Author: Sudarshan (sudarshan_g@outlook.com)
Date Created: 19th September 2023

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 1:

- Optimal Lambda for Ridge: `6.0` , Lasso: `0.0001`
- Changes in model statistics when it's doubled -

  Lasso with Lambda 0.0002
  `RSS (Training): 8.3493 (⬆️ Increased)`
  `RSS (Test): 6.3004 (⬇️ Decreased)`
  `Training R-Squared: 0.9414 (⬇️ Decreased)`
  `Test R-Squared: 0.8974 (⬆️ Increased)`
  `Training MSE: 0.0082 (⬆️ Increased)`
  `Test MSE: 0.0144 (⬇️ Decreased)`
  `Training RMSE: 0.0904 (⬆️ Increased)`
  `Testing RMSE: 0.1199 (⬇️ Decreased)`

  Lasso with Lambda 0.0001
  `RSS (Training): 8.0312`
  `RSS (Test): 6.4541`
  `Training R-Squared: 0.9437`
  `Test R-Squared: 0.8949`
  `Training MSE: 0.0079`
  `Test MSE: 0.0147`
  `Training RMSE: 0.0887`
  `Testing RMSE: 0.1214`

  Ridge with Lambda 12
  `RSS (Training): 9.0344 (⬆️ Increased)`
  `RSS (Test): 6.3872 (⬆️ Increased)`

`Training R-Squared: 0.9366 (⬇️ Decreased)`
`Test R-Squared: 0.8960 (⬇️ Decreased)`
`Training MSE: 0.0088 (⬆️ Increased)`
`Test MSE: 0.0146 (⬆️ Increased)`
`Training RMSE: 0.0941 (⬆️ Increased)`
`Testing RMSE: 0.1208 (⬆️ Increased)`

Ridge with Lambda 6

`RSS (Training): 8.4770`
`RSS (Test): 6.2901`
`Training R-Squared: 0.9405`
`Test R-Squared: 0.8975`
`Training MSE: 0.0083`
`Test MSE: 0.0144`
`Training RMSE: 0.0911`
`Testing RMSE: 0.1198`

- Top 10 variables for Ridge with 2x Lambda with coefficients

```
OverallQual_9           0.105803
GrLivArea               0.089565
Neighborhood_Crawfor    0.089380
OverallCond_9           0.080461
OverallQual_8           0.079830
TotalBsmtSF             0.060841
Exterior1st_BrkFace     0.057032
OverallCond_7           0.056596
Neighborhood_StoneBr    0.053598
CentralAir_Y            0.050557
```

- Top 10 variables for lasso with 2x Lambda with coefficients

```
OverallCond_9           0.205079
OverallQual_9           0.198687
Neighborhood_Crawfor    0.120103
OverallQual_8           0.120058
GrLivArea               0.106497
OverallCond_7           0.100082
OverallCond_8           0.099157
Neighborhood_StoneBr    0.096573
Exterior1st_BrkFace     0.084060
SaleType_ConLD          0.071804
```

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 2:

Lasso would be preferable with alpha 0.0001 for the following reasons -

- Significantly less number of features leading to a simpler model.Lasso got rid of `173` features out of `205`.
- Minimal differences in (& between) test/train R2 when compared with Ridge.
- Minimal differences in (& between) MSE and RMSE when compared with Ridge.

Therefore, not much of a downside but a huge upside for model simplicity due to reduced number of features.

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 3:

New 5 predictor variables for Lasso are -

- `OverallCond_9 0.217272`
- `OverallQual_9 0.203869`
- `Neighborhood_Crawfor 0.132758`
- `SaleType_ConLD 0.129297`
- `OverallQual_8 0.120210`

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 4:

- Make sure we don't overfit the model so that it has less bias. This can be done via regularization methods such as Ridge and Lasso regularization to keep model complexity and variance low. Allowing it to predict and adapt for unseen data.
- Accuracy will drop slightly as a trade-off for generalization and robustness. A robust/generalized model is not aware of all aspects/patterns of the dataset therefore it will have some bias to it.
- Models with high variance will be accurate but be susceptible to smallest changes in the underlying data. Implying, they are not performant on unseen data and are overfitted. Basically, they have learnt even the noise within the dataset.
- Models that are excessively simplistic will be highly robust and performant on all types of dataset but accuracy goes for a toss. Implying that they are not good at predicting properly and are underfitted.
- These implications relate to the bias-variance trade off & the need to strike a balance between underfitting and overfitting.