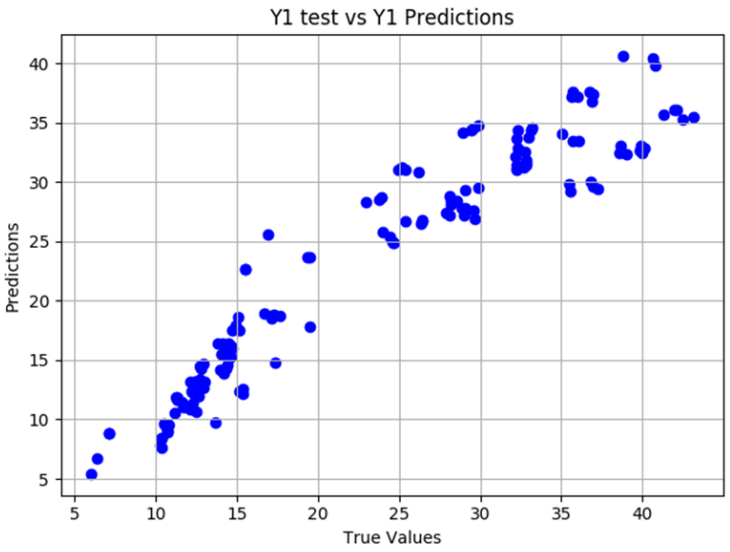
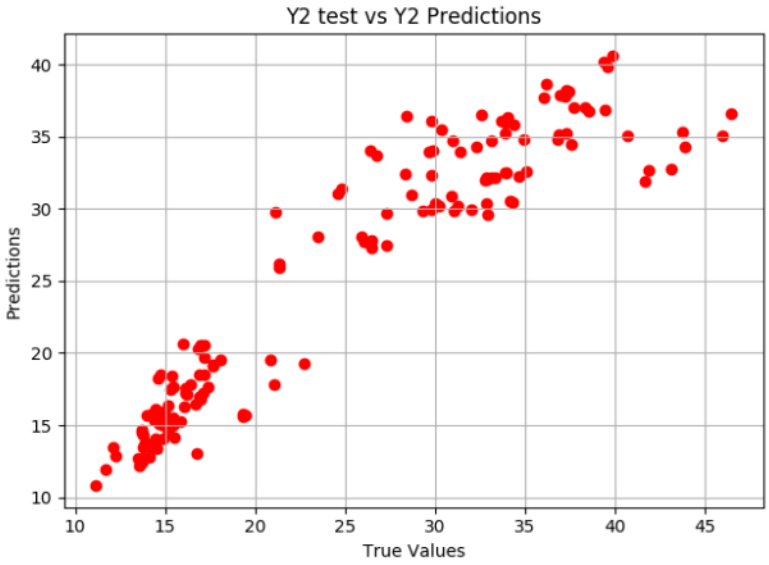
Kendra Maggiore
3/26/19

Machine Learning Homework 1: Linear and Logistic regression

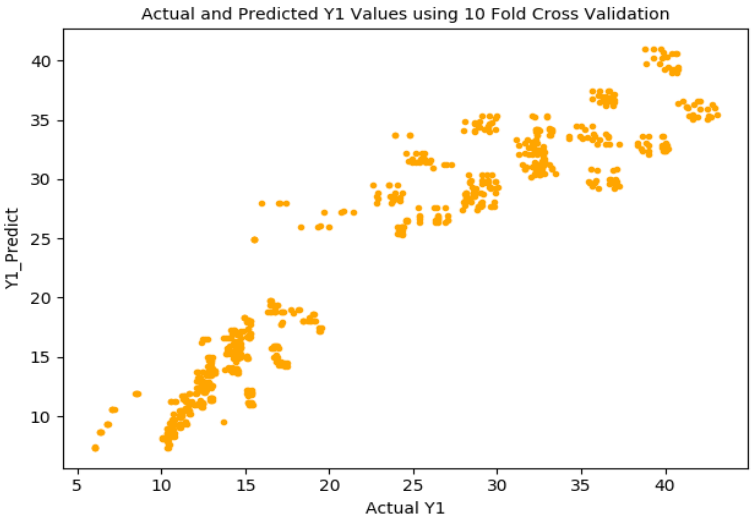
For our homework 1 in machine learning we had to preform linear and logistic regression on energy analysis using 12 different building shapes simulated in Ecotect. For my linear regression I set up this problem by using various imports such as matplotlib to do my graphs, pandas to handle my data input, and sklearning to do the linear regression. The first thing I did after my imports was to separate the eight X data points from the two Y data points. I then separated these datasets into X, Y1 and, Y2 test/intermediate/train/validation sets. I deleted the intermediate set after I introduced the validation set, because the intermediate set was only meant to help separated the data into test/train/validate sets.

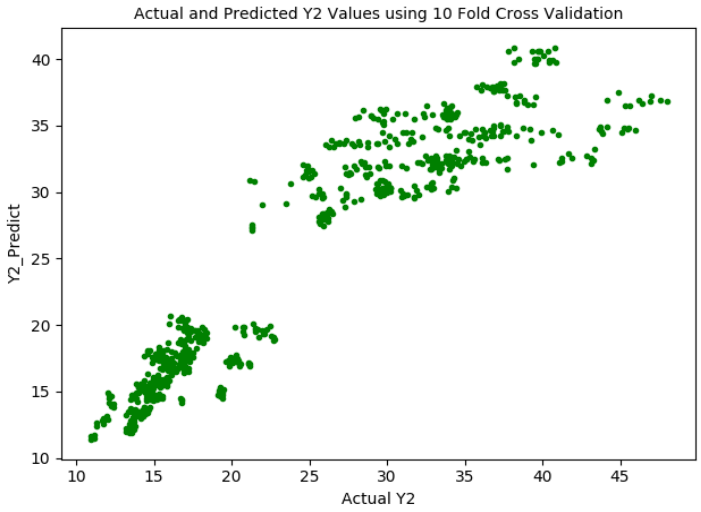
Following this I used standard scaler to help scale the data to have the same decimal cut offs. I then implemented the linear regression using sklearning for my Y1 train, and my Y2 train. Then I predicted my data, and its accuracy. Following this I ran Kfold cross validations on my training, and validation data. My data was split into 60% train, 20% validation, and 20% testing. My predicted accuracy for Y1 was about 91%, and Y2 was about 85%. My cross-validation R2 accuracy for Y1 was about 90%, and Y2 was about 88%.





Kendra Maggiore
3/26/19



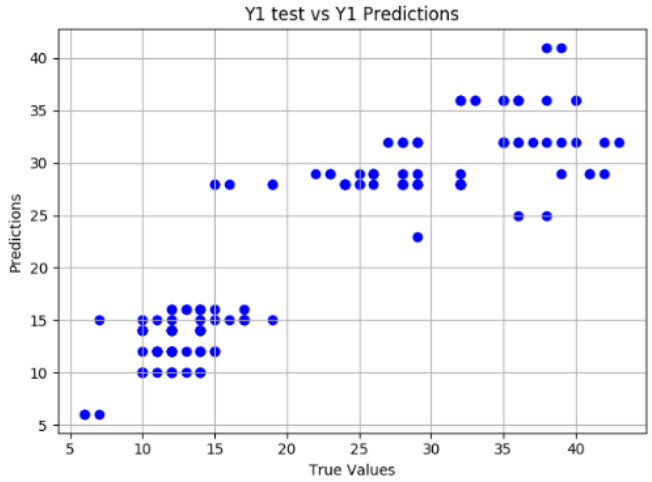


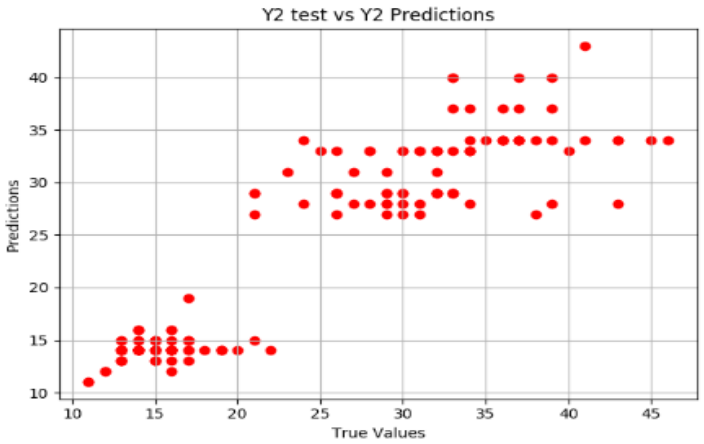
Above are four of my graphs from my linear regression. They all show a linear regression upward in data. That is common among all the data sets.

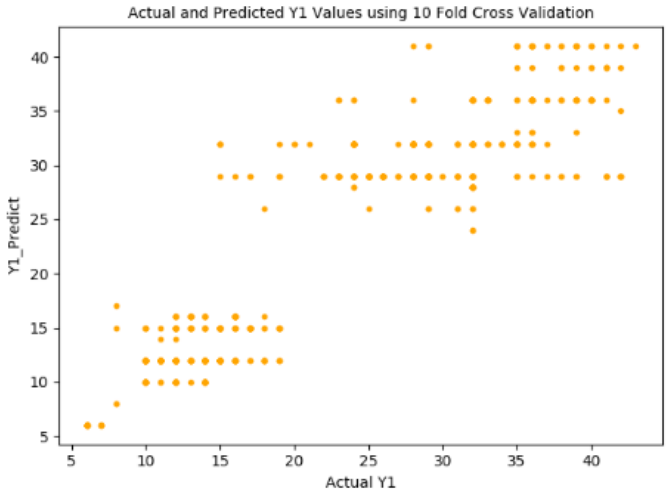
For my logistic regression I did much of the same thing I did for my linear regression except before the dataset was processed, I used astype to round my data set to integer sets instead of using floats. Since logistic regression only takes whole number inputs. I split the data the same way into test/train/validation sets. Then when it came to using the data in a logistic regression model I used sklearning logistic regression. I set my model to 'lbfgs', my multi-class to 'ovr', max integration to 1000, and my random sate to 0. I also did Kfold cross-validation. My cross-validation ran into issues with some of the classes only having one data point, and this did not allow that iteration to run.

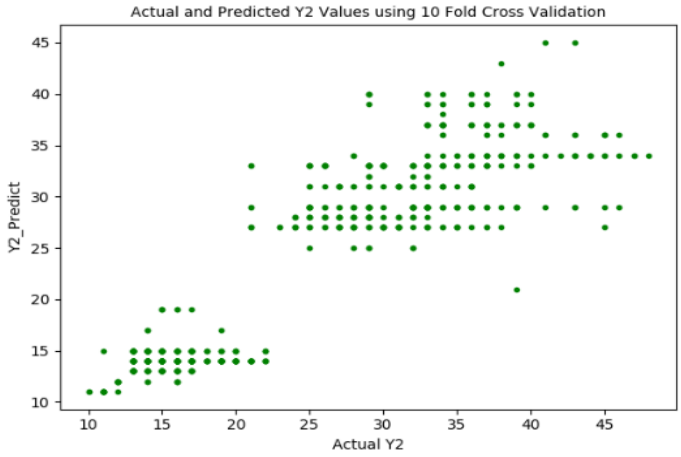
My data for logistical regrssion was split into 60% train, 20% validation, and 20% testing. My predicted accuracy for Y1 was about 27%, and Y2 also about 27%. My cross-validation R2 accuracy for Y1 was about 17%, and Y2 was about 1%.

Kendra Maggiore
3/26/19









Above the four graphs again show the same linear regression upwards, but you can visualize the different classifications based on how the graph is grouped.

In conclusion, I believe that this data ran best as a linear regression, as shown by the more accurate results. It is my option that the logistic regression did not perform as well because there were to many classes to classify and this hurt the resulting accuracy. I better improve the logical regression accuracy, I believe that grouping the classification into smaller groups could help for insurance all number from 0 to 10 in a group and so on.