Question 1

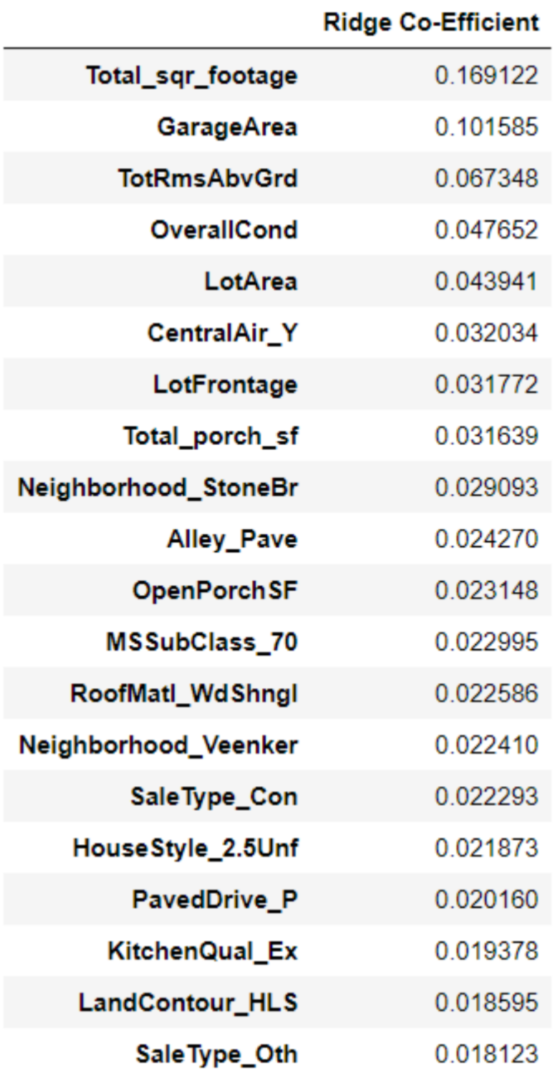
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans: In the Jupiter notebook, two regularization techniques were applied to a predictive model: Ridge and Lasso. Initially, the optimal alpha values for Ridge and Lasso were determined to be 2 and 0.001, respectively, resulting in an R-squared value of approximately 0.83, indicating a good fit to the data.

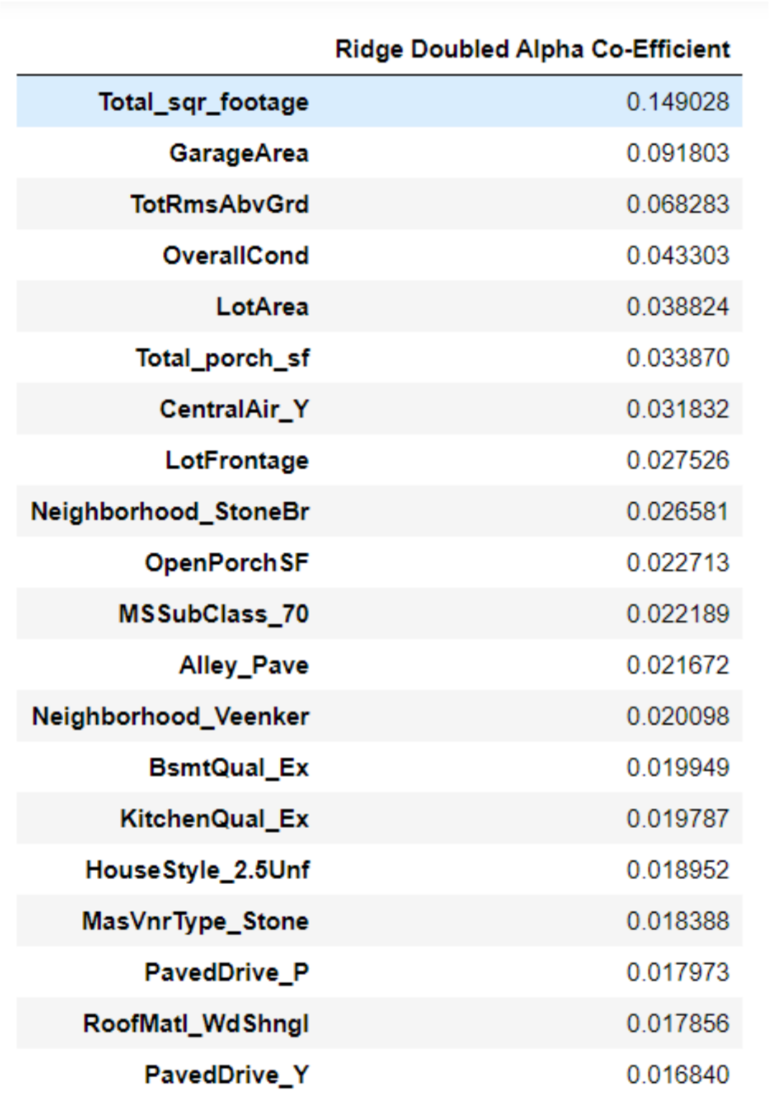
Subsequently, the alpha values for both Ridge and Lasso were doubled, but the prediction accuracy remained relatively stable at around 0.82. Despite the consistent prediction accuracy, there were minor changes observed in the coefficient values. The updated model with doubled alpha values was presented and analyzed in the notebook.

.

Ridge Regression Model

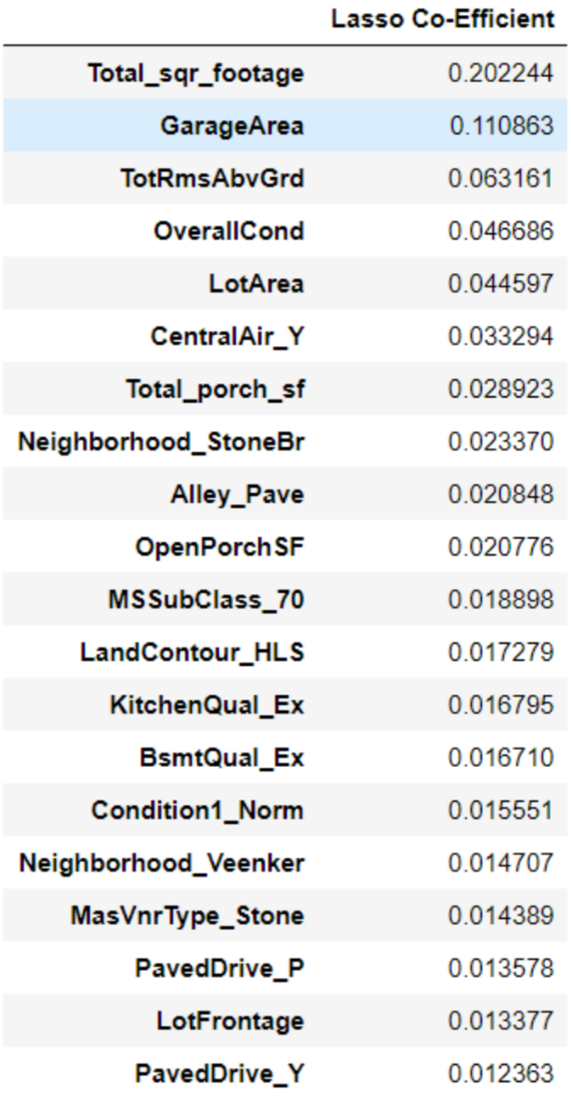


	Ridge Co-Efficient
Total_sqr_footage	0.169122
GarageArea	0.101585
TotRmsAbvGrd	0.067348
OverallCond	0.047652
LotArea	0.043941
CentralAir_Y	0.032034
LotFrontage	0.031772
Total_porch_sf	0.031639
Neighborhood_StoneBr	0.029093
Alley_Pave	0.024270
OpenPorchSF	0.023148
MSSubClass_70	0.022995
RoofMatl_WdShngl	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

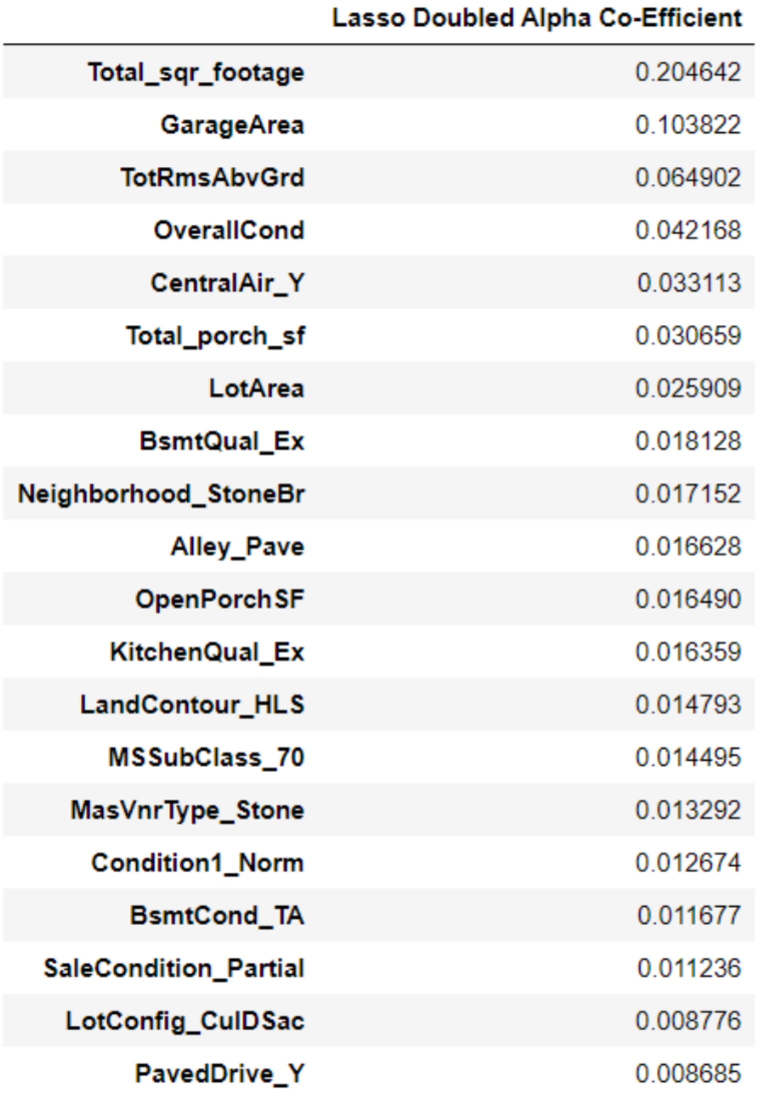


	Ridge Doubled Alpha Co-Efficient
Total_sqr_footage	0.149028
GarageArea	0.091803
TotRmsAbvGrd	0.068283
OverallCond	0.043303
LotArea	0.038824
Total_porch_sf	0.033870
CentralAir_Y	0.031832
LotFrontage	0.027526
Neighborhood_StoneBr	0.026581
OpenPorchSF	0.022713
MSSubClass_70	0.022189
Alley_Pave	0.021672
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 Regression Mode



	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?

Ans : The optimal lambda values for Ridge and Lasso regularization techniques are found to be 2 and 0.0001, respectively. Upon evaluation, the Mean Squared Error for Ridge is approximately 0.0018396, while for Lasso, it is approximately 0.0018634. Both models exhibit very similar performance in terms of Mean Squared Error.

However, Lasso stands out due to its additional capability of feature reduction. As Lasso can drive some feature coefficients to exactly zero, it effectively performs feature selection, leading to a more interpretable and potentially simpler model. Given this advantage, Lasso is recommended as the final model choice over Ridge, especially when interpretability and feature importance are important considerations.

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: The five most important predictor variables in the current lasso model is:-

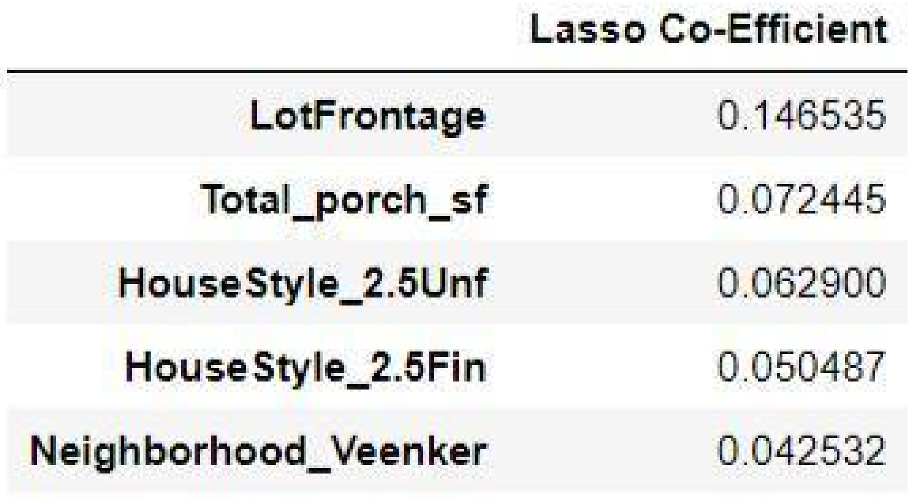
- Total_sqr_footage
- GarageArea
- TotRmsAbvGrd
- OverallCond
- LotArea

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 .73

The Mean Squared Error increases to 0.002857567090648253 8

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?

As per Occam's Razor, when two models demonstrate comparable performance on finite training or test data, we should select the one with fewer assumptions or complexities. This preference is driven by the notion that simpler models are more favorable for several reasons:

- Simpler models are generally more universal and have broader applicability across different scenarios.
- Training simpler models requires a smaller number of samples for effective performance compared to more complex models, making them easier to train.
- Simpler models often exhibit greater robustness, as they are less sensitive to variations in the training data.
- Complex models tend to exhibit highly volatile changes in response to variations in the training dataset.
- Simple models typically have low variance and high bias, while complex models have low bias and high variance.
- Simpler models may make more errors on the training set, whereas complex models are prone to overfitting—performing exceptionally well on the training data but poorly on other unseen test samples.

Hence, it is essential to maintain a level of simplicity in the model without making it overly simplistic, as that would render it ineffective.

To achieve this balance, regularization techniques can be employed. Regularization plays a crucial role in controlling the complexity of the model, preventing it from becoming too naive or overly complex. In the context of regression, regularization entails incorporating an additional term to the cost function that accounts for the absolute values or squares of the model's parameters. This helps in achieving a well-balanced model that is both interpretable and capable of generalizing effectively to new data.

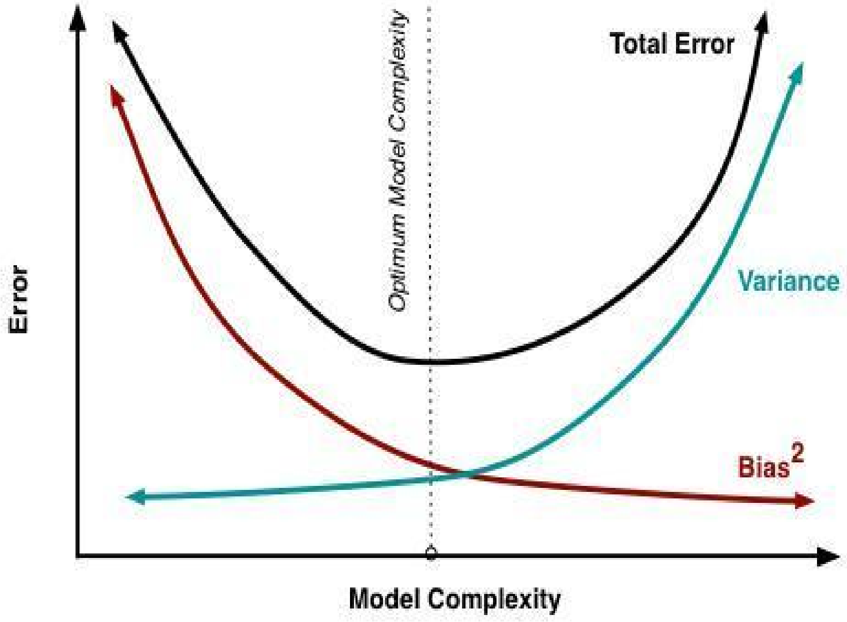
Also,Simplifying a model results in a trade-off between bias and variance.

- A complex model exhibits high instability and extreme sensitivity to even minor changes in the dataset, necessitating frequent adjustments as the training data varies.
- A simpler model, which captures the underlying patterns followed by the data points, is less likely to undergo drastic changes even when additional points are added or some are removed from the dataset..

"Bias quantifies the model's likelihood of accurately predicting test data. A complex model, given sufficient training data, can achieve accurate predictions. However, overly simplistic models, such as those that consistently yield the same output for all test inputs without discrimination, exhibit high bias, leading to significant expected errors across all test inputs.

Variance, on the other hand, describes how much the model's predictions vary concerning changes in the training data.

Maintaining an optimal balance between Bias and Variance is crucial for preserving the model's accuracy. The goal is to minimize the total error, as illustrated in the graph below."


Optimum Model Complexity
Total Error
Variance
Bias^2
Error
Model Complexity