Question 1

a) What is the optimal value of alpha for ridge and lasso regression?

b) What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

c) What will be the most important predictor variables after the change is implemented?

Answer :

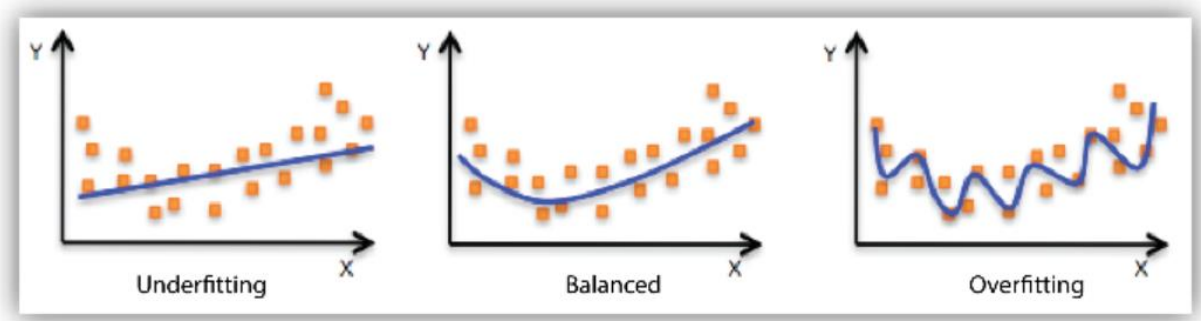
a) The optimal value of alpha for ridge regression is 7
The optimal value of alpha for Lasso regression is 0.0001

b) If we double the value for alpha the coefficients values reduce as indicated in the table below. This is because when the alpha value increases the model tends to do underfitting.

	Ridge	
Features	Alpha = 7	Alpha = 14
OverallQual_Excellent	0.070613	0.057242
OverallQual_Very Excellent	0.063893	0.048727
GrLivArea	0.058499	0.046514
2ndFlrSF	0.057919	0.044729
Neighborhood_NoRidge	0.054106	0.030824
TotRmsAbvGrd	0.049153	0.044538
GarageCars	0.048333	0.040311
1stFlrSF	0.042782	0.034491
FullBath	0.040248	0.036643
Neighborhood_NridgHt	0.032972	0.030824

	Lasso	
Features	Alpha = 0.0001	Alpha = 0.0002
GrLivArea	0.296687	0.273382
OverallQual_Very Excellent	0.126305	0.125255
OverallQual_Excellent	0.112040	0.119114
GarageCars	0.065736	0.065924
Neighborhood_NoRidge	0.053513	0.054597
OverallQual_Very Good	0.039895	0.042686
Neighborhood_NridgHt	0.035272	0.036108
BsmtExposure_Gd	0.026243	0.029650
Neighborhood_Crawfor	0.031178	0.029386
BsmtFullBath	0.029294	0.028212

Please see the below graph to see the behaviour of underfitting distinguished from overfitting and balanced fitting. Ideal model shall be balanced as seen in the middle graph.



c) The most important predictor variables (features) after doubling the Alpha is shown below for both Ridge and Lasso regression models.

	Lasso
GrLivArea	0.273382
OverallQual_Very Excellent	0.125255
OverallQual_Excellent	0.119114
GarageCars	0.065924
Neighborhood_NoRidge	0.054597

	Ridge
OverallQual_Excellent	0.057242
OverallQual_Very Excellent	0.048727
Neighborhood_NoRidge	0.046575
GrLivArea	0.046514
2ndFlrSF	0.044729

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?

I will choose Lasso regression because it provides non zero coefficients of the relevant variables and removes the irrelevant variables by nullifying their coefficients. In my approach I have decided to feed in all the predictor variables for model creation because I didn't want to remove any variables to avoid risk of removing those relevant variables with my limited domain knowledge to solve the business problem. I did apply RFE technique to remove the irrelevant variables on a first pass and additionally relied on the Lasso model to finally give me only the relevant dependent variables.

Moreover, the evaluation metrics like mse, rmse, r squared shows that Lasso is better than Ridge.

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.891671	0.906330
1	R2 Score (Test)	0.868984	0.863113
2	RSS (Train)	1.365929	1.181090
3	RSS (Test)	0.717317	0.749462
4	MSE (Train)	0.036684	0.034112
5	MSE (Test)	0.040561	0.041460

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 re-performed the Lasso regression by removing the top 5 predictors. The python implementation is present in the code file uploaded in GitHub.

The updated top 5 predictor variables and its Lasso coefficients are as shown in the table below.

	Lasso
1stFlrSF	0.291342
2ndFlrSF	0.188718
RoofMatl_WdShngl	0.100376
LotArea	0.056454
MasVnrArea	0.049157

Question 4

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

Answer :

a)

Model is robust and generalizable when :

1. It performs well on unseen data
2. It is simple (or not complex)
3. Its coefficients does not change significantly when the training data points changes slightly
4. Hyperparameters can be used to finetune or regularize the model to keep it optimally complex

Regularization using techniques like Lasso and Ridge help to reduce complexity by adding a penalty term to the cost function used by OLS.

Ridge regularization technique allows some bias to get significant decrease in variance thus pushing the model coefficient towards zero (but not equal to zero)

Whereas Lasso technique makes some of the coefficients to zero thus resulting in model selection and hence easy to interpret when the number of coefficients is very large.

b) Accuracy of the model can be maintained by balancing the bias and variance such that the total error can be minimized.

Bias quantifies the accuracy of the model on future test data

Variance refers to the degree of changes to the model with respect to the changes in the training data.

To improve the accuracy of a model we need to reduce both bias and variance because the expected total error of a model is the sum of errors in bias and variance.