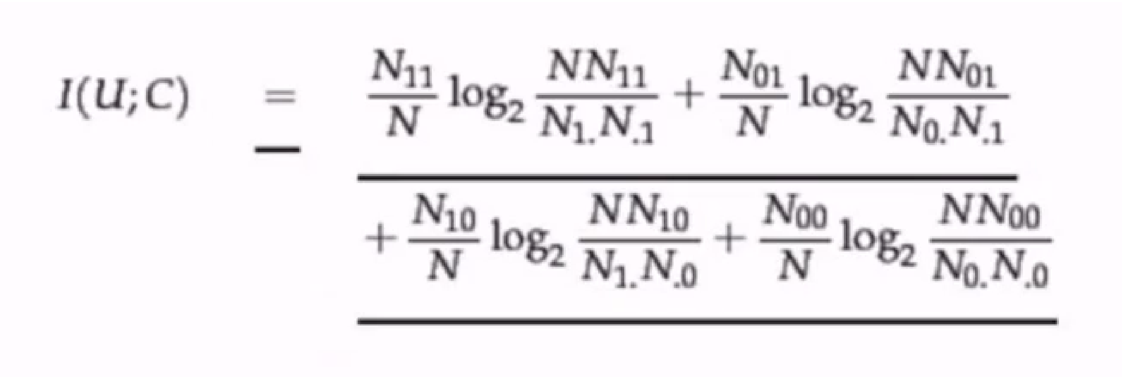
Ryan_Plain_HW5_Report

March 20, 2022

1 1. Conceptual questions

a.) Which keyword is more informative for deciding whether or not the email is a spam The formula is based off of the confusion matrix from the classification output.



$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}}$$

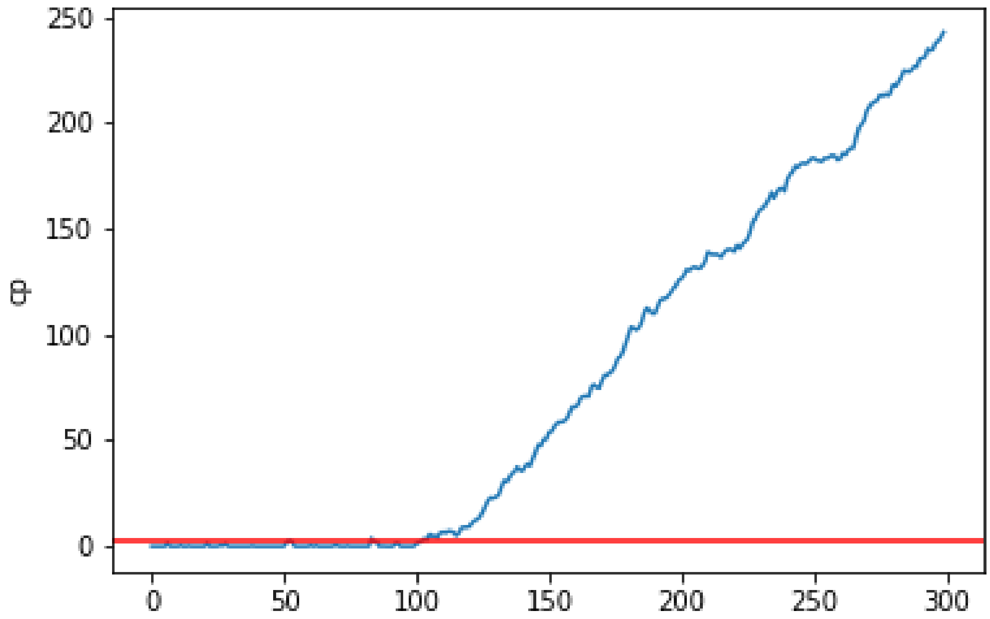
Deriving the value for both *hello* and *prize*, we get ~0.0002 and ~0.033 respectively. This would lead us to know that **prize** contains more information as a keyword for predicting spam based on mutal information.

b.) Given two distributions, f0 = N (0, 1), f1 = N (1.5, 1.1), explicitly derive what the CUSUM statistic should be (i.e., write down the CUSUM detection statistic). Plot the CUSUM statistic for a sequence of randomly generated samples, x1, . . . , x100 are are i.i.d. (independent and identically distributed) according to f0 and x101,...,x200 that are i.i.d. according to f1. To solve this, generate independtly distributed values from a normal distribution with parameters for $f_0(x)$ of length 100 and $f_1(x)$ of length 200. Append the outputs together and perform the CUSUM evaluation to determine the changepoint.

$$f_0(x) = N(0,1)$$

$$f_1(x) = N(1.5,1.1)$$

$$W_t = max(W_{t-1} + log(\frac{f_1(x)}{f_0(x)}), 0)$$

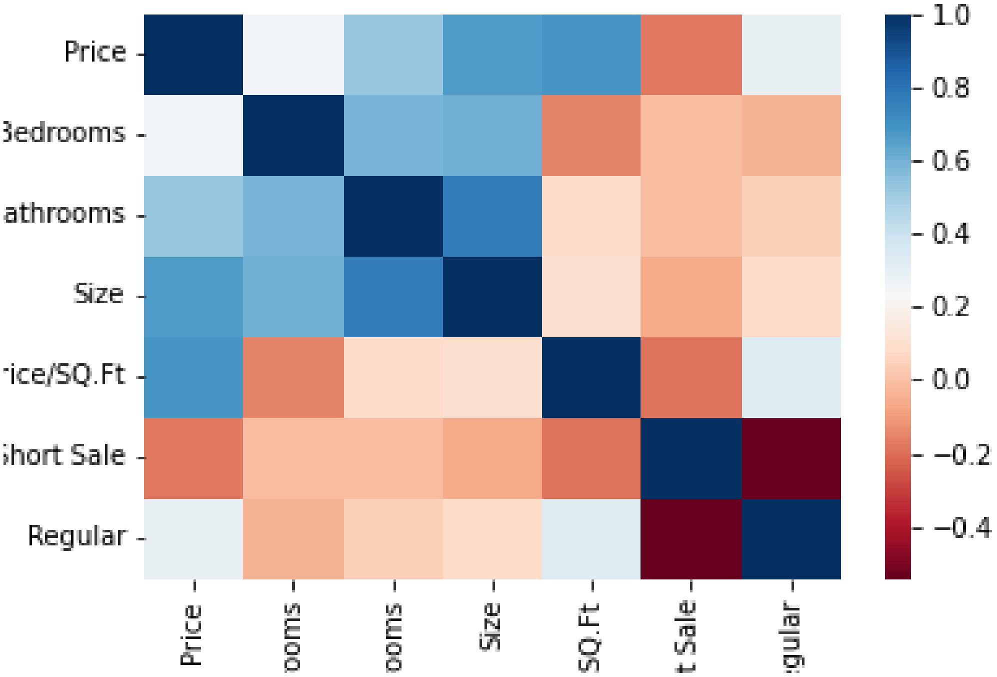


The CUSUM changepoint detection > 3 happened just after point 101. With the random seed set to 6704, the detection point happened at index 103. This aligns with the probelm, as the last 200 inputs from $f_1(x)$ have a different sampling distribution.

---

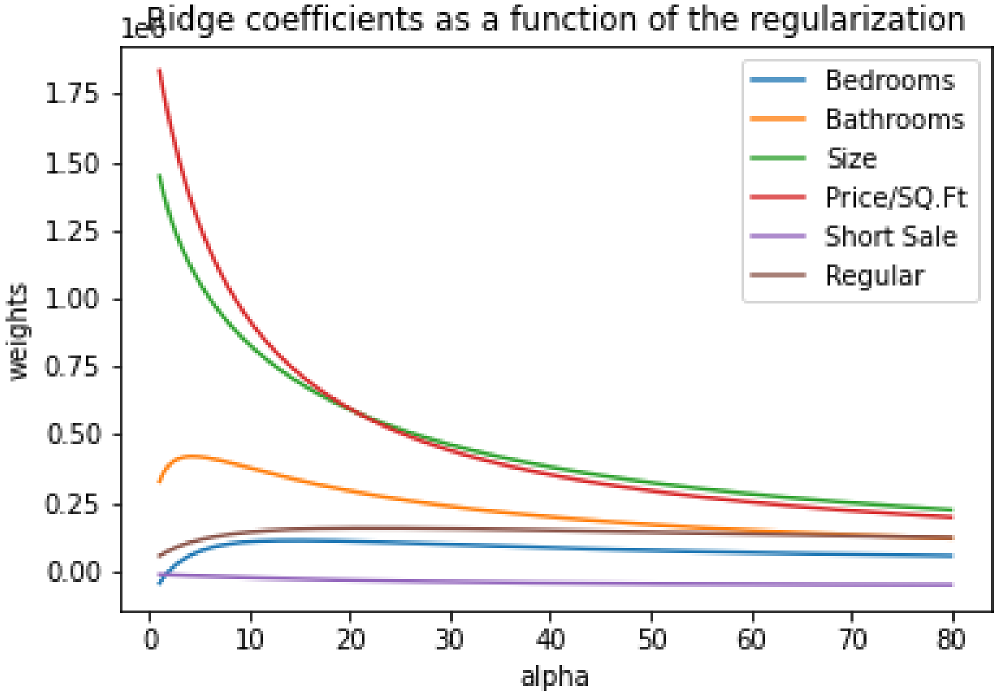
2 2. House price dataset

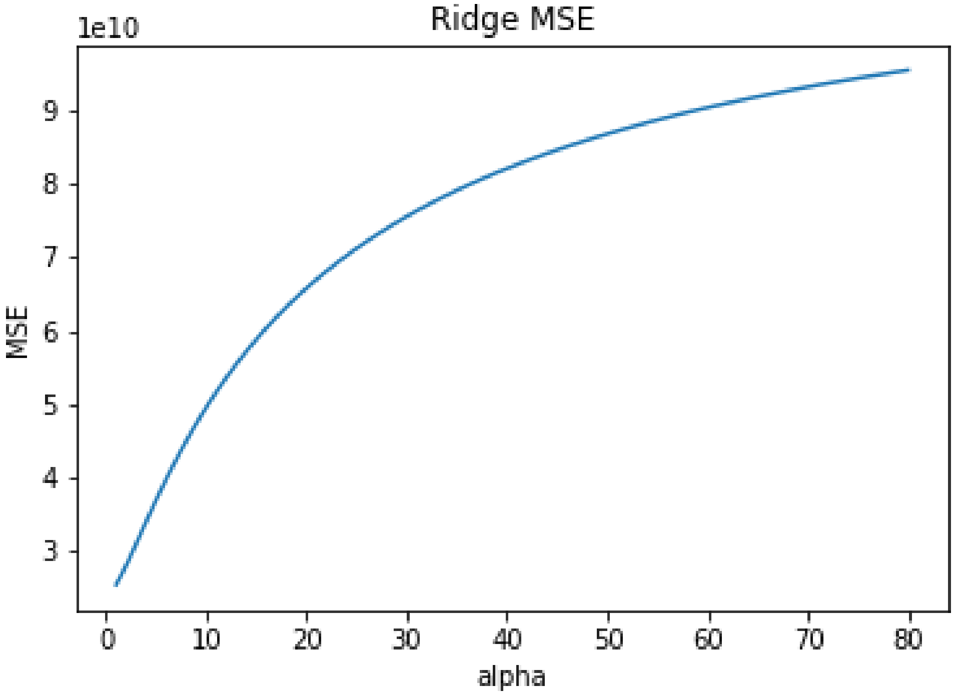
The objective is to use Ridge and LASSO regression to perform feature selection on the house price dataset. Shown below, this dataset has several highly correlated variables. `Price/Sq Ft.` is the target variable `Price` divided by `Size`. The number of `Bedrooms` and `Bathrooms` are directly correlated with `Size` as well. What is shown out through both models is that `Price/Sq Ft.` and `Size` can describe the target variable best if we were to only use 2 features, which makes sense due to the correlation with `Price`. Reasonably, you could not use `Price/Sq Ft.` in a model to determine price because that information would not be available at the time to assess the price of a new observation.

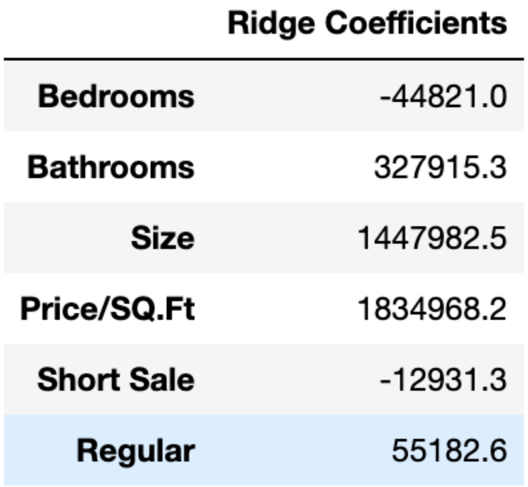


Ridge Regression Ridge is based off of the L2 norm, and compresses the values close to 0. The downside of ridge is that it is not sparse and typically does not do feature selection automatically.

What we see below is that the MSE is lowest at $\alpha = 1$. This keeps all the coefficients near their original value and utilizes all of them. Since the features are both correlated with each other and the target variable, it makes sense that including all of them would result in the lowest MSE. As you increase α, features begin to shrink and this is going to cause a higher bias in the model. For most datasets, this would be preferred. If we were to apply knowledge that `Price/Sq Ft.` is not valid, and other feature engineering, it is possible that there is a differet sweet spot for MSE to balance the bias/variance tradeoff.


Ridge coefficients as a function of the regularization
1e6
weights
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
50
60
70
80
alpha
Bedrooms
Bathrooms
Size
Price/SQ.Ft
Short Sale
Regular


Ridge MSE
1e10
MSE
9
8
7
6
5
4
3
0
10
20
30
40
50
60
70
80
alpha

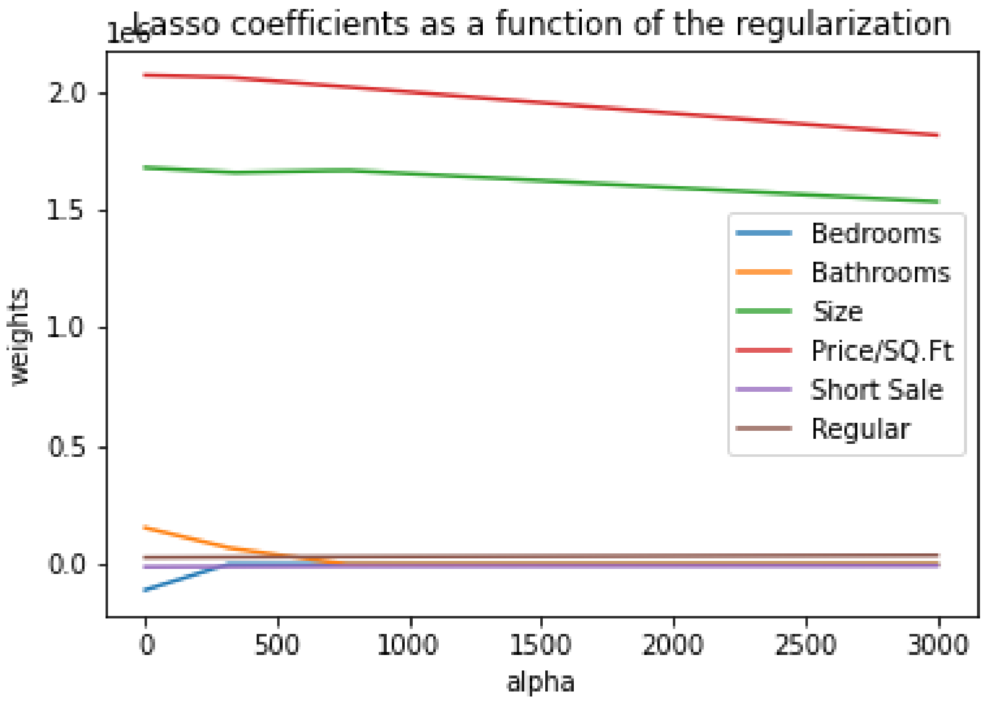


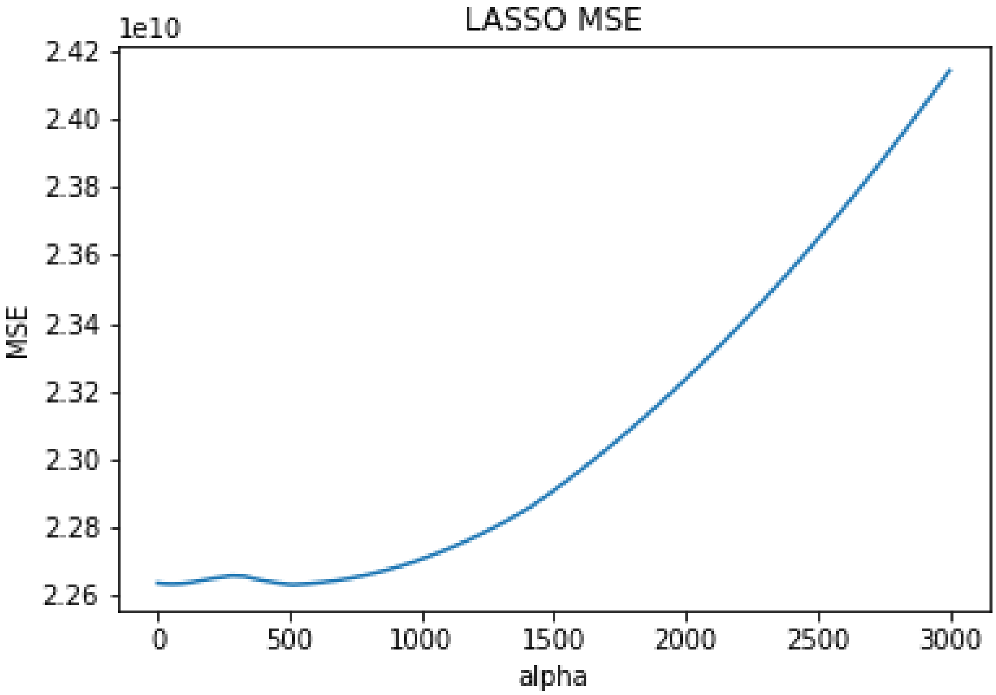
	Ridge Coefficients
Bedrooms	-44821.0
Bathrooms	327915.3
Size	1447982.5
Price/SQ.Ft	1834968.2
Short Sale	-12931.3
Regular	55182.6

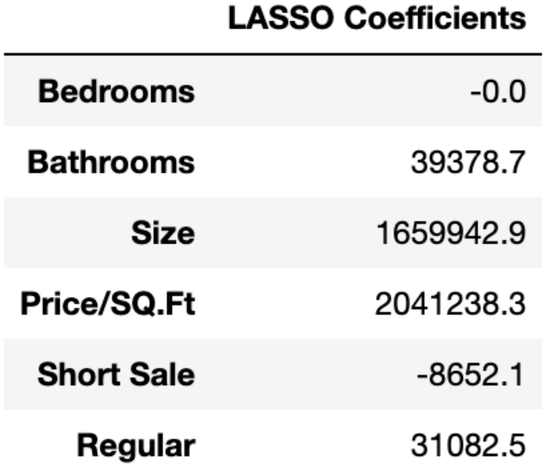
LASSO Regression LASSO is based off of the L1 norm, and creates a sparse vector that does in fact compress values to 0.

The output shows again that `Price/Sq Ft.` and `Size` are the two dominating features, at a higher value of α those would be the only feautres selected.

The lowest MSE for the LASSO regression was at $\alpha = 509$. This leaves out `Bedrooms` as a coefficient.

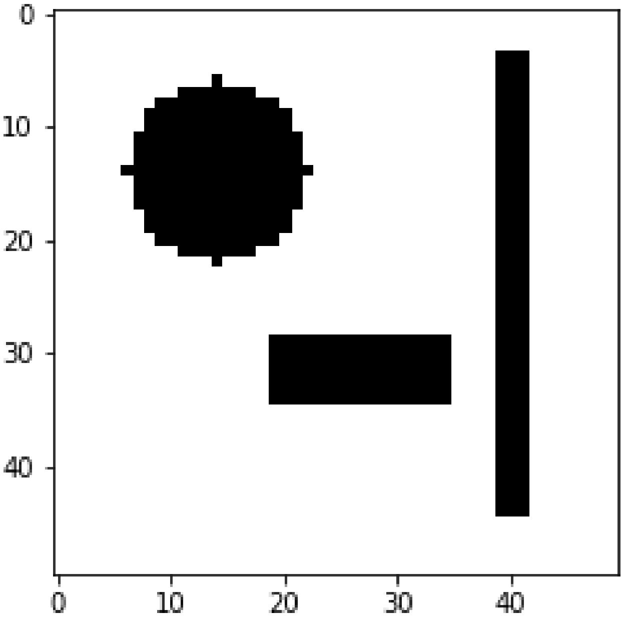






	LASSO Coefficients
Bedrooms	-0.0
Bathrooms	39378.7
Size	1659942.9
Price/SQ.Ft	2041238.3
Short Sale	-8652.1
Regular	31082.5

3 3. Medical imaging reconsturction



The MRI data provides a sparse Matrix that creates the image above. The objective is to run Ridge and LASSO regressions to see how well the coefficents can recreate the original image. To do so, 10 fold cross validation was performed to find the best α based on the lowest MSE value.

To create the dataset:

$$m = 1300$$

$$n = 2500$$

$$A = N \sim (0,1)(m,n)$$

$$\epsilon = N\ (0,5)(m,)$$

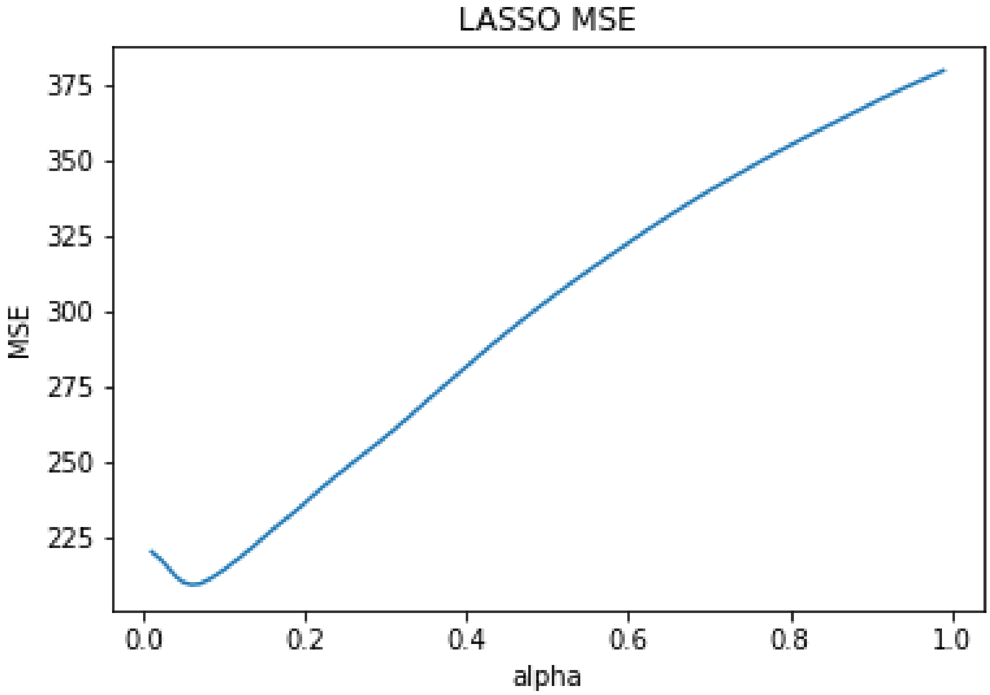
x is the actual values from the image

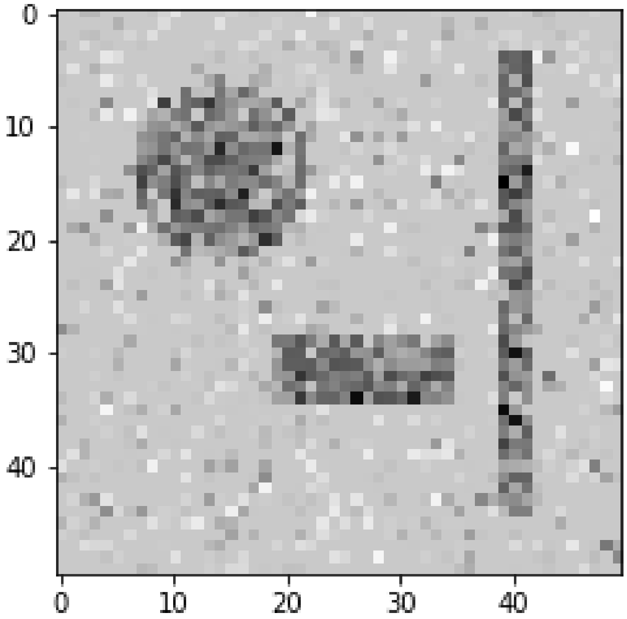
$$Y = Ax + \epsilon$$

Once Y is created, the regressions were fitted with (A, Y)

3.0.1 LASSO

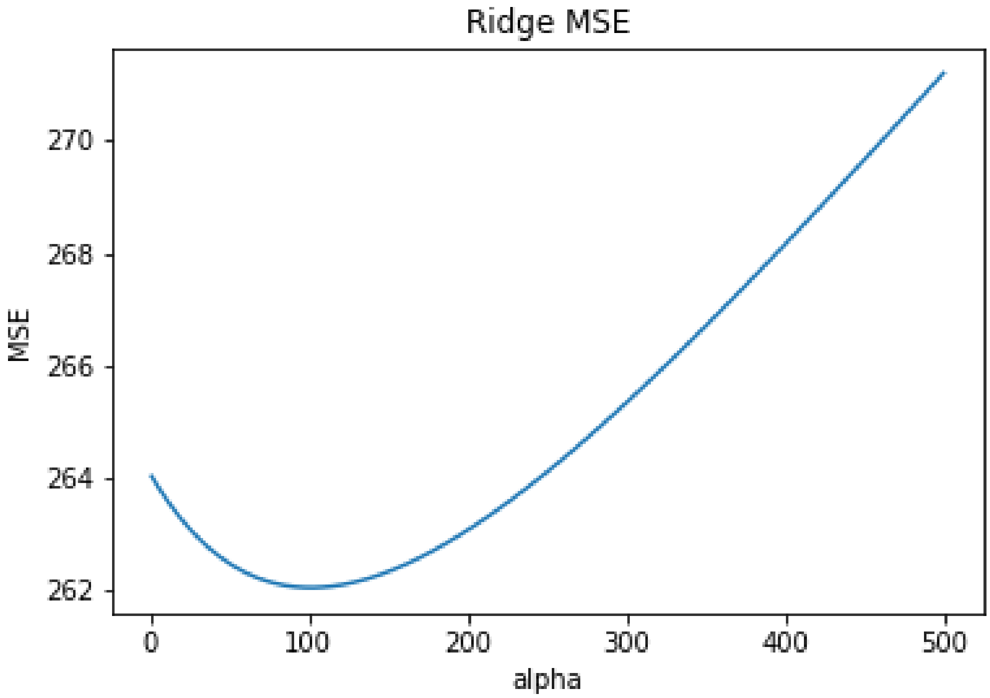
The LASSO coefficients did much better recreating the image than Ridge. This is expected, since the image was a sparse matrix and the LASSO coefficents are a sparse vector. The best performing input was $\alpha = 0.06$ from the cross validated model. Using the best α, the LASSO regression coefficients alone produced the recreated image.

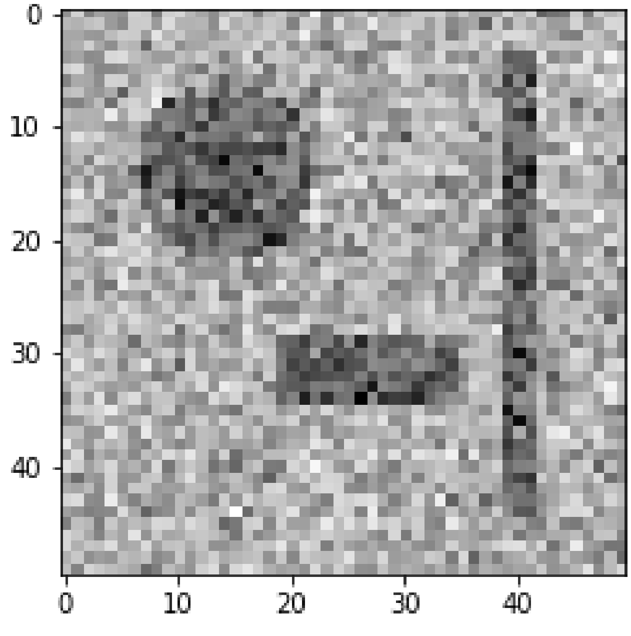




3.0.2 Ridge

The Ridge model did not perform as well depending on what criteria you look at. The image does indeed show the correct shapes, location and size, but does not leave out any empty space similar to the original image. This is because Ridge compresses the values close to 0 but not 0. Meaning there will be values throughout the matrix, which does not do well for the sparse matrix provided.





It is safe to say with the known structure of the original image being a true sparse matrix, that LASSO would perform better objectively. The lower MSE value and visualizations support the LASSO being the better model for the image recreation.