**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**

Optimal Alpha values as detected by the model are

- Ridge – 2.0

- Lasso – 0.001

If we double the value of alpha in both Ridge and Lasso regression there will be increase the regularization strength, which can help in reducing overfitting and handling multicollinearity. However, it may also lead to simpler models with smaller coefficients (Ridge) or even some coefficients being set to zero (Lasso)

Model score metrics after doubling the alpha as 4.0 for ridge & 0.002 for lasso.

*Note: Working for this is in Jupyter notebook at last*

A table with numbers and text

Description automatically generated

Most important predicator variables after change are as follow

A screenshot of a table

Description automatically generatedA screenshot of a table

Description automatically generated

**Lasso with alpha 0.001 Lasso with alpha 0.002**

After doubling the alpha, most important predictor variable mostly remains same, however the order of coefficients impacting has changed.

**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:**

**A table with numbers and symbols

Description automatically generated**

Both models showed very close R2 scores on test data prediction, however we can see Lasso scores are slightly better.

I choose to use Lasso model in this scenario as with lasso we have another advantage of that it actually reduces coefficients to absolute 0 while Ridge never

eliminates the features it reduces the all coefficients are reduced but never to make them absolute 0.

Due to lesser variables from Lasso we will have less complex model too.

**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:**

Removing the initial model’s best 5 variables, below are new best 5 variables suggested by model

1. 1stFlrSF : First Floor square feet
2. MSZoning\_FV : Properties belonging to zone Floating Village Residential
3. MSZoning\_RL : Properties belonging to zone Residential Low Density
4. MSZoning\_RH : Properties belonging to zone Residential High Density
5. 2ndFlrSF : Second floor square feet

*Note: Working for this is in Jupyter notebook at last*

A table with numbers and letters

Description automatically generated

**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:**

Bias (underfitting) is error in model, when the model is weak to learn from the data. High bias means model is too simple & is unable to capture or learn underlying patterns in the data. Model performs poor on training and testing data.

Variance (overfitting) is error in model, when model tries to over learn from the data by making itself complex. High variance means model is too sensitive to noise in training data & performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.

It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.

Even though a model is less accurate, it should be as simple as possible, but more robust and generalized.

To make the model robust, we should

* reduce the complexity of model, use less predictor features. Simpler the model, the greater the bias, but the less the variance and the more generalized it is.
* make sure training data scores are as close to test data
* drop more correlated feature, and not relying on lasso only
* If possible have more data available so that outliers treatment can be significant
* Apply regularization
* Apply bias, variance trade off rules

In terms of accuracy, its implication is that a robust, generalizable model works equally well on both training and test data. Accuracy does not change significantly between training and test data.