Assignment: Surprise Housing
Student Name: Kajal Kaspate
Batch ID: C61

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:

	Ridge	**Lasso**
Optimal Value of Alpha	500	0.01
R-squared on Train	0.88	0.8899
R-squared on Test	0.854	0.8405
Important Variables	OverallQual GrLivArea NeighborHood_NoRideg 1stFlrSF Neighborhood_NridgeHt 2ndFlrSF	GrLivArea OverallQual Neighbourhood_NridgHt Neighbohood_NoRidge GarageCars RoofMatl_WdShngl

	Ridge	**Lasso**
Doubled Alpha	1000	0.02
R-squared on Train	0.8586	0.8746
R-squared on Test	0.8443	0.837
Important Variable	OverallQual GrLivArea NeighborHood_NoRideg 1stFlrSF Neighborhood_NridgeHt 2ndFlrSF	GrLivArea OverallQual Neighbourhood_NridgHt Neighbohood_NoRidge GarageCars RoofMatl_WdShngl

As we see by doubling the alpha for both Ridge and Lasso, have affected the R-squared values. We see that Lasso is doing better on the train test when alpha is doubled. But Ridge looks more stable with not much reduction in the R-squared for the test data.

For both, important features remain same when the alpha is doubled.

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:

In terms of choosing alpha for Ridge and Lasso, I will choose initial values.
Ridge = 500 and Lasso = 0.01

These values have better R-squared on both train and test set than any other alpha values.
Also, in terms of model, I would choose Ridge here as we see Ridge is slightly more stable on test data for different params.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.914085	0.880160	0.889995
1	R2 Score (Test)	0.806266	0.854669	0.840570
2	RSS (Train)	1.057190	122.357144	112.315315
3	RSS (Test)	1.053101	65.549197	71.908080
4	MSE (Train)	0.032178	0.346180	0.331670
5	MSE (Test)	0.049034	0.386854	0.405184

With Ridge, we can also see comparable MSE on Train and Test data. Ridge includes all the features in the final model, however, we can still select the features which are significant based on their coefficients.

In the current dataset, we have multiple factors that can affect the SalesPrice e.g. square feet of the house, locality of the house, quality of the building etc.
When we have multiple factors affecting SalesPrice equally, it is advisable to use Ridge regression over Lasso.

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:

Below are the 5 most important predictor variables

1. GrLivArea
2. OverallQual
3. Neighborhood_NridgHt
4. Neighborhood_NoRidge
5. GarageCars

After dropping these:

	Lasso	**Ridge**
Alpha	0.01	500
R-squared on Train	0.88	0.85
R-squared on Test	0.817	0.83

After removing these and creating lasso model again, we get the same alpha value for Lasso which is 0.01.

In this case, R-squared for train is 0.88 and for test it is 0.817. We see that model is becoming a bit more unstable in terms of it is tending to overfit on the train data.

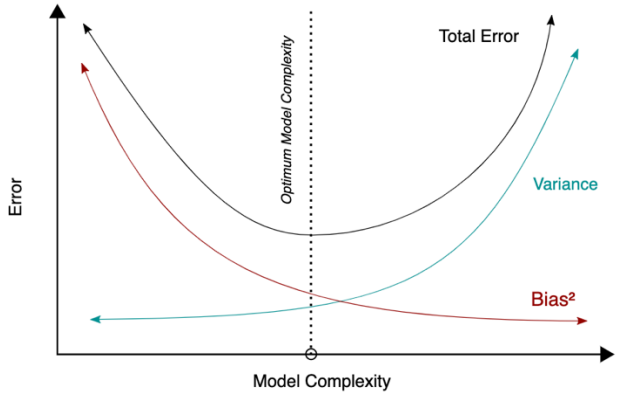
On the other hand, when we create Ridge model again, we get same alpha which is 500. For Ridge, the R-squared on train is 0.85 and on test this is 0.83. Based on these values, we see that Ridge is performing better over Lasso.

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?

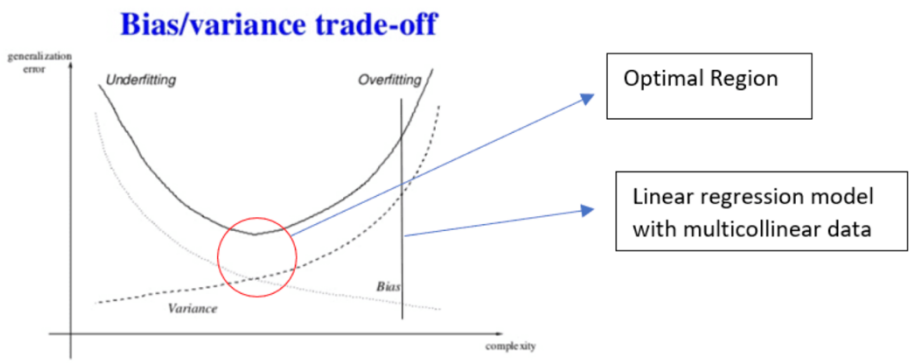
Answer:

To make the model robust and generalizable, we need to make it perform well on both train and test data sets. This is where the bias variance trade off comes into the picture.



As the model complexity increases, we have high variance and low bias. In this case, model tries to fit every possible data point which is called overfitting. For such models, when an unseen data is given, they do not fair well and will have high error.

When we build simple models, they may not see all the variations in data and can miss some data patterns that exist in data. In this case, such models underfit. They have high bias and low variance.



To avoid instability of the model, we need to find the optimal region for the variance and bias where the total error is the lowest. We can achieve this by introducing regularization in the model.

With regularization, we can control the complexity of the model. Higher the alpha, higher is the regularization and hence simpler is the model. When we reduce the alpha, model complexity increases. Hence by finding optimum alpha, we could control the stability of the model.