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:

1) **Optimal Value of Alpha:**

   The optimal value of alpha for Ridge Regression: *100*

   The optimal value of alpha for Lasso Regression: *0.001*

2) **The Changes in the model if you choose double the value of alpha for both ridge and lasso regression:**

   For Ridge Regression:

   - ***Original Model for Ridge Regression (alpha*=100)**

     Ridge Regression r2_score for train data: 0.9337403868472463
     Ridge Regression r2_score for test data: 0.9211596109707338
     RMSE: 0.11512462314745651

   - ***Doubled Alpha Model for Ridge Regression (alpha*=200)**

     Ridge Regression r2_score for train data: 0.9284220779749317
     Ridge Regression r2_score for test data: 0.9195075399474633
     RMSE: 0.1163245664717595

   Observation:

   - The test r2 score of the ridge regression model (alpha=100) is slightly higher in comparison to the test r2 score of the doubled alpha model (doubled alpha=200).
   - MSE test score is slightly smaller for the single alpha model than the doubled alpha model.
   - Ridge Regression Original model seems to perform better on the train and test data in comparison to the doubled alpha Ridge Regression model.

   From above results, the original (single) alpha model is a better choice.

   For Lasso Regression:

   - ***Original Model for Lasso Regression (alpha*=0.001)**

     Lasso Regression r2_score for train data: 0.938625005282869
     Lasso Regression r2_score for test data: 0.9234354892050745
     RMSE: 0.11345080929836424

   - ***Doubled Alpha Lasso for Lasso Regression (alpha*=0.002)**

     Lasso Regression r2_score for train data: 0.9323464630311609
     Lasso Regression r2_score for test data: 0.9219088018253238
     RMSE: 0.11457632425909463

   Observation:

   - The test r2 score of the Lasso regression model (alpha=0.001) is slightly higher in comparison to the test r2 score of the doubled alpha model (doubled alpha=0.002).

- MSE test score is slightly smaller for the single alpha model than the doubled alpha model.
- Lasso Regression Original model seems to perform better on the train and test data in comparison to the doubled alpha Lasso Regression model.

From above results, the original (single) alpha model is a better choice.

3) The most important predictor variables after the change is implemented. Top 10 features are as follows:

Top features for *Ridge Regression (The* doubled alpha model - *alpha*=200):
['GrLivArea', 'OverallQual', 'TotalBsmtSF', 'OverallCond', 'GarageArea', 'BsmtFinSF1', 'HouseAge', 'SaleCondition_Normal', 'LotArea', 'FullBath']

Top features for *Lasso Regression (The* doubled alpha model - *alpha*=0.002):
['GrLivArea', 'OverallQual', 'HouseAge','OverallCond', 'TotalBsmtSF', 'SaleType_New', 'SaleCondition_Normal', 'GarageArea', '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:

Optimal Value of Lambda (Alpha):

The optimal value of alpha for Ridge Regression: *100*

The optimal value of alpha for Lasso Regression: *0.001*

- ***Original Model for Ridge Regression (alpha*=100)**
  Ridge Regression r2_score for train data: 0.9337403868472463
  Ridge Regression r2_score for test data: 0.9211596109707338
  RMSE: 0.11512462314745651
- ***Original Model for Lasso Regression (alpha*=0.001)**
  Lasso Regression r2_score for train data: 0.938625005282869
  Lasso Regression r2_score for test data: 0.9234354892050745
  RMSE: 0.11345080929836424

From above observations, **Lasso Regression** model would be better choice with optimal alpha =0.001 as the R2_score of train data and test data are slightly greater than the R2_score of R2_ score of train data and test data of Ridge Regression model with optimal alpha=100. This mean s Lasso Regression Model is performing slightly better than Ridge Regression model on unseen data.
Also RMSE value of Lasso Regression model of test data is slightly less than RSME of Ridge R egression model on test data. This means Lasso Regression Model is performing better on test data. Lasso Regression also helps in feature elimination and the model will be more robust.

Question-3:

After building the model, you realized 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:

After creating another model excluding the five most important predictors from Original Lasso Regression Model, the five most important predictor variables now are:

['TotalBsmtSF', '2ndFlrSF', 'BsmtCond_None', 'OverallCond', 'GarageArea'].

Question-4:

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

Answer:

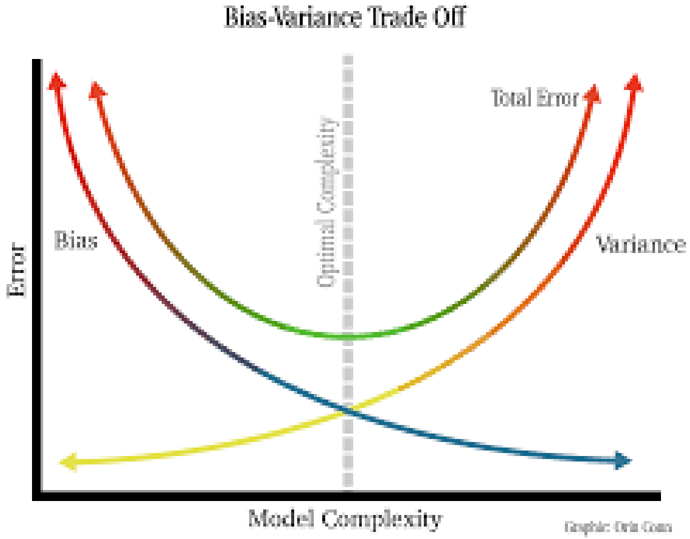
A model should be robust, so that the model the model performs well with enough stability and is not impacted by outliers in the dataset. A model should be generalizable, so that the test accuracy is less than the training accuracy. Thus, the robustness (or generalizability) of a model is a measure of its successful application to data sets other than the one used for training and testing.

By the implementing regularization techniques, we can control the trade-off between model complexity and bias which is directly connected the robustness of the model. Regularization, helps in penalizing the coefficients for making the model too complex; thereby allowing only the optimal amount of complexity to the model. It helps in controlling the robustness of the model by making the model optimal simpler. Therefore, in order to make the model more robust and generalizable, one need to make sure that there is a delicate balance between keeping the model simple and not making it too naive to be of any use.

Making a model simple leads to Bias Variance Trade-off:

A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.

A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.



Graphic: Orin Conn

Bias helps you quantify, how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there has to be enough training data. Models that are too naïve, for e.g., one that gives same results for all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high. Variance is the degree of changes in the model itself with respect to changes in the training data.

Thus, accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the above graph.