Subjective Questions

Clara Regalado

April 2023

1. First 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?

   The optimal value for Ridge is 0.1
   The optimal value for Lasso is 0.00001

```
Lambda =  0.0001
r2 score train = 0.7100208541706636
r2 score test = 0.6933648253569545
Lambda =  0.001
r2 score train = 0.7100208337240133
r2 score test = 0.6933841251242017
Lambda =  0.01
r2 score train = 0.7100188839409469
r2 score test = 0.6935634017687486
Lambda =  0.1
r2 score train = 0.7098904930136483
r2 score test = 0.6944294990115378
Lambda =  1
r2 score train = 0.708017451585097
r2 score test = 0.6916883962375391
Lambda =  10
r2 score train = 0.6816967846583395
r2 score test = 0.667631299461408
```

```
Lambda =  1e-05
r2 score train = 0.7099916008448957
r2 score test = 0.693594158201421
Lambda =  0.0001
r2 score train = 0.7079283052949407
r2 score test = 0.6924814846814056
Lambda =  0.001
r2 score train = 0.6745411231247468
r2 score test = 0.6629184846763108
Lambda =  0.01
r2 score train = 0.12951763318774512
r2 score test = 0.11571763203747287
Lambda =  0.1
r2 score train = 0.0
r2 score test = -0.016481916426241305
Lambda =  1
r2 score train = 0.0
r2 score test = -0.016481916426241305
```

   When the values of lambda are duplicated for Ridge and Lasso, the values are as follow:

```
---
r2 score train = 0.7096595760506534
[ 0.20066086  0.10295632  0.00503561  0.16148441  0.07377896  0.1155477
  0.04941754 -0.0053379  -0.02215114 -0.06280861  0.02597756  0.04703419
  0.01066491  0.03827804  0.11222456  0.16405917  0.10102787  0.13500927
  0.03824684 -0.01584373 -0.03587175  0.05592756  0.01186116  0.02439369
 -0.05137796 -0.07868098 -0.00580306  0.12854637  0.02945597]
r2 score test = 0.6944499723167701
[ 0.20066086  0.10295632  0.00503561  0.16148441  0.07377896  0.1155477
  0.04941754 -0.0053379  -0.02215114 -0.06280861  0.02597756  0.04703419
  0.01066491  0.03827804  0.11222456  0.16405917  0.10102787  0.13500927
  0.03824684 -0.01584373 -0.03587175  0.05592756  0.01186116  0.02439369
 -0.05137796 -0.07868098 -0.00580306  0.12854637  0.02945597]
```

Figure 1: Ridge

```
r2 score train = 0.7099147083132833
[ 0.26624044  0.10234476  0.00488531  0.16191234  0.07292003  0.11408805
  0.04668716 -0.00313024 -0.02008933 -0.06551105  0.02527676  0.04542961
  0.00838537  0.038292   0.11222387  0.16411214  0.10125902  0.13528538
  0.03360872 -0.01411006 -0.03469219  0.05565153  0.          0.02167809
 -0.05161466 -0.07833006 -0.         0.12639648  0.02736136]
r2 score test = 0.6937741625847809
[ 0.26624044  0.10234476  0.00488531  0.16191234  0.07292003  0.11408805
  0.04668716 -0.00313024 -0.02008933 -0.06551105  0.02527676  0.04542961
  0.00838537  0.038292   0.11222387  0.16411214  0.10125902  0.13528538
  0.03360872 -0.01411006 -0.03469219  0.05565153  0.          0.02167809
 -0.05161466 -0.07833006 -0.         0.12639648  0.02736136]
```

Figure 2: Lasso

The R2 score for Ridge when lambda is double gets worse for train and set. However, the R2 score value for Lasso when lambda is double improves for both sets.

The most important predictors after the change is implemented are : LotArea, 2ndFlrSF, Neighborhood_NoRidge, Neighborhood_Somerst, and SaleType_Con.

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?

   I will still chose 0.1 for Ridge and 0.00001 for Lasso. The difference between those and their doubles is not remarkable, so there are no reason to change them.

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?

   The most important are : 1stFlrSF, 2ndFlrSF, GarageType_No Garage, Neighborhood_OldTown, Neighborhood_Edwards

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?

   The model is explaining 69% of the data for the test set, which is ok, not the optimal as I wished, but it is still good. It can be generalisable as it is, but it is better to make more changes for improving the values. The accuracy will be affected due to the R2 value because it is a bit low. This will affect the predictions by providing a low price for a property that worth more or the opposite, given a high sale price when the property is less value.