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?

Ans::

Optimal value of alpha for ridge regression => 0.5

Optimal value of alpha for lasso regression => 0.0001

Here are the metrics by having the above alpha values getting used.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score(Train)	0.879305	0.930700	0.922711
1	Residual Sum of Squares(Train)	2.049619	1.176835	1.312506
2	Mean Squared Error(Train)	0.002007	0.001153	0.001286
3	Root Mean Squared Error(Train)	0.044805	0.033950	0.035854
4	R2 Score(Test)	0.867496	0.875827	0.881931
5	Residual Sum of Squares(Test)	1.010423	0.946896	0.900349
6	Mean Squared Error(Test)	0.002307	0.002162	0.002056
7	Root Mean Squared Error(Test)	0.048030	0.046496	0.045339

And data w.r.to number of features is as below

```
Features selected by Lasso: 138
Features present in Ridge: 225
```

Here are the coefficients of few features in the given data set (Top 10 features)

Ride	Lasso

	Ridge
GrLivArea	0.222645
LotArea	0.092923
MSZoning_RH	0.092363
MSZoning_FV	0.091813
MSZoning_RL	0.089512
MSZoning_RM	0.081468
GarageCars	0.076604
OverallQual_9	0.075961
SaleType_ConLD	0.064487
2ndFlrSF	0.059266

	Lasso
GrLivArea	0.325122
OverallQual_9	0.075870
GarageCars	0.074656
MSZoning_RL	0.048867
MSZoning_RH	0.047990
MSZoning_FV	0.046979
LotArea	0.046148
BsmtFullBath	0.045149
FullBath	0.044166
Neighborhood_Crawfor	0.043726

By doubling the alpha value for both ridge and lasso, magnitude of the model coefficients will be getting lower and model is more regularized. (A large lambda implies a simpler model)

we get the data as below on the coefficient side (Top 10 features)

Ride	Lasso

	Ridge
GrLivArea	0.178873
GarageCars	0.074148
LotArea	0.073237
OverallQual_9	0.072223
MSZoning_RH	0.071407
MSZoning_FV	0.069686
MSZoning_RL	0.069203
2ndFlrSF	0.065934
MSZoning_RM	0.061373
FullBath	0.055840

	Lasso
GrLivArea	0.343422
OverallQual_9	0.080505
GarageCars	0.076475
OverallQual_8	0.046500
FullBath	0.045071
BsmtFullBath	0.041883
Neighborhood_Crawfor	0.038118
Neighborhood_Somerst	0.032219
Neighborhood_NridgHt	0.031410
Exterior1st_BrkFace	0.028809

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?

Ans::

Optimal value of alpha for ridge regression => 0.5

Optimal value of alpha for lasso regression => 0.0001

And data w.r.to number of features is as below

```
Features selected by Lasso: 138
Features present in Ridge: 225
```

From the above data, using Lasso number of features got reduced.

And scores w.r.to R-square and other related parameters we have the below data.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score(Train)	0.879305	0.930700	0.922711
1	Residual Sum of Squares(Train)	2.049619	1.176835	1.312506
2	Mean Squared Error(Train)	0.002007	0.001153	0.001286
3	Root Mean Squared Error(Train)	0.044805	0.033950	0.035854
4	R2 Score(Test)	0.867496	0.875827	0.881931
5	Residual Sum of Squares(Test)	1.010423	0.946896	0.900349
6	Mean Squared Error(Test)	0.002307	0.002162	0.002056
7	Root Mean Squared Error(Test)	0.048030	0.046496	0.045339

R-square score for test data set in both Ridge and Lasso is closeby.

Going with the number of features getting used for predicting the target variable, selecting Lasso seems to be good option (As we are using 138 features only)

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?

Ans::

By dropping columns which are coming in the top most five ('GrLivArea','OverallQual','GarageCars','MSZoning') have generated new lasso model and got the below data.

Note :: In the above step have dropped only 4 columns as MSZoning is a categorical column.

	Lasso
TotalBsmtSF	0.242764
2ndFlrSF	0.134364
FullBath	0.079302
Neighborhood_StoneBr	0.061663
LotArea	0.056344
BedroomAbvGr	0.054770
Neighborhood_NridgHt	0.054755
Neighborhood_NoRidge	0.053005
BsmtFullBath	0.044133
Neighborhood_Somerst	0.043259

Question 4

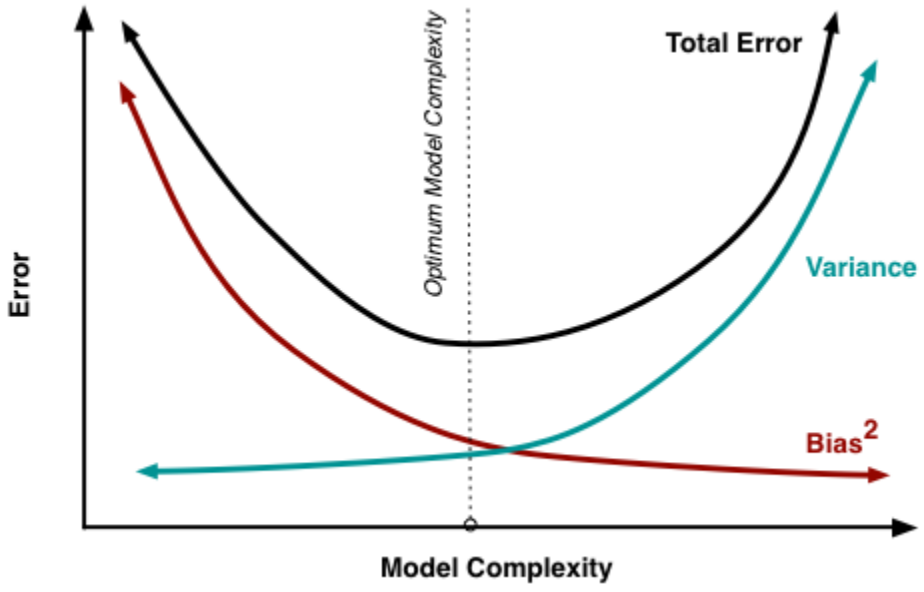
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Ans::

A model can be called as more robust and generalisable when it performs pretty well on the unseen data. When a model becomes complex, it might perform well on the data that is used to train it. But doesn't perform well on the unseen data. This problem is called as overfitting. So we need to choose a model which have low bias and high variance. To get this "Regularization" concept will be applied to find the right model. As part of this "Regularization" process a balance between bias and variance is made out. We need to select right set of variables for getting a good model.

Similarly, we have another problem called underfitting, it occurs when our model neither fits the training data nor generalizes on the new data.

The below image talks about how to handle balancing of bias and variance for getting an optimal or robust model.



(Above picture taken from https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/)

In linear regression, model complexity depends on the following

1) Magnitude of co-efficients
2) Number of co-efficients

The more extreme values of the coefficients the more complex the model is and hence the higher are the chances of overfitting.

Regularization works by shrinking the coefficient towards Zero. In current module learnt about below techniques that are used for Regularization

1) Ridge
2) Lasso

With regularization we compromise by allowing a little bias for a significant gain in variance.