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 Alpha values as detected by the model are
- Ridge – 2.0
- Lasso – 0.001

If we double the value of alpha in both Ridge and Lasso regression there will be increase the regularization strength, which can help in reducing overfitting and handling multicollinearity. However, it may also lead to simpler models with smaller coefficients (Ridge) or even some coefficients being set to zero (Lasso)

Model score metrics after doubling the alpha as 4.0 for ridge & 0.002 for lasso.
Note: Working for this is in Jupyter notebook at last

	Metric	Ridge Regression (alpha 2)	Ridge Regression (alpha 4)	Lasso Regression (alpha 0.001)	Lasso Regression (alpha 0.002)
0	R2 Score (Train)	0.943879	0.939026	0.925700	0.910669
1	R2 Score (Test)	0.916203	0.916884	0.913429	0.899682
2	RSS (Train)	6.742969	7.326116	8.927197	10.733238
3	RSS (Test)	4.702481	4.664219	4.858151	5.629559
4	RMSE (Train)	0.086945	0.090626	0.100040	0.109694
5	RMSE (Test)	0.110806	0.110355	0.112625	0.121238

Most important predicator variables after change are as follow

	Feature Name	Coefficient	Absolute Coefficient
14	GrLivArea	0.845116	0.845116
3	OverallQual	0.428539	0.428539
11	TotalBsmtSF	0.305025	0.305025
4	OverallCond	0.226871	0.226871
28	houseAge	-0.212380	0.212380
8	BsmtFinSF1	0.158066	0.158066
2	LotArea	0.112989	0.112989
21	GarageCars	0.111497	0.111497
204	SaleCondition_Partial	0.098197	0.098197
50	Neighborhood_Crawfor	0.089401	0.089401

Lasso with alpha 0.001

	Feature Name	Coefficient	Absolute Coefficient
14	GrLivArea	0.808472	0.808472
3	OverallQual	0.427105	0.427105
11	TotalBsmtSF	0.245095	0.245095
8	BsmtFinSF1	0.164093	0.164093
4	OverallCond	0.155187	0.155187
28	houseAge	-0.121032	0.121032
21	GarageCars	0.116971	0.116971
204	SaleCondition_Partial	0.082381	0.082381
31	MSZoning_RL	0.082024	0.082024
65	Neighborhood_Somerst	0.054062	0.054062

Lasso with alpha 0.002

After doubling the alpha, most important predictor variable mostly remains same, however the order of coefficients impacting has changed.

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:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.950891	0.943879	0.925700
1	R2 Score (Test)	0.808349	0.916203	0.913429
2	RSS (Train)	5.900490	6.742969	8.927197
3	RSS (Test)	10.754899	4.702481	4.858151
4	RMSE (Train)	0.081332	0.086945	0.100040
5	RMSE (Test)	0.167573	0.110806	0.112625

Both models showed very close R2 scores on test data prediction, however we can see Lasso scores are slightly better.

I choose to use Lasso model in this scenario as with lasso we have another advantage of that it actually reduces coefficients to absolute 0 while Ridge never
eliminates the features it reduces the all coefficients are reduced but never to make them absolute 0.

Due to lesser variables from Lasso we will have less complex model too.

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:

Removing the initial model's best 5 variables, below are new best 5 variables suggested by model

1. 1stFlrSF : First Floor square feet
2. MSZoning_FV : Properties belonging to zone Floating Village Residential
3. MSZoning_RL : Properties belonging to zone Residential Low Density

4. MSZoning_RH : Properties belonging to zone Residential High Density
5. 2ndFlrSF : Second floor square feet

Note: Working for this is in Jupyter notebook at last

	Feature Name	Coefficient	Absolute Coefficient
9	1stFlrSF	0.632327	0.632327
24	MSZoning_FV	0.544138	0.544138
26	MSZoning_RL	0.519942	0.519942
25	MSZoning_RH	0.488336	0.488336
10	2ndFlrSF	0.475181	0.475181
27	MSZoning_RM	0.466806	0.466806
6	BsmtFinSF1	0.421492	0.421492
8	BsmtUnfSF	0.262236	0.262236
49	Neighborhood_MeadowV	-0.198378	0.198378
2	LotArea	0.179715	0.179715

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:

Bias (underfitting) is error in model, when the model is weak to learn from the data. High bias means model is too simple & is unable to capture or learn underlying patterns in the data. Model performs poor on training and testing data.

Variance (overfitting) is error in model, when model tries to over learn from the data by making itself complex. High variance means model is too sensitive to noise in training data & performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.

It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.

Even though a model is less accurate, it should be as simple as possible, but more robust and generalized.

To make the model robust, we should

- reduce the complexity of model, use less predictor features. Simpler the model, the greater the bias, but the less the variance and the more generalized it is.
- make sure training data scores are as close to test data
- drop more correlated feature, and not relying on lasso only
- If possible have more data available so that outliers treatment can be significant
- Apply regularization
- Apply bias, variance trade off rules

In terms of accuracy, its implication is that a robust, generalizable model works equally well on both training and test data. Accuracy does not change significantly between training and test data.