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 and lasso is –
Ridge - 0.05
Lasso – 0.001

The changes in model if we choose to double the value of alpha for both ridge and regression –

Changes in metric:

Out[1083]:

	Metric	Linear Regression	Ridge Regression	Lasso Regression	Ridge twice alpha	Lasso twice alpha
0	R2 Score (Train)	0.925479	0.919114	0.874561	0.911571	0.862241
1	R2 Score (Test)	0.833722	0.847808	0.860581	0.853923	0.866140
2	RSS (Train)	1.265498	1.373579	2.130165	1.501671	2.339376
3	RSS (Test)	1.267967	1.160549	1.063149	1.113924	1.020762
4	MSE (Train)	0.035206	0.036679	0.045677	0.038351	0.047867
5	MSE (Test)	0.053804	0.051475	0.049267	0.050430	0.048275

1. By making alpha double test r2 score increased slightly. No major difference.
2. By making alpha double test RSS score decreased slightly. No major difference.
3. By making alpha double test MSE score decreased slightly. No major difference.

Changes in Coefficients:

By making alpha double for Ridge and lasso, the coefficient magnitude is reduced as you can see in below table. Also theoretically if we increase alpha value , coefficient needs to be decreased in order to reduce cost function. Refer below table as result from the experiment.

	Linear	Ridge	Lasso	Ridge_twice_alpha	Lasso_twice_alpha
LotFrontage	0.050915	0.021587	-0.000000	0.008876	-0.000000
LotArea	0.224126	0.182154	0.070041	0.160474	0.047230
OverallQual	0.214421	0.230769	0.266482	0.239115	0.278518
OverallCond	0.133435	0.130715	0.127747	0.129341	0.122958
BsmtFinSF1	0.236117	0.193761	0.109418	0.173324	0.095614
BsmtFinSF2	0.033116	0.033789	0.025234	0.033977	0.018318
BsmtUnfSF	0.017682	0.012763	-0.000000	0.010457	-0.000000
TotalBsmtSF	0.232858	0.192014	0.121349	0.172300	0.101489
GrLivArea	0.514304	0.484624	0.426225	0.468012	0.405543
KitchenAbvGr	-0.091091	-0.078169	-0.042923	-0.071533	-0.032722
GarageCars	0.073789	0.088005	0.111587	0.095037	0.113186
houseAge	-0.118729	-0.116562	-0.115373	-0.115493	-0.114799
MSZoning_FV	0.144334	0.137904	0.070454	0.133048	0.031944
MSZoning_RH	0.134809	0.131376	0.063061	0.127843	0.019158
MSZoning_RL	0.132834	0.131995	0.075462	0.129874	0.041059
MSZoning_RM	0.111934	0.104479	0.036386	0.099169	0.000000
LandSlope_Sev	-0.074113	-0.033951	0.044254	-0.014230	0.028796
Condition2_PosN	-0.650088	-0.572329	-0.407083	-0.523694	-0.281469
Condition2_RRAe	-0.200084	-0.143735	-0.000000	-0.112592	-0.000000
RoofStyle_Shed	0.173006	0.128060	0.000000	0.103455	0.000000
RoofMatl_CompShg	1.082466	0.705442	0.078807	0.527064	0.019284
RoofMatl_Membran	1.231607	0.772700	0.004424	0.556309	0.000000
RoofMatl_Metal	1.214644	0.758667	0.000000	0.543605	0.000000
RoofMatl_Roll	1.079329	0.665170	0.000000	0.471000	-0.000000

The most important predictor variables after the change is implemented :
GrLivArea, OverallQual, OverallCond, **GarageCars**, **TotalBsmtSF**, **BsmtFinSF1**, **LotArea**

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: I will choose lasso based on the Mean squared error difference between train and test. In this particular problem statement there are high chances of overfitting because data is very less. One way of identifying overfitting is the high different in training and test error. With Ridge the difference has been reduced but with lasso it is reduced significantly and both train and test have almost similar error and R2 score of lasso is slightly higher compared to Ridge. Please refer below table.

Out[1083]:

	Metric	Linear Regression	Ridge Regression	Lasso Regression	Ridge twice alpha	Lasso twice alpha
0	R2 Score (Train)	0.925479	0.919114	0.874561	0.911571	0.862241
1	R2 Score (Test)	0.833722	0.847808	0.860581	0.853923	0.866140
2	RSS (Train)	1.265498	1.373579	2.130165	1.501671	2.339376
3	RSS (Test)	1.267967	1.160549	1.063149	1.113924	1.020762
4	MSE (Train)	0.035206	0.036679	0.045677	0.038351	0.047867
5	MSE (Test)	0.053804	0.051475	0.049267	0.050430	0.048275

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:

Earlier the five most important predictor variables were –

Predictor Coefficient magnitude
GrLivArea : 0.4262252109969818
OverallQual: 0.2664821230292393
OverallCond: 0.12774730924956318
TotalBsmtSF: 0.12134874297047783
GarageCars : 0.11158742036906844

After removing those the five most important predictor variables are –

Predictor Coefficient magnitude
BsmtFinSF1 : 0.5716306902750273
BsmtUnfSF : 0.225234685423552
LotArea : 0.20283134293103527
RoofMatl_WdShngl: 0.19202914353698017
RoofMatl_WdShake: 0.1754133582644311

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:
Model can be considered robust and generalisable when it does not remember the noise along with the useful pattern. The model should not overfit on training data. Model should not be sensitive with the outliers.
Few points can be taken care: -
Use of transformation when data is assumed to have outliers and right tailed
Consider adjusted r2 square metric.
Use of regularization technique.

Implications on accuracy could be like the training accuracy will go down when you try to make model more generalized, because we are trying to make model learn more varieties of data and handle more generalized data. Accuracy is compromised when we try to make model more generalisable. It's a kind of trade-off we need to decide.