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 $\lambda = 20$
- Lasso $\lambda = 0.001$

Ridge:

- There's very little decrease in the R^2 values (train: 0.001935 and test: 0.000159) after increasing the λ.
- There's a slight increase in MSE for train set (0.0013) but menial change in test set (0.000091) after λ increased.
  Top predictors: `'1stFlrSF', '2ndFlrSF', 'OverallQual', 'SaleCondition_Normal', 'BsmtFinSF1'. 'SaleType_New'` at rank 5 was replaced by `'BsmtFinSF1'` after λ update.

Lasso:

- There's very little decrease in the R^2 values (train: 0.004461 and test: 0.000959) after increasing the λ.
- There's a slight increase in MSE for train set (0.003143) but menial change in test set (0.000539) after λ increased.
- Top predictors: `'1stFlrSF', '2ndFlrSF', 'OverallQual', 'SaleType_New', 'OverallCond'. 'SaleCondition_Normal'` at rank 5 was replaced by `'OverallCond'` after λ update.

	Metric	Linear Regression	Ridge (λ = 20.0)	Lasso (λ = 0.001)
0	R2 Score (Train)	0.926355	0.923917	0.922023
1	R2 Score (Test)	0.868791	0.874700	0.873337
2	RSS (Train)	13.608746	14.059268	14.409434
3	RSS (Test)	6.115941	5.840508	5.904017
4	MSE (Train)	0.108313	0.110091	0.111454
5	MSE (Test)	0.145222	0.141914	0.142684

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:

Ridge & Lasso models have very little difference in values of R-squared or MSE in both test & train data sets. As the R-squared value hasn't gone up after feature selection using the Lasso model, it's evident that most predictors do impact the response variable here. So *Ridge* is a better fit.

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:

Before drop:

```
'1stFlrSF', '2ndFlrSF', 'SaleType_New', 'OverallQual', 'SaleCondition_Normal'
```

After drop:

These are the top 5 predictors - `'BsmtFinSF1', 'BsmtUnfSF', 'TotRmsAbvGrd', 'FullBath', 'OverallCond'`

Notice the R^2 dipped while the RSS and MSE increased.

	Lasso
BsmtFinSF1	0.091572
BsmtUnfSF	0.061481
TotRmsAbvGrd	0.061242
FullBath	0.053369
OverallCond	0.050753

Metric	Lasso: λ = 0.002	Lasso: drop 5 var
R2 Score (Train)	0.917562	0.885933
R2 Score (Test)	0.872378	0.841309
RSS (Train)	15.233700	21.078349
RSS (Test)	5.948760	7.396927
MSE (Train)	0.114597	0.134800
MSE (Test)	0.143223	0.159708

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:

A simple model would usually have high bias and low variance whereas a complex model would have low bias and high variance. In either case, the total error would be high. What we need is lowest total error, i.e., *low bias* and *low variance*.

To achieve this we use regularization where we add a *penalty term* to the model's cost function that shrinks the magnitude of the model coefficients towards 0, preventing the risk of overfitting.

If λ, the shrinkage penalty, is too small the model would remain *overfit* but if it's too high, it may end up with an *underfit* model.

With regularization, maybe the coefficients are more biased but the variance of the model may see a marked reduction i.e. trading bias for significant reduction in variance.

Regularization does not improve the accuracy on the data set but it can improve the *generalization performance*, i.e., the performance on new, unseen data, which is exactly what we want.