Advanced Regression Assignment

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:

The Optimal Value for alpha is 10 for Lasso and 1.022942 for Ridge.
After doubling the value of alpha for Ridge and Lasso:

- For Ridge,
  - Alpha = 2.045884
  - Metrics:
    - R2 : 0.893 (same as before upto 3 decimal places)
    - RMSE Score: 24650.68 (slight improvement)
    - MAPE : 10.78% (slight improvement)

	Feature_Ridge	Coefficient_Ridge	Feature_Ridge_2x	Coefficient_Ridge_2x
0	1stFlrSF	129396.477619	1stFlrSF	115787.340649
1	2ndFlrSF	117980.661449	2ndFlrSF	108022.642752
2	OverallQual	97659.869722	OverallQual	92131.006593
3	BsmtFinSF1	59409.705967	BsmtFinSF1	57791.302603
4	Neighborhood_StoneBr	34869.068795	Neighborhood_StoneBr	33693.648512

The top 5 features remain the same with slight decrease in coefficient value as the alpha value is small for ridge, doubling doesn't really a substantial change

- For Lasso,
  - Alpha = 20
  - Metrics:
    - R2 : 0.892 (same upto 2 decimal places)
    - RMSE : 24864.80 (marginal improvement)
    - MAPE : 11.11% (marginal improvement)
  - Total Features whose coefficient become zero (before and after):
    - Before doubling: 1 feature removed (coefficient = 0)
    - After doubling: 4 features removed (coefficient = 0)

	Feature_Lasso	Coefficient_Lasso	Feature_Lasso_2x	Coefficient_Lasso_2x
0	1stFlrSF	152737.901941	1stFlrSF	156272.506559
1	2ndFlrSF	129979.539320	2ndFlrSF	128301.121079
2	OverallQual	104567.528135	OverallQual	105798.853696
3	BsmtFinSF1	57248.975659	BsmtFinSF1	53178.168078
4	Exterior1st_BrkComm	-37422.594500	Neighborhood_StoneBr	34267.530391

Since the alpha value for lasso is big, doubling it changes the top 4 features along with the coefficients associated with them.

We see that overall, there is a very marginal improvement when the regularization parameters are doubled. But we cannot consider this model as the new best model as we are changing the parameter after seeing the results on the test data (vs. during actual training we are selecting the best params based on 10 Fold CV on training data and then applying the results to test data)

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:

- The optimum lambda value in case of Ridge and Lasso is as follows:-
  - Ridge: 1.022942
  - Lasso: 10
- The Root Mean Squared Error in case of Ridge and Lasso are:
  - Ridge: 24740.78
  - Lasso: 24995.91
- The MAPE for Lasso and ridge are:
  - Ridge: 10.96%
  - Lasso: 11.25%
- The R2 Score for Lasso and ridge are:
  - Ridge: 0.893
  - Lasso: 0.890
- Considering all the metrics, Ridge performs better than Lasso, also lasso removes just one feature during modelling (coefficient = 0), hence, Ridge is chosen as the best model.

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?

Ans: The five most important predictor variables in the current lasso model is:-

1. 1stFlrSF (First Floor Sq. feet)
2. 2ndFlrSF (Second Floor Sq. feet)
3. OverallQual (Overall Quality of the Property)
4. BasementFinSF1 (Sq. feet area of basement)
5. Neighbourhood_StoneBr (Boolean, whether or not neighbourhood is in Stone Brook)

We removed the features from the data, performed preprocessing and RFE feature selection again, and ran the models. The performance decreases to:

- Best Model : Ridge
- MAPE Test: 13.73% (increased from 10.96%)
- RMSE Score: 31739.64 (increased from 24740.78)
- R2 Score : 0.824 (decreased from 0.89)

The top 5 most important variables now are:

1. BsmtQual (Basement Quality)
2. GarageArea
3. FullBath (Full bathrooms above grade)
4. Neighborhood_Gilbert (Boolean, whether or not the neighbourhood is in Gilbert)
5. ExterQual (Exterior Quality)

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 sure a model is both robust and generalisable, we make sure it follows the two priciples:

1. Occam's Razor
2. Regularization

Occam's Razor:

Occam's Razor principle can be used to pick models which are simpler yet robust enough. It states that a model should be simple, but not simple enough!

The term simplicity can refer to any or all of these factors:

- Lesser number of independent variables required to map the dependent variable
- Lesser degree of the function if it's a polynomial
- Size of the best possible representation of the model (y = x1+2x2 is a simpler model compared to y = x1 + 2.0423x2)
- The depth of the decision tree.

Given two or more models that show similar performance in training or test data, we should pick the one model that makes fewer assumptions and follows at least any/some of the points mentioned above.

Advantages:

- Simpler models are usually more 'generic' and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.

Disadvantages:

- Simpler models make more errors in the training set. Complex models lead to overfitting. Simpler models have a high bias and low variance, complex models have a low bias and high variance.

Therefore, while choosing a model, we should make sure the model is simple, but not simple enough that it's performance is too bad and it cannot be used.

Regularization:

Regularization can be used to make the models simpler. Regularization is the process of introducing an additional penalty term in the cost function (essentially increasing the bias) to decrease the variance in the model (the

coefficients will not take higher values, thereby decreasing the variance; for Lasso more coefficients will be getting to 0). Regularization helps in maintaining the balance between bias and variance, such that the model is simpler than a vanilla model, but not simple enough to be unusable.

Making a model simpler leads to better Bias-Variance Trade-off:

- A simpler model usually has high bias in the model (the fit on training data will be poorer) and low variance (error for validation data will be lower if training data is changed slightly)
- A complex model has low bias in the model (perfect fit on training data) and high variance (error for validation data will be higher if training data is changed slightly)

Thus as we move from complex to simpler model, the accuracy will start decreasing, but change in accuracy with slight change in training data will reduce. The optimum model complexity is where there is a right balance between bias and variance, such that both are as low as possible.