Final Project

Graedon Beeler
Luke Valerio

CPSC 392

Introduction

Description of the data (e.g. What variables do you have? Where is this data from? How much data do you have? Anything notable about the data).

Our data came from a Strava user by downloading a zip file of every activity they posted on the site. One note about the data is that when we imported the data it included all possible metrics for all possible strava activities when we only cared about the biking activity so to get around this we made a new data set that only included the columns with variables we were interested in. In the end for the majority of our model we had about 4000 usable rows to work with.

List of Variables:

["Max Heart Rate", "Elapsed Time", "Distance", "Bike Weight", "Average Heart Rate", "Elevation Gain", "Average Watts" , "Average Cadence", "Distance" , "Moving Time", "Elapsed Time", "Average Grade", "hour", "day_of_week", "Relative Effort.1"]

Question #1: Are we able to develop a model that replicates the data 'relative effort' (with 90%R2) to understand how this value is predicted, and be able to predict it for future activities?

Methods

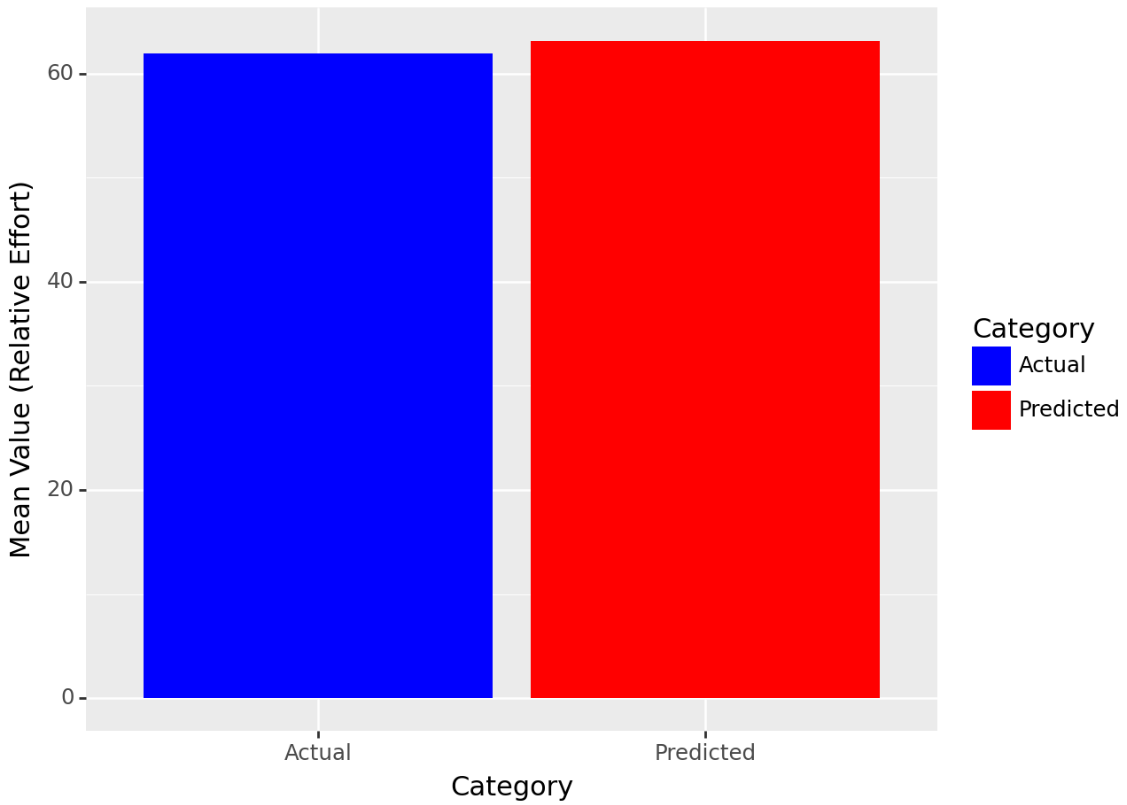
The first steps of developing this model was to find the missing values in the data and remove them. The source we got our data from included a lot of variables that weren't stored so those were removed. Our final list of variables was then put through a linear regression model with an 80/20 train test split. Finally the metrics MSE, MAE, MAPE, and R2 values were listed for the train and test set to evaluate performance.

Results

The result from this model is that using the predictors we were able to develop a model that was able to predict the relative effort metric in Strava. The test R2 value was 0.85. The figures we created for this showed the mean relative effort measurement was 63.65, and our predicted mean was 62.85. Relative effort is unitless and is just meant to be compared between activities.

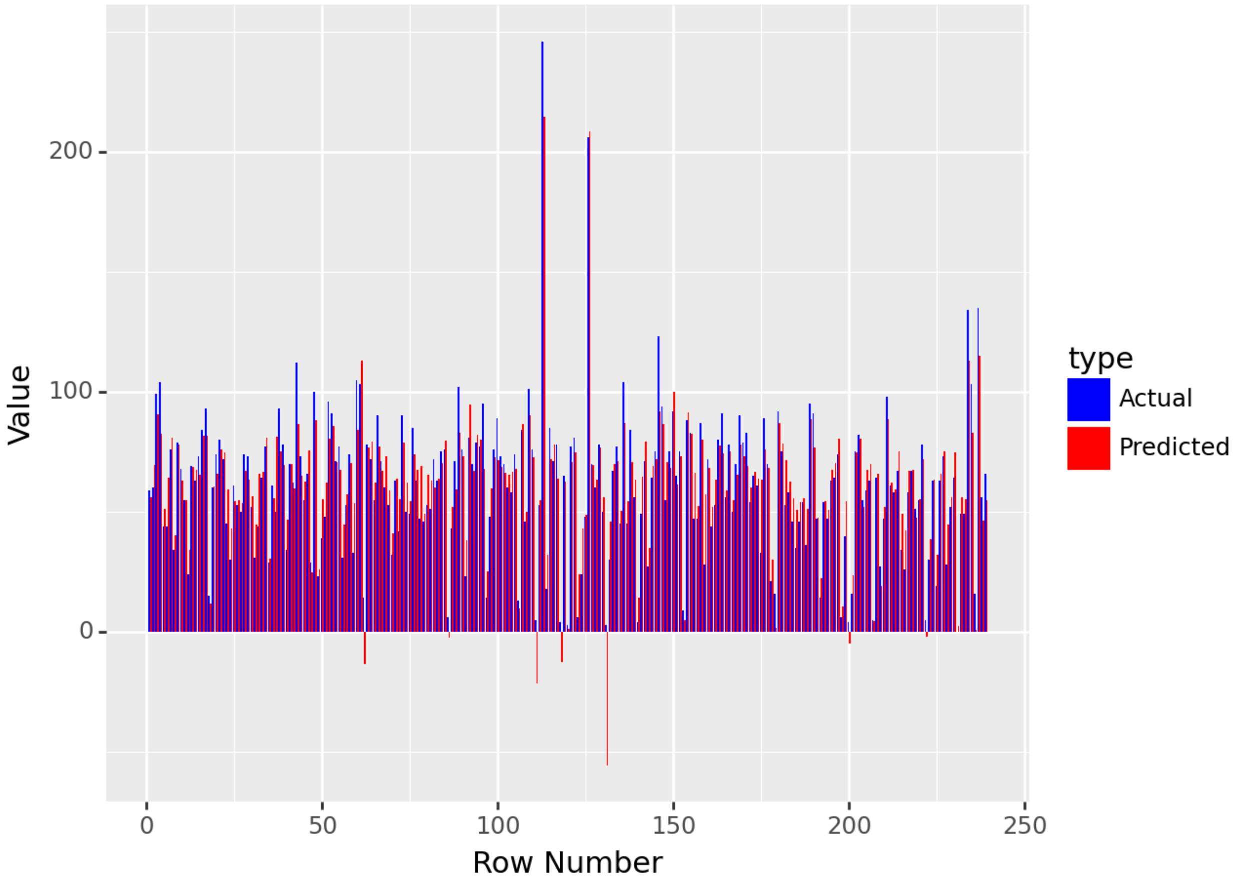
Discussion

Here is my very detailed discussion of the answer to this question, as well as how I found the answer (e.g. describing the models, presenting the graphs, discussing the implication).
Overall a linear regression model was able to create a very similar model to the one that the actual Strava site uses. The answer to our initial question was that we were able to make a model however not quite the high standards of an R2 value of greater than 0.90. The two figures we made were one showing all of our predictions vs the actual values and created a bar graph that shows how closely we were able to match these values. The other one was a bar graph comparing the mean of all of the relative effort values of the actual data set vs the means of all of our predicted values. These were even closer which potentially indicates that our model may be able to get the activities in the middle of the data set correct but was incorrect on the either very high, or very low values of relative effort.
Figure 1



A bar graph with an actual and a predicted category. With a value of the mean relative effort of both sets. They are very close to the same

Figure 2



A bar graph with the actual and predicted values for each of the rows. They are colored to show the two different categories and we can see the predictions following the strava calculated value very closely.

Question #2: Is LASSO or PCA a more accurate dimensionality reduction?

Methods

Both the LASSO and PCA were applied to the linear regression model previously described. Z-scoring was the only preprocessing for doing both the LASSO and PCA. The linear regression model was run twice more, one with the LASSO, and the other with the principal components. We defined accuracy to mean the ability to keep the highest R2 value while reducing the complexity of our model.

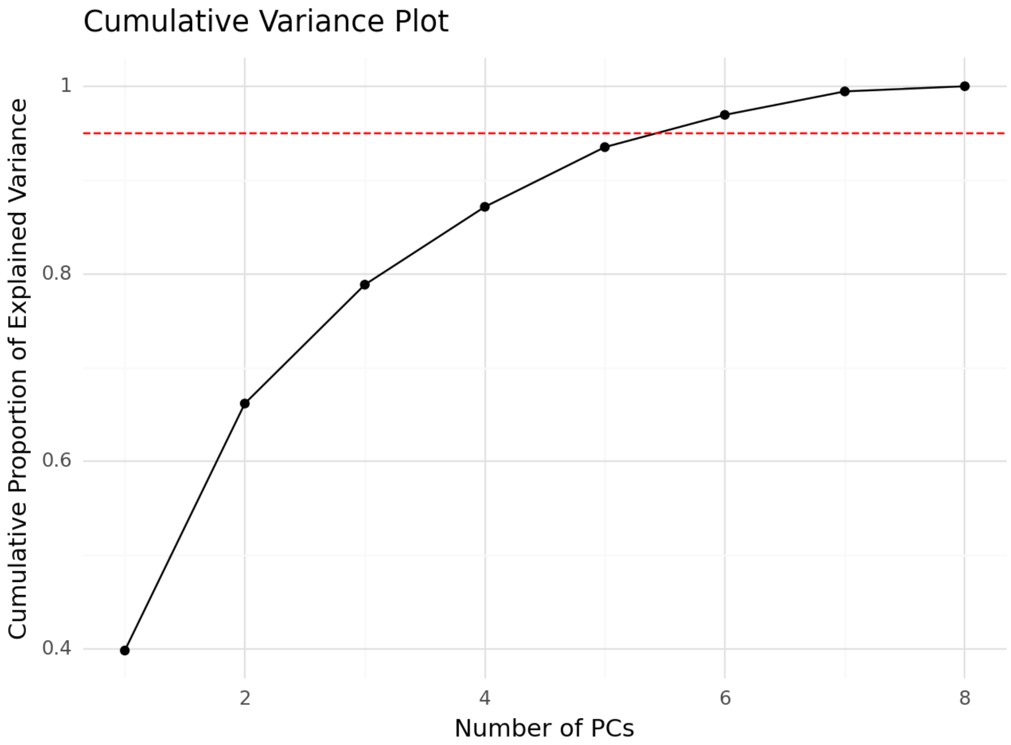
Results

What were the results of your model(s) and analysis?
As expected both models performed a little bit worse with the LASSO having an R2 value of 0.84, and the PCA with 6 principle components (this was chosen with the elbow method on a cumulative variance plot)had an R2 = 0.8. Of course this is expected with dimensionality reduction, the goal of this question was a comparison between the PCA and LASSO so the actual R2 values were not of great concern.

Discussion

Here is my very detailed discussion of the answer to this question, as well as how I found the answer (e.g. describing the models, presenting the graphs, discussing the implication).

Based on our results we saw that LASSO had less of an effect on the R2 value for our linear regression model indicating that based on accuracy it would be the superior dimensionality reduction technique. An important consideration though is that this data was not very complex with only having 9 predictors. It is possible that PCA would perform better on much higher dimensional data, in order to confirm this we would need to add more predictors to our model. However on this lower dimensional data LASSO performed better. The figures that helped with the answer to this question was a cumulative variance plot of the principal components. Then a change from our initial plan was to include a table with the performance metrics of the three types of graphs instead of the scree plot because it would have just been redundant information.



A cumulative variance plot with the number of principal components on the x axis and the percentage of variance explained on the left. A red dashed line is placed at 95% variance explained.

Type of Model	MSE	R2
Linear Regression	151.09	0.858
PCA	220.572	0.805
Lasso	172.98	0.835

Table comparing metrics of the test set of the initial linear regression, and the linear regression with the dimensionality reduction techniques.

Question #3: Could we find certain classes of activities (hard, easy, long day) based on performance statistics?

Methods

We created a pipeline and z scored our predictors then used a k means clustering algorithm with 3 clusters. We decided on this number by iterating through a few different k values and found the silhouette score was the best when we used 3. To show off the clusters we did a PCA and plotted the first two on a ggplot.

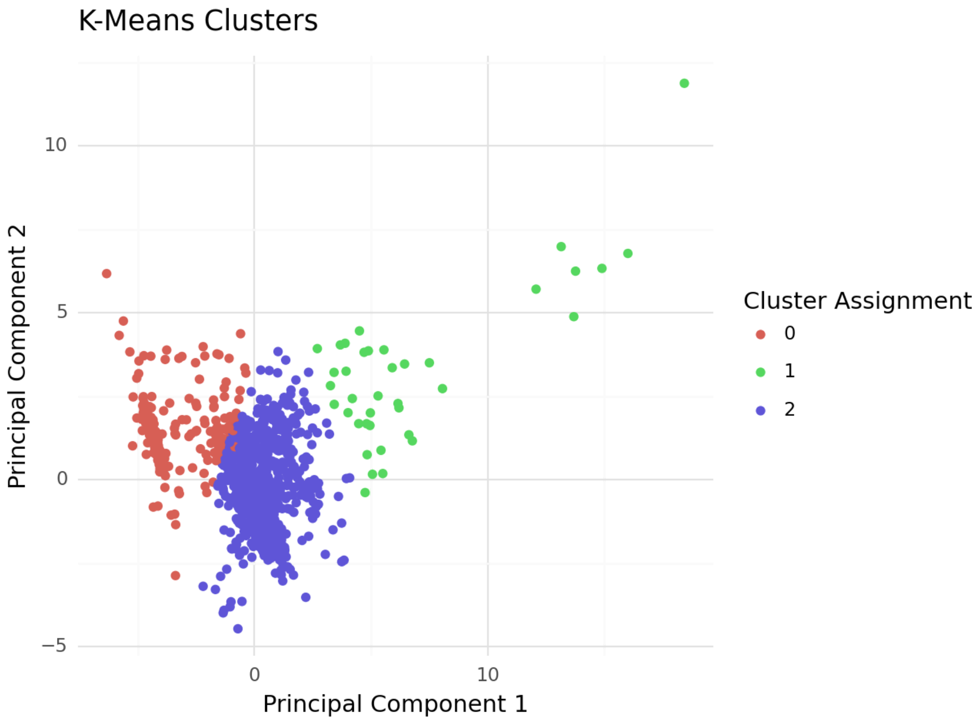
Results

What were the results of your model(s) and analysis?
The overall result is that we were able to make some clusters that indicated an ability to classify activities into a few different categories. The best silhouette score we came up with was 0.463.

Discussion

Here is my very detailed discussion of the answer to this question, as well as how I found the answer (e.g. describing the models, presenting the graphs, discussing the implication).

In conclusion, by looking at the graph of the two PCs, and by looking at the summary table, we were able to make distinctions between the different types of rides that each cluster represents. It was helpful to personally know the athlete to verify that the results of our cluster make sense. The red cluster represents an “easy day”, a 10 mile ride at 15 miles an hour, which takes well under an hour. The max heart rate and average watts is the lowest of the 3, the cadence is the lowest meaning the pedaling is slower, and there is 128m, or under 500ft of climbing, which is relatively flat. The second cluster is the most dense, and it looks to be the quick and hard weekday lunch ride. It is an hour and a half of hard training, getting up to 180 bpm (for a middle-aged man), and putting out the highest wattage. Finally, the last cluster represents a “long day”. These would probably be the occasional big weekend rides, as they average over 60 miles, taking over 4.5 elapsed hours from start to finish. It is done at a slower pace than the cluster 2 and slightly faster than cluster 1, but the calories burned are over double, the elevation gain and steepest climb are the greatest by far (which makes sense when you are riding for that long in a hilly terrain), and the average power and maximum heart rate are in the middle as well.



A graph of the clusters our model made by plotting the first two principal components of our PCA on the x and y axis respectively. Clusters were colored to show the separation.

Distance (miles)	Avg MPH	Calories	Elapsed Time	Moving Time	Max Heart Rate	Average Watts	Average Cadence	Elevation Gain	Max Grade
15.5 (9.6)	15.1	286	00:43:48	00:38:24	149.4	105.3	78.7	128.5	41.3
40.6 (25.2)	18.4	1010	01:31:47	01:23:07	179.7	184.6	87.1	413.6	45.0
97.4 (60.5)	16.2	2266	04:33:26	03:45:11	177.6	156.3	82.9	1296.5	787.3

Data table showing the averages of each variable for each cluster. Each row represents a different cluster, and the columns represent the variables and corresponding values.

Question #4: Question: Can we train a Logistic Regression Model to accurately predict (0.8 accuracy or higher) whether the given athlete's bike ride was a commute or not.

Methods

The initial steps to this model were to set up the data frame we were going to use for this model by creating the one that uses all of the predictors we wanted. Then separating the list of categorical and continuous variables and making an 80/20 train test split. Our preprocessing included z scoring the continuous variables. Then OneHotEncoding the variables that needed it. Finally we displayed the train and test accuracy, as well as the ROC AUC. The plots for this included a confusion matrix, a calibration curve, as well as the ROC curve.

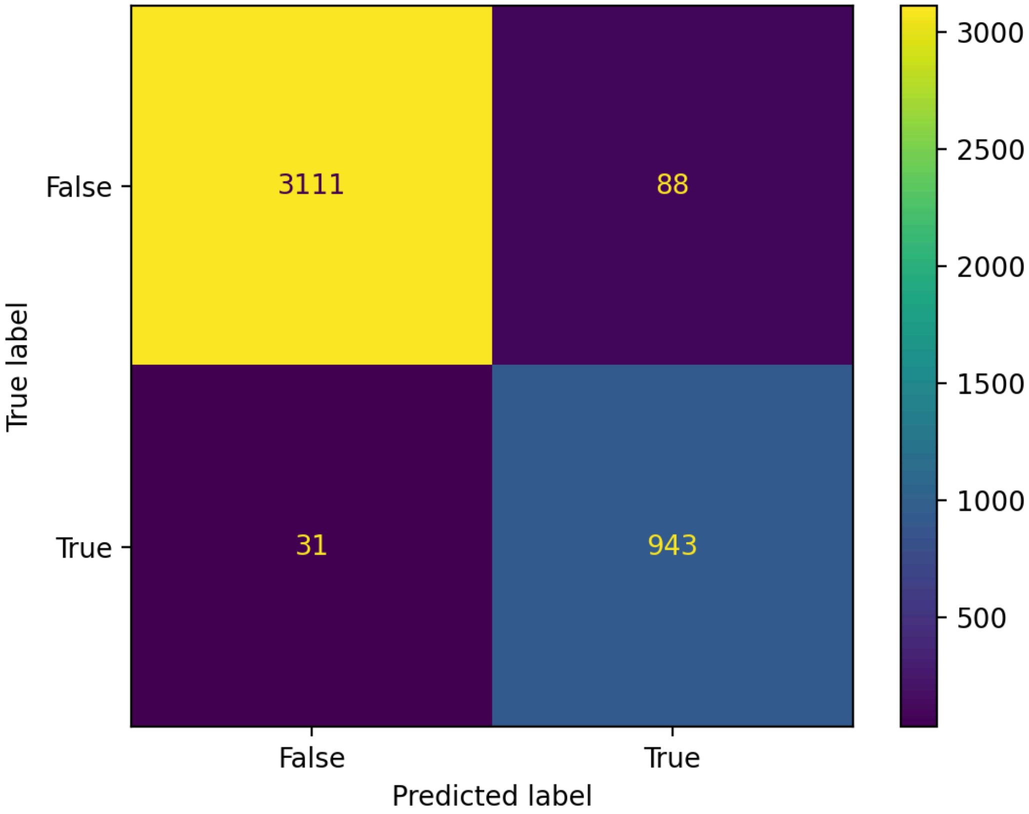
Results

The model was very good at classifying if the activity was a commute with a test accuracy of 95.6, and a test ROC AUC of 98.47. This would give us good reason to believe that we could accurately classify whether an activity was a commute or not on any user based on our predictors.

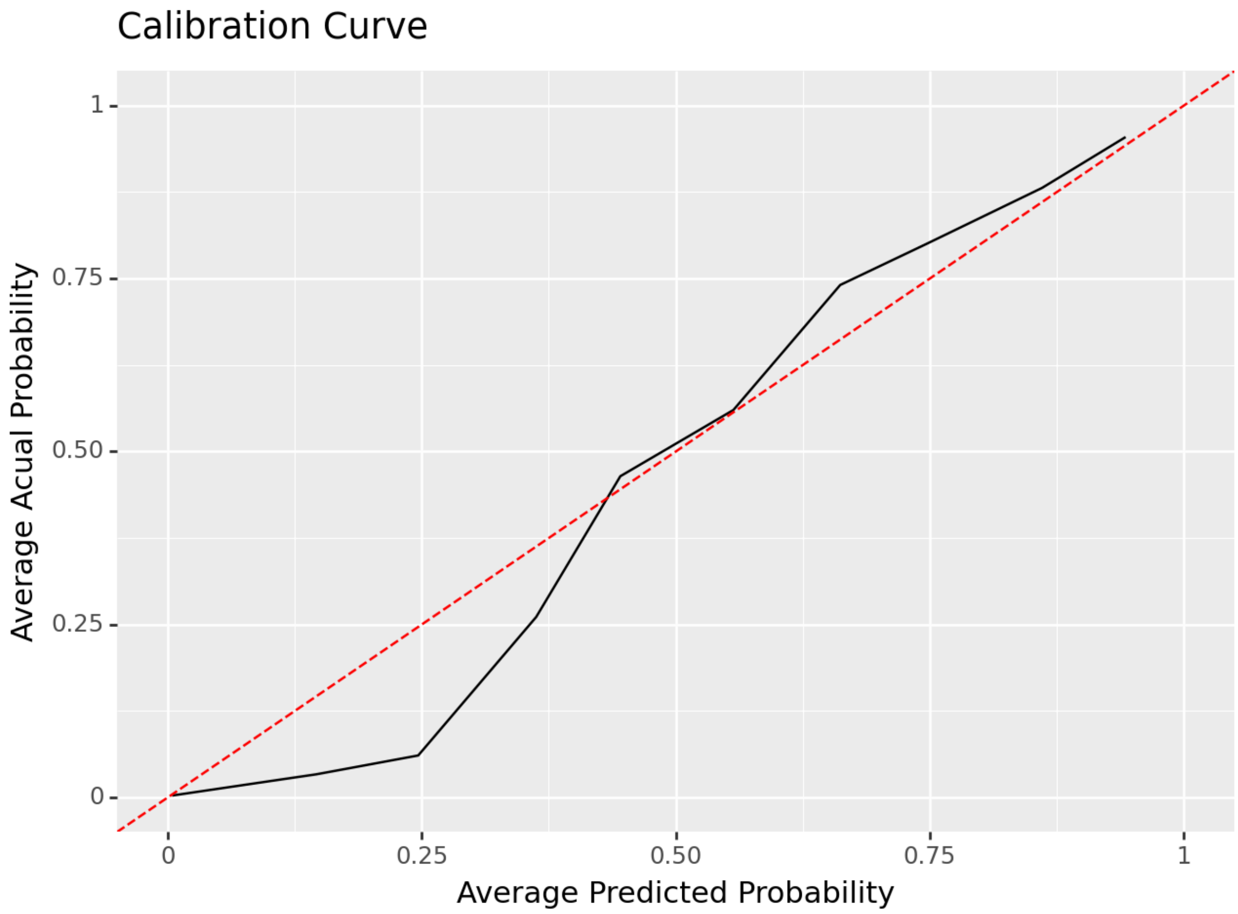
Discussion

Here is my very detailed discussion of the answer to this question, as well as how I found the answer (e.g. describing the models, presenting the graphs, discussing the implication).

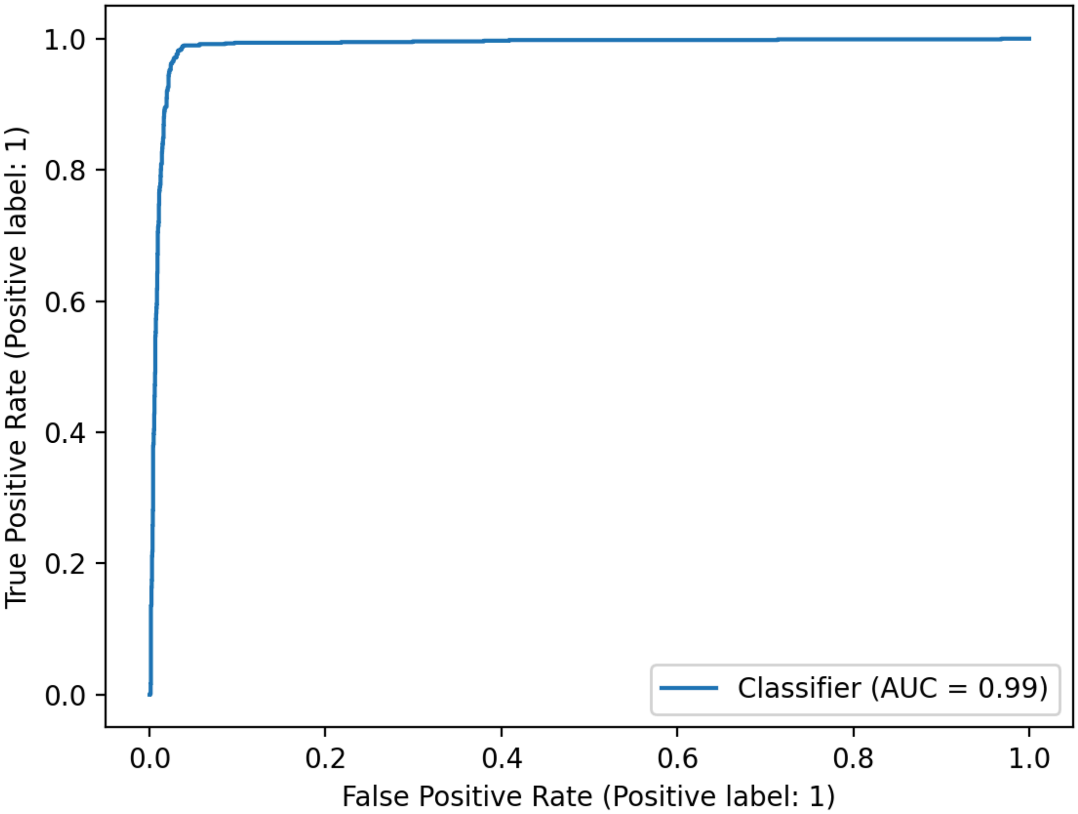
The answer to the question is that yes we were able to train the model with an even higher accuracy than we anticipated. We confirmed this by creating the logistic regression model then looking at the accuracy of the model. We made a confusion matrix to better illustrate how effective our model was. Then also created a calibration curve and a ROC curve to demonstrate that our model was well calibrated and had a very good ROC AUC score. Our model would be able to provide information on how often users are using strava as a way to post commutes as opposed to the typical use of strava which is for training. This could potentially inform features added to the app for people who do commute.



A standard confusion matrix with our two labels of ‘true’ meaning it was a commute, or ‘false’ indicating it was some other activity. We see that for over 4000 results only 119 of them were incorrectly classified. The graph is colored to show magnitudes of each category.



A graph with our predicted probability on the axis and the actual probability on the y access. With the perfect one to one line shows as a red dotted line.



A nearly perfect ROC curve that shows the false positive rate on the x axis and the true positive rate on the y axis.