Surprise housing 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 double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Below is the optimal value of Ridge and Lasso regression

Ridge alpha – 100

Lasso alpha – 0.01

Model after alpha is doubled

- Ridge

Ridge	alpha = 100	alpha = 200
R2 Train	0.93	0.93
R2 Test	0.87	0.87
RSS Train	51.5	56.63
RSS Test	47.98	46.6
RMSE Train	0.06	0.07
RMSE Test	0.14	0.14

- Lasso

Lasso	alpha = 0.01	alpha = 0.02
R2 Train	0.92	0.9
R2 Test	0.88	0.88
RSS Train	60.27	73.3
RSS Test	44.24	43.96
RMSE Train	0.08	0.09
RMSE Test	0.13	0.13

There is no difference in R2 score for Ridge and Lasso, however there is change in other metrics in both.

Top 10 Predictor variable after the change

1. GrLivArea
2. OverallQual
3. TotalBsmtSF
4. LotArea
5. NumYearsBuilt
6. NumYearRemodAdd
7. GarageArea
8. BsmtFinSF1
9. Neighborhood
10. OverAllCond

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?

Both Ridge and Lasso provide similar R2 score for the data provided, but Lasso also provides feature selection. Hence, will consider Lasso regression in this case.

Question 3

After building the model, you realized 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?

Top 5 Feature original

1. GrLivArea
2. OverallQual
3. TotalBsmtSF
4. LotArea
5. NumYearsBuilt

Top 5 Features after removing top 5 features

1. BsmtFinSF1
2. BsmtUnfSF

3. 2ndFlrSF
4. GarageArea
5. NewYearRemodAdd

Question 4

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

It is crucial for the model to be robust and generalizable, so that the model performs well on not only training data, but also on unseen data. In case if the model is not generalized then there can be chances of overfitting. In order to avoid overfitting and created a robust as well as generalized model below steps should be considered:-

1. Split data into Training and Validation set -
   In this case the validation set is used for tuning the model created on training data. Once the model is finalized, then the model is executed on test data, which is unseen by the model.
   Implication – Provide robustness, as present of validation set, keeps testdata unseen till model is finalized.

2. Cross Validation -
   The training data is divided into k parts, such that k-1 part is used for training and 1 part is used for testing. This process is done k times, with different set as test set. The average of all these is considered as resulting r2 value.
   Implication - Provide robustness, as model is tested on different combination of data.

3. Regularization –
   Regularization technique like Ridge and Lasso are used to prevent overfitting by adding penalty to large coefficients.
   Implication- Prevent overfitting, overfitting by adding penalty to large coefficients.

4. Hyperparameter Tuning -
   It is used to keep the complexity of the model under control, by fixing certain properties of the model. For example for decision tree fixing the height of tree, or limiting number of values in leaf nodes.
   Implication- Prevent overfitting, by regulating complexity of the model

5. Evaluate on multiple metrics –
   Multiple metrics like F1-score, ROC, AUC etc. should be considered to evaluate the model
   Implication- Provide robustness and prevent overfitting, as combination of different metric take care of both.