

Shane Lafollette

Predicting Daily Calorie Usage








Problem:

- Fitbit sells over 10 million activity trackers annually. Fitbit collects data on steps, distance, activity intensity, heart rate, sleep time, and quality of sleep. With all this data collected, they have created one of the best estimators of calories burned.
- Can we find what data was used for the estimation?
- How important is the data to the estimation ?
- Can we build a model that generates close estimates ?

The Data:

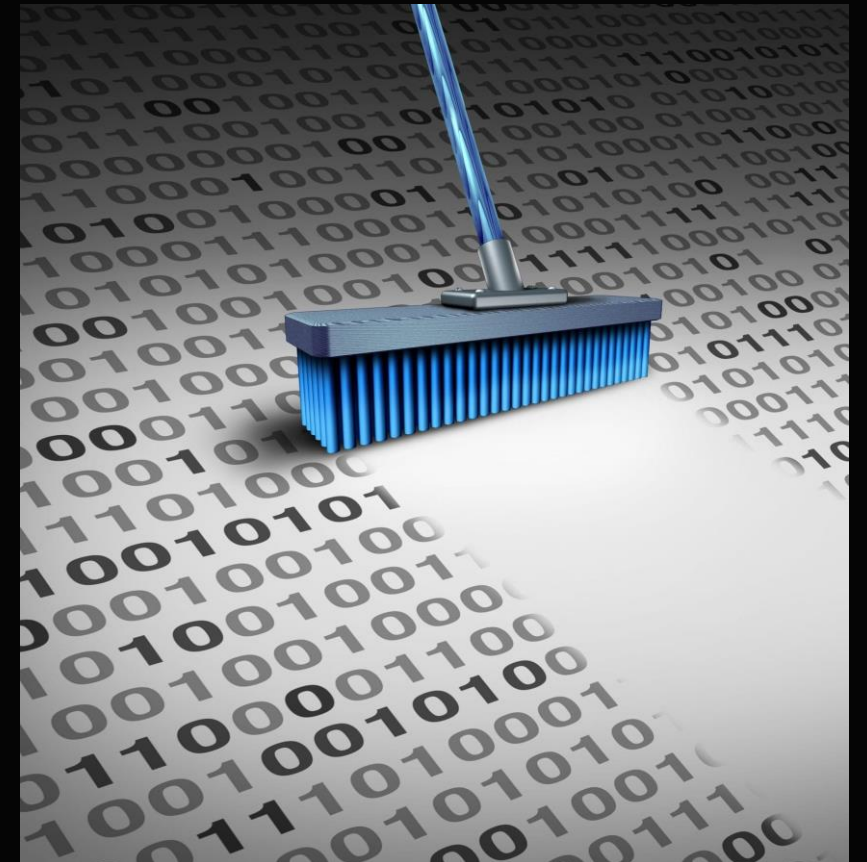
- The Fitbit dataset was downloaded from Kaggle.
 - The data was taken from 33 users over a 1-month period.
 - The data was had several data columns on distance, steps, heart rate(HR), intensity, and sleep.
 - Most columns were duplicate data broken down by different time intervals.

 Fitness Devices / Apps	 Steps	 Heart Rate	 Speed	 Sleep
Fitbit Device (via Fitbit app on iOS and Android)	✓	✓	✓	✓

*Depend on the device models and features.

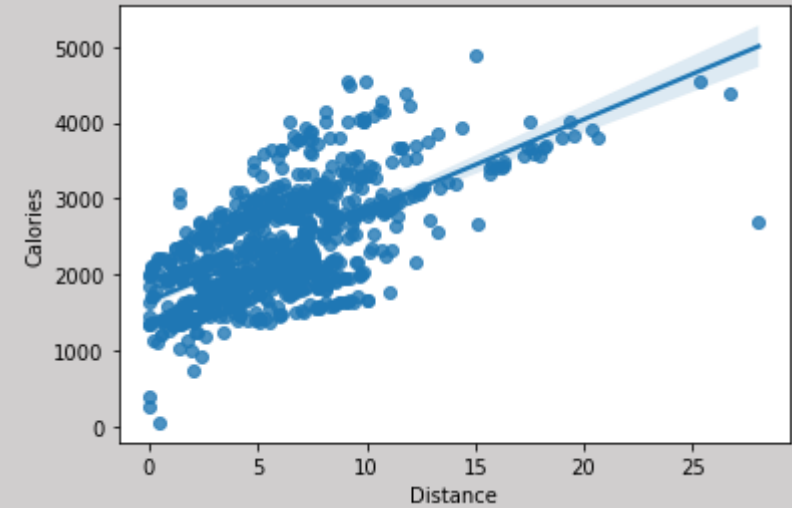
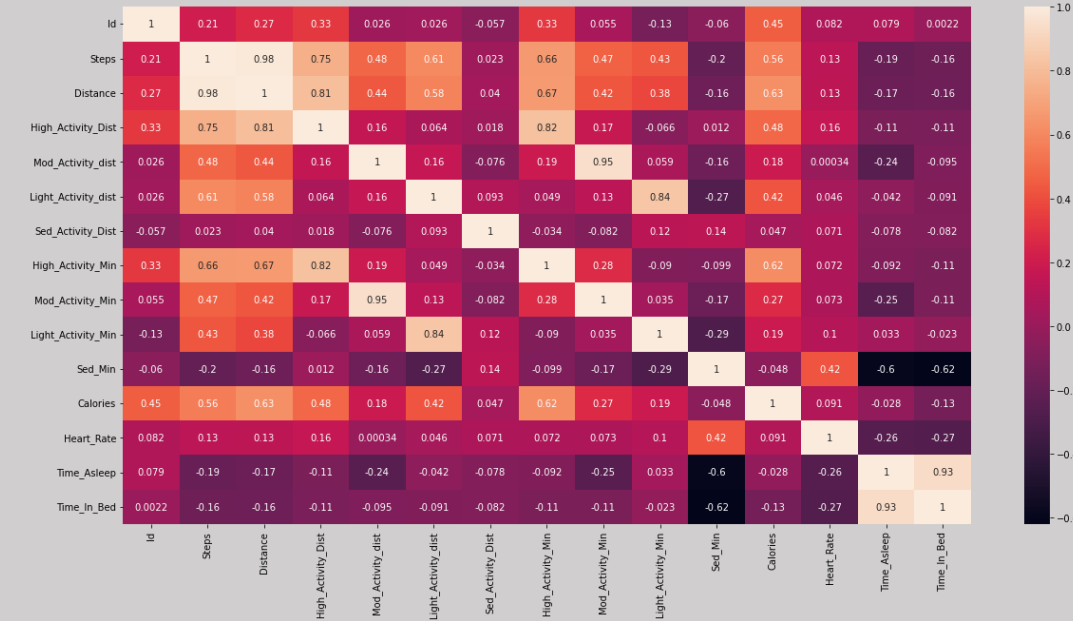
Data cleaning:

- The minutes time interval was chosen over day or hour totals.
- The HR and sleeping data was dropped due to large amounts of missing data. The decision was made to not impute values due to the small size of dataset.
- Rows that were found to have large amounts of missing data were dropped.
- The user-id was dropped, and the all rows were treated equally regarding how we estimated calories. This is going to lose some data for the estimation, but with a small dataset, combining all data should help the overall model rather than creating a model for each user.
- After cleaning and dropping columns, the data set consists of 9 columns of independent variables that will predict the dependent calorie variable and has 844 rows of user data.

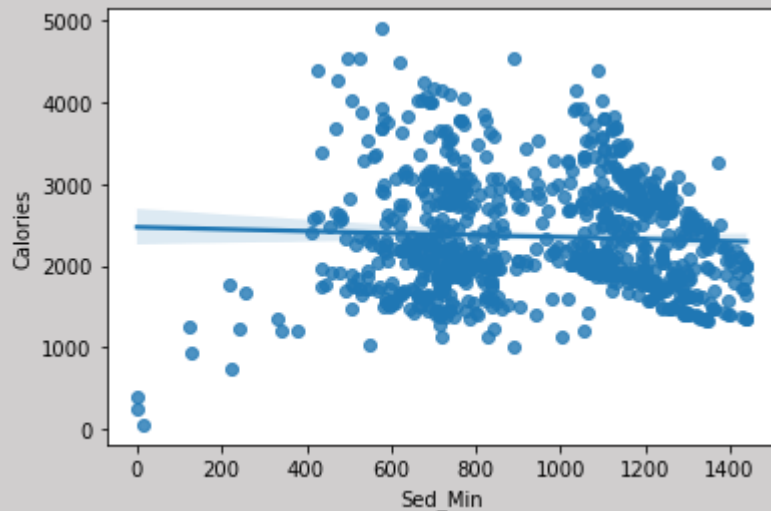


EDA:

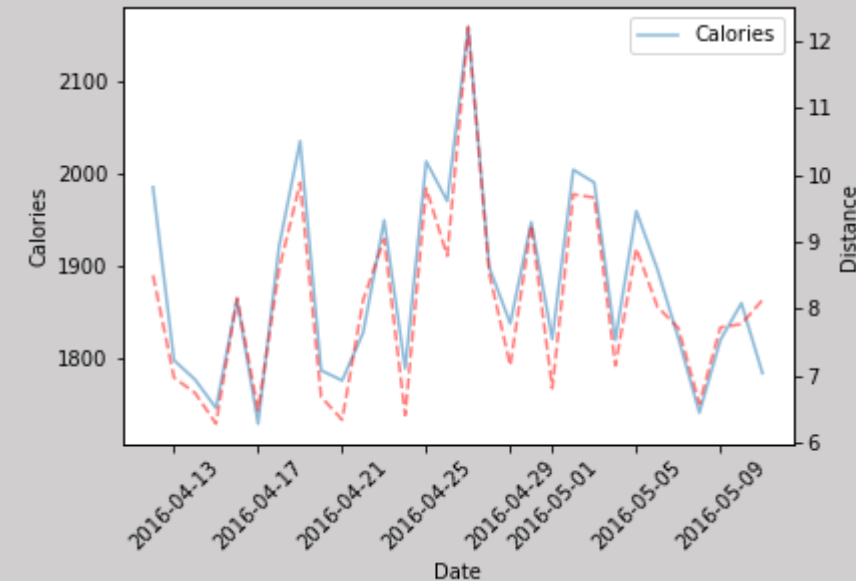
- Check for correlation between variables and the target.
- Also, variables that correlate very well with each other, may show variables that can be removed. Since the dataset was so small, the steps and distance column were kept.
- The sedentary minutes column is a bit confusing. It seems to show no good pattern but may provide some data that will be informative.



- Distance and calories have a good positive correlation.



- Calories vs. Distance tracked by date shows how both variables follow very closely with each other.



Models:

The models will be looking for a continuous variable, so a linear regression model will be used.

This model finds the best line of fit for the data given. R-squared(r^2) is the scoring metric to determine how well the line fits the data.

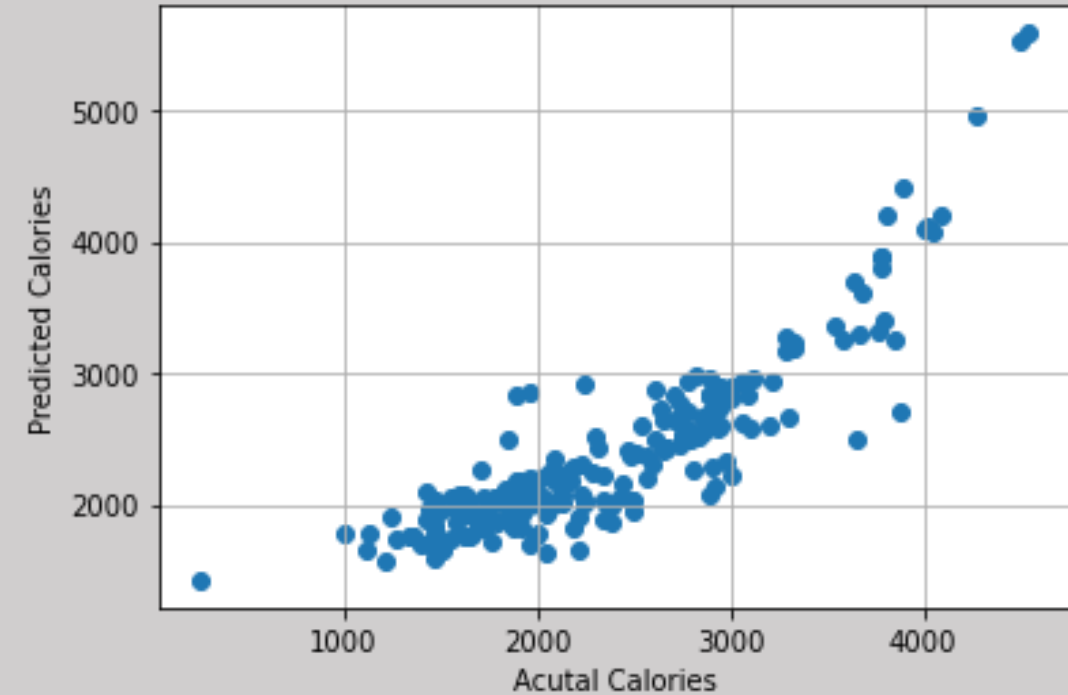
This model should provide a good starting point for predictions

The next models will be linear regression, but with some hyper-parameter tuning to make better predictions



OLS Linear Regression:

Actual VS. Predicted



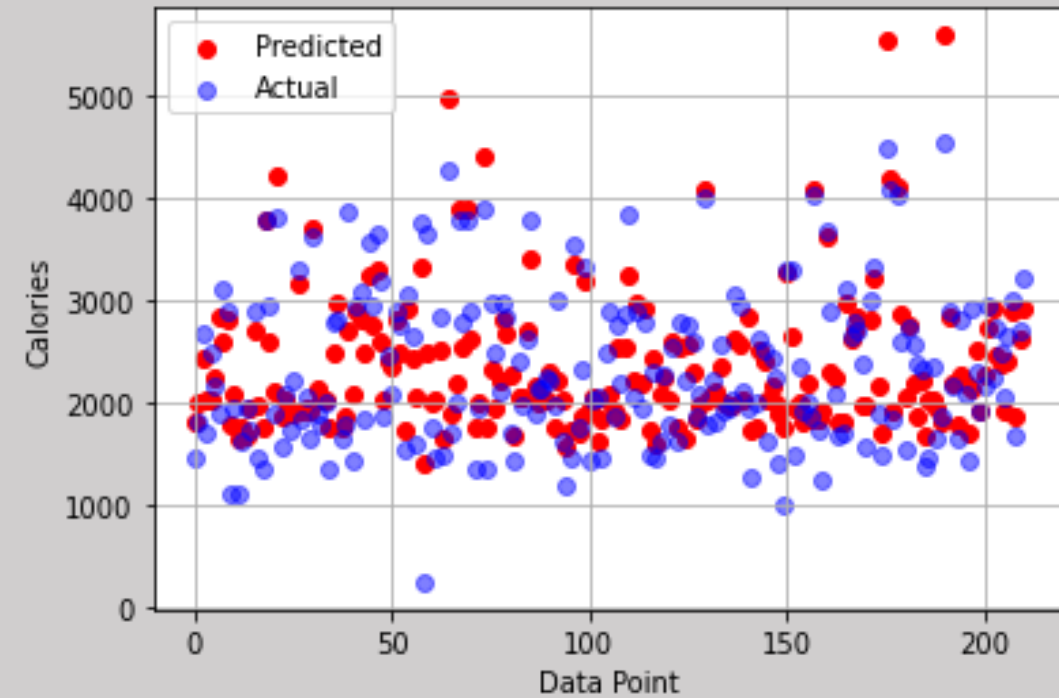
r2 score on testing dataset:

0.768

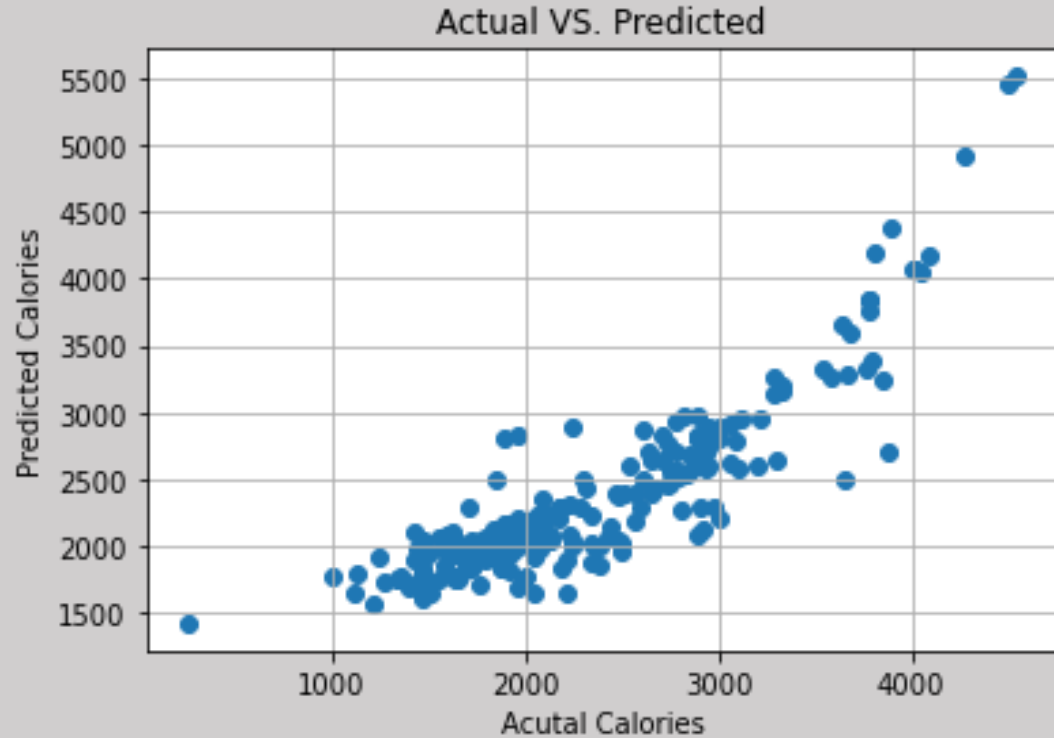
5-fold cross-validation r2 score range on training dataset:

0.722 – 0.791

Actual VS. Predicted



Lasso Model:

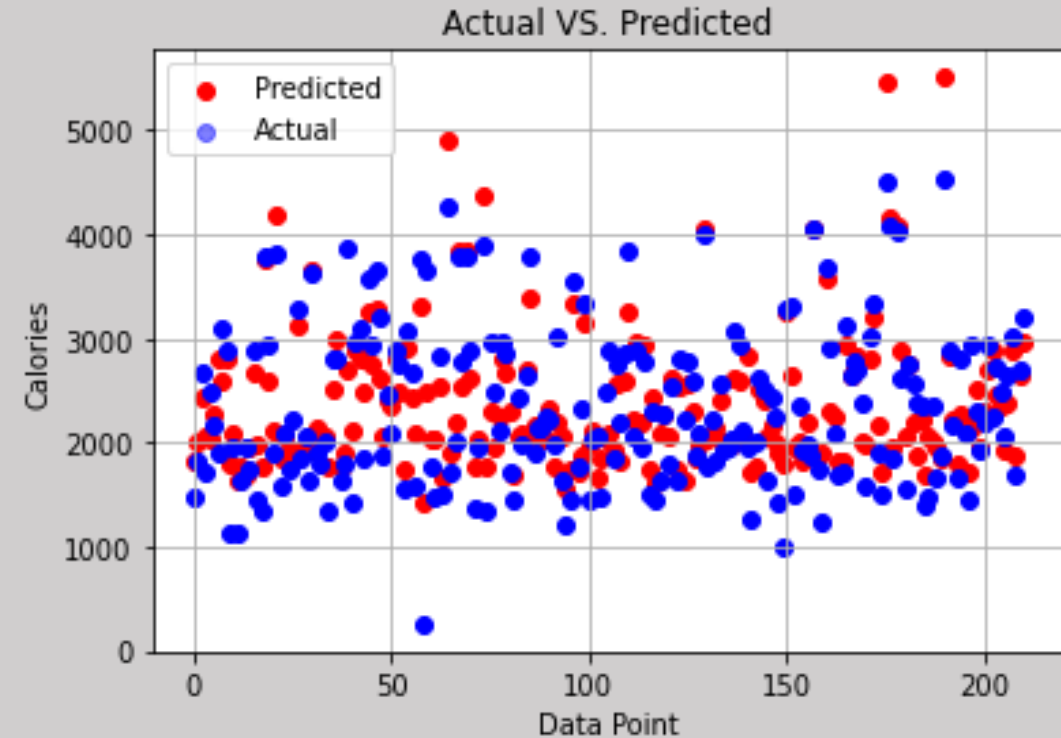


r2 score on testing dataset:

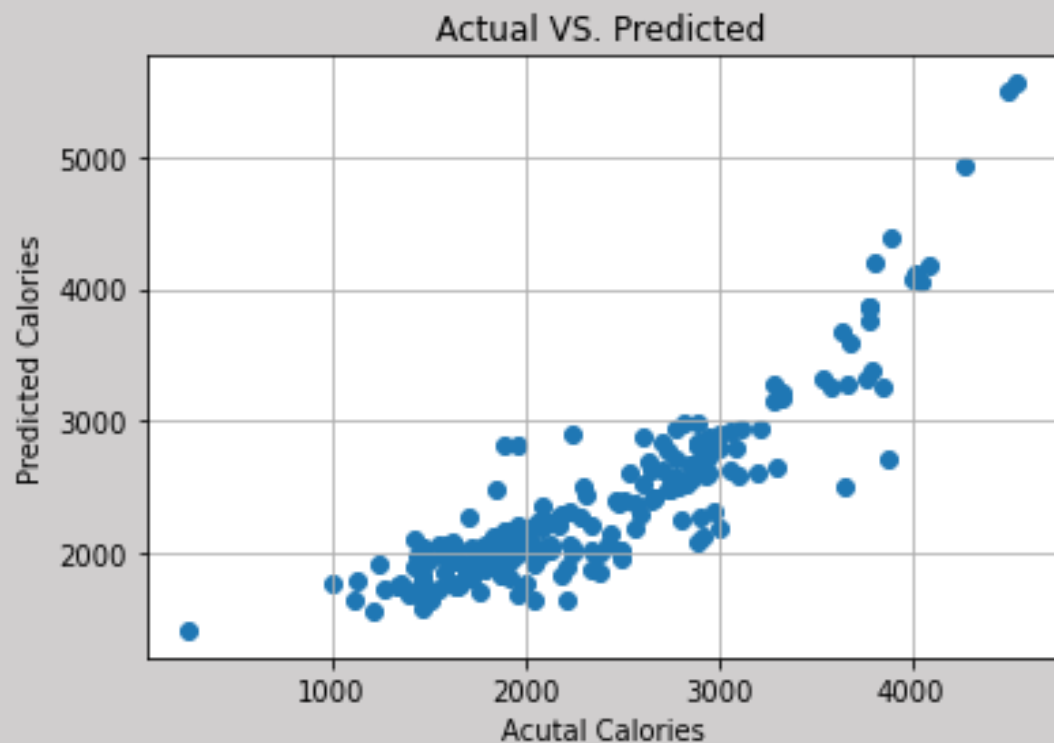
0.769

5-fold cross-validation r2 score range on training dataset:

0.728 – 0.788



Ridge Model:

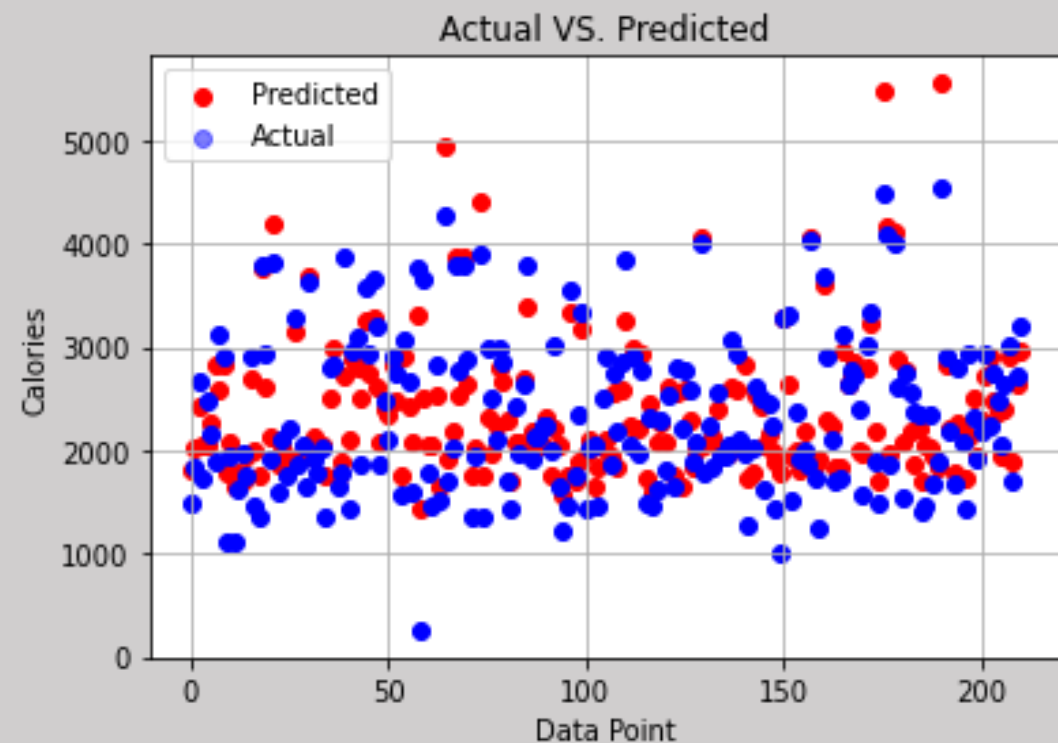


r2 score on testing dataset:

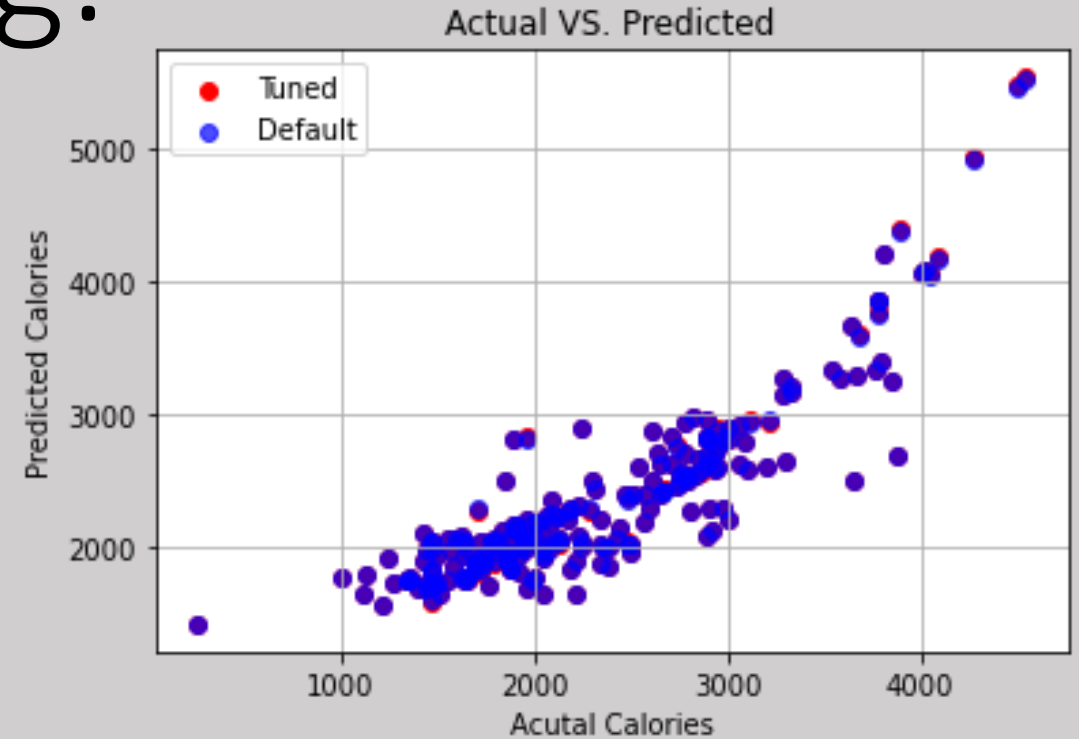
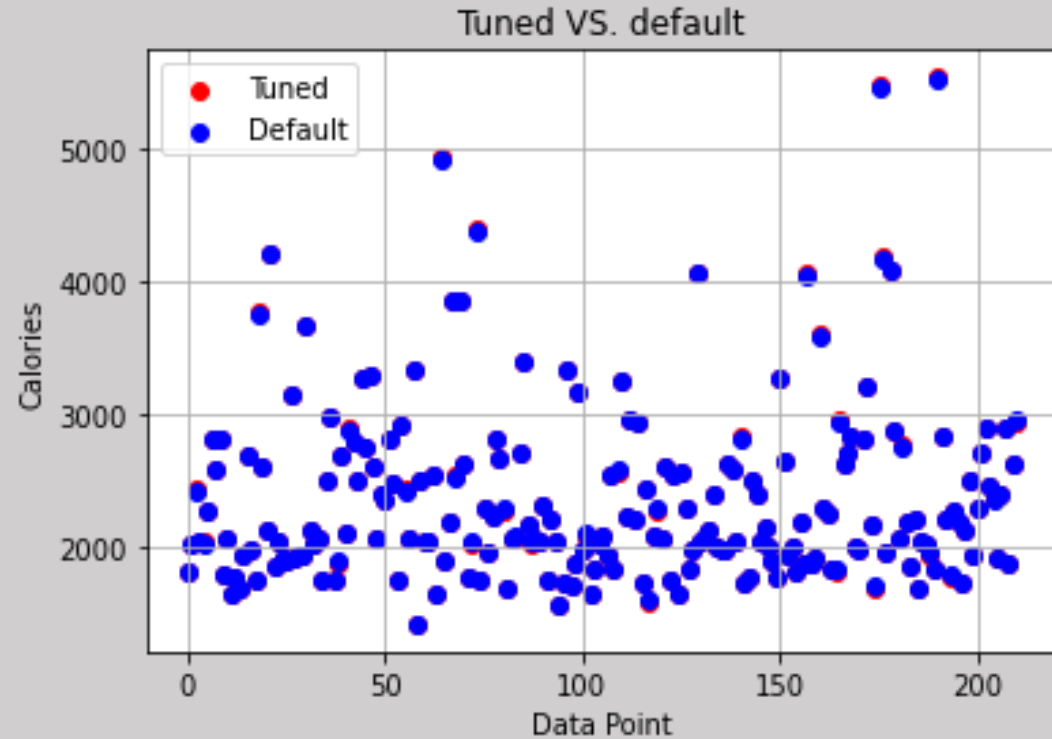
0.766

5-fold cross-validation r2 score range on training dataset:

0.712 – 0.778



Tuning:



We found the best combination of hyper-parameters using BayesSearchCV. This search ran cross-validation scores on the data using different combinations narrowing down and finding the best model to fit the data. As you can see above, the **Lasso** model did well from the default settings, with the tuned model slightly changing predictions at the higher and lower estimates of the target variable.

Final Model and Summary:

The model chosen after tuning and using cross validation was the Lasso Regression model. Final scores were **0.7692** on the test set and the best score of **0.7557** on the training set using cross validation. The hyperparameters helped on the slightest of scales, and the Lasso Regression outscored the other models by less than **0.005**.

The finished model, with a line of fit with an r^2 score of 0.7692, is still off by an average of **367 calories**. These calories are equivalent to one McDonald's Filet-O-Fish sandwiches. So, while we have a decent scoring model, improvement is still needed to create a model that can be trusted when used for dieting or exercising. The Steps, Distance, High-intensity distance and minutes were the most important variables for the predictions.

Using only the data we started with, this model does a fair job of finding a good fit. I would recommend more data. User data such as sex, height, weight, Heart-rate, and age would greatly increase our prediction accuracy. I would also recommend a larger collection of the data. A six-month collection, unlike the 1-month collection I worked with would also be very beneficial.

How close are we to an optimal model?

1 Fish Sandwich

