Writeup

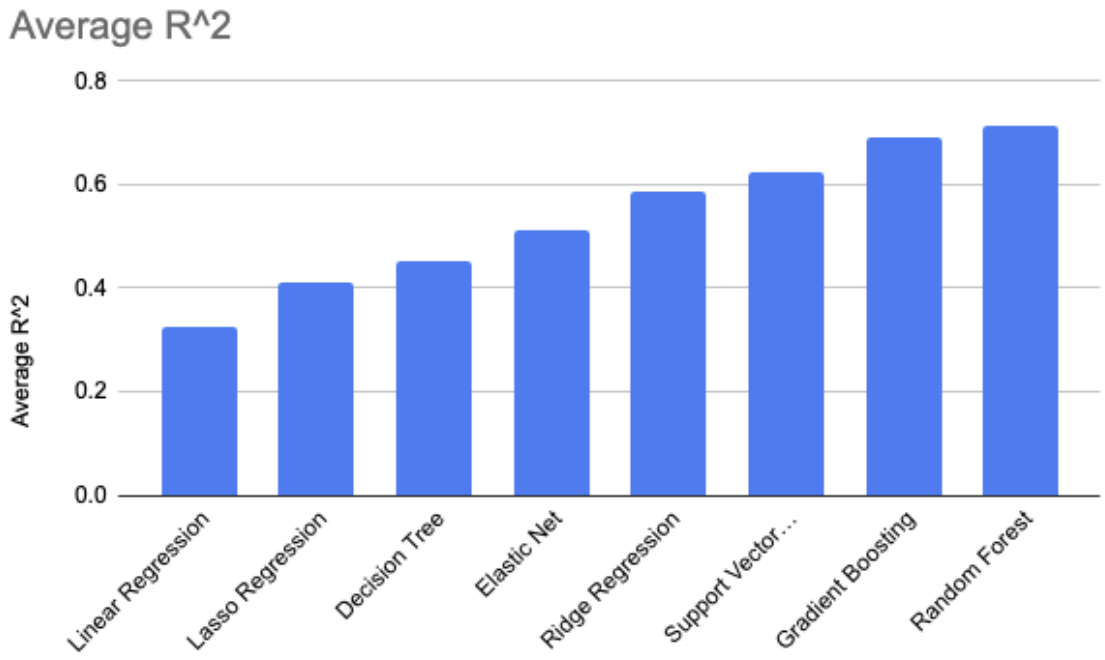
Code: https://github.com/munir-v/ml-final-project

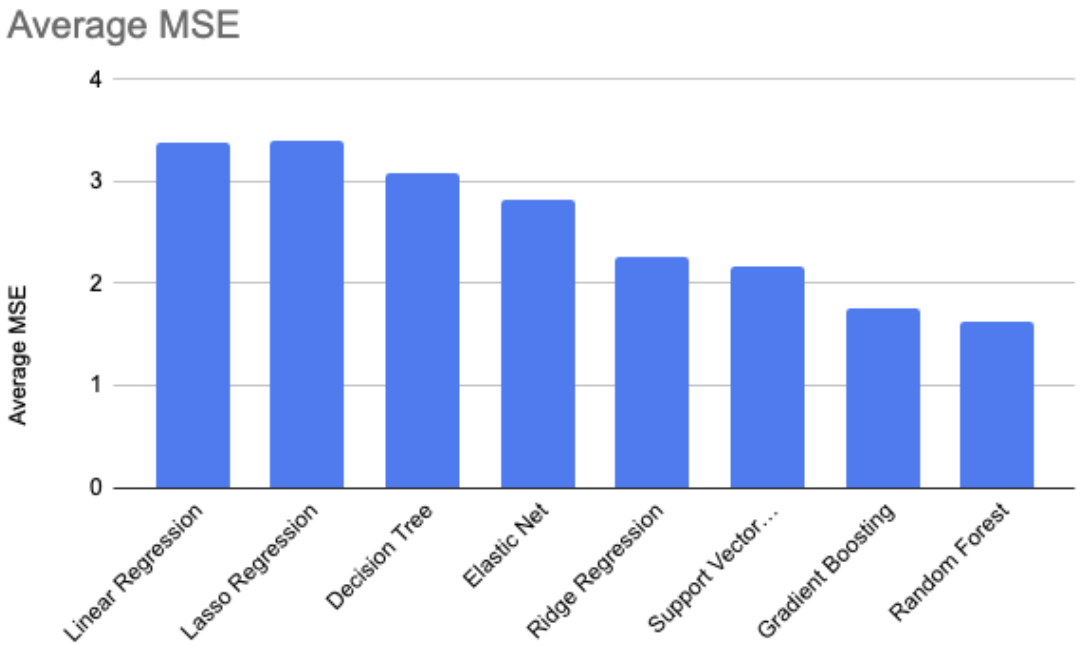
Introduction: My original plan was to make predictions involving my sleep and exercise data, using features like daily exercise, daily steps, resting heart rate, etc. I ultimately settled on trying to predict my resting heart rate (RHR) for a given day, using my sleep and exercise measures as part of the feature set. RHR is a discrete variable and lends itself better to prediction than a continuous variable like total sleep time or a measure of sleep quality. I wanted to test several models to see which performed the best at predicting my RHR.

Experimental setup: I used almost a year's worth of my own Garmin fitness watch data, which included features for each day such as total steps; daily minutes of moderate/vigorous exercise; amount of time spent in REM, deep, and light sleep; among other metrics.
Since my data had around 23 features, not all of which were potentially good predictors of RHR, I performed recursive feature elimination with cross-validation using a scikit-learn package to remove the less relevant features. This removed the "moderateIntensityMinutes", "averageRespiration", "lowestRespiration", and "awakeCount" features. I then tested eight models from scikit-learn, using 10-fold cross validation. I used a two-tailed t-test on the r^2 and mean squared error (MSE) values to evaluate my results.

Results: I used average measures of the MSE and the r^2 values to compare the models (see charts below). Random forest had the highest average r^2 value of 0.76 and the lowest MSE of 1.00. However, running a t-test comparing it to gradient boosting yielded 0.06 for the r^2 value and 0.08 for MSE, so this test does not show a statistically significant difference between the two models. Running a t-test comparing gradient boosting to support vector regression does yield a significant difference, though, with values of 0.01 and 0.02, respectively.

Conclusions: The gradient boosting and random forest models performed the best at predicting my RHR based on the supplied features, and I did not find a statistically significant difference between them.


Average R^2
0.8
0.6
0.4
0.2
0.0
Average R^2
Linear Regression
Lasso Regression
Decision Tree
Elastic Net
Ridge Regression
Support Vector…
Gradient Boosting
Random Forest


Average MSE
4
3
2
1
0
Average MSE
Linear Regression
Lasso Regression
Decision Tree
Elastic Net
Ridge Regression
Support Vector…
Gradient Boosting
Random Forest