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 1:

The optimal value of alpha for ridge and lasso regression is as below:

- Ridge - 12.5,
- Lasso - 0.002

After doubling, model performance slightly decreased in both cases as below:

	Optimal Alpha	Double Alpha
r2_train_lasso	0.89	0.87
r2_test_lasso	0.87	0.86
mse_test_lasso	0.02	0.02
r2_train_ridge	0.92	0.91
r2_test_ridge	0.89	0.88
mse_test_ridge	0.0188	0.0189

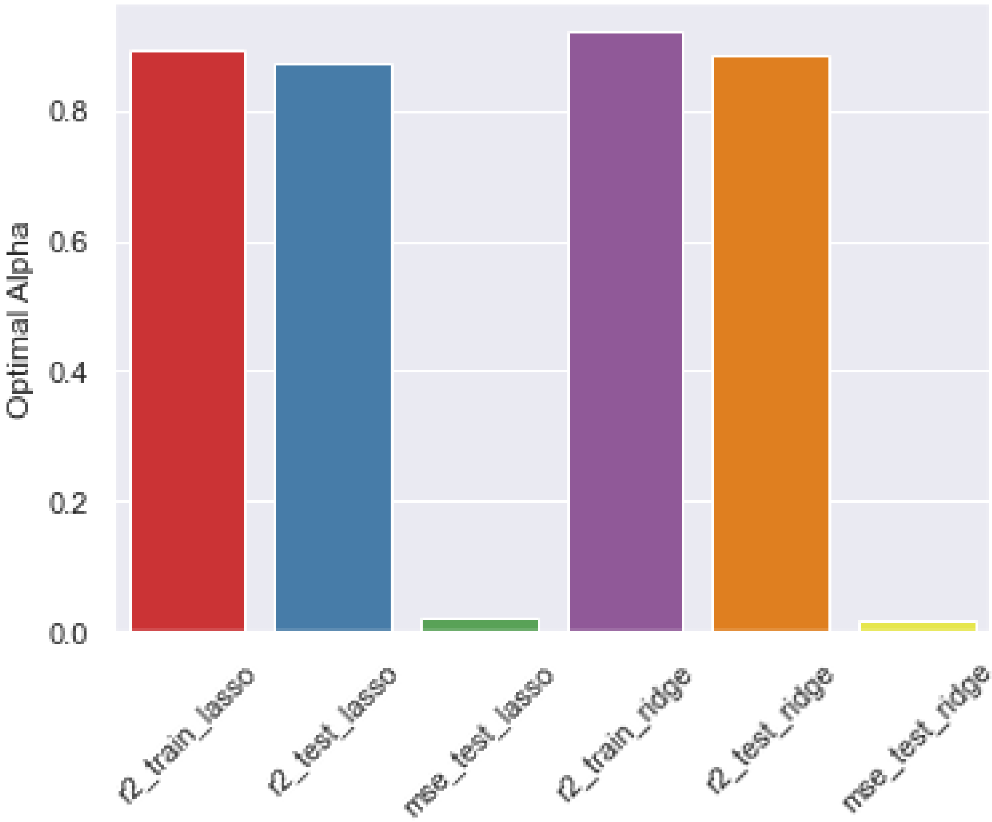
Question 2
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer 2:

The optimal value of alpha for ridge and lasso regression is as below:

- Ridge - 12.5
- Lasso - 0.002

The model performance by Ridge Regression is slightly better in terms of R2 values and Mean Square Error of Train and Test as below. However, it would be better to use Lasso since helps in feature selection (as the coefficient value of the features can become 0).



In the ridge, the coefficients of the linear transformation are normal distributed and in the lasso they are Laplace distributed. In the lasso, this makes it easier for the coefficients to be zero and therefore easier to eliminate some of your input variable as not contributing to the output.

It is advisable to use a simple yet robust model. So we would go with the Lasso Model here.

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 3:

The five most important predictor variables in the lasso model are:

No.	Features	Lasso Coefficient
1	Neighborhood_Crawfor	0.074
2	OverallQual	0.072
3	Neighborhood_Somerst	0.063
4	BsmtFullBath	0.054
5	Neighborhood_NridgHt	0.053

If the five most important predictor variables in the lasso model are not available in the incoming data. The five most important predictor variables in the lasso model now are:

No.	Features	Lasso Coefficient
1	Condition1_Norm	0.051
2	Neighborhood_Edwards	-0.051
3	GarageCars	0.046
4	OverallCond	0.042
5	FireplaceQu_none	-0.042

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 4:

According to Occam's Razor, if two models exhibit comparable "performance" in finite training or test data, we should choose the one that makes fewer mistakes because of the following reasons:

1. Simpler models are typically more "generic" and have a wider range of applications.
2. Less complex models can be effectively trained using fewer training data, making simpler models easier to train.
3. More reliable models are simpler ones.
    a. Complex models frequently experience drastic changes when the training data set is altered.
    b. Simple models have high bias and low variance, while complicated models have high bias and low variance.
4. More mistakes are made in training set by simpler models. Overfitting occurs as a result of complex models, which perform admirably on training samples but horribly on test data.

The idea is to make the model simple but not simpler in order to increase the model's robustness and generalizability.

The model can be made simpler by regularization. Regularization aids in striking the difficult balance between keeping the model straightforward and preventing it from being overly simplistic and useless. Regression regularization entails multiplying the squares or absolute values of the model's parameters by a regularization term, which is added to the cost function.

Additionally, a model's bias-variance trade-off results from simplification.

1. A complicated model is particularly unstable and highly sensitive to any changes in the training data because it must be changed for any tiny change in the dataset.
2. Even if additional data points are added or subtracted, a more basic model that abstracts out any pattern revealed by the provided data points is unlikely to change drastically.

How accurate the model is expected to be on test data is quantified by bias. If there is enough training data, a complicated model can provide an accurate job forecast. Too naive models, such as those that respond the same to every test input and make no distinction at all, have a very big bias since their predicted error for every test input is quite high. Variance describes how much the model has changed compared to how the training set has changed.

As a result, maintaining a balance between bias and variance can help to preserve the model's accuracy because doing so reduces overall error, as illustrated in the graph below.