Advanced Linear Regression Assignment
Part II: Subjective Questions
By: Kris Laumann

Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

The optimal values are about 120 for ridge and 490 for lasso regression. Doubling them has slightly decreased the R2 on the training set from 0.9359 to 0.9228 for ridge and from 0.9389 to 0.9273 for lasso. This is unsurprising since we optimised alpha on the training set.
However R2 on the test set improved from 0.73 to 0.78 for ridge and 0.53 to 0.57 for lasso.
This suggests that we are still overfitting on the training set.
Strongest predictors are now 'above ground living area' and 'pool quality'.

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:

Ridge regression with a lambda of 120 (alpha in sklearn terms) turned out to perform best on the test set in all metrics measured.

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:
By beta:

- Roof material: Standard (Composite) Shingle
- 2nd floor square footage
- 1st floor square footage
- Roof material: Wood Shakes

- Roof material: Gravel & Tar

By category:

- Roof material
- 2nd floor square footage
- 1st floor square footage
- Kitchen Quality
- Neighbourhood

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

Cross validation and the test set can help us determine the model's performance beyond the train data. As we saw the model hopelessly overfits if we don't regularise it with bridge or lasso regression. This however comes with a tradeoff between bias and variance. We stabilise the model but hurt its maximum (overfitted performance). In the assignment we were able to improve the model but increasing lambda too far would ultimately hurt accuracy.