Subjective Questions – Advanced Regression

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?**

The optimal value of alpha obtained are as follows:
Ridge: 1.0
Lasso: 0.0001

The following changes will occur in the model if we double the value of alpha:

- **Smaller Coefficients:** Doubling the value of alpha increases the strength of the regularization term. This will lead to smaller coefficient values for all features. The influence of each feature on the prediction will be reduced.
- **Increased Bias:** Higher the value of alpha, higher will be the model's bias. It will tend to potentially underfit the training data more than the original model with the optimal alpha.
- **Reduces Overfitting:** If the original model was found to be overfitting, doubling the alpha can help mitigate overfitting by reducing the complexity of the model.
- **Feature Selection (Lasso):** Lasso regularization forces some coefficients to be exactly zero, effectively performing feature selection. Doubling the alpha in Lasso will increase the likelihood of more features being eliminated from the model which might be good for models used for prediction.

In our use case, the doubled value of alpha are as follows:
Ridge: 2.0
Lasso: 0.0002

In the code file 'JeniferSam_SurpriseHousing', the code to calculate updated coefficients is available under **Subjective Questions**.

Using the doubled value of alpha,
The most important predictor variables using Ridge are as follows:

Predictor Variable	**Coefficient**
OverallQual	0.15
GrLivArea	0.14
TotalSF	0.139
YearBuilt	-0.099
Neighborhood_StoneBr	0.0867
Neighborhood_NridgHt	0.0865
LotArea	0.083
TotRmsAbvGrd	0.066
OverallCond	0.065
GarageArea	0.0649

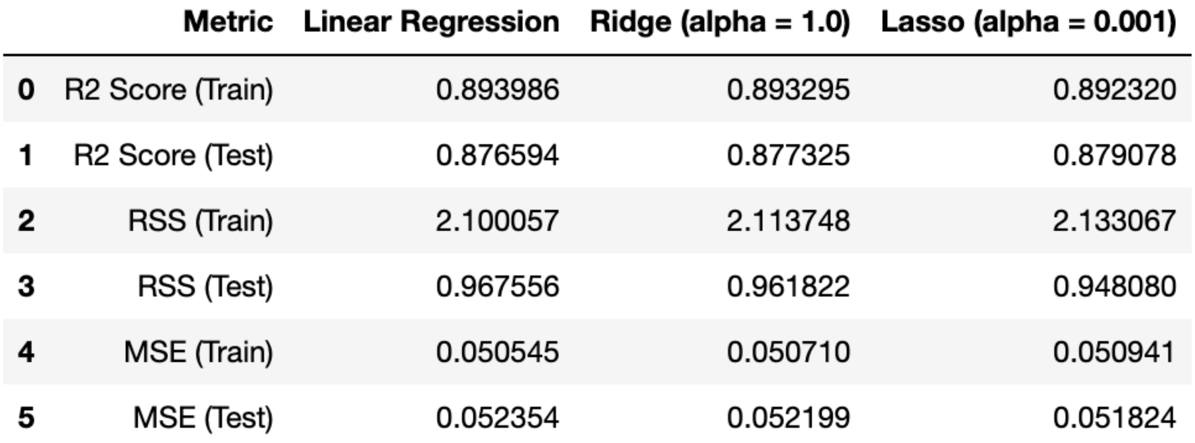
The most important predictor variables using Lasso are as follows:

Predictor Variable	Coefficient
OverallQual	0.17
TotalSF	0.157
GrLivArea	0.156
YearBuilt	-0.096
Neighborhood_NridgHt	0.086
LotArea	0.0836
Neighborhood_StoneBr	0.0825
OverallCond	0.0631
GarageArea	0.0630
KitchenQual	0.0622

The important predictor variables have not changed with doubling of alpha, however, the significance of the predictor variables (coefficients) have become smaller.

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?**

We will choose Lasso Regression for this problem statement based on the comparison of regression metrics tabulated in the table below for optimal alpha:



	Metric	Linear Regression	Ridge (alpha = 1.0)	Lasso (alpha = 0.001)
0	R2 Score (Train)	0.893986	0.893295	0.892320
1	R2 Score (Test)	0.876594	0.877325	0.879078
2	RSS (Train)	2.100057	2.113748	2.133067
3	RSS (Test)	0.967556	0.961822	0.948080
4	MSE (Train)	0.050545	0.050710	0.050941
5	MSE (Test)	0.052354	0.052199	0.051824

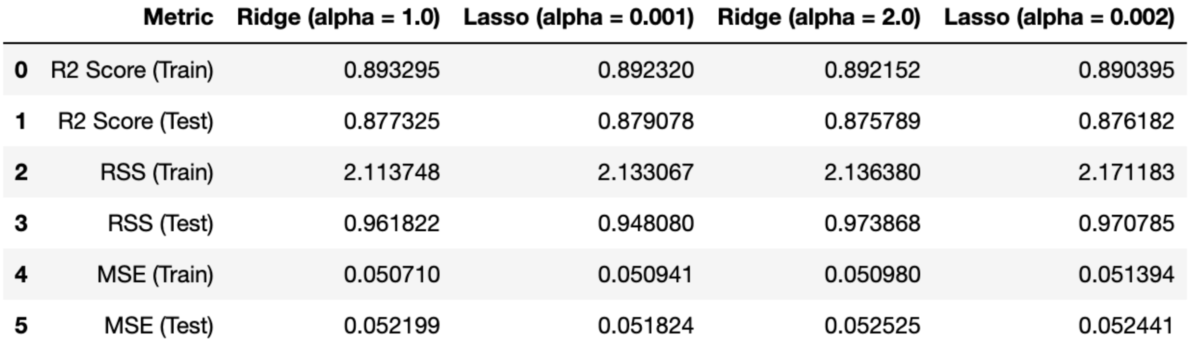
The above table is the output of metric evaluation performed under model interpretation in Step 7 in the code file.

We can see that Lasso has the:

- Highest R2-score on the test data
- Lowest RSS score on the test data
- Lowest MSE score on the test data

In addition, the advantage of using Lasso over Ridge is that Lasso will also perform feature selection by forcing some coefficients to become zero. Besides, getting a more accurate model (based on metrics above), we also get a simpler model since lesser predictors are considered in model building.

If we compare the regression metrics obtained for the optimal value of alpha and the doubled value of alpha, we get the following results:



	Metric	Ridge (alpha = 1.0)	Lasso (alpha = 0.001)	Ridge (alpha = 2.0)	Lasso (alpha = 0.002)
0	R2 Score (Train)	0.893295	0.892320	0.892152	0.890395
1	R2 Score (Test)	0.877325	0.879078	0.875789	0.876182
2	RSS (Train)	2.113748	2.133067	2.136380	2.171183
3	RSS (Test)	0.961822	0.948080	0.973868	0.970785
4	MSE (Train)	0.050710	0.050941	0.050980	0.051394
5	MSE (Test)	0.052199	0.051824	0.052525	0.052441

Here again, we can clearly see that Lasso with optimal alpha = 0.001 has better metrics in comparison to when alpha has doubled. We will therefore, choose **Lasso Regression with optimal alpha = 0.001** for SalePrice prediction.

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?**

The top 5 variables that I eliminate are:

OverallQual
TotalSF
GrLivArea
YearBuilt
Neighborhood_StoneBr

After eliminating these and rebuilding the model (steps performed in the codefile under Subjective questions), the top 5 predictor variables are:

Predictor	**Coefficients**
TotRmsAbvGrd	0.191494
BsmtQual	0.156904
KitchenQual	0.112498
ExterQual	0.109967
GarageArea	0.103737

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?**
We can make sure that a model is robust and generalizable by considering the following:

(i) **Data quality** – It is important to get the training dataset from trustworthy data sources. Missing values need to be removed or imputed. Data cleanup measures have to be performed considering business value in mind.

(ii) **Outlier treatment** – Visualize data using visual plots to understand the outliers in the dataset. Treat them by removing or flooring/capping them. Interquartile ranges are a common way to detect and remove outliers in the dataset. This technique has been performed on the dataset in this assignment too.

(iii) **Data preparation** –

- Remove irrelevant features and wherever possible, perform feature engineering to create derived metrics.
- Scale the data using min-max scalar or normalization.
- Categorical data needs to be modified based on whether they are ordinal or nominal variables.

(iv) Cross-validation – Use techniques such as GridSearchCV to split training and test datasets so that the model is tested on unseen data strictly.

(iv) **Hyperparameter tuning** – Very important to choose an optimal value for hyperparameter in order to avoid overfitting and underfitting data.

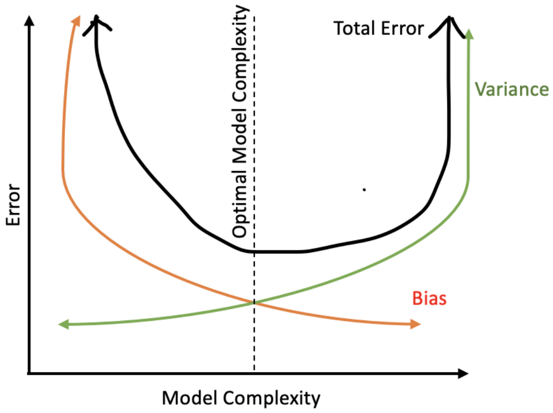
By generalizing a model, we affect the accuracy of the model.

(i) **Model accuracy decreases** - When a model is generalizable, the model is less complex which may reduce the accuracy of the model on the training data. But preferring model generalization over model accuracy helps the model perform better on unseen data.

(ii) **Reliable predictions** – Since the model is generalized, the predictions although less accurate are more reliable since the prediction is based on learning.

(iii) **Avoiding Overfitting –** Generalizable a model helps avoid overfitting a model.

This process of striking a balance between model robustness and model accuracy is called bias-variance trade-off.



As model complexity increases, bias reduces. That is, initially the model identifies the pattern in the data and then it begins to identify the noise in the data as well. A high bias model fails to fit well on the training data. A high variance model fails to fit well on testing data. This is a trade-off relationship. It is essential to have a model at optimal model complexity.