ASSIGNMENT 2

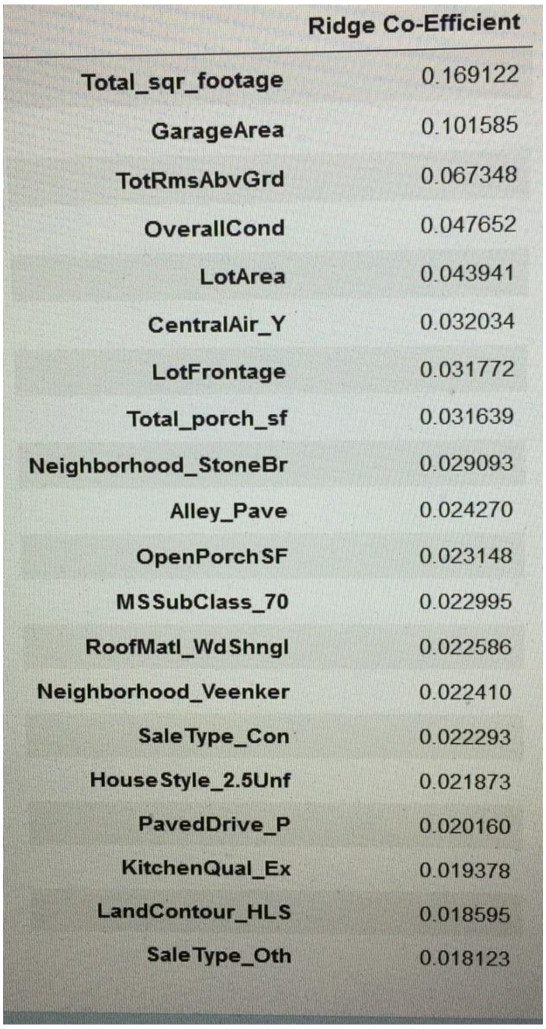
Question 1

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?

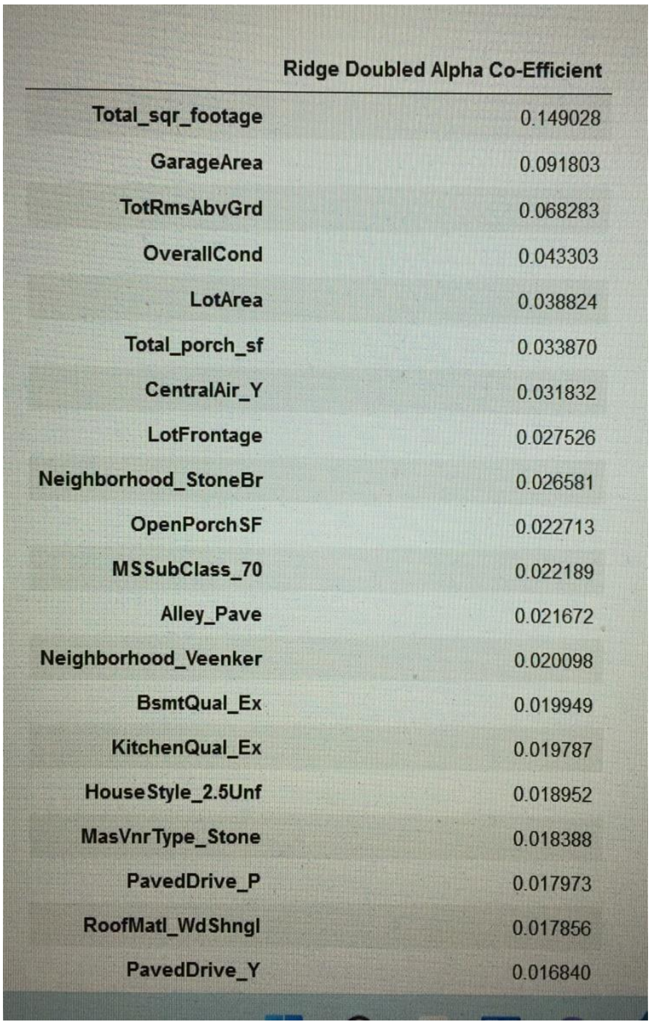
SOLUTION-:

The optimal value of alpha for Ridge is 2 and R2 Score of the model is 0.82811
The optimal value of alpha for Lasso is 0.0001 and R2 Score of the model is 0.825894

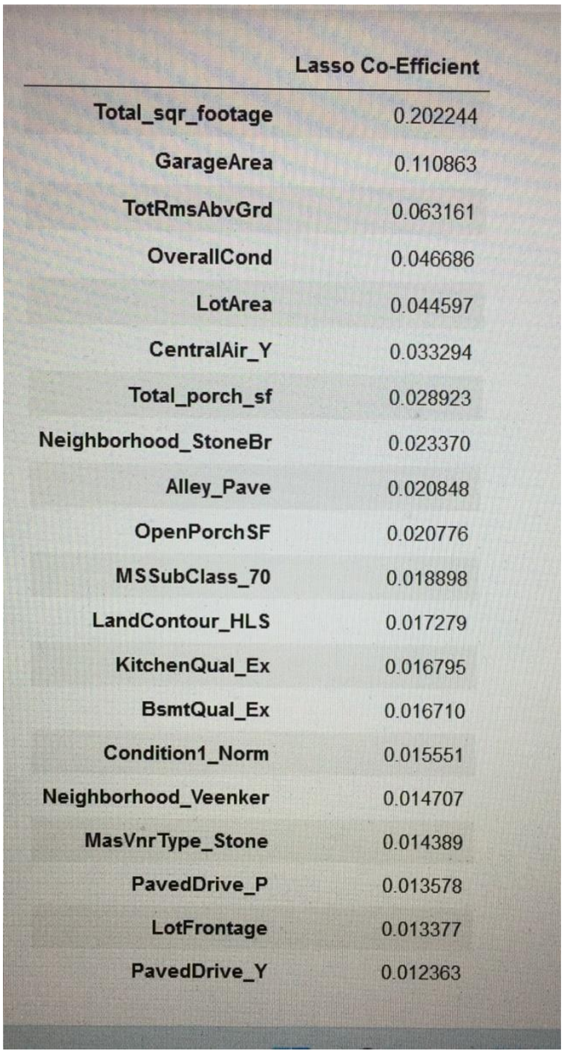
After doubling the alpha values in the Ridge and Lasso, the prediction accuracy for Ridge is 0.82599 and for lasso is 0.823779 respectively. But there is small change in the co-efficient values. The new model is created and demonstrated in the Jupiter notebook. Below are the changes in the co-efficient.



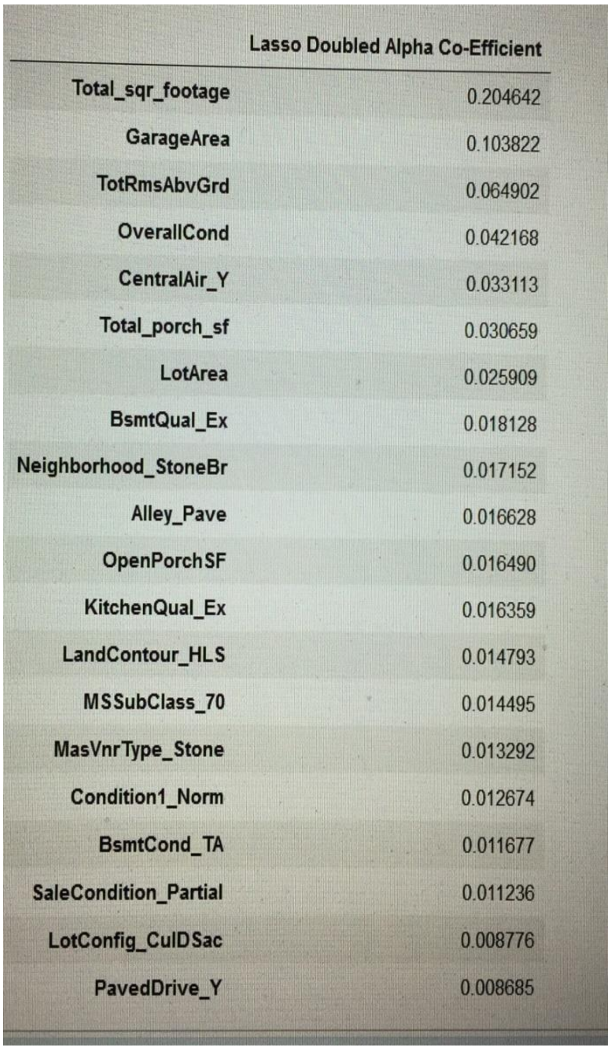
	Ridge Co-Efficient
Total_sqr_footage	0.169122
GarageArea	0.101585
TotRmsAbvGrd	0.067348
OverallCond	0.047652
LotArea	0.043941
CentralAir_Y	0.032034
LotFrontage	0.031772
Total_porch_sf	0.031639
Neighborhood_StoneBr	0.029093
Alley_Pave	0.024270
OpenPorchSF	0.023148
MSSubClass_70	0.022995
RoofMatl_WdShngl	0.022586
Neighborhood_Veenker	0.022410
SaleType_Con	0.022293
HouseStyle_2.5Unf	0.021873
PavedDrive_P	0.020160
KitchenQual_Ex	0.019378
LandContour_HLS	0.018595
SaleType_Oth	0.018123



	Ridge Doubled Alpha Co-Efficient
Total_sqr_footage	0.149028
GarageArea	0.091803
TotRmsAbvGrd	0.068283
OverallCond	0.043303
LotArea	0.038824
Total_porch_sf	0.033870
CentralAir_Y	0.031832
LotFrontage	0.027526
Neighborhood_StoneBr	0.026581
OpenPorchSF	0.022713
MSSubClass_70	0.022189
Alley_Pave	0.021672
Neighborhood_Veenker	0.020098
BsmtQual_Ex	0.019949
KitchenQual_Ex	0.019787
HouseStyle_2.5Unf	0.018952
MasVnrType_Stone	0.018388
PavedDrive_P	0.017973
RoofMatl_WdShngl	0.017856
PavedDrive_Y	0.016840



	Lasso Co-Efficient
Total_sqr_footage	0.202244
GarageArea	0.110863
TotRmsAbvGrd	0.063161
OverallCond	0.046686
LotArea	0.044597
CentralAir_Y	0.033294
Total_porch_sf	0.028923
Neighborhood_StoneBr	0.023370
Alley_Pave	0.020848
OpenPorchSF	0.020776
MSSubClass_70	0.018898
LandContour_HLS	0.017279
KitchenQual_Ex	0.016795
BsmtQual_Ex	0.016710
Condition1_Norm	0.015551
Neighborhood_Veenker	0.014707
MasVnrType_Stone	0.014389
PavedDrive_P	0.013578
LotFrontage	0.013377
PavedDrive_Y	0.012363



	Lasso Doubled Alpha Co-Efficient
Total_sqr_footage	0.204642
GarageArea	0.103822
TotRmsAbvGrd	0.064902
OverallCond	0.042168
CentralAir_Y	0.033113
Total_porch_sf	0.030659
LotArea	0.025909
BsmtQual_Ex	0.018128
Neighborhood_StoneBr	0.017152
Alley_Pave	0.016628
OpenPorchSF	0.016490
KitchenQual_Ex	0.016359
LandContour_HLS	0.014793
MSSubClass_70	0.014495
MasVnrType_Stone	0.013292
Condition1_Norm	0.012674
BsmtCond_TA	0.011677
SaleCondition_Partial	0.011236
LotConfig_CulDSac	0.008776
PavedDrive_Y	0.008685

Overall since the alpha values are small, we do not see a huge change in the model after doubling the alpha.

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?

SOLUTION-:

▪ The optimum lambda value in case of Ridge and Lasso is as follows:-

- Ridge – 2
- Lasso – 0.0001

▪ The Mean Squared Error in case of Ridge and Lasso are:

- Ridge - 0.0018396090787924262
- Lasso - 0.0018634152629407766

- The Mean Squared Error of both the models are almost same.
- Since Lasso helps in feature reduction (as the coefficient value of some of the features become zero), Lasso has a better edge over Ridge and should be used as the final 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?

SOLUTION-:

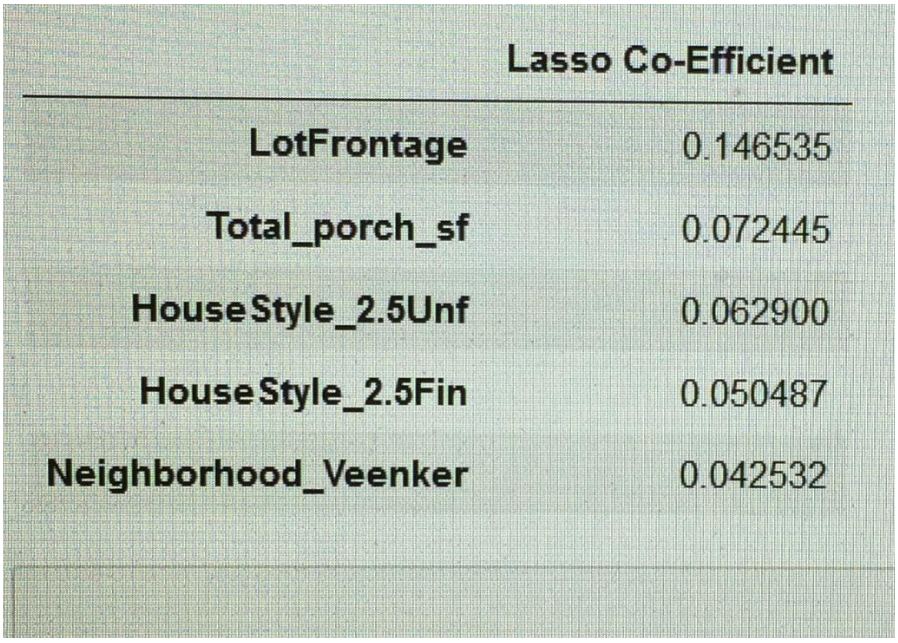
Ans: The five most important predictor variables in the current lasso model is:-
1. Total_sqr_footage
2. GarageArea
3. TotRmsAbvGrd
4. OverallCond
5. LotArea

We build a Lasso model in the Jupiter notebook after removing these attributes from the dataset.

The R2 of the new model without the top 5 predictors drops to 0.733007796426846

The Mean Squared Error increases to 0.0028575670906482546

The new Top 5 predictos are:-



	Lasso Co-Efficient
LotFrontage	0.146535
Total_porch_sf	0.072445
HouseStyle_2.5Unf	0.062900
HouseStyle_2.5Fin	0.050487
Neighborhood_Veenker	0.042532

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?

SOLUTION-:

1. A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data.

2. The model should also be generalisable so that the test accuracy is not lesser than the training score.

3. The model should be accurate for data sets other than the ones which were used during training.

4. Too much weightage should not given to the outliers so that the accuracy predicted by the model is high.

5. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained.

6. Those outliers which it does not make sense to keep must be removed from the dataset.

7. This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used (typically 3-5 standard deviations).

8. This would help standardize the predictions made by the model. If the model is not robust , it cannot be trusted for predictive analysis

Also, Making a model simple leads to Bias-Variance Trade-off

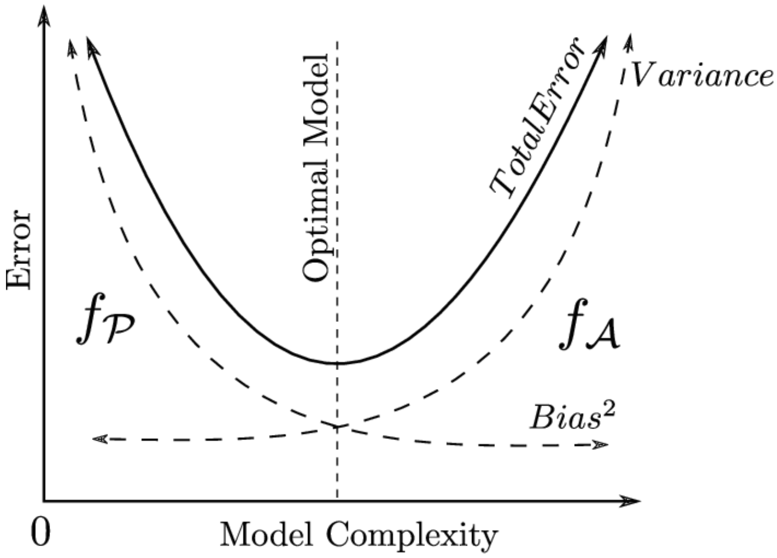
- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for

e.g.,-:one that gives same answer to 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 refers to 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 below graph**.**


Error
Optimal Model
$Total Error$
$Variance$
$f_{\mathcal{P}}$
$f_{\mathcal{A}}$
$Bias^2$
0
Model Complexity