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 is the penalty term which penalizes the coefficient values for regression model. In the current assignment the best optimal values for Ridge and Lasso are as follows.

Alpha for Ridge = 1
Alpha for Lasso = 0.0001

If we choose or double the value of alpha in both Ridge and Lasso regression the penalty term will increase and the model complexity will be reduced. In current assignment if we double the alpha following changes are happened with respect to coefficient values.

Ridge regression:

	Top 7 features with Alpha=1	Coefficient values with Alpha=1	Top 7 features with Alpha=2	Coefficient values with Alpha=2
1	1stFlrSF	0.243769	1stFlrSF	0.227859
2	OverallQual	0.215656	OverallQual	0.204147
3	2ndFlrSF	0.192918	2ndFlrSF	0.181200
4	TotalBsmtSF	0.154438	TotalBsmtSF	0.147520
5	BsmtUnfSF	-0.097500	GarageCars	0.095666
6	GarageCars	0.096226	BsmtUnfSF	-0.091321
7	OverallCond	0.088677	BsmtQual	0.082081

Lasso regression:

	Top 7 features with Alpha=0.0001	Coefficient values with Alpha=0.0001	Top 7 features with Alpha=0.0002	Coefficient values with Alpha=0.0002
1	1stFlrSF	0.264703	1stFlrSF	0.265229
2	OverallQual	0.239660	OverallQual	0.243165
3	2ndFlrSF	0.196574	2ndFlrSF	0.190603
4	TotalBsmtSF	0.142910	TotalBsmtSF	0.131567
5	BsmtUnfSF	-0.100121	BsmtUnfSF	-0.096670
6	GarageCars	0.088778	GarageCars	0.083122
7	OverallCond	0.086817	YearBuilt	0.079776

MSE with Alpha:

	Ridge	Lasso
Alpha	0.003143	0.003157
2 * Alpha	0.003163	0.003164

From the tables we can clearly observe that there is a change/decrease in coefficient values form Alpha to 2XAlpha, some times the order of variables also changes.

Also there is a slight increase in mean square error from Alpha to 2XAlpha, It means due to the complexity got reduced bias of the model increased.

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 regression, because

The optimal value of alpha for Ridge is 1, and for Lasso is 0.0001, And we can observe that the MSE value for Lasso regression in slightly higher then Ridge regression, But the difference in MSE from train and test data is low for Lasso regression.

But in terms of complexity Ridge is more complex then Lasso because Lasso regression penalizes the coefficients up to '0' where as Ridge not. So Lasso can be much use full for feature elimination and complexity reduction, By compromising a little amount of MSE we can remove multiple number of independent variables from prediction. So that model will be more robust and simple.

In current assignment with Alpha=0.0002 we are able to eliminate 16 independent variables using Lasso the difference in some of coefficient values are shown in below table.

Feature name	Ridge coefficient alpha=2	Lasso Coefficient alpha=0.0002
Exterior2nd_Stone	0.019813	0.000000
MiscFeature_Othr	0.002638	0.000000
MasVnrType_BrkFace	0.012519	0.000000
SaleCondition_AdjLand	0.015249	0.000000
MiscFeature_Shed	0.014012	0.000000
Exterior1st_CemntBd	-0.007679	0.000000
MiscFeature_No	0.013342	0.000000
Foundation_Slab	0.019596	0.000000

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:

I have removed to five features from the data without changing the alpha value and created another models with both Ridge and Lasso following are the top 5 independent variables after change is implemented.

New key features after removing actual top five features:

	Top five features Lasso, alpha=0.0001	Top five features Ridge, alpha=1
1	BsmtQual	BsmtQual
2	GarageCars	GarageCars
3	ExterQual	ExterQual
4	LotArea	LotArea
5	BedroomAbvGr	BedroomAbvGr

Top five features removed:

	Top 5 features in Ridge and Lasso
1	1stFlrSF
2	OverallQual
3	2ndFlrSF
4	TotalBsmtSF
5	BsmtUnfSF

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:

There are several key points to be followed to make sure that our model is more robust and generalisable, as follows.

- The first impotent factor is data, any machine learning algorithm or model performance mainly depends on input training data. If our data is more diverse and clean without much outliers we can get good model.
- Model selection is one of the most key factor to make our model more robust and generalised. Choosing proper model depends on use case. For example building classification model for continues target variable will never work as expected.
- After model selection fine tuning the parameters are most impotent for example choosing proper lambda value for Ridge or Lasso regression. For this we can use cross validation to get optimal lambda value.
- Avoiding collinearity and choosing significant variables in data is one of the most important step to make sure our model is more robust.
- Model complexity, has very high importance to get robust models. A very high complex models can perform better in limited incoming test data but fails at un-seen diverse test data. So making simpler model can lead to reduction in accuracy but simpler models are more generalized. Bias variance trad off: Choosing optimal point between bias and variance helps in building simpler and generalised models.