House Prediction Subjective Questions

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 to 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 value of alpha for ridge regression is 2
The optimal value of alpha for lasso regression is 0.001

when we increase the value of alpha the model tries to penalize more and try to make most of the coefficient value zero. When we double the value of alpha for our ridge regression, we will take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set .from the graph we can see that when alpha is 10 we get more error for both test and train.

Similarly, when we increase the value of alpha for lasso, we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases. The most important variable after the changes has been implemented for ridge regression are as follows:

- BsmtFinSF1
- BsmtUnfSF
- 2ndFlrSF
- FullBath
- BedroomAbvGr
- LotArea
- BsmtFullBath
- Fireplaces
- BsmtQual_Ex
- BsmtFinSF2

The most important variable after the changes has been implemented for ridge regression are as follows:

- BsmtFinSF1
- BsmtUnfSF
- 2ndFlrSF

- FullBath
- Fireplaces
- BedroomAbvGr
- BsmtQual_Ex
- BsmtFullBath
- OverallQual_8
- BsmtFinSF2

Question2:
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:
It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance and making the model interpretably.

During the assignment, I got the r2-score for ridge regression for train set as 85% and for test set as 82% by using all the variables in the dataset. In the same way, got the r2-score for lasso regression for train set as 85% and for test set as 81% by using few variables in the dataset.

I chose Lasso Regression over Ridge Regression as it would help in feature elimination and the model will be more robust.

Question3:
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:
Those five most important predictor variables that will be excluded are –

- BsmtFinSF1
- BsmtUnfSF
- 2ndFlrSF
- FullBath
- Fireplaces

Question4:
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:
The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e., the accuracy does not change much for training and test data.

Outlier analysis needs to be done and just those which are relevant to the dataset need to be retained. Those outliers which does not make sense to keep should be removed from the dataset. This would assist with increasing the accuracy of the predictions made by the model. Confidence intervals can be utilized (typically 3-5 standard deviations). This would assist with normalizing the expectations made by the model. If the model is not robust, it cannot be trusted for predictive analysis.