FINAL REPORT

Calorie Usage

Shane Lafollette

PROBLEM:

As technology has advanced and we have learned how to use large data to help answer tough questions, Healthcare data has exploded in growth. Any device that is usually carried on your person, usually has a health tracking app today, like your phone or smart watch. You can also buy specialty activity trackers that accumulate even more data points. Things such as steps, elevation, distance traveled, heart rate, oxygen saturation, temperature, sleep time, and stages of sleep are monitored. Many users like these devices because they help them keep track of exercise and energy output throughout the day while keeping fit or dieting. Along with dieting, one main data point that people look at are calories burned. This data gives the user an estimate of how much energy they've used and helps them to estimate how much food they should intake. This data point is an estimate that is calculated from other data points. As Fitbit has devices that have been tested to be within 3 to 4 calories of actual calorie usage, can we use their data points to identify which data is most important and how each contributes to the calorie estimation.

Data:

The Fitbit data I reviewed was somewhat small but should provide the insights that we're looking for. The data was from 33 Fitbit users over a one-month period. The data was divided into columns segmented by minutes, hours, and days. I dropped the measurements by day and hours, as using the information by minute should give us the closest exact calorie measurements. The calories burned column is a daily total. Heart rate data was recorded every 5 to 10 seconds. To deal with this large amount of heart rate data, it was grouped by user and day and averaged over the day time period. I dropped rows where the user forgot to wear their device and no activity data was logged. The activity or steps is measured by quantity and time and was divided into columns of sedentary, light, moderate, and heavy activity. After cleaning and removing data not needed and data that had large rows of missing values, our dataset is 844 rows and 16 columns.

EDA:

Since the calories column is an estimation based on data in other columns, correlation between our data and target variable should be high. Our analysis is to verify which data points are the most important. Figure 1 shows the correlation between each variable in our dataset.

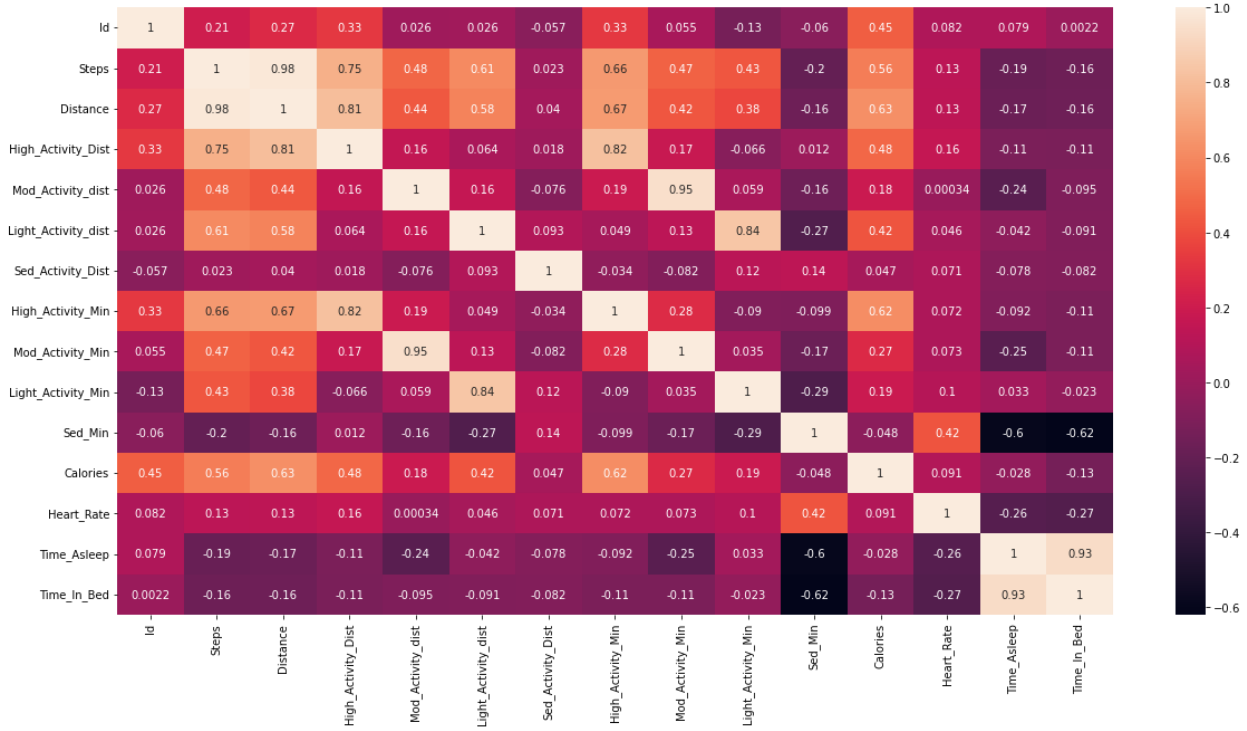

Figure 1. Correlation between variables

The highest correlations between our variables the target variable is Distance (0.63), High activity minutes (0.62), and Steps (0.56). The activity metrics have a positive correlation, while the sedentary minutes, time in bed and time asleep show a very small negative correlation.

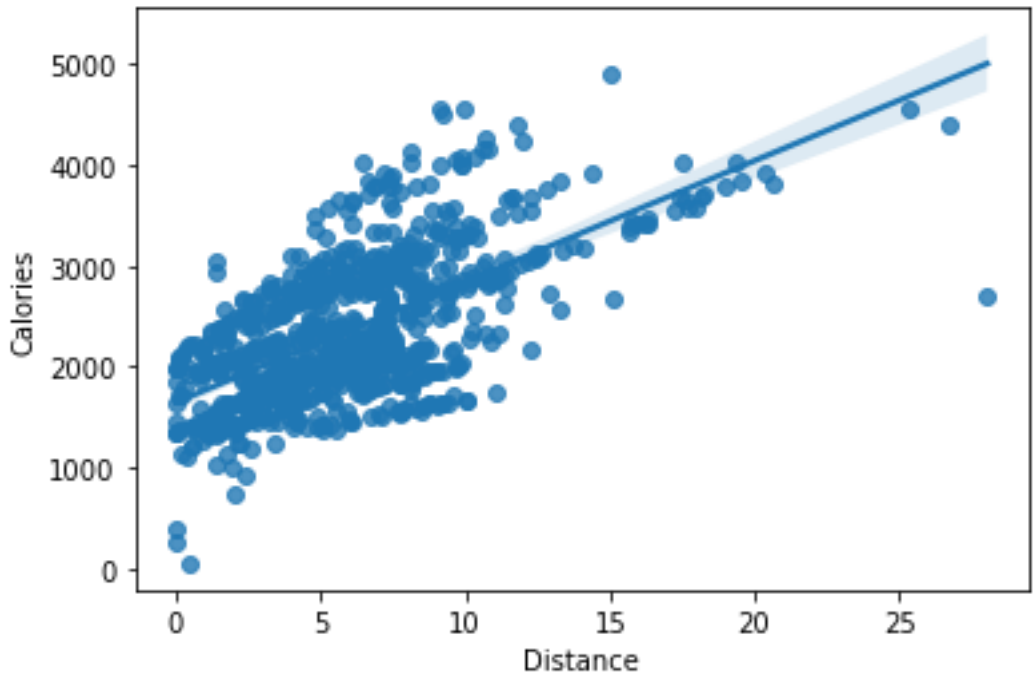

Figure 2. Distance vs. Calories with line of regression

Figure 2. shows the correlation between Distance and Calories. While the line of regression seems to follow with larger distance data points, there seems to be great variation at the lower end. Figure 3. below is a line graph showing a user's distance vs. calories burned over the dataset. We can see by this line graph that the target variable follows very well with Distance traveled.

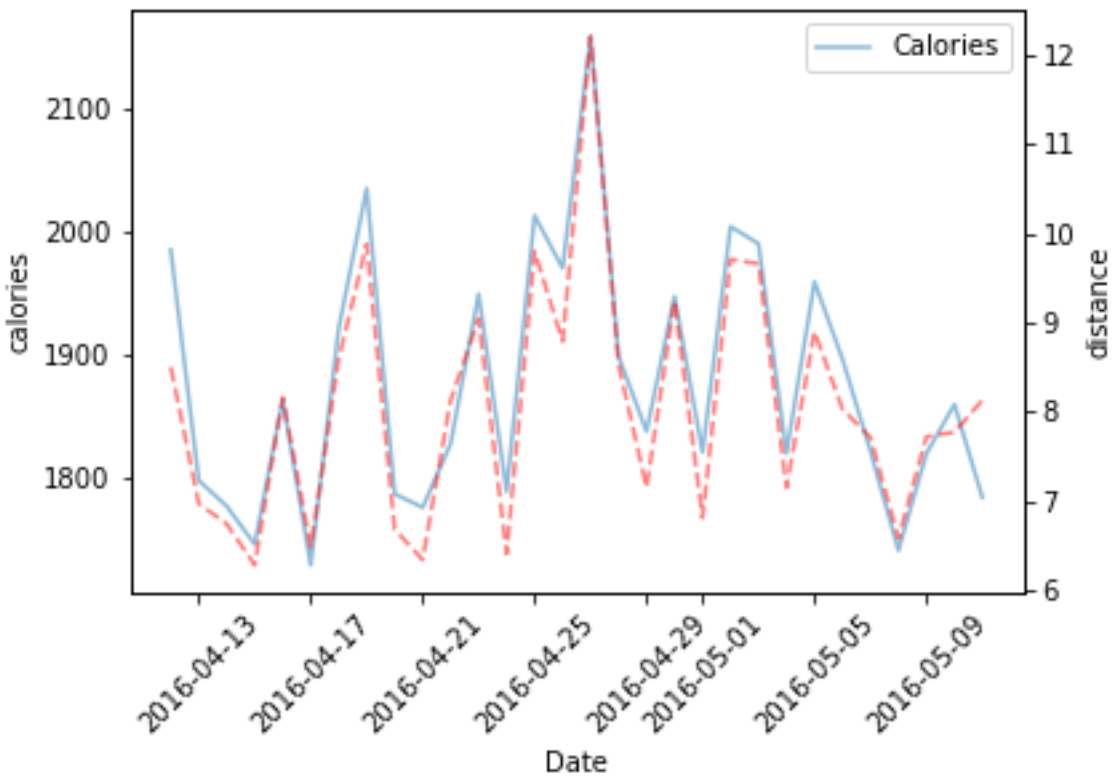

Figure 3. Calories and Distance across Month

The heart rate data was missing in 60% of the data, so unfortunately, I had to cut the column from the data, fearing the model would get skewed by the data points in such a small dataset. I also didn't want to fill the missing values as they could vary greatly depending on every other variable. The sleep data was also missing large amounts of data, so I cut it from the modeling process. This loss of data in these columns could have occurred for several reasons, but with both occurring simultaneously and the user still logging steps and distance, it is likely that some user devices were not capable of reading this type of data and some users did not wear their devices to bed.

Modeling:

As the analysis is looking for the number of calories burned in a day, a regression model will be needed and we have nine variables that show correlation, so it'll be a multivariable regression model. Using scikit learn, several regression models were tried.

These models used all nine of the variables we've been able to keep:

- **Steps**
- **Distance**

- **High Activity Distance**
- **Moderate Activity Distance**
- **Light Activity Distance**
- **High Activity Minutes**
- **Moderate Activity Minutes**
- **Light Activity Minutes**
- **Sedentary Minutes**

The data was divided into a train/test split. 25% or 211 testing data points were set aside. The training data was standardized to get the data on a relative scale. I started with OLS Linear regression, then proceeded to Lasso and Ridge regression to see if I could improve the fit of the model. Below are the R-squared results for each model.

R-squared scores against test set

OLS: 0.768812

Lasso: 0.769058

Ridge: 0.766663

The first model, an OLS model, finding the best line of fit by least squares of residuals, was fit using the training data and 5-fold cross validation. R-squared values are a metric that represents how well the model line matches the data given (1.0 being a perfect fit). The cross-validation scores ranged between 0.722 and 0.791. After standardizing the test set, we predicted from and scored the model on the test set. The R-squared score of 0.768 is a good starting score before searching for the best hyper-parameters. Figure 4. And Figure 5. show how the model's predictions did against the test set. Figure 4. Gives a good indication where our model does a good job and where it could improve. The model does well around the 2,000-calorie mark, where the large amount of the data lies, while it starts to lose predictive accuracy at the ends of the scale. Figure 5. shows each data point and the model's prediction for that data point. You can see there seems to be a floor to our model, as it doesn't seem to make predictions below around 1,400 calories. It also seems to over-estimate above 4,000 calories.

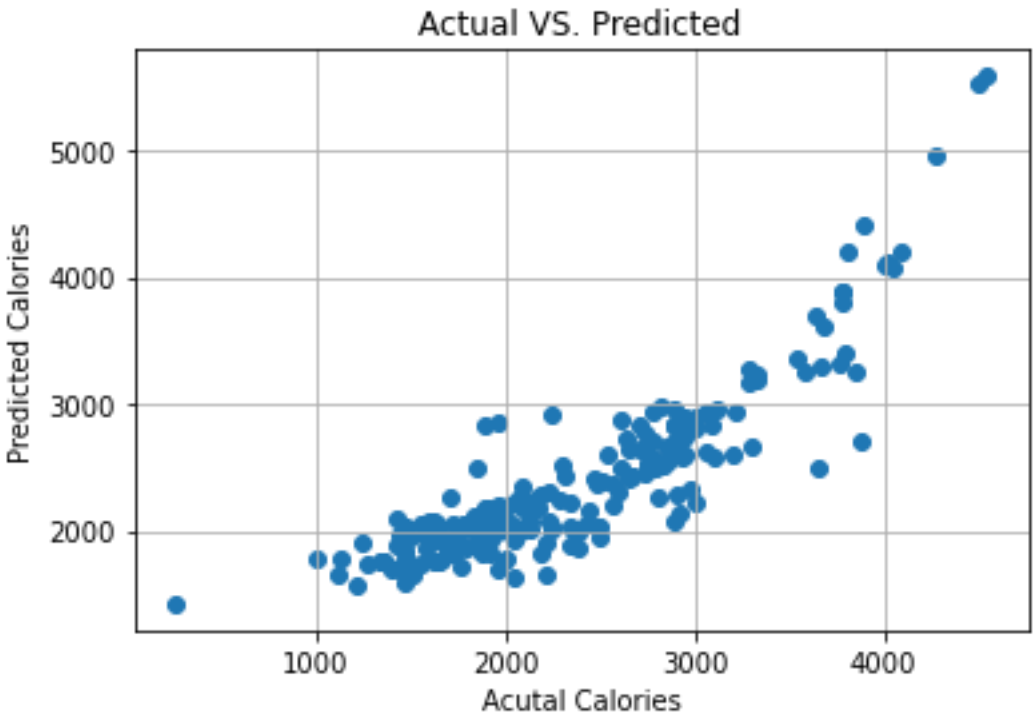

Figure 4. Actual calories vs. predicted calories (Above)

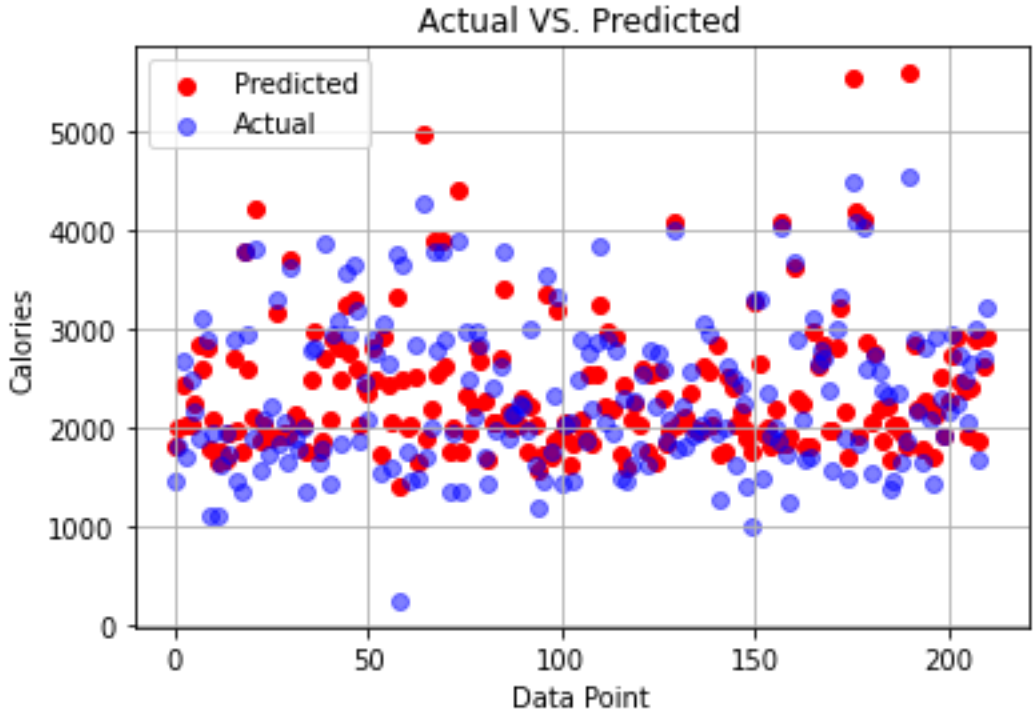

Figure 5. Actual data points vs. predicted data points

The next model tested was a Lasso Regression model. This model penalizes coefficients as they grow larger and zero some coefficients that aren't important to our model. This model used the same standardized training and testing sets with cross-validation. Training scores ranged from 0.728 to 0.788. The R-squared score is basically the same at 0.769 on the test set. Figure 6. shows the predictions from the OLS model plotted over the Lasso predictions. Where we see RED is where changes were made with the model.

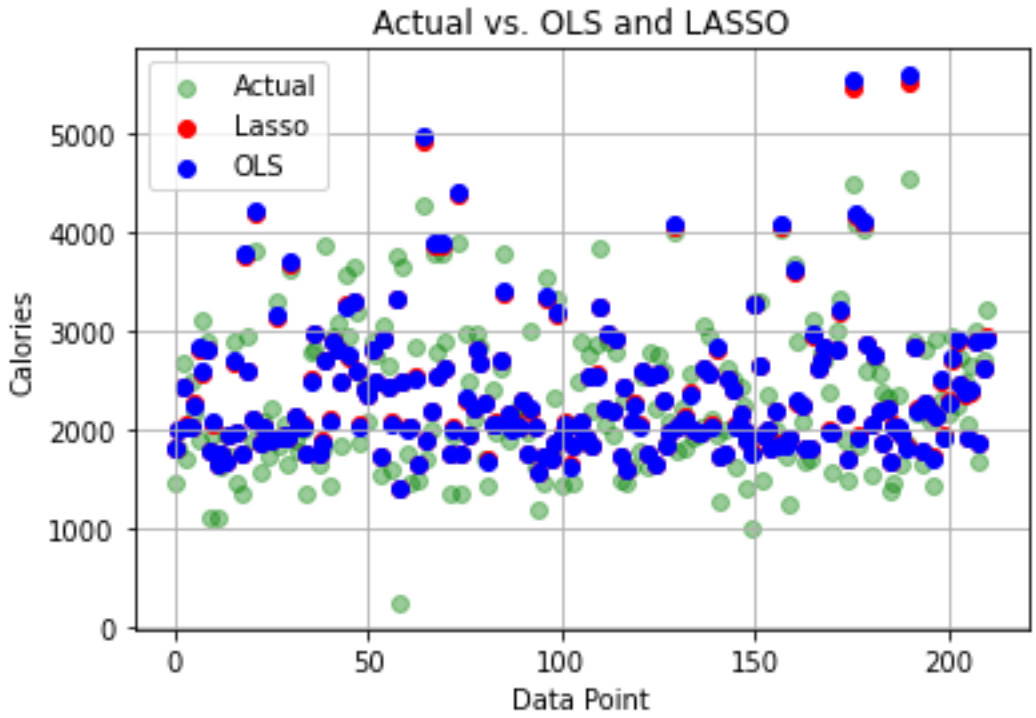

Figure 6. OLS and Lasso predictions against actual

The lasso model pulled our high predictions down and pulled lower predictions slightly closer to center. No large gains were made with this model.

A Ridge regression model was fit next. Ridge regression also penalizes our coefficients but uses a different equation for the penalty. Scores on the training data ranged between 0.712 to 0.778. The R-squared score was 0.766, slightly lower than other models. After graphing these predictions against Lasso predictions in Figure 7., this model doesn't give any better information.

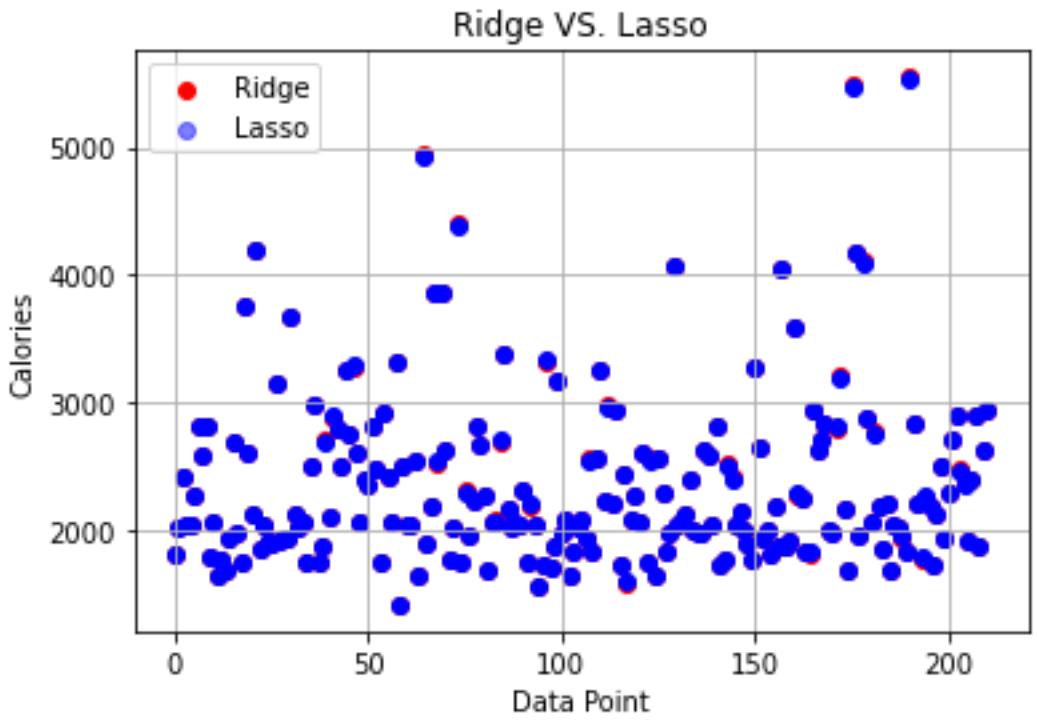

Figure 7. Ridge vs. Lasso predictions

Tuning:

As we had only nine variables in the models, I doubted that simplifying our model by dropping some of our models would help our goodness of fit. I looped through each model and a # of features using a minimum of three but found no improvement with the R-squared scores. The four most important features found was Steps, Distance, High activity distance, and high activity minutes. With these four features alone, we could acquire an average R-squared value across all models of 0.74 on the training set. The other variables included doesn't hold much weight, but do make the predictions better overall.

SKOPT BayesSearchCV was used to search through the hyper-parameters of the Lasso and Ridge models to find a combination of parameters to get the best predictions from the model. The main hyper-parameter in these models is the Alpha parameter. The Alpha parameter determines the size of penalty put upon the coefficients. With the Alpha set to the optimal number, the Lasso model R-squared test result moved slightly to 0.76926, while the Ridge model went to 0.76908. This shows that even with optimization, the predictions barely moved.

Summary:

With these last steps, the conclusion must be drawn that an R-squared value between 0.75 – 0.77 is what can be expected from this small collection of data. The score may get better with a greater number of records, but I believe our list of variables will help the most. Receiving data with large amounts of the heart rate and sleep data intact will help greatly, as well as certain fields that contain categories that will help draw conclusions about the individual burning the calories. This Fitbit data doesn't show it, but some of the most important factors in calorie consumption is Age, Sex, Height, and Weight. These four categories set the base for how many calories a person burns throughout a day if they are completely stationary. Fitbit asks the user for this data when the device is activated to set a baseline for their estimate. This measurement (BMR) would give the model a better starting point or y-intercept dependent on these data points.

Without these important fields, the models created have done very well in predicting calories burned. Without knowing sex, age, or someone's size, which the participants probably varied greatly, the lasso model had the best average prediction score, mere tenths of a point higher than the OLS model. The difference between the two, when adjusted to calories by the RMSE of each model:

OLS RMSE: 367.86

Lasso RMSE: 367.50

This RMSE tells us our Lasso model's average distance from the actual data point was 367.5 calories while the OLS model is at 367.86. These models are performing almost identical on this small dataset. It would be recommended that after acquiring more data points, to revisit these metrics to identify if one model consistently outperforms the other. The 370-calorie average error is equivalent to one McDonald's Filet-o-Fish sandwich. Probably not the best model yet for people dieting, as it's error could cost them almost a whole meal worth of calories. This error could also be above or below our actual number. While this model is not ready to make accurate predictions, it has given us a great starting point and helped us to understand what information we are lacking in our data.