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?

The optimal value (alpha) for Ridge regression is 50 with train accuracy of 87% and test accuracy of 86.4%, and for Lasso regression is 0.001 with train accuracy of 93.6% and test accuracy of 65.5%.

Metric	Linear Regression	Ridge Regression	Lasso Regression
R2 Score (Train)	9.496532e-01	0.870067	0.936221
R2 Score (Test)	-1.041645e+18	0.864195	0.655466
RSS (Train)	5.140408e+01	132.661862	65.118276
RSS (Test)	4.708995e+20	61.393567	155.754475
MSE (Train)	2.243809e-01	0.360463	0.252545
MSE (Test)	1.035695e+09	0.373963	0.595646

When the alpha value is doubled for both Ridge and Lasso regression from 50 to 100 and 0.001 to 0.002 respectively, we get the below output. In the table we notice that the test accuracy for Lasso regression has a significant improvement of 10.1 points. The train accuracy however, saw a slight drop from 93.6% to 92%. There has been slight drop in scores for both train and test scores for Ridge regression, however, they are both within the acceptable range.

Metric	Linear Regression	Ridge Regression	Lasso Regression
R2 Score (Train)	9.496532e-01	0.855908	0.920600
R2 Score (Test)	-1.041645e+18	0.858775	0.756291
RSS (Train)	5.140408e+01	147.118168	81.067069
RSS (Test)	4.708995e+20	63.843959	110.174249
MSE (Train)	2.243809e-01	0.379595	0.281779
MSE (Test)	1.035695e+09	0.381353	0.500965

After doubling the alpha values, below are top 5 features based on Lasso regression

PoolQC_Gd	Pool quality - Good	-4.702317
Condition2_PosN	Condition2 - Near positive off-site feature--park, greenbelt, etc.	-2.521328
Neighborhood_NoRidge	Neighborhood - Northridge	0.423450
RoofMatl_WdShngl	Roof material - Wood Shingles	0.397797
GrLivArea	Above grade (ground) living area square feet	0.343817

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?

The initial alpha value that was calculated by the script for ridge was 50 and for Lasso was 0.001, and they give the below scores.

Metric	Linear Regression	Ridge Regression	Lasso Regression
R2 Score (Train)	9.496532e-01	0.870067	0.936221
R2 Score (Test)	-1.041645e+18	0.864195	0.655466
RSS (Train)	5.140408e+01	132.661862	65.118276
RSS (Test)	4.708995e+20	61.393567	155.754475
MSE (Train)	2.243809e-01	0.360463	0.252545
MSE (Test)	1.035695e+09	0.373963	0.595646

The alpha value for Lasso regression was very small and if we double the values for both Ridge and for Lasso we see the test accuracy of Lasso improving greatly.

Metric	Linear Regression	Ridge Regression	Lasso Regression
R2 Score (Train)	9.496532e-01	0.855908	0.920600
R2 Score (Test)	-1.041645e+18	0.858775	0.756291
RSS (Train)	5.140408e+01	147.118168	81.067069
RSS (Test)	4.708995e+20	63.843959	110.174249
MSE (Train)	2.243809e-01	0.379595	0.281779
MSE (Test)	1.035695e+09	0.381353	0.500965

When alpha is set to 50 for Ridge regression and 0.004 for Lasso regression, we get the below.

Metric	Linear Regression	Ridge Regression	Lasso Regression
R2 Score (Train)	9.496532e-01	0.870067	0.875084
R2 Score (Test)	-1.041645e+18	0.864195	0.847576
RSS (Train)	5.140408e+01	132.661862	127.539066
RSS (Test)	4.708995e+20	61.393567	68.906665
MSE (Train)	2.243809e-01	0.360463	0.353434
MSE (Test)	1.035695e+09	0.373963	0.396185

The business objective here is to identify the predictors that are used by the model and in our data, we had 245 predictors after creating dummies for all the columns. Ridge regression would make the co-efficient values of the predictors very close to 0 but not 0, because of which, the final model for Ridge has all the predictors. Lasso regression, on the other hand, moves the coefficients to 0 and thus making predictor selection very easy. When the alpha value for Lasso was set to 0.003, we had 69 predictor variables that the important and these can be used to make business decisions.

If we fine tune Lasso regression, we can get better accuracy along with a smaller set of predictors, and since this is what the business needs, I would be using this model.

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?

Below were the top 5 predictors for Lasso regression with alphas = 0.004

PoolQC_Gd	-1.069250
Neighborhood_NoRidge	0.450275
GrLivArea	0.307067
Neighborhood_NridgHt	0.291969
BsmtQual_Gd	-0.237376

After removing the above predictors from the model, we get the below new predictors for the model

2ndFlrSF	Second floor square feet	0.31598220
KitchenQual_Gd	Kitchen quality - Good	-0.28674930
OverallQual	Rates the overall material and finish of the house	0.25196500
1stFlrSF	First Floor square feet	0.23151170
KitchenQual_TA	Kitchen quality - Typical/Average	-0.22963490

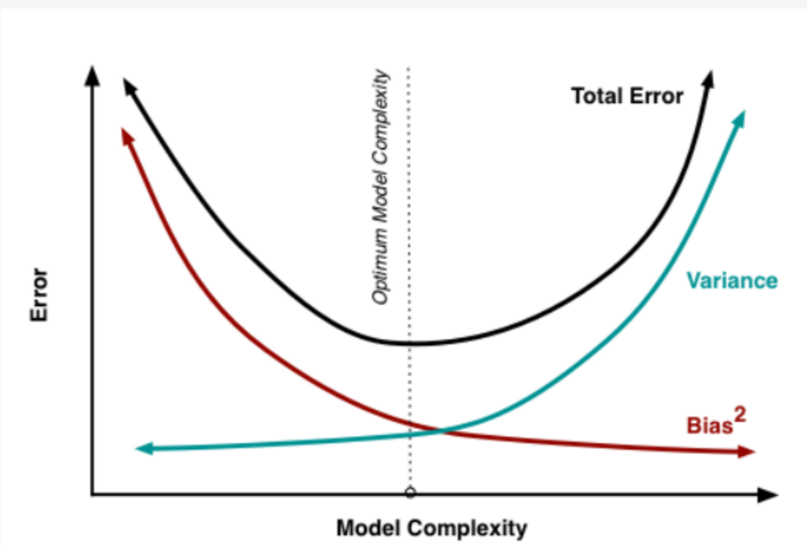
Question 4

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?

When creating a model, we need to check the train accuracy as and while making sure that this is high, we should not compromise on the test accuracy. Creating a model with high train accuracy and a low test accuracy means that the model has learnt the training data and so it will predict correctly within the training set, but since it has not seen the test data, it will perform poorly. Such models when deployed, would not work correctly in the real-world data.

In order to avoid such scenarios, we need to make the model robust and generalized so that it performs correctly on the test data as well. A case where the gap between train and test accuracy is high is called overfitting and this could be checked by plotting the residuals, which is the gap between the actual values and the predicted values.

There are many ways in which overfitting can be removed, and this could be done by restricting the predictor variables or by using either Ridge or Lasso regression. Ridge and Lasso regression adds Lambda to the equation which helps in reducing overfitting.



In the above figure we how the variance and bias changes as the model complexity changes. Higher the model complexity, higher is the variance and lower is the bias. Regularization techniques like Ridge and Lasso adds in a Lambda value such that as we increase the lambda value, the complexity reduces and so the variance reduces with a slight increase in bias. This lambda value should not be kept too high as it might lead to underfitting.