Advanced Linear Regression

Assignment-based Subjective Questions

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 value of alpha for Ridge Regression: **3.9**
     Optimal value of alpha for Lasso Regression: **0.001**

   - Metrics with different values of alpha:
     Ridge Regression:

	Metric	lambda=3.9	lambda=7.8
0	R-Squared (Train)	0.940374	0.934230
1	R-Squared (Test)	0.917801	0.915332
2	RSS (Train)	9.115041	10.054296
3	RSS (Test)	5.424146	5.587098
4	MSE (Train)	0.009207	0.010156
5	MSE (Test)	0.012763	0.013146
6	RMSE (Train)	0.095954	0.100776
7	RMSE (Test)	0.112972	0.114657
8	No. of features	282.000000	282.000000

When alpha value is doubled, we see that R-squared value on Test set almost remains the same whereas, R-squared on training dataset is reduced by 0.01

Lasso Regression:

	Metric	lambda=0.001	lambda=0.002
0	R-Squared (Train)	0.924489	0.904761
1	R-Squared (Test)	0.919767	0.898826
2	RSS (Train)	11.543362	14.559232
3	RSS (Test)	5.294461	6.676318
4	MSE (Train)	0.011660	0.014706
5	MSE (Test)	0.012458	0.015709
6	RMSE (Train)	0.107981	0.121270
7	RMSE (Test)	0.111613	0.125335
8	No. of features	94.000000	70.000000

When alpha value is doubled, we see that R-squared value on both training and test set is reduced by 0.02.

- Most important predictor variables after alpha is doubled
  Ridge Regression:

	Features	Coefficient	Mod
10	GrLivArea	1.240037	1.240037
7	1stFlrSF	1.238971	1.238971
6	TotalBsmtSF	1.170897	1.170897
17	TotRmsAbvGrd	1.136139	1.136139
13	FullBath	1.134210	1.134210
3	BsmtFinSF1	1.131998	1.131998
8	2ndFlrSF	1.128294	1.128294
123	OverallQual_9	1.110147	1.110147
1	LotArea	1.108795	1.108795
19	GarageCars	1.104082	1.104082

Lasso Regression:

	Features	Coefficient	Mod
10	GrLivArea	2.751944	2.751944
6	TotalBsmtSF	1.343755	1.343755
19	GarageCars	1.158459	1.158459
3	BsmtFinSF1	1.128484	1.128484
274	SaleType_New	1.127588	1.127588
123	OverallQual_9	1.119972	1.119972
124	OverallQual_10	1.100456	1.100456
219	CentralAir_Y	1.095522	1.095522
71	Neighborhood_Crawfor	1.092312	1.092312
20	GarageArea	1.088490	1.088490

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:

- Optimal value of alpha for Ridge Regression: **3.9**
  Optimal value of alpha for Lasso Regression: **0.001**

- From the model evaluation, we see that test performance of both Ridge and Lasso regression is almost same. However, we see that r2-score of Ridge is higher than lasso for training data. Also, the number of features selected by lasso is 94

as compared to ridge which has 282 features; thus, lasso has removed unwanted features from the model without affecting the model accuracy. This makes the model more generalized and simpler. Hence, we can go with Lasso regression.

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?

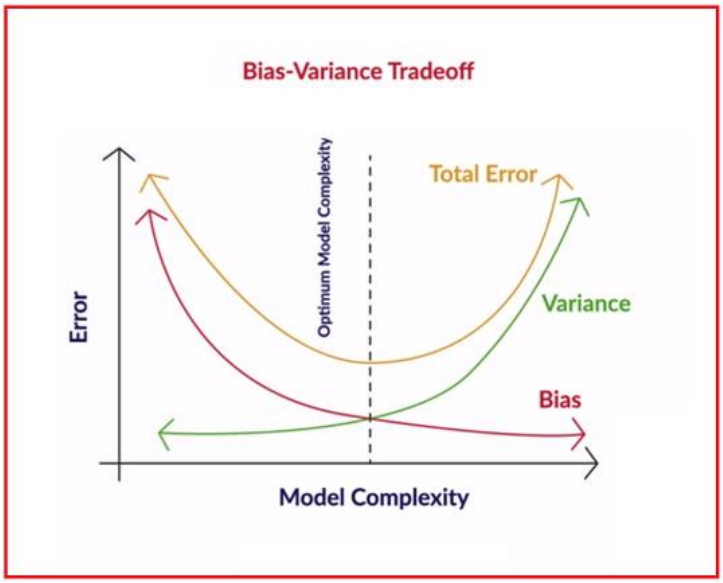
   Answer:
   After removing the top 5 features of lasso model, below are the 5 important predictor variables:
   1) **1stFlrSF** - For an increase in 1 square foot of first floor, the price of the house will increase by 2.4 to 2.5 times
   2) **2ndFlrSF** - For an increase in 1 square foot of first floor, the price of the house will increase by 1.5 to 1.6 times
   3) **BsmtFinSF1** - For an increase in 1 square foot of basement finished area, the price of house will increase by 1.1 to 1.2 times
   4) **GarageArea** - For an increase in 1 square foot of garage, the price of house will increase by 1.1 to 1.2 times
   5) **GarageCars** - For every unit increase in garage car capacity, the price of house will increase by 1.1 to 1.2 times

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:
   - A simpler model is usually more generic than a complex model. It is also robust and does not change significantly if the training data points undergo small changes.
   - On one hand, simplicity is generalisable and robust, and on the other hand, some problems are inherently complex in nature. There is a trade-off between the two, which is known as the bias-variance trade-off.

   - **Variance**: Variance refers to the degree of changes in the model itself with respect to changes in the training data.
   - **Bias**: Bias quantifies how accurate the model will be on the unseen data (test data). Extremely simple models will fail to predict complex real-world problems.

   - Ideally, we want a model with low variance and low bias which is not possible practically.
   - We need to find the optimal complexity of the model, where the total error is minimal (sum of variance and Bias).
   - With a little trade-off on bias, we can get significant reduction in variance.
   - With this optimal model complexity, model will be more generic and robust .


Bias-Variance Tradeoff
Optimum Model Complexity
Total Error
Variance
Bias
Error
Model Complexity