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?

For Lasso model the optimal value is 0.0001

n [107]:
```
lasso = Lasso(alpha=0.0001)
lasso.fit(X_train,y_train)

y_train_pred = lasso.predict(X_train)
y_test_pred = lasso.predict(X_test)

print(r2_score(y_true=y_train,y_pred=y_train_pred))
print(r2_score(y_true=y_test,y_pred=y_test_pred))
```

```
0.9253921472169352
0.858440420123422
```

For Ridge model the optimal value is 5

In [104]:
```
ridge = Ridge(alpha = 5)
ridge.fit(X_train,y_train)

y_pred_train = ridge.predict(X_train)
print(r2_score(y_train,y_pred_train))

y_pred_test = ridge.predict(X_test)
print(r2_score(y_test,y_pred_test))
```

```
0.9101202551420328
0.8798104854055782
```

When Alpha is doubled

For Lasso Model the Train and Test R2score is as below:

In [111]:
```
double_alpha=0.0001 *2
lasso = Lasso(alpha=double_alpha)
lasso.fit(X_train,y_train)

y_train_pred = lasso.predict(X_train)
y_test_pred = lasso.predict(X_test)

print(r2_score(y_true=y_train,y_pred=y_train_pred))
print(r2_score(y_true=y_test,y_pred=y_test_pred))
```

```
0.9116670610263857
0.8635040740503076
```

For Ridge Model the Train and Test R2score is as below:

In [112]:
```
double_alpha=5.0 *2
ridge = Ridge(alpha=double_alpha)
ridge.fit(X_train,y_train)

y_train_pred = ridge.predict(X_train)
y_test_pred = ridge.predict(X_test)

print(r2_score(y_true=y_train,y_pred=y_train_pred))
print(r2_score(y_true=y_test,y_pred=y_test_pred))
```

```
0.897053419987246
0.8699693054805355
```

Conclusion

When Alpha is doubled, we find that the R2 score tend to decrease.

Post change there is no change interms of top5 variables though its noticed that coefficients are slightly different.

	Featuere	Coef
12	BsmtFullBath	0.288767
3	OverallCond	0.138575
6	BsmtUnfSF	0.073790
8	1stFlrSF	0.056768
22	GarageArea	0.044258
4	MasVnrArea	0.042642
82	Neighborhood_NridgHt	0.041768
2	OverallQual	0.037886
5	BsmtFinSF1	0.036054
9	2ndFlrSF	0.035631

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?

For Lasso model the optimal value is 0.0001 and for Ridge model the optimal value is 5.

The R2square error is nearly the same. So the conclusion is drawn based on how the Lasso vs Ridge model works.

We need to first understand that our dataset post cleansing has around 200 columns, thus Lasso is more efficient as it helps shrink some coefficients towards zero for both variable reduction and model simplification. Lasso model tends to eliminate variables thus focusing on significant features.

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?

The Top5 features are BsmtFullBath , OverallCond , BsmtUnfSF , 1stFlrSF , GarageArea

Once we drop these

The R2 score has dropped for

- Train Set from 0.9253921472169352 to 0.8486945460009021
- Test Set from 0.858440420123422 to 0.8197883201643478

Also the Top5 features now are:

- BsmtHalfBath - Basement half bathrooms
- MasVnrArea - Masonry veneer area in square feet
- TotalBsmtSF - Total square feet of basement area
- GarageQual - Garage quality
- 2ndFlrSF - Second floor square feet

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?

Robustness of a model implies, either the testing error of the model is consistent with the training error, the model performs well with enough stability even after adding some noise to the dataset. 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.

Bias is one type of error that occurs due to wrong assumptions about data such as assuming data is linear when in reality, data follows a complex function. On the other hand, variance gets introduced with high sensitivity to variations in training data. This also is one type of error since we want to make our model robust against noise. There are two types of error in machine learning. Reducible error and Irreducible error. Bias and Variance come under reducible error.

Bias

Bias is simply defined as the inability of the model because of that there is some difference or error occurring between the model's predicted value and the actual value. These differences between actual or expected values and the predicted values are known as error or bias error or error due to bias.

- **Low Bias:** Low bias value means fewer assumptions are taken to build the target function. In this case, the model will closely match the training dataset.
- **High Bias:** High bias value means more assumptions are taken to build the target function. In this case, the model will not match the training dataset closely.

The high-bias model will not be able to capture the dataset trend. It is considered as the underfitting model which has a high error rate. It is due to a very simplified algorithm.

Ways to reduce high bias in Machine Learning:

- **Use a more complex model:** One of the main reasons for high bias is the very simplified model. it will not be able to capture the complexity of the data. In such cases, we can make our mode more complex by increasing the number of hidden layers in the case of a deep neural network.
- **Increase the number of features:** By adding more features to train the dataset will increase the complexity of the model. And improve its ability to capture the underlying patterns in the data.
- **Reduce Regularization of the model:** Regularization techniques such as L1 or L2 regularization can help to prevent overfitting and improve the generalization ability of the model. if the model has a high bias, reducing the strength of regularization or removing it altogether can help to improve its performance.
- **Increase the size of the training data:** Increasing the size of the training data can help to reduce bias by providing the model with more examples to learn from the dataset.

Variance

Variance is the measure of spread in data from its mean position. In machine learning variance is the amount by which the performance of a predictive model changes when it is trained on different subsets of the training data. More specifically, variance is the variability of the model that how much it is sensitive to another subset of the training dataset. i.e. how much it can adjust on the new subset of the training dataset.

- **Low variance:** Low variance means that the model is less sensitive to changes in the training data and can produce consistent estimates of the target function with different subsets of data from the same distribution. This is the case of underfitting when the model fails to generalize on both training and test data.
- **High variance:** High variance means that the model is very sensitive to changes in the training data and can result in significant changes in the estimate of the target function when trained on different subsets of data from the same distribution. This is the case of overfitting when the model performs well on the training data but poorly on new, unseen test data. It fits the training data too closely that it fails on the new training dataset.

Ways to Reduce the reduce Variance in Machine Learning:

- **Cross-validation:** By splitting the data into training and testing sets multiple times, cross-validation can help identify if a model is overfitting or underfitting and can be used to tune hyperparameters to reduce variance.
- **Feature selection:** By choosing the only relevant feature will decrease the model's complexity. and it can reduce the variance error.
- **Regularization:** We can use L1 or L2 regularization to reduce variance in machine learning models
- **Ensemble methods:** It will combine multiple models to improve generalization performance. Bagging, boosting, and stacking are common ensemble methods that can help reduce variance and improve generalization performance.
- **Simplifying the model:** Reducing the complexity of the model, such as decreasing the number of parameters or layers in a neural network, can also help reduce variance and improve generalization performance.
- **Early stopping:** Early stopping is a technique used to prevent overfitting by stopping the training of the deep learning model when the performance on the validation set stops improving.

Different Combinations of Bias-Variance

There can be four combinations between bias and variance.

- **High Bias, Low Variance:** A model with high bias and low variance is said to be underfitting.
- **High Variance, Low Bias:** A model with high variance and low bias is said to be overfitting.
- **High-Bias, High-Variance:** A model has both high bias and high variance, which means that the model is not able to capture the underlying patterns in the data (high bias) and is also

too sensitive to changes in the training data (high variance). As a result, the model will produce inconsistent and inaccurate predictions on average.

- **Low Bias, Low Variance:** A model that has low bias and low variance means that the model is able to capture the underlying patterns in the data (low bias) and is not too sensitive to changes in the training data (low variance). This is the ideal scenario for a machine learning model, as it is able to generalize well to new, unseen data and produce consistent and accurate predictions. But in practice, it's not possible.

Error (Model) = Variance + Bias + Irreducible Error