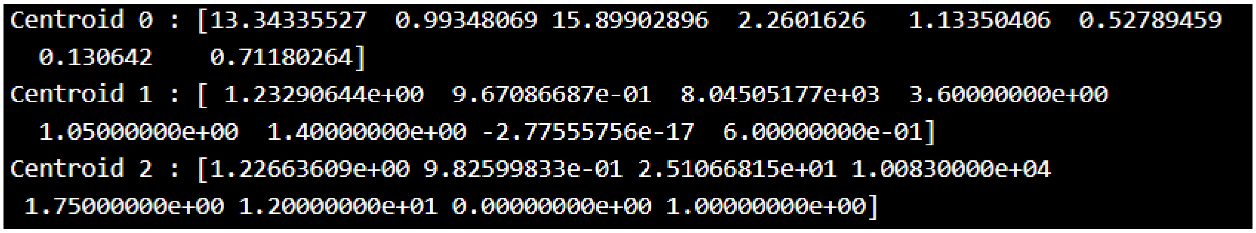
Problem 3:

For problem three, we will use the two behaviors chosen in problem two to predict a student's performance on a particular in-video quiz. we need to find the nearest center average score s. To do this, we need to calculate the distance from the data to all the centroids by using the distance to the centroid of zeros divide by the distance to the centroid of ones plus the distance to the centroid of zeros and adding all the distances and return the list. After that, we predicted the score for a student based on the Kmeans model and ridge regression model. We use the fractional distance between each data and all the centroids to perform prediction. Then, use the distances from each centroid to the data to predict the score. We then find out all the predictions for the student and then find the average correctness or accuracy of the predictions for both Kmeans and ridge methods. Finally, we will return the accuracy of the prediction. Just like question two, we will ask the user to enter the number of the students you want to predict and show the predicted students' performance in a different video.

4. Analysis

Problem 1:

To answer question one, we grouped and clustered the data by their behavior using the method of k means. In order to find the center point of different clusters, we used the silhouette score and find out 3 center points from three clusters as figure 1 showed. However, when we use the methods of Gaussian Mixture Models, we found out the best number of clusters is 10. The way of Gaussian Mixture model is a more accurate method. However, since the best number of clusters we found for Gaussian Mixture Models is a little too large, we decided to use Kmean to cluster the video as we expected.

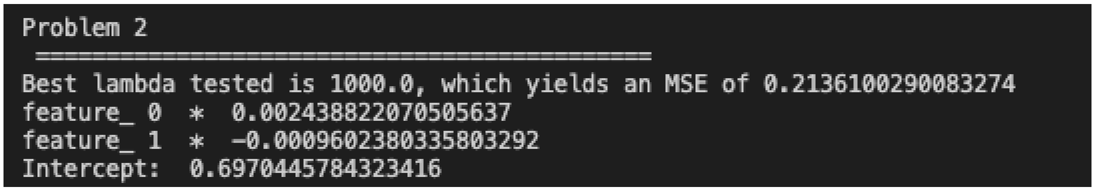


```
Centroid 0 : [13.34335527  0.99348069 15.89902896  2.2601626   1.13350406  0.52789459
  0.130642    0.71180264]
Centroid 1 : [ 1.23290644e+00  9.67086687e-01  8.04505177e+03  3.60000000e+00
  1.05000000e+00  1.40000000e+00 -2.77555756e-17  6.00000000e-01]
Centroid 2 : [1.22663609e+00 9.82599833e-01 2.51066815e+01 1.00830000e+04
 1.75000000e+00 1.20000000e+01 0.00000000e+00 1.00000000e+00]
```

Figure 1: The centroids found using the Kmean method.

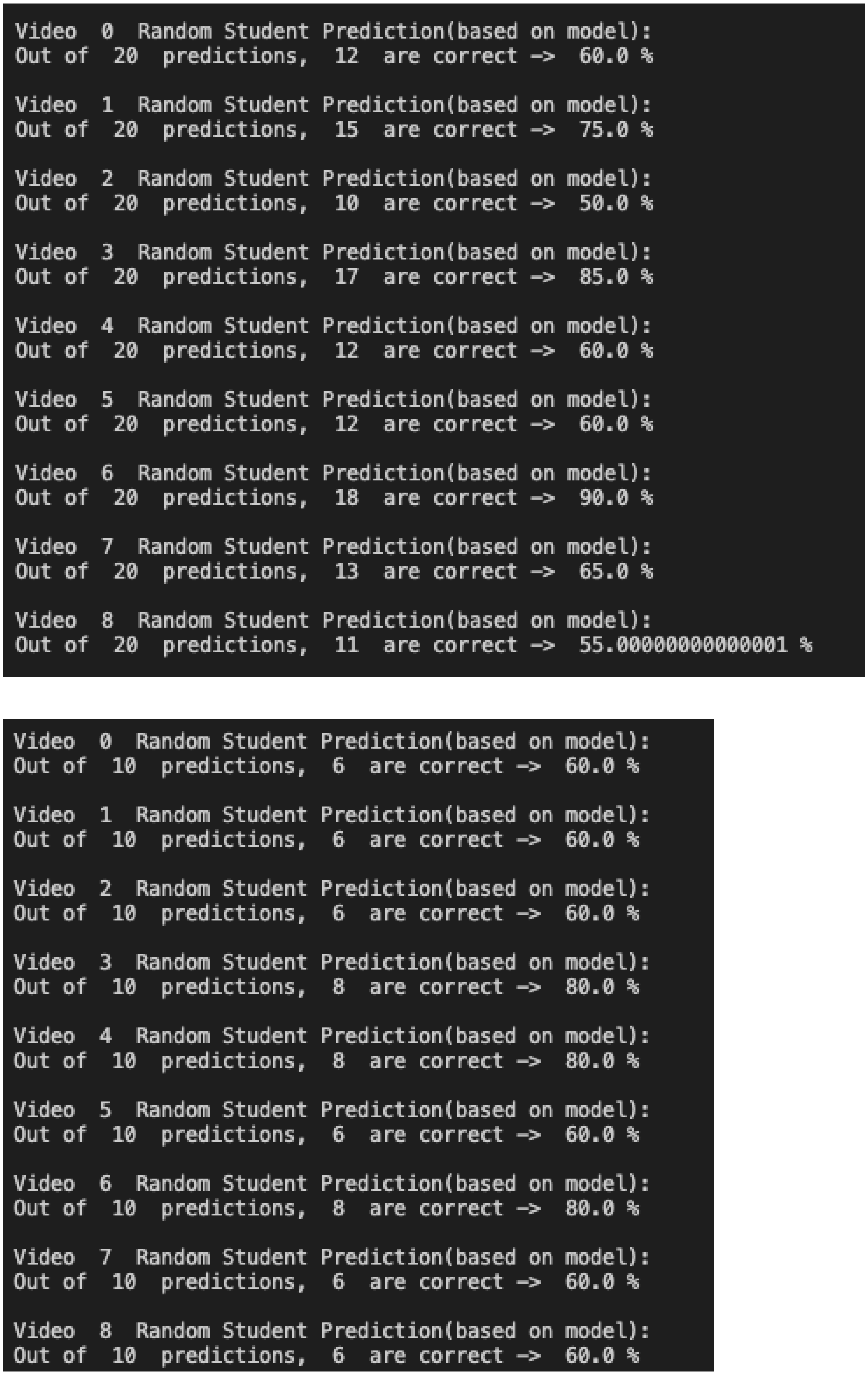
Problem 2:

As discussed in the approach section, a student's video-watching behavior can be used to predict a student's performance. As figure 2 shows, the values of X_ 0 are the parameter for fracComp, the values of X_ 1 are the parameter for numRWs, and the values of Intercept are the constant of the model function. As shown in figure 3, If we enter the number of the students you want to predict as 1, the prediction will be 100% for every video. This is because we only have one data to act as trained and tested data. If we use a number other than 1 the values of percentage will obviously be between 0 and 100.



```
Problem 2
 ============================================
Best lambda tested is 1000.0, which yields an MSE of 0.2136100290083274
feature_ 0  *  0.002438822070505637
feature_ 1  *  -0.0009602380335803292
Intercept:  0.6970445784323416
```

Figure 2: The parameter for the ridge regression model. The feature is the list of behaviors we chose, feature 0 = featureList[0].



```
Video  0  Random Student Prediction(based on model):
Out of  20  predictions,  12  are correct ->  60.0 %

Video  1  Random Student Prediction(based on model):
Out of  20  predictions,  15  are correct ->  75.0 %

Video  2  Random Student Prediction(based on model):
Out of  20  predictions,  10  are correct ->  50.0 %

Video  3  Random Student Prediction(based on model):
Out of  20  predictions,  17  are correct ->  85.0 %

Video  4  Random Student Prediction(based on model):
Out of  20  predictions,  12  are correct ->  60.0 %

Video  5  Random Student Prediction(based on model):
Out of  20  predictions,  12  are correct ->  60.0 %

Video  6  Random Student Prediction(based on model):
Out of  20  predictions,  18  are correct ->  90.0 %

Video  7  Random Student Prediction(based on model):
Out of  20  predictions,  13  are correct ->  65.0 %

Video  8  Random Student Prediction(based on model):
Out of  20  predictions,  11  are correct ->  55.00000000000001 %
```

```
Video  0  Random Student Prediction(based on model):
Out of  10  predictions,  6  are correct ->  60.0 %

Video  1  Random Student Prediction(based on model):
Out of  10  predictions,  6  are correct ->  60.0 %

Video  2  Random Student Prediction(based on model):
Out of  10  predictions,  6  are correct ->  60.0 %

Video  3  Random Student Prediction(based on model):
Out of  10  predictions,  8  are correct ->  80.0 %

Video  4  Random Student Prediction(based on model):
Out of  10  predictions,  8  are correct ->  80.0 %

Video  5  Random Student Prediction(based on model):
Out of  10  predictions,  6  are correct ->  60.0 %

Video  6  Random Student Prediction(based on model):
Out of  10  predictions,  8  are correct ->  80.0 %

Video  7  Random Student Prediction(based on model):
Out of  10  predictions,  6  are correct ->  60.0 %

Video  8  Random Student Prediction(based on model):
Out of  10  predictions,  6  are correct ->  60.0 %
```

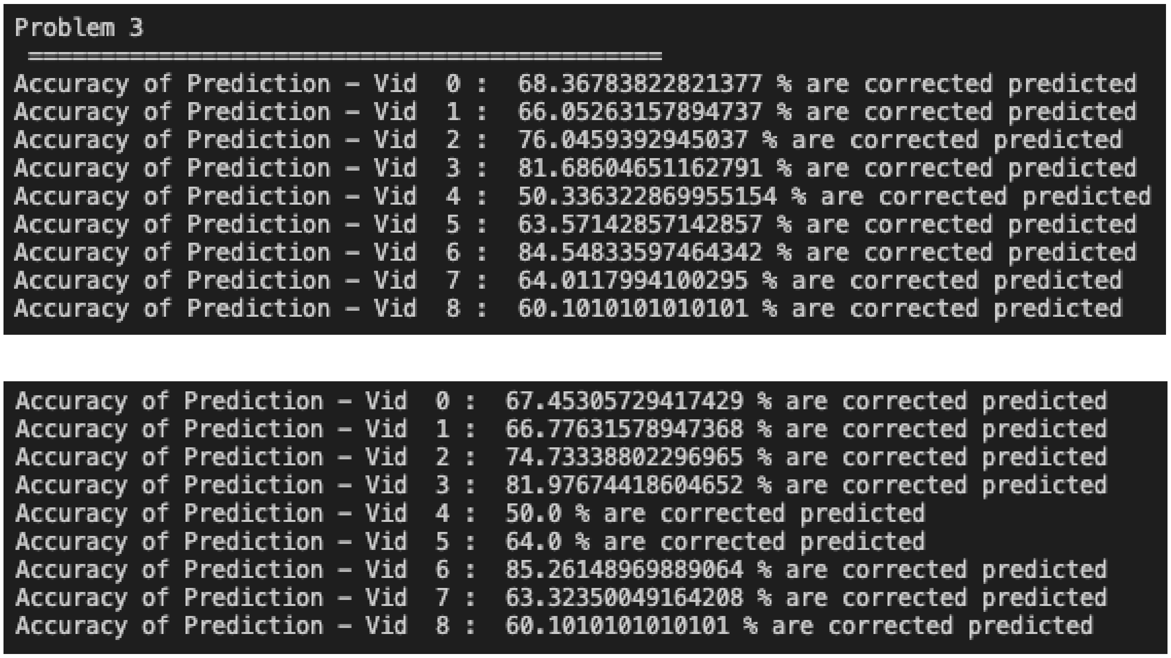
Figure 3: Random student prediction. We randomly selected students from the model with the same video ID to perform prediction. The accuracy of prediction results are close to the accuracies in figure 2.

Problem 3:

To answer problem 3, we chose to use the two behaviors (fracComp and numRWs) to predict a student's performance on a particular in-video quiz question.

Figure 4 shows the accuracy of the prediction in different videos when considering all student-video pairs in our analysis. For some videos, the accuracies are high, which informs that the chosen behaviors are more likely to affect the test results. While for videos with low accuracies, we can know that the chosen behaviors are not good features to perform a prediction.

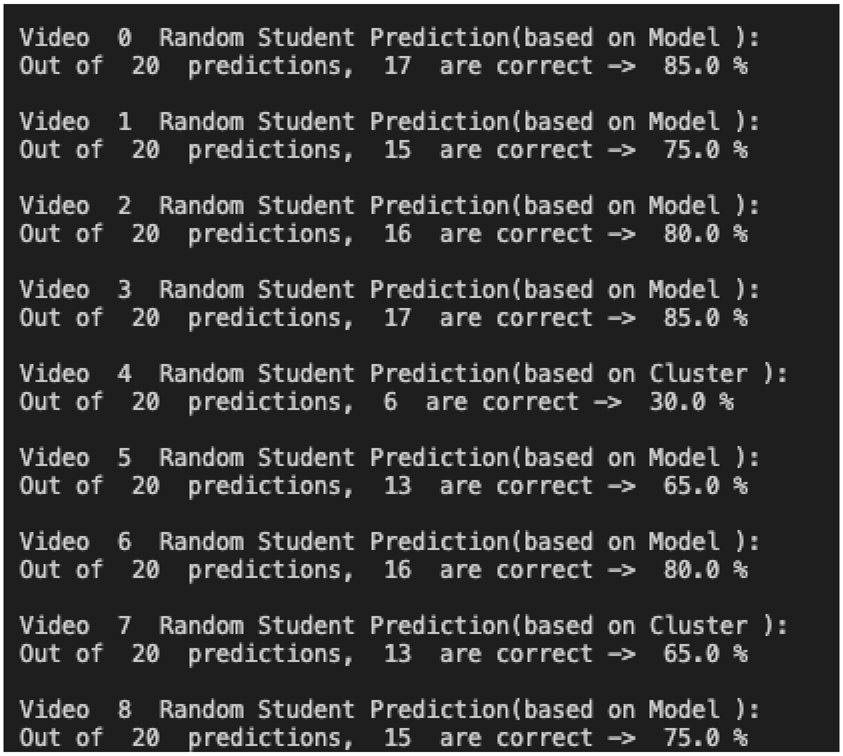
As shown in figure 5, if we enter the number of the students you want to predict 20 random students' test scores, we can predict students' performance on a particular in-video quiz question.



```
Problem 3
 ============================================
Accuracy of Prediction - Vid  0 :  68.36783822821377 % are corrected predicted
Accuracy of Prediction - Vid  1 :  66.05263157894737 % are corrected predicted
Accuracy of Prediction - Vid  2 :  76.0459392945037 % are corrected predicted
Accuracy of Prediction - Vid  3 :  81.68604651162791 % are corrected predicted
Accuracy of Prediction - Vid  4 :  50.336322869955154 % are corrected predicted
Accuracy of Prediction - Vid  5 :  63.57142857142857 % are corrected predicted
Accuracy of Prediction - Vid  6 :  84.54833597464342 % are corrected predicted
Accuracy of Prediction - Vid  7 :  64.0117994100295 % are corrected predicted
Accuracy of Prediction - Vid  8 :  60.1010101010101 % are corrected predicted
```

```
Accuracy of Prediction - Vid  0 :  67.45305729417429 % are corrected predicted
Accuracy of Prediction - Vid  1 :  66.77631578947368 % are corrected predicted
Accuracy of Prediction - Vid  2 :  74.73338802296965 % are corrected predicted
Accuracy of Prediction - Vid  3 :  81.97674418604652 % are corrected predicted
Accuracy of Prediction - Vid  4 :  50.0 % are corrected predicted
Accuracy of Prediction - Vid  5 :  64.0 % are corrected predicted
Accuracy of Prediction - Vid  6 :  85.26148969889064 % are corrected predicted
Accuracy of Prediction - Vid  7 :  63.32350049164208 % are corrected predicted
Accuracy of Prediction - Vid  8 :  60.1010101010101 % are corrected predicted
```

Figure 4: Accuracy of the prediction in different videos. The higher the score shows that the data we choose are likely to be indicators for the test score performance, and vice versa.



```
Video  0  Random Student Prediction(based on Model ):
Out of  20  predictions,  17  are correct ->  85.0 %

Video  1  Random Student Prediction(based on Model ):
Out of  20  predictions,  15  are correct ->  75.0 %

Video  2  Random Student Prediction(based on Model ):
Out of  20  predictions,  16  are correct ->  80.0 %

Video  3  Random Student Prediction(based on Model ):
Out of  20  predictions,  17  are correct ->  85.0 %

Video  4  Random Student Prediction(based on Cluster ):
Out of  20  predictions,  6  are correct ->  30.0 %

Video  5  Random Student Prediction(based on Model ):
Out of  20  predictions,  13  are correct ->  65.0 %

Video  6  Random Student Prediction(based on Model ):
Out of  20  predictions,  16  are correct ->  80.0 %

Video  7  Random Student Prediction(based on Cluster ):
Out of  20  predictions,  13  are correct ->  65.0 %

Video  8  Random Student Prediction(based on Model ):
Out of  20  predictions,  15  are correct ->  75.0 %
```

Figure 5: The figure shows the percentage of the accuracy if there are 20 students. The accuracy is usually around the numbers in figure 4. Based on ___ indicates which way is a better way for prediction according to the score for prediction accuracy.