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

Optimal value of Alpha for Ridge is 55.
Optimal value of Alpha for Lasso is 0.001

Changes in the model on choosing to double the value of alpha for both Ridge and Lasso can be seen in the below table, in terms of metrics that are used to evaluate the models:

	Metric	Ridge(Alpha = 55)	Ridge(Alpha=110)	Lasso(Alpha=0.001)	Lasso(Alpha=0.002)
0	R2 Score (Train)	0.942303	0.937966	0.943446	0.939561
1	R2 Score (Test)	0.898049	0.898049	0.895143	0.894960
2	RSS (Train)	5.665090	6.090934	5.552873	5.934285
3	RSS (Test)	4.050169	4.050170	4.165617	4.172906
4	MSE (Train)	0.088826	0.092104	0.087942	0.090912
5	MSE (Test)	0.114673	0.114673	0.116296	0.116398

After the change is implemented, the most important predictors are:

For Ridge with Alpha doubled:
OverallQual, GrLivArea, AgeOfHouse, 1stFlrSF, OverallCond, BsmtFinSF1, GarageArea

For Lasso with Alpha doubled:
GrLivArea, AgeOfHouse, OverallQual, BsmtFinSF1, OverallCond, GarageArea

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

To choose one model over other we need to compare R squared, RSS and MSE. A higher R squared explains more variation of the dependent variable and is thus better. While a lower RSS and MSE is preferable as these define errors. Also, the difference between the R squared of train and test data should be low. A higher difference, with R squared for train data higher than that of test data, indicates overfitting.

We can see from the table pasted in the previous answer that for optimal alphas values, R squared values for both Ridge and Lasso are almost the same, and decently near to 1. Also the difference btween R squared of train and test data is equally low.

RSS and MSE values of the two are also considerably close to each other.

Both models are good for this case.

I would still choose Lasso, because we have significant number of variables that do not affect the model much, and Lasso would automatically drop them by making their coefficients as Zero.

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
Five most important predictor variables in the lasso model with optimal alpha are:
- GrLivArea
- AgeOfHouse
- GrLivArea
- OverallQual
- BsmtFinSF1

After dropping these variables, and rebuilding the lasso model, the optimal alpha is again 0.001, and the new most important five predictor variables are:
- 1stFlrSF
- 2ndFlrSF
- BsmtFinType2_TA
- GarageArea
- OverallCond

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 model can be called robust and generalizable if it performs good even on unseen data and not just on the train data. In order to increase the accuracy we might end up making the model overfit, which means it would learn the train data too well i.e. it would try to fit through most data points and in doing so it would become useless when it encounters unseen data.

Removing the outliers helps in removing the data that adds no real value to the model and contributes in overfitting. In the assignment outlier treatment has been done by retaining only those values whose z score is between -3 and 3.

Another technique that helps in making the model robust is regularization. Higher values of model coefficients and higher number of variables can also result in overfitting. Regularization shrinks the coefficients towards Zero, by adding a penalty term to the cost function. In doing so we compromise on the bias to get significant reduction in variance. Bias is the error in training data, and variance is the error on test data. Too less bias would result in very high variance, making the model useless for real unseen data - this would be a complex model. In the assignment we have used Ridge and Lasso methods for getting regularized models.

Cross Validation is another technique that can be brought into practice to utilize the full data in the best way possible for training. All variations in the data would be traversed by the model, as opposed to not using cross validation where we have fixed test, train and validation sets and the model loses out on seeing some data.

The implication for making the model robust would be somewhat less accuracy on train data, but at the same time it would perform well on unseen data.