Housing Price Prediction

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 of alpha for Ridge and Lasso Regression are

- Ridge = 100
- Lasso = 1000

When the value of alpha is doubled the model slightly under fits the data and thus r2_score reduces for both Ridge and Lasso.

Incase of Ridge, the rank of alpha doubled(alpha=200) is next to the optimal value of alpha(100), which means there is a very slight difference in the model rank and performance when lambda is doubled.

Top features for alpha=100 and alpha=200 are the same but are ordered differently

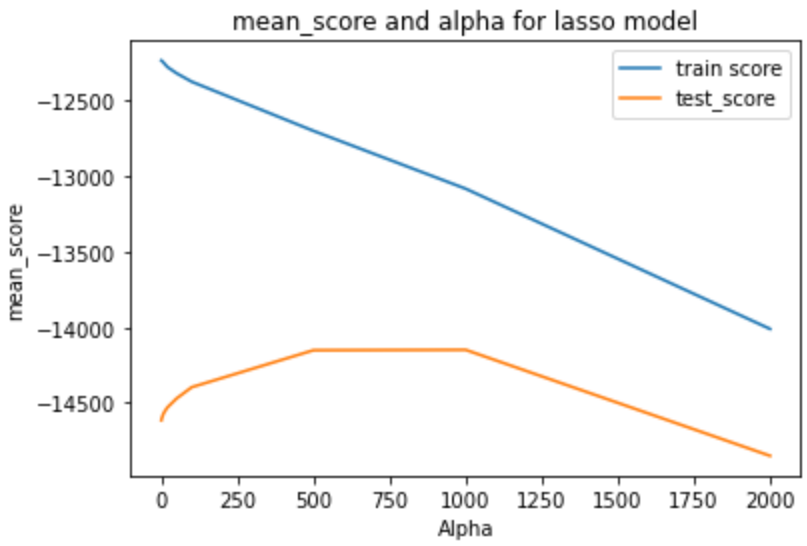
	Feature
3	OverallQual
11	GrLivArea
8	TotalBsmtSF
9	1stFlrSF
6	BsmtFinSF1
10	2ndFlrSF
17	GarageCars
64	Neighborhood_StoneBr
31	BsmtQual_cd
25	BldgType_cd

Fig1. for alpha=200

	Feature
3	OverallQual
11	GrLivArea
8	TotalBsmtSF
10	2ndFlrSF
9	1stFlrSF
6	BsmtFinSF1
64	Neighborhood_StoneBr
17	GarageCars
25	BldgType_cd
31	BsmtQual_cd

Fig2. for alpha=100

In the case of Lasso, the optimal value of alpha is 1000 and if it is doubled the accuracy of the model will take a hit on both the train and test set



As we can see in the figure above the test score of the model keeps up increasing until alpha=1000 and then falls suddenly until 2000. The mean test score rank for alpha=1000 is 1, whereas the mean test score rank for alpha=2000 is last.

Important features for alpha=1000 and alpha=2000 are respectively

	Feature
11	GrLivArea
3	OverallQual
8	TotalBsmtSF
6	BsmtFinSF1
17	GarageCars
31	BsmtQual_cd
21	BuiltOrRemAge
64	Neighborhood_StoneBr
35	KitchenQual_cd
2	LotArea

	Feature
3	OverallQual
11	GrLivArea
8	TotalBsmtSF
6	BsmtFinSF1
17	GarageCars
21	BuiltOrRemAge
31	BsmtQual_cd
29	ExterQual_cd
35	KitchenQual_cd
2	LotArea

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 mean train test score for Ridge and Lasso are as follows

- Ridge
  - Train r2_score = 0.92
  - Test r2_score = 0.89
- Lasso
  - Train r2_score = 0.91
  - Test r2_score = 0.88

Both Ridge and Lasso differ by very small values thus both are performing well for this dataset, But Ridge is giving this score by selecting all the features(107) whereas Lasso has selected 46 features selected for its model out of 107.

The Ridge regressor has a complex model with slightly better performance whereas Lasso has a simpler model with the almost same performance as the Ridge regressor, Hence from *Occam's razor* principle, **Lasso** seems a better model to choose between these two.

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?

Answer:

On excluding the top five derived columns from the data and then building the model, the optimal value of alpha changes, the model is now selecting 61 columns, initially, it was only selecting 46 features.

Now the top features of Lasso are: OverallQual, GrLivArea, TotalBsmtSF, BsmtFinSF1, GarageCars

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?

Answer:

As per *Occams Razor* principle given two models of similar performance for finite data of train and test data, we should choose the one less complex and fewer features compared to the one with more features and complex model for the following reason

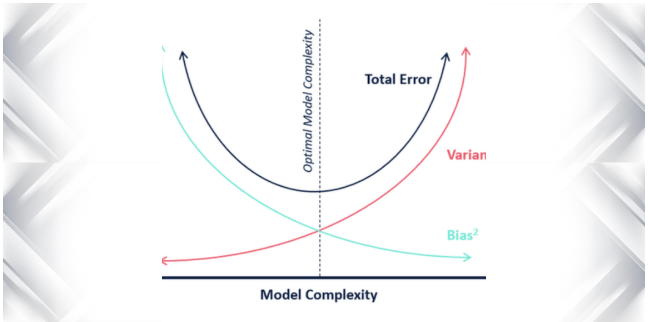
- Simpler models are more 'generic' and are widely applicable
- Simpler models are easy to train as they require lesser data to train as compared to complex models
- Simpler models are more robust,
    - Simpler models have high bias and low variance whereas complex models have high variance and low bias
    - Complex models tend to change wildly when there's a change in training data
- Complex models overfit the data thus work very well for training data but fail miserably on the test data. Simpler models make more errors on training data

"Make the model simple but not too simple leading into underfitting"

Regularization can be used to avoid making the model simple but not making it too naive to be of any use. In regression, Regulaizartion shrinks the coefficient estimates towards zero, thus discouraging the model from becoming a complex model. ƛ is the tuning hyperparameter used for regularization which decides how much we want to penalize the flexibility of the model

While making the model simpler we may come across a tradeoff called *"Bias-Variance tradeoff"*

- A complex model tends to overfit model thus for very little change in the data changes wildly thus having high variance and low bias
- The simpler model tends to underfit the data thus leading to high bias and low variance



Thus the best model would be the one that balances off both Bias and Variance leading to a model which is neither too complex nor is too simple