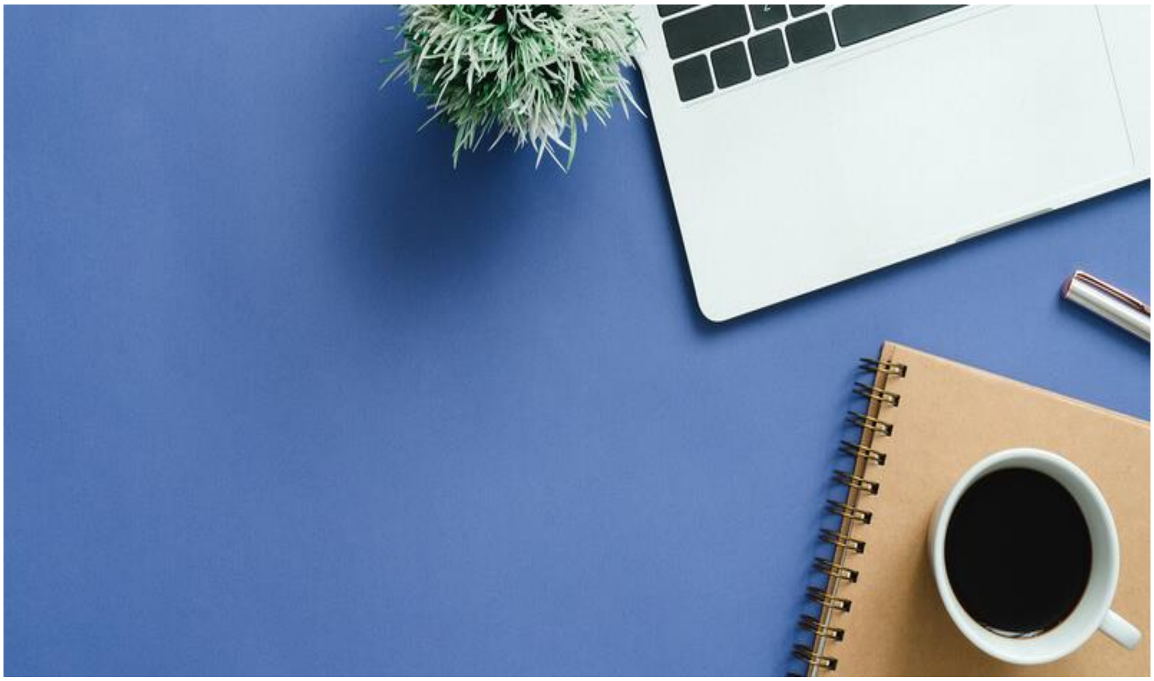
TABLE OF CONTENTS



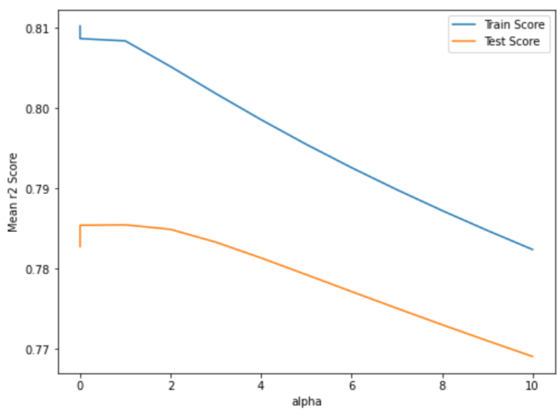
ASSIGNMENT-BASED SUBJECTIVE

Que 1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans. We choose the following optimal values of alpha by plotting the mean train score and mean test score vs alpha value.

The optimal value of alpha for **Ridge Regression** is: 1.

The optimal value of alpha for **Lasso Regression** is: 0.0005



Then we build respective models and evaluated their performance and found the coefficients of most important predictors.

Then, **doubled the value of alpha** for both Ridge and Lasso Models and observed the following changes in performance of models.

	Metric	Linear Regression	RFE Linear Regression	Ridge Regression	Ridge With Double Alpha Regression	Lasso Regression	Ridge With Double Alpha Regression	Lasso With Double Alpha Regression
0	R2 Score (Train)	0.871269	0.846188	0.844983	0.842913	0.829715	0.842913	0.812117
1	RSS (Train)	2.495047	2.981155	3.004516	3.044623	3.300430	3.044623	3.641510
2	MSE (Train)	0.046358	0.050673	0.050871	0.051210	0.053317	0.051210	0.056005
3	R2 Score (Test)	0.870260	0.852937	0.854958	0.853623	0.840927	0.853623	0.818300
4	RSS (Test)	0.689940	0.782060	0.771315	0.778413	0.845929	0.778413	0.966256
5	MSE (Test)	0.048692	0.051841	0.051484	0.051720	0.053916	0.051720	0.057623

- Comparative value of R2 score for Train as well as test data decreased slightly, on doubling the value of Alpha.
- Decrease in R2 score is comparatively more in case of Lasso.
- Comparative value of RSS and MSE increased for Train as well as test data, on doubling the value of Alpha.

Most Important Predictors after Changes:

- Top 10 predictors are same, but coefficients have changed little.
- For Ridge, Coefficients reduced for all predictors.
- For Lasso, Number of Predictors with coefficient value 0, increased.

The below mentioned variables are Top 10 significant in predicting the price

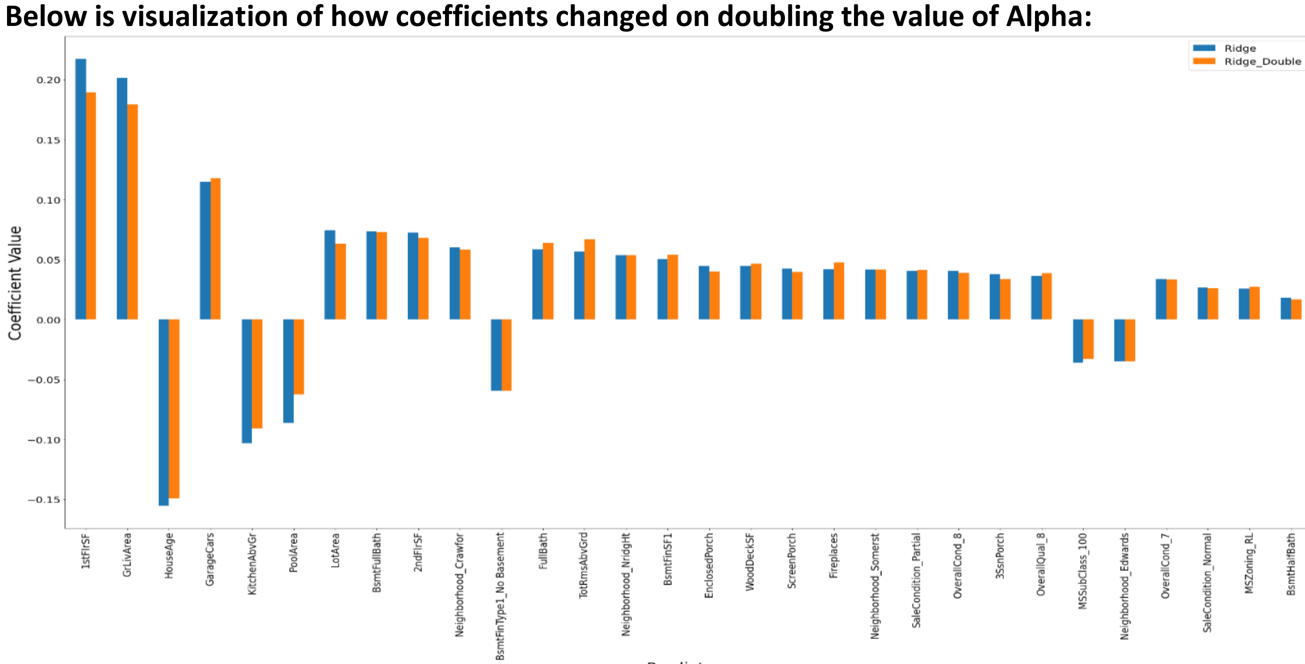
Postive Features

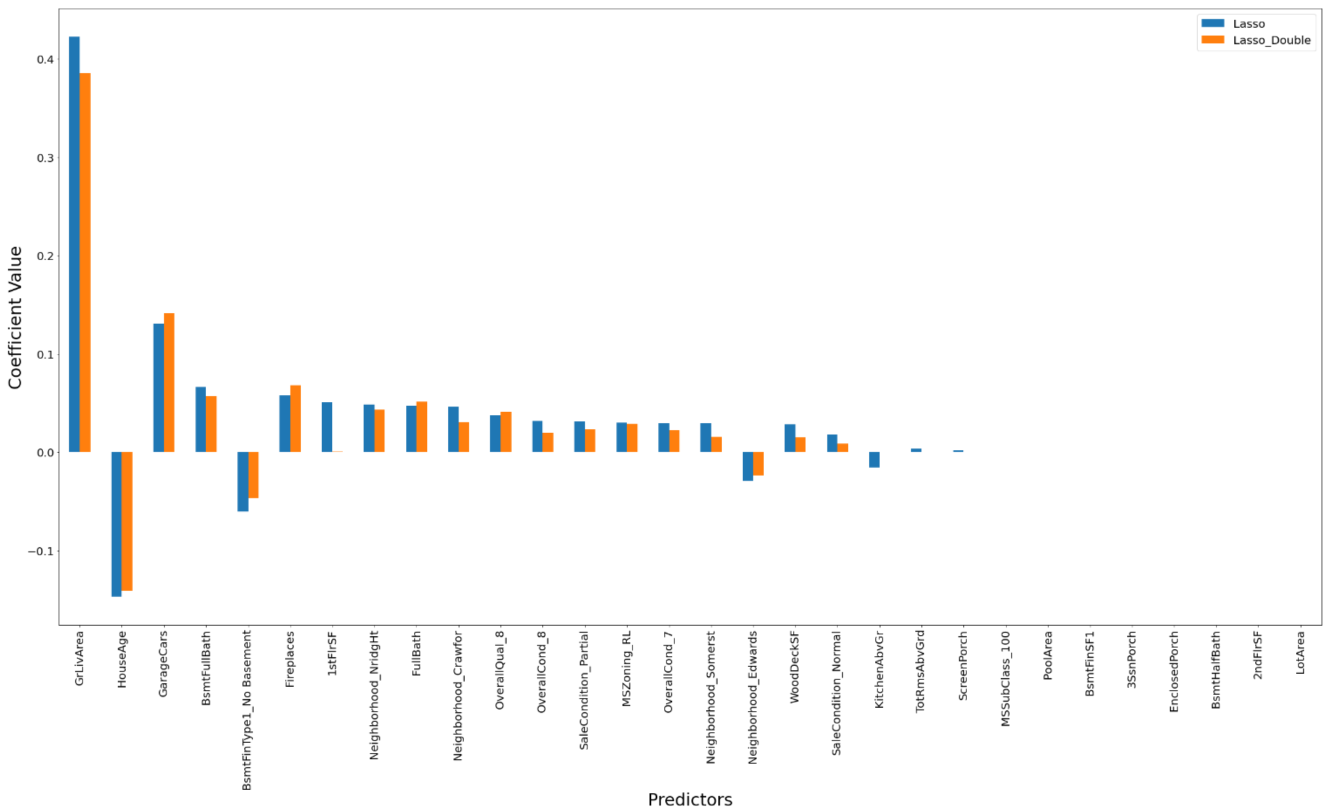
- GrLivArea : Above grade (ground) living area square feet
- GarageCars : Size of garage in car capacity
- BsmtFullBath : Basement full bathrooms
- Fireplaces : Number of fireplaces
- KitchenAbvGr : Kitchens above grade (Negative Relation)
- 1stFlrSF : First Floor square feet
- Neighborhood_NridgHt : Physical locations within Ames city limits: Northridge Heights
- FullBath : Full bathrooms above grade
- Neighborhood_Crawfor : Physical locations within Ames city limits: Crawford

Negative Features

- HouseAge : Difference of Year when house is sold and the year it was built.(Negative Relation)
- BsmtFinType1_No Base : Rating of basement finished area: No Basement (Negative Relation)

Below is visualization of how coefficients changed on doubling the value of Alpha:





Que 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 For respective optimal values of alpha let's observe how both models perform.

How well different Models describe the price of a house.

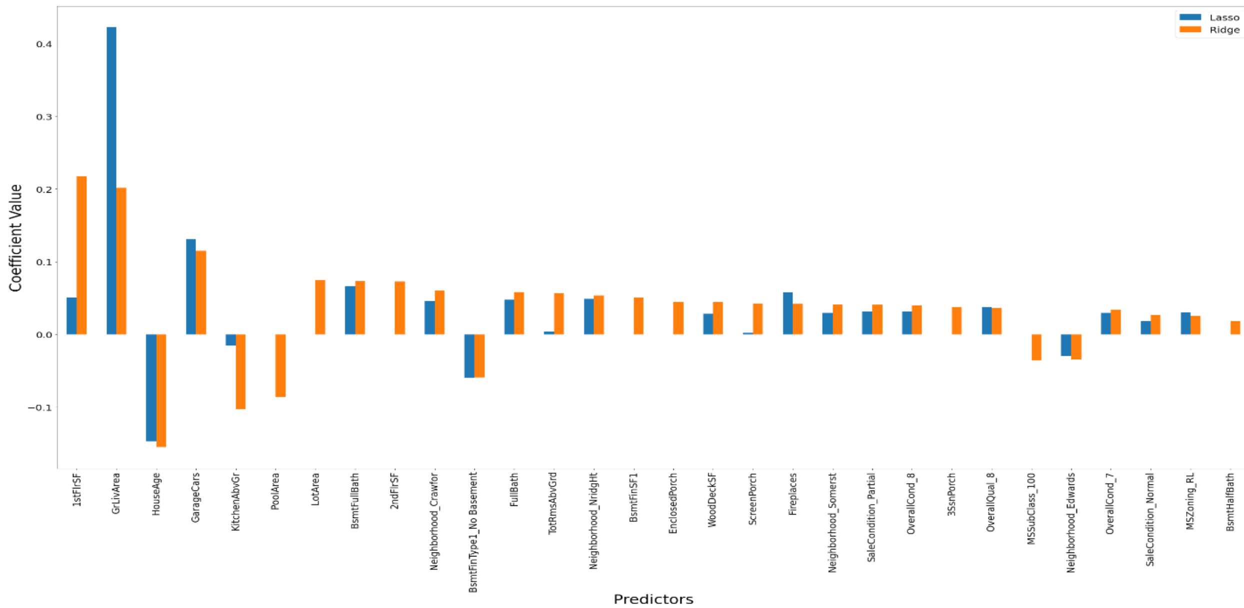
	Metric	Linear Regression	RFE Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.871269	0.846188	0.844983	0.829715
1	RSS (Train)	2.495047	2.981155	3.004516	3.300430
2	MSE (Train)	0.046358	0.050673	0.050871	0.053317
3	R2 Score (Test)	0.870260	0.852937	0.854958	0.840927
4	RSS (Test)	0.689940	0.782060	0.771315	0.845929
5	MSE (Test)	0.048692	0.051841	0.051484	0.053916

We can observe from above table that, R2 score for Lasso Regression is lesser than Ridge for test as well as Train data. Also, MSE and RSS increased a bit for Lasso as compared to Ridge.

But we will choose **Lasso Regularized Model** over Ridge here, because of following reasons:

- **Reducing Features** when we have LOTS of PARAMETERS in the model: In **Ridge**, when we increase the value of LAMBDA, the most important parameters might shrink a little bit and the **less important parameter stay at high value**. Whereas with **LASSO** when we increase the value of LAMBDA the most important parameters shrink a little bit, and the **less important parameters goes closed to ZERO**.

Observe below the Ridge and Lasso Regularization, reduced the coefficients of Predictors.



- **Reduced Multicollinearity**: We observed that for many features chosen by RFE the VIF value was very high. Then we also observed that the LASSO regularization, reduced the coefficient values of most of these predictors to 0, thus eliminated those from our model.

	Features	VIF
4	GrLivArea	861.70
2	1stFlrSF	442.62
3	2ndFlrSF	161.49
9	TotRmsAbvGrd	33.45
8	KitchenAbvGr	23.77
7	FullBath	19.71
11	GarageCars	11.54
28	SaleCondition_Normal	8.49
19	MSZoning_RL	6.80
17	HouseAge	6.01
1	BsmtFinSF1	4.44
5	BsmtFullBath	3.34
10	Fireplaces	2.95
0	LotArea	2.12
29	SaleCondition_Partial	2.12
12	WoodDeckSF	1.90
23	Neighborhood_Somerst	1.49
13	EnclosedPorch	1.45
24	OverallQual_8	1.38
25	OverallCond_7	1.36
22	Neighborhood_NridgHt	1.33
21	Neighborhood_Edwards	1.24
26	OverallCond_8	1.19
6	BsmtHalfBath	1.18

Let's Check Features Eliminated by Lasso

```
FeaturesToEliminate=coef_df[coef_df['abs_value_
FeaturesToEliminate
```

```
['LotArea',
 'BsmtFinSF1',
 '2ndFlrSF',
 'BsmtHalfBath',
 'EnclosedPorch',
 '3SsnPorch',
 'PoolArea',
 'MSSubClass_100']
```

Que 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?

Ans. We dropped the 5 Topmost predictors from incoming data and build the model again with Lasso with the same alpha value. And observed the following changes. The most important five predictors are now:

	Metric	Lasso Regression	Lasso With Reduced Columns Regression
0	R2 Score (Train)	0.829715	0.765447
1	RSS (Train)	3.300430	4.546056
2	MSE (Train)	0.053317	0.062575
3	R2 Score (Test)	0.840927	0.791167
4	RSS (Test)	0.845929	1.110545
5	MSE (Test)	0.053916	0.061776

	Lasso_Updated
1stFlrSF	0.472614
2ndFlrSF	0.169707
BsmtFinSF1	0.157022
FullBath	0.128565
KitchenAbvGr	-0.115175

R2 score for the new model decreased, also RSS and MSE increased.

Que 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?

Ans.
We may have problems with too complex model that it may really perform well on training data but might not perform well with unseen data. This is **overfitted** model. Such a model will have very low bias, but variance will be too high. So, we need a **Generalizable** model.
Similarly, we may have problems with too simple model that it may not perform well on the training data itself. This is **underfitted** model. Such a model will have very high bias but may show less variance with unseen data. So, we need a **Robust** model also.
To make sure that the model is robust and generalizable, we need to consider **Bias Variance Tradeoff**.

A model can not be less complex and more complex at the same time. So, we need to trade off with complexity of model to keep the balance of biasness and variance such that it minimizes the total error.

We need to find the appropriate balance without underfitting and overfitting the data.

To manage the model complexity a technique called Regularization is used. It helps in managing model complexity by suppressing the model coefficients towards 0. This discourages the model from becoming too complex, thus avoiding the risk of overfitting.

Regularization uses cost function as : Cost = RSS + Penalty.

Adding the penalty term in the cost function helps in shrinking the magnitude of the model coefficients.
To conclude, we use regularization to build a generalizable and robust model to gain the overall high accurate model. For this, we must make a compromise by allowing a little bias for a significant reduction in variance.