Project: Student Behavior Performance

Dataset:

The features related to the fractions for students watching the videos include: the fraction of time the student spent watching the video with a range of 0 to 18,215, the fraction of the video the student watched with a range of 0 to 9,846,341, as well as the fraction of time the student spent paused on the video with a range of 0 to 15,957. Additional features related to interactions with the video include: the number of times the student paused the video with a range of 0 to 10,083 clicks, the average playback rate that the students used while watching the video with a range of 0 to 2 speed, the number of times the students skipped backwards in the video with a range of 0 to 2,237 rewinds, as well as the number of times the student skipped forward in the video with a range of 0 to 309. Finally, a standard output feature was whether or not the student was correct on their first attempt at answering the video quiz. There are a total of 93 different videos, 3,976 different users, and 29,304 different samples. The source of the dataset comes from an IEEE article where they had a similar goal of predicting whether a user will be correct on the first attempt in answering a question.

Methods:

The first analysis question asks for how well the students can be naturally grouped or clustered by their video-watching behavior. The method that was chosen was the K-means clustering algorithm. The features that were used are the same as the reduced dataset. The data points were reduced down to students who completed at least five videos. The model was verified via the elbow method. The number of clusters was plotted against the squared error between the point and the assigned cluster, and the optimal cluster was analyzed by looking at the point for the elbow of the curve.

The second analysis question asks for how the student's video-watching behavior can be used to predict a student's average score s across all quizzes. The method that was chosen was Ridge regression with a 5-fold cross validation. The features are the same as the reduced dataset. The data points were reduced down to students who completed at least half of the quizzes. The model was verified by looking at the minimum mean squared error for the testing data set based on a range of lambda values.

The third analysis question asks for how well one can predict a student's performance on a particular in-video quiz question based on their video-watching behaviors. The method that was chosen was Logistic regression with a 5-fold cross validation. The features are the same as the reduced dataset. All of the users and videos were included in this analysis. The model was verified by looking at the accuracy score from the logistic regression model.

Results:

For the first analysis question, the K-means algorithm provided the elbow for the curve at around k = 11 or k = 12 clusters since it has the steepest point of the plot. This means that the video-watching behavior of the student can generally be clustered into 11 or 12 groups with around a total distance from point to assigned cluster of 1044.

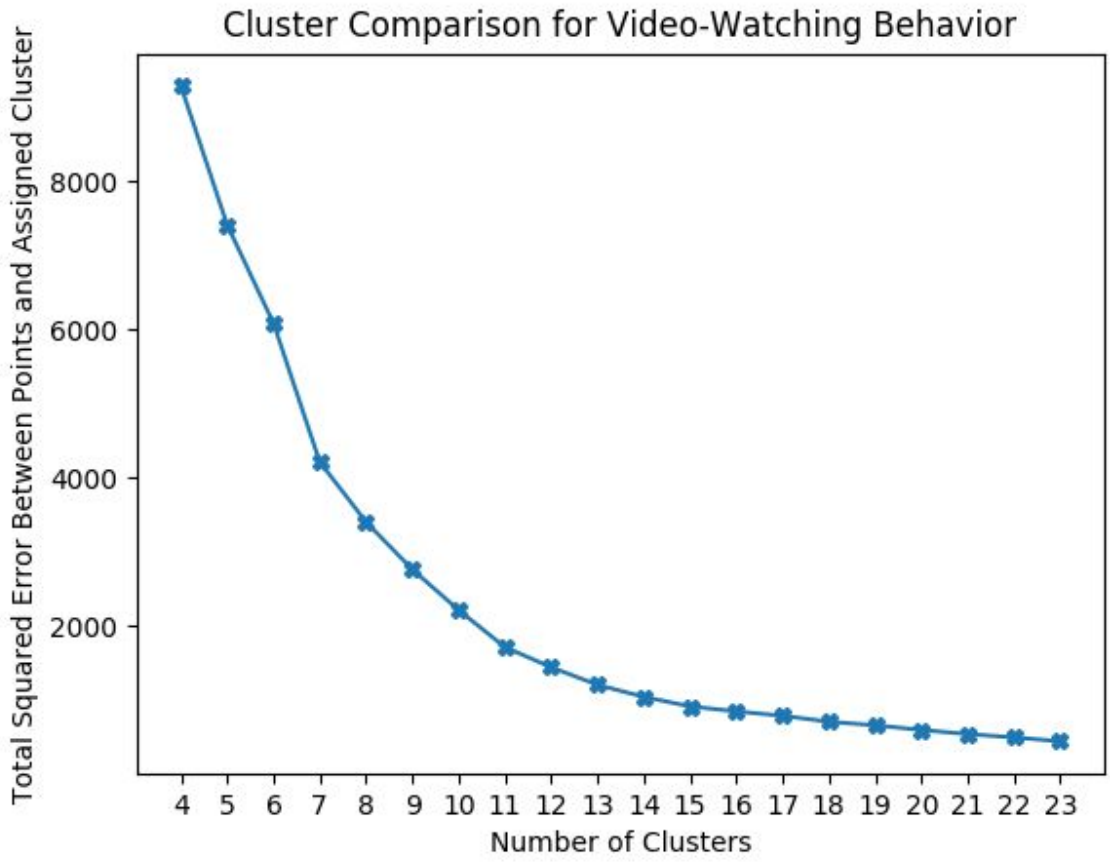

Figure 1: K-Means Cluster Comparison for Video-Watching Behavior

For the second analysis question, the student's performance (average score across all quizzes) can be predicted pretty well via a 5-fold cross validation and ridge regression. When evaluating the Mean Squared Error for various lambda values, the model with the lambda parameter of 97.72372 resulted in the lowest mean squared error of 0.024563.

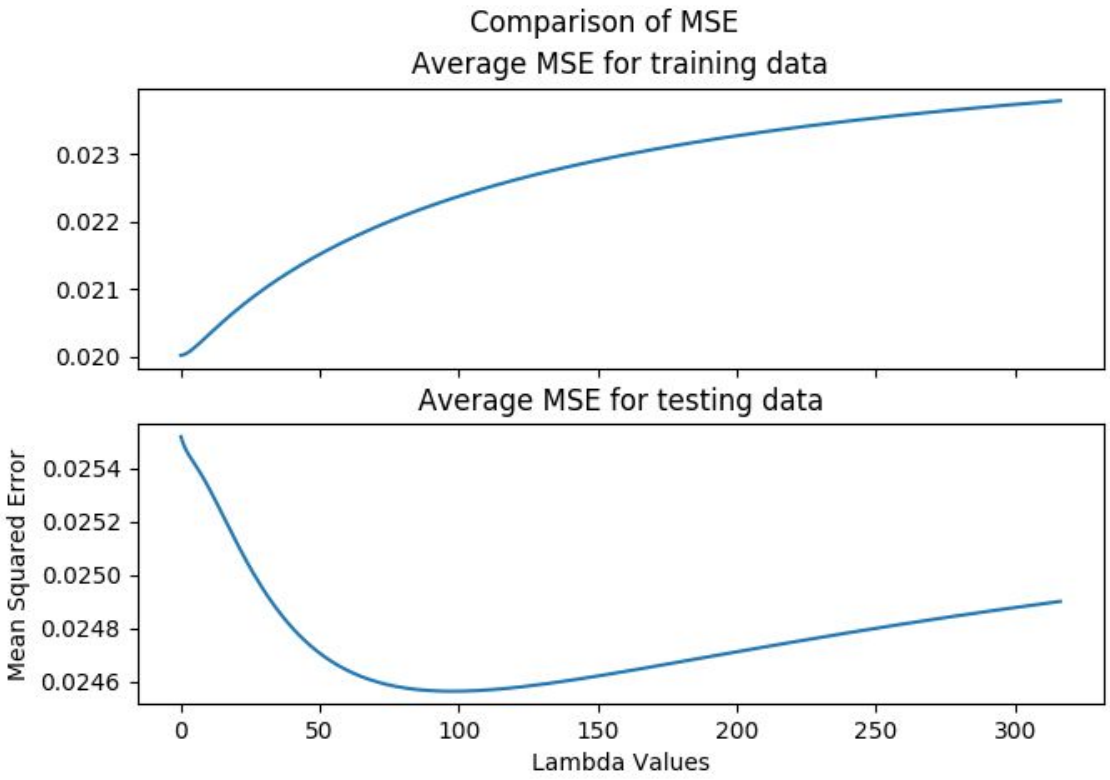

Figure 2: 5-Fold Ridge Regression Comparison of Training and Testing MSE for various lambda values

For the third analysis question, a student's performance for a particular in-video quiz was predicted with Logistic Regression.When evaluating the accuracy for various C (regularization) values, there is a definitive optimum parameter of a C value of 20.962; however, the accuracy for this C value is only 51.25% which is essentially guessing the quiz score. One reason could be that the sample data may not have been varied enough to cover all cases. Another reason could be that there truly is no relationship between the features and the in-video quiz score which means the score is typically random.

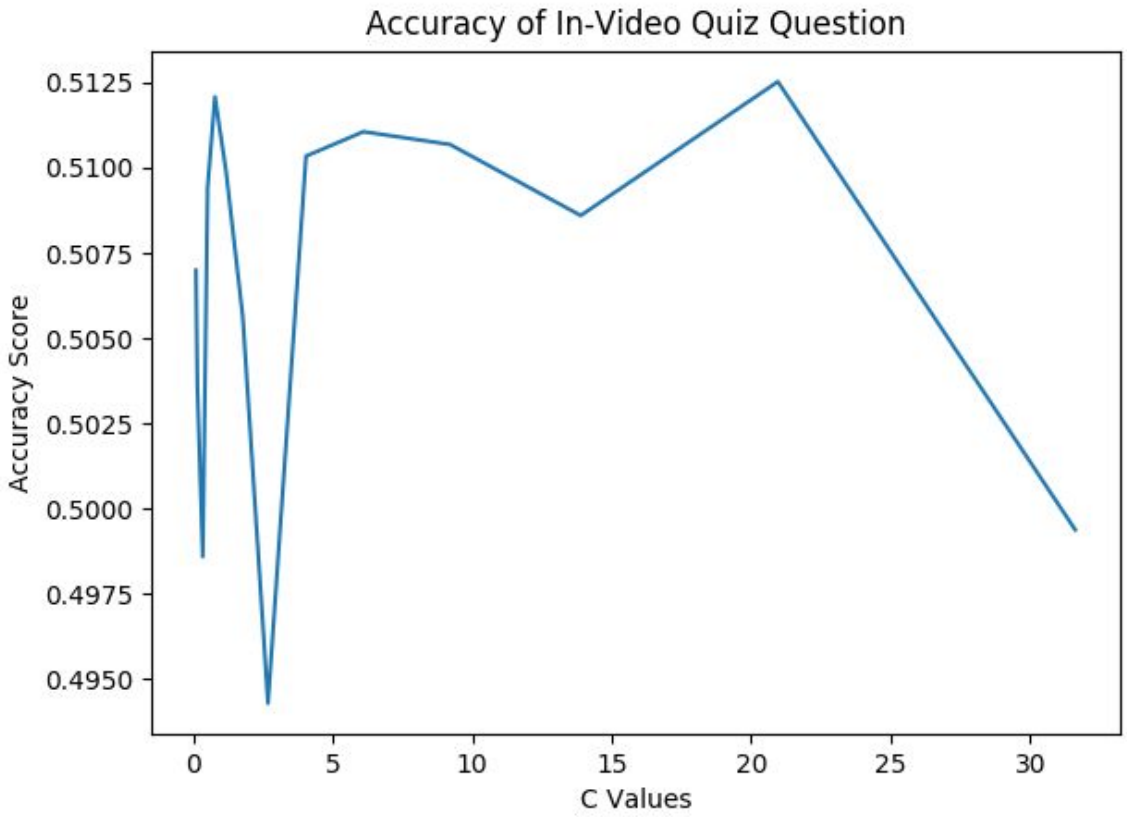

Figure 3: 5-Fold Logistic Regression Testing Accuracy for various C values