COSC 2673 Machine Learning

Assignment 1

Introductory Machine Learning

Joshua Hansen

s3589185

Introduction

The aim of the assignment was to use linear regression to help predict target death rate of cancer sufferers. The data set was taken from a city in USA and is freely available to download. SciKitLearn was used for the different linear models to allow us to focus on tuning the regression models to make them more accurate rather than coming up with the algorithms ourselves.

Approach/Methodology

I first started of by loading the data set into Jupiter Notebooks and run a few basic tests to better understand the data. I also used some basic visual representation of the data using histograms and a relationship diagram. Here I was able to view the data more easily.

Next I plotted each feature against the target death rate using linear regression. I did this to better understand how the features where related to the target death rate. I also calculated the mean squared error for each of these relations. As there are 32 features it is impossible to visualize using them all in the one plot so by using a single uni variant I was able to understand the relationship between them and the target death rate a lot clearer.

After understanding how the data was related I was able to re-run the linear regression model against all the features and calculate the coefficients and mean squared error. After calculating the mean squared error I realized that the model needed to be refined as some features were not as important as others. These features caused the mean square error to be a lot higher than desired.

Reading through the Scikit-learn regression functions I decided on trying the following 3 different linear regression models, Ridge, Lasso and Elastic Net. From my analysis of the data earlier than well as the documentation I believe that Ridge Regression will give the best results and a lower mean absolute error. This is because it imposes a penalty on the size of the coefficients. I believe that this will help reduce the impact the same features had on the target death rate that aren’t as related.

I believe that Lasso will not help improve the mean absolute error as it generally prefers fewer parameters and our model has 32. This would be a useful model if we used feature selection to remove certain features that are below a given threshold.

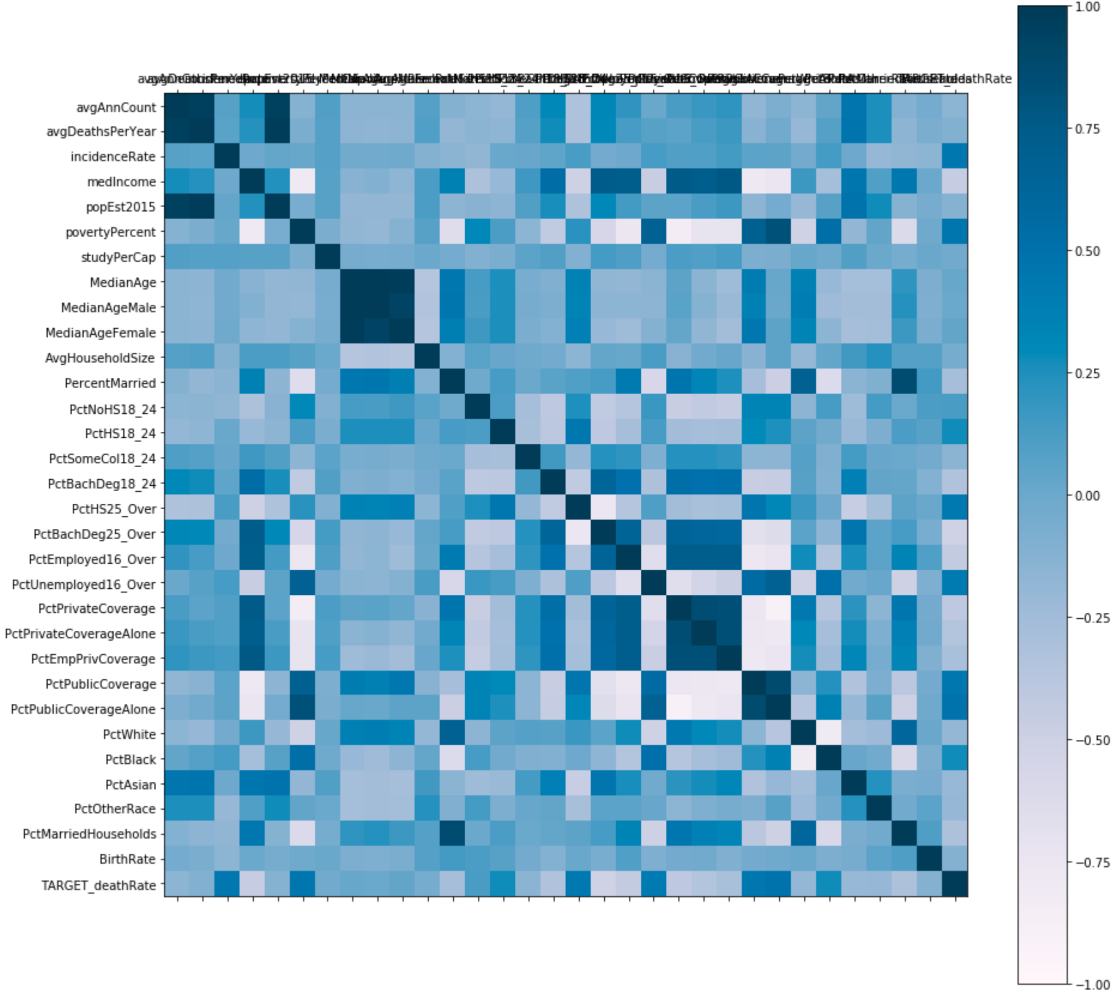
Elastic Net Regression could in fact be better than Ridge Regression but I believe that because it is a combination between Lasso and Ridge it will be affected by the number of features that our data has.

I also explored using Polynomial Features and K-Fold’s to better fit the training data and thus hopefully improve the prediction.

Evaluation/Results

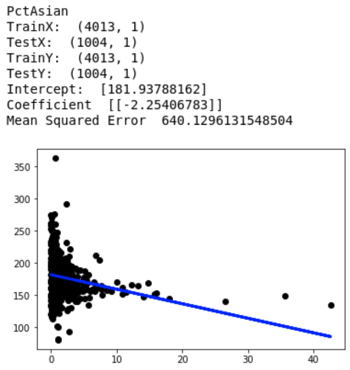
When I was running the basic visual representations I noticed that the data was not represented clearly in the box plots as there are outlining data that causes the box plots to be compressed and hard to read.

Next I plotted the correlations between the features. I found this really useful to help picture which features impacted the target death rate more and which features are less important.

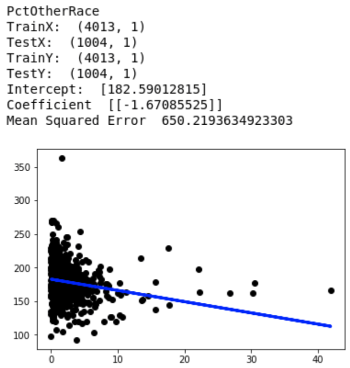


Feature correlations

Next I wanted to see how accurately each feature plotted against the target death. I used linear regression to plot each feature against the target death rate and also to calculate the coefficient of the hypothesis and the mean squared error. Doing this allowed me to better understand and also really helped me visualize the relationships the target death rate has with each feature.

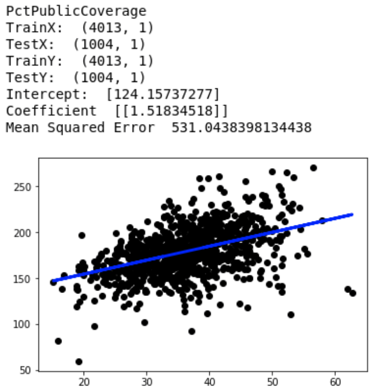


```
PctAsian
TrainX:  (4013, 1)
TestX:  (1004, 1)
TrainY:  (4013, 1)
TestY:  (1004, 1)
Intercept:  [181.93788162]
Coefficient  [[-2.25406783]]
Mean Squared Error  640.1296131548504
```

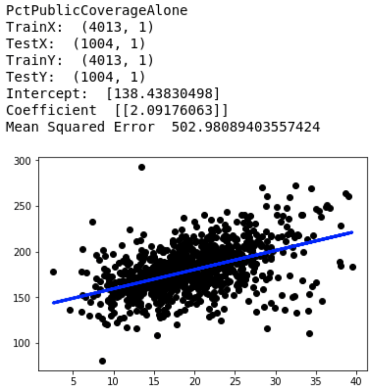


```
PctOtherRace
TrainX:  (4013, 1)
TestX:  (1004, 1)
TrainY:  (4013, 1)
TestY:  (1004, 1)
Intercept:  [182.59012815]
Coefficient  [[-1.67085525]]
Mean Squared Error  650.2193634923303
```

Higher mean squred error.
Features less related to target.



```
PctPublicCoverage
TrainX:  (4013, 1)
TestX:  (1004, 1)
TrainY:  (4013, 1)
TestY:  (1004, 1)
Intercept:  [124.15737277]
Coefficient  [[1.51834518]]
Mean Squared Error  531.0438398134438
```



```
PctPublicCoverageAlone
TrainX:  (4013, 1)
TestX:  (1004, 1)
TrainY:  (4013, 1)
TestY:  (1004, 1)
Intercept:  [138.43830498]
Coefficient  [[2.09176063]]
Mean Squared Error  502.98089403557424
```

Lower mean squared error.
Features more related to target.

After plotting them individually I ran the linear regression model over the entire data set using all features against the target death rate. As there are so many features it is impossible to plot the graph. I was able to see that by using all the data the mean square error was generally half what it was when I plotted them individually. I also decided to use the mean absolute error to have a more meaningful number as the result isn't so large.

I then created a function for each of the regression models I had chosen. I set up the functions to allow testing of different alpha values before I decided on the final value.

I split the data into train, validate and test groups and ran all the linear models over the same data to make sure that they were all being training to the same training data. I was able to calculate the mean absolute error and as I mentioned in the above Methodology the Ridge Regression performed better with a lower mean absolute error.

To improve the accuracy and try to remove limit any bias from the training I used K-Fold to randomly split the data into 5 group. I then trained each model and calculated their mean absolute error for each split. I added them to an array and then to the average across the 5 K-Fold splits. This slightly improved the mean absolute error, this demonstrated that there was some bias but not a lot from the other train test split.

Throughout my evaluation I was able to run multiple different training tests to see if I could fine tune the alpha parameters of the regression function to better improve the mean absolute error. This lead to some interesting outcomes especially when I increase the polynomial features above 3. My computer was able to handle 4 degrees but took about 30 minutes to compute. This resulted in a worse mean absolute error because the hypothesis was over fitted to the training data. As a further personal experiment I was curious to see if I could compute to a polynomial degree of 5. This cause the program to crash as it used all 24Gb of me memory and had a memory error.

I was able to compare the different alpha values across the regression models and chose a low value of 0.01 as if gave an overall better mean absolute error value across the different models. I was very surprised that that Elastic Net regression model was out performed by the Lasso Regression and in some cases the standard Linear Regression. As it is a combination of Lasso and Ridge I thought it would have been almost the same or better than ridge regression alone.

Conclusion

Throughout this analysis I was able to learn and visualize how each of the regression models learn the training data and were able to hypothesis and predict the outcomes of the target death rates. As I was using all the features this caused the mean absolute error and mean square error to be higher. I feel like using feature selection to removes some of the less relevant terms would greatly improve the prediction accuracy as some of the features weren’t that relevant to the target feature.