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. Optimal value of alpha for Ridge and Lasso is .0001 and 10.0 respectively. Below table shows predictor variable

Changes in model if alpha value is doubles
Increasing alpha makes regularization stricter resulting in reduction in variance, increase in bias, and model becomes simpler. For lasso more and more beta coefficients will be marked as zeros and for Ridge beta coefficient becomes more and more smaller almost near to zero for less significant features.

Lasso: if alpha is doubled from .0001 then below are the changes in the model

1. Top five predictor variable changes and beta coefficient values also changes
2. In the Lasso regression model, the beta coefficients of less significant features are identified as zero. Notably, when the alpha parameter is doubled, there is a corresponding increase in the number of zero beta coefficients.
   at an alpha value of 0.0001, zero beta coefficient count is 77
   at an alpha value of 0.0001, zero beta coefficient count is 108
3. R2 score of training and test has decreased slightly, while RSS, MSE and RMSE values has increased slightly
4. At alpha .01 model becomes to simple to capture relevance. This can be observed in scatter plot

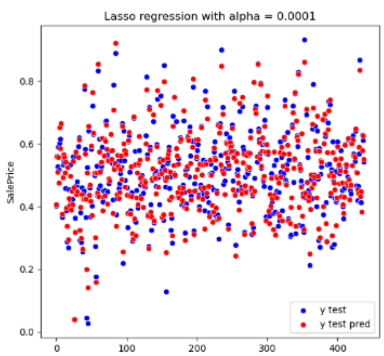
Most Important predictor after change implemented

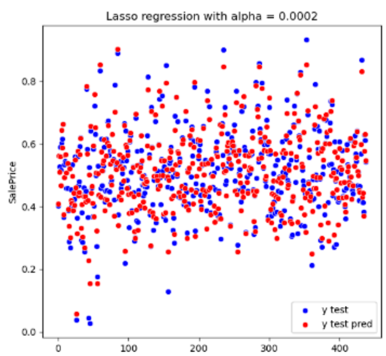
Lasso Alpha=.0001		**Lasso double alpha =.0002**	
Beta	**Predictor Variables**	**Beta**	**Predictor Variables**
0.308287	GrLivArea	0.302146	GrLivArea
0.162438	OverallQual	0.180248	OverallQual
0.094896	OverallCond	0.091497	OverallCond
0.078638	GarageCars	0.078666	GarageCars
0.056504	MSZoning_RL	0.052211	BsmtFullBath

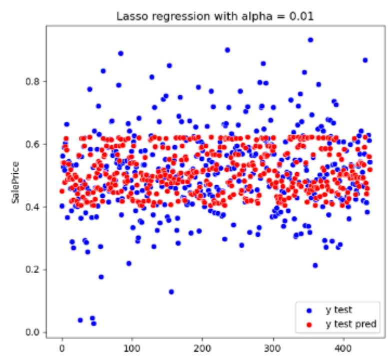
Lasso Alpha=.0001		**Lasso double alpha =.0002**	
Metric	**Values**	**Metric**	**Values**

MSE Test	0.001999	MSE Test	0.002030
MSE Train	0.001383	MSE Train	0.001549
R2_score Test	0.885180	R2_score Test	0.883405
R2_score Train	0.916848	R2_score Train	0.906895
RMSE Test	0.044710	RMSE Test	0.045055
RMSE Train	0.037189	RMSE Train	0.039352
RSS Test	0.875573	RSS Test	0.889112
RSS Train	1.412076	RSS Train	1.581093

Alpha .0001 (best) Alpha= .0002 Alpha= .01







Ridge: if alpha is doubled from the 10 to 20 the below are the changes in model.

1. Complexity of model gets reduced, variance gets decreased
2. Beta coefficient of predictor variable changes and even top five predictor also changes
3. There is not difference in the R2 score ,RSS ,MSE, RMSE but R2 Score decreases and RSS, MSE and RMSE increases

Most Important predictor after change implemented

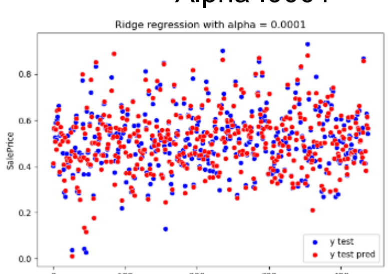
Ridge Alpha= 10		**Ridge double alpha = 20**	
Beta	**Predictor Variables**	**Beta**	**Predictor Variables**
0.087912	OverallQual	0.067063	OverallQual
0.060083	GrLivArea	0.048695	TotRmsAbvGrd
0.055074	TotRmsAbvGrd	0.047204	GrLivArea

0.053297	OverallCond	0.040141	FullBath
0.049739	1stFlrSF	0.039021	OverallCond

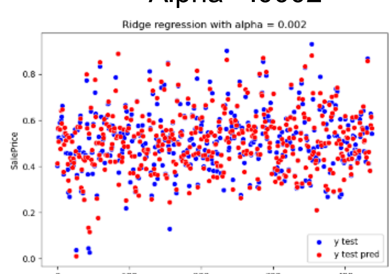
	Ridge alpha 10	Ridge alpha 20
MSE Test	0.002251	0.002476
MSE Train	0.001586	0.001831
R2_score Test	0.870710	0.857778
R2_score Train	0.904645	0.889933
RMSE Test	0.047444	0.049760
RMSE Train	0.039825	0.042787
RSS Test	0.985915	1.084531
RSS Train	1.619305	1.869132

The reason for observing a significant variation in alpha to observe a significant reduction in model performance, could be due to the dataset containing many features, some of which may not be highly relevant. By increasing alpha, Ridge regression reduces the magnitude of coefficients associated with less important features, which can help in reducing model variance without substantially increasing bias. However, it's important to note that while the coefficients for less relevant features are reduced, they are not set to zero.
with increase of alpha r2 score and other parameter also changes, out of these parameters alpha 10 is the best

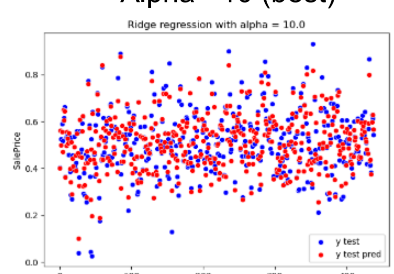
Alpha .0001

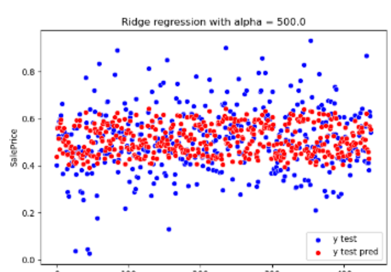


Alpha= .0002



Alpha= 10 (best)





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
Below is the table based on which model is selected
Best alpha value for Lasso is .0001 and Ridge is 10

	Linear Regression	Lasso	Ridge
MSE Test	1.127104e+18	0.001999	0.002251
MSE Train	1.480541e-03	0.001383	0.001586
R2_score Test	-6.473859e+19	0.885180	0.870710
R2_score Train	9.109850e-01	0.916848	0.904645
RMSE Test	1.061652e+09	0.044710	0.047444
RMSE Train	3.847780e-02	0.037189	0.039825
RSS Test	4.936716e+20	0.875573	0.985915
RSS Train	1.511633e+00	1.412076	1.619305

Based on the provided table, the Lasso regression model appears to be the most suitable choice for several reasons:

1. **MSE (Mean Squared Error)**: The Lasso model has the lowest MSE values for both Test (0.001999) and Train (0.001383) datasets. Lower MSE indicates better model performance with fewer errors.

2. **R2 Score**: The R2 score reflects the proportion of the variance in the dependent variable that is predictable from the independent variable(s). For both Test (0.885180) and Train (0.916848) datasets, the Lasso model scores higher than the Ridge model and significantly outperforms the Linear Regression model, indicating a better fit to the data.

3. **RMSE (Root Mean Squared Error)**: Similar to MSE, lower RMSE values indicate better model performance. The Lasso model has lower RMSE values for both Test (0.044710) and Train (0.037189) datasets compared to the Ridge model, suggesting it predicts with lesser error.

4. **RSS (Residual Sum of Squares)**: The Lasso model has a lower RSS for the Test dataset (0.875573) compared to the Ridge model (1.084531), indicating better predictive accuracy. Although the Linear Regression model shows extreme values due to its poor fit, making it an unreliable comparison.

The Lasso model is the best choice due to its lowest MSE and RMSE, highest R2 scores, and reduced RSS, indicating superior accuracy, fit, and predictive efficiency. It effectively balances overfitting reduction and interpretability by nullifying less relevant features, making it optimal for scenarios with many potentially irrelevant variables.

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?

Answer

Based on the Lasso regression as it was the best model, the five most important predictor variables are determined by their coefficient values (which represent the strength and nature of the relationship with the dependent variable), are:

Below are the top five predictor with their beta coeeficients

At best alpha .0001

	beta	independentVariable
14	0.308287	GrLivArea
3	0.162438	OverallQual
4	0.094896	OverallCond
22	0.078638	GarageCars
31	0.056504	MSZoning_RL

After removing above top five predictor below are new top five predictors according to Lasso at alpha .0002. Although removing top parameters did reduce the R2 Score

1. **1stFlrSF (0.248370):** First-floor square footage, indicating that larger first floors are associated with an increase in the (SalesPrice) dependent variable.
2. **2ndFlrSF (0.131027):** Second-floor square footage, suggesting that larger second floors also contribute significantly to the (SalesPrice) dependent variable.
3. **GarageArea (0.079007):** The area of the garage in square feet, which shows a substantial positive impact on the (SalesPrice) dependent variable.
4. **TotRmsAbvGrd (0.060334596):** Total rooms above grade (excluding bathrooms), indicating that a higher number of rooms is positively associated with the (SalesPrice) dependent variable.
5. **Neighbourhood_crawfor(0.050909):** A dummy variable indicating whether the property is in the Crawford neighborhood, suggesting a positive effect on the (SalesPrice) dependent variable when the property is located in this area.

These variables represent the most significant predictors in the model, highlighting the importance total room above grade , of property size (as measured by square footage and the number of rooms) and Full Bathoom condition in influencing the dependent variable.

At best alpha .0002

	beta	independentVariable
10	0.248370	1stFlrSF
11	0.131027	2ndFlrSF
19	0.079007	GarageArea
16	0.060334	TotRmsAbvGrd
40	0.050909	Neighborhood_Crawfor

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?

Answer
Ensuring a model is robust and generalizable involves several strategies and considerations, each aiming to make the model perform well on unseen data, rather than just on the specific data it was trained on.

1. **Data Quantity, Quality and Diversity:** The model should be trained on sufficient high-quality, diverse data that represent the problem space well. This diversity helps in building a model that can generalize better across different situations, improving its overall accuracy.

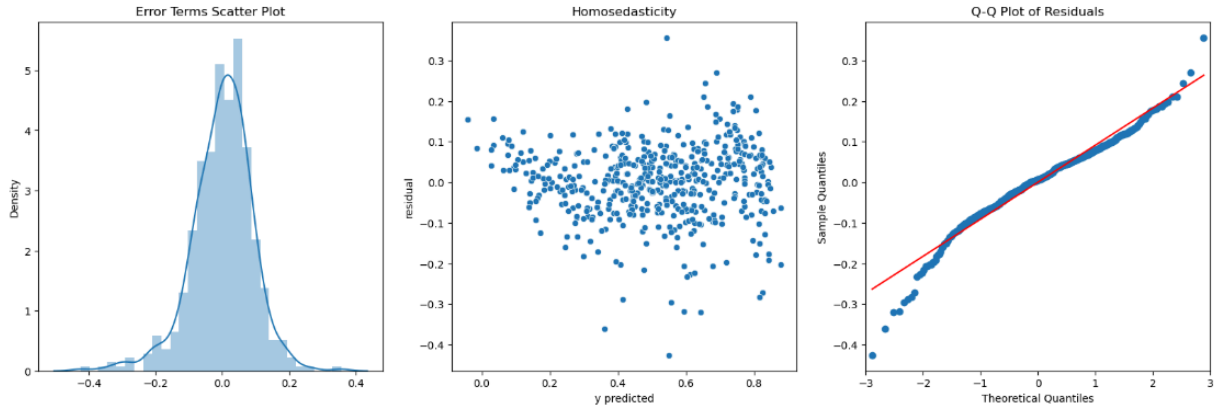
2. **Cross-validation:** Utilize cross-validation techniques during the training process, such as **k-fold cross-validation**. This involves dividing the data into k sets and training the model k times, each time using a different set as the validation data and the remaining as the training data. This helps in assessing the model's performance across different subsets of data, enhancing its generalizability and providing a more accurate estimate of its performance on unseen data.

3. **Regularization**: Regularization is done using methods such as Lasso and Rigid. In Lasso beta coefficients of insignificant features become zero, resulting in reduced model complexity and variance, While in Rigid Beta coefficient of insignificant features becomes as possible near to zero. This prevents the model from becoming too complex and overfitting to the training data, which can impair its performance on new, unseen data.

4. **Model Complexity**: (**Larger the lambda simpler the model)-** Choose the hyper parameter to appropriately. We can use metric such as MSE,RMSE,RSS and R2_Score to identify which lambda results in better metric values. For overfittrf models there is a huge reduction in R2_Score from training to test.

5. **Feature Selection and EDA**: Carefully select and engineer features that are relevant to the problem and are likely to be predictive across different contexts. This can involve removing redundant features, creating new features that capture important patterns, and normalizing or standardizing data. Effective feature engineering can improve both the model's accuracy and its ability to generalize.

6. **Linear Regression assumption verification**: Once a model is build, we can plot error terms (residuals) to identify is there is homocedasticty (constance variance and no pattern in error terms). Error terms should follow normal curve and centered around zero.



7. **Monitoring and Updating**: Continuously monitor the model's performance in real-world applications and update it with new data. This ensures the model remains relevant and accurate over time as the underlying data distributions change.

The implications of these strategies for model accuracy are nuanced. While some techniques, like regularization and selecting an appropriately complex model, might slightly reduce the model's accuracy on the training data by preventing overfitting, they are crucial for ensuring the model performs well on unseen data, which is a better indicator of its true accuracy and utility. Balancing the trade-off between fitting the training data well and maintaining the model's ability to generalize is key to developing robust and accurate models.