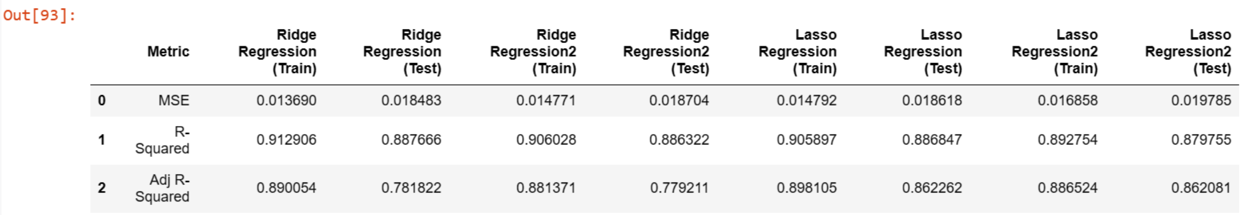
Answer 1:

Ridge Regression: Optimal Alpha = 20

Lasso Regression: Optimal Alpha = 0.001

Below are the observations for doubling the values of alpha for Ridge and Lasso -

The r-squared and adjusted r-squared have dropped and MSE has slight increase, in both Train and Test.



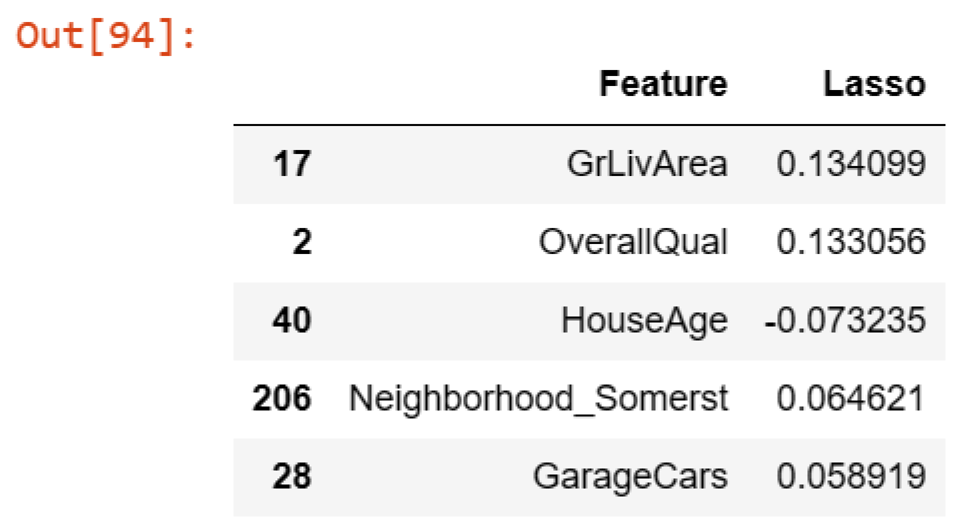
Out[93]:

	Metric	Ridge Regression (Train)	Ridge Regression (Test)	Ridge Regression2 (Train)	Ridge Regression2 (Test)	Lasso Regression (Train)	Lasso Regression (Test)	Lasso Regression2 (Train)	Lasso Regression2 (Test)
0	MSE	0.013690	0.018483	0.014771	0.018704	0.014792	0.018618	0.016858	0.019785
1	R-Squared	0.912906	0.887666	0.906028	0.886322	0.905897	0.886847	0.892754	0.879755
2	Adj R-Squared	0.890054	0.781822	0.881371	0.779211	0.898105	0.862262	0.886524	0.862081

Among the top 5 features, the top 2 features remained same in both the cases, but with an increase in coefficient value for the doubled alpha. The next 3 features got modified with decrease in coefficients.

The number of features remained same for Ridge, but has dropped in case of Lasso.

- Top 5 predictors for Lasso



Out[94]:

	Feature	Lasso
17	GrLivArea	0.134099
2	OverallQual	0.133056
40	HouseAge	-0.073235
206	Neighborhood_Somerst	0.064621
28	GarageCars	0.058919

• GrLivArea: Above grade (ground) living area square feet

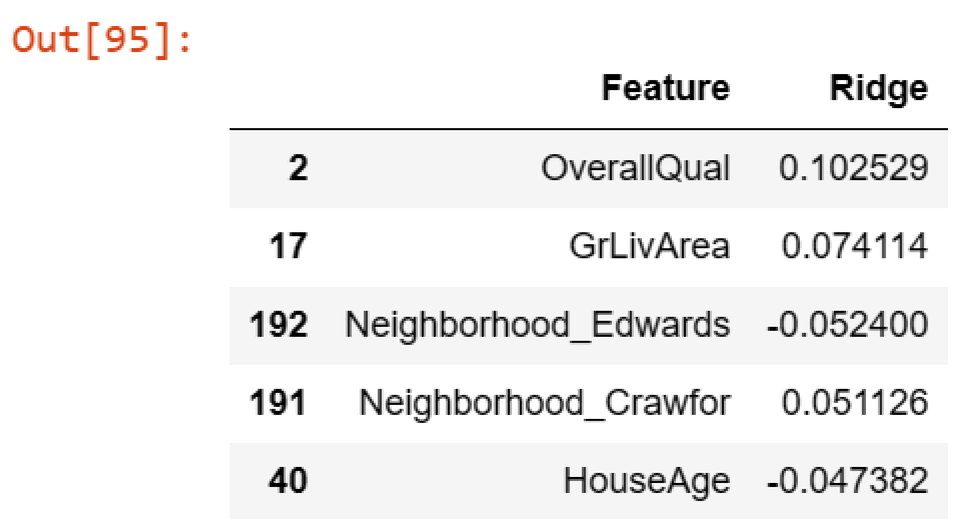
• OverallQual: Rates the overall material and finish of the house

- HouseAge: Age of the house [Sold Year – Construction Year]

- Neighborhood_Somerst: Physical locations within Ames city limits – Somerset

- GarageCars: Size of garage in car capacity

- Top 5 predictors for Ridge



Out[95]:

	Feature	Ridge
2	OverallQual	0.102529
17	GrLivArea	0.074114
192	Neighborhood_Edwards	-0.052400
191	Neighborhood_Crawfor	0.051126
40	HouseAge	-0.047382

- OverallQual: Rates the overall material and finish of the house

- GrLivArea: Above grade (ground) living area square feet

- Neighborhood_Edwards: Physical locations within Ames city limits – Edwards

- Neighborhood_Crawfor: Physical locations within Ames city limits – Crawford

- HouseAge: Age of the house [Sold Year – Construction Year]

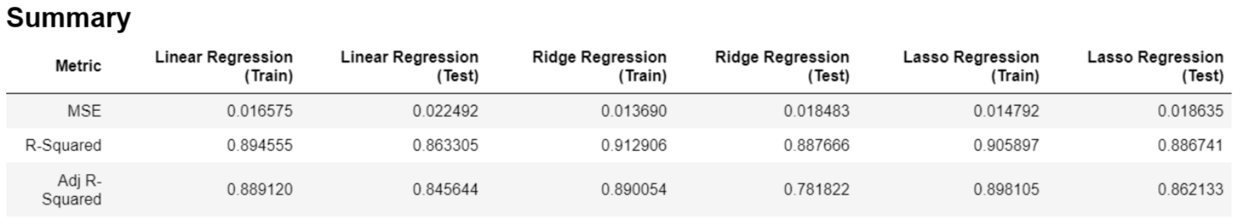
By doubling the lambda values for Ridge and Lasso regression, the r-squared and adjusted r-squared have dropped and MSE has slight increase.

Among the top 5 features, the top 2 features remained same in both the cases, but with an increase in coefficient value for the doubled alpha. The next 3 features got modified with decrease in coefficients.

The number of features remained same for Ridge, but has dropped in case of Lasso.

Answer 2:

Lasso is better considering the explainability. Lasso gives better adjusted r-squared by selecting less number of features and is robust. The difference between Test and Train accuracy for lasso is less compared to Ridge. If feature explainability is not a constraint and need to look for accuracy, ridge can be selected.



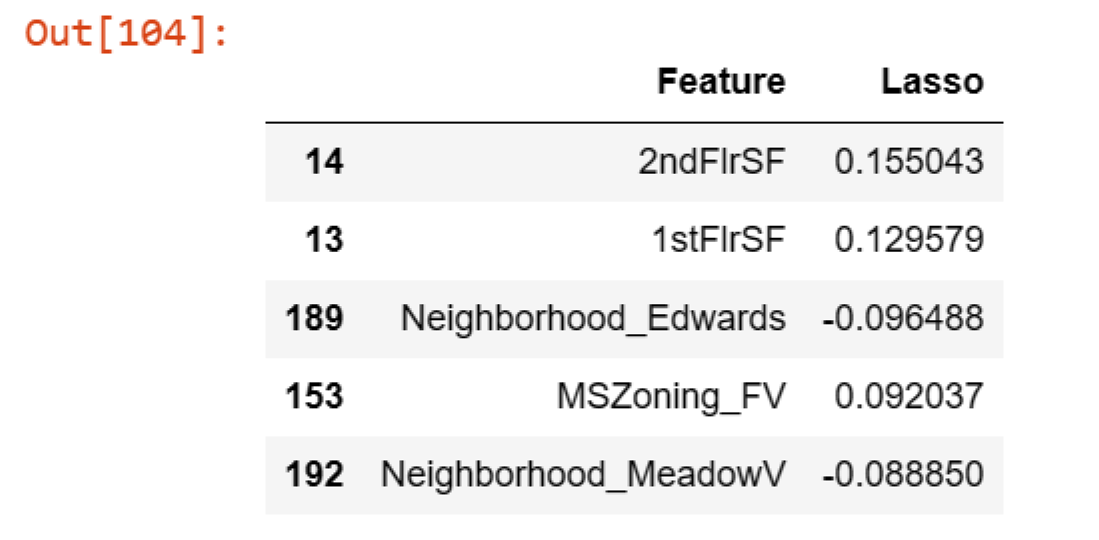
Summary

Metric	Linear Regression (Train)	Linear Regression (Test)	Ridge Regression (Train)	Ridge Regression (Test)	Lasso Regression (Train)	Lasso Regression (Test)
MSE	0.016575	0.022492	0.013690	0.018483	0.014792	0.018635
R-Squared	0.894555	0.863305	0.912906	0.887666	0.905897	0.886741
Adj R-Squared	0.889120	0.845644	0.890054	0.781822	0.898105	0.862133

Answer 3:

After removing the top-5 predictors in lasso model, the top 5 features got modified. The number of features selected got increased to 84.

Below are the new top-5 predictors:



Out[104]:

	Feature	Lasso
14	2ndFlrSF	0.155043
13	1stFlrSF	0.129579
189	Neighborhood_Edwards	-0.096488
153	MSZoning_FV	0.092037
192	Neighborhood_MeadowV	-0.088850

- 2ndFlrSF: Second floor square feet

- 1stFlrSF: First Floor square feet

- Neighborhood_Edwards: Physical locations within Ames city limits – Edwards

• MSZoning_FV: Identifies the general zoning classification of the sale. - Floating Village Residential

• Neighborhood_MeadowV: Physical locations within Ames city limits – Meadow Village

Answer 4:

Robust and generalizable Model can be ensured by:

1. **Diverse and representative dataset**: Model trained on a diverse and representative dataset is more likely to be robust and generalizable. Such a dataset exposes the model to a wide range of data examples, allowing it to learn patterns and relationships that can be applied to new data.
2. **Cross-validation**: Cross-validation is a technique that involves splitting the dataset into multiple subsets and training the model on each subset while testing it on the others. This helps to identify and **mitigate overfitting**, which can occur when the model becomes too specialized to the training data and does not perform well on new data.
3. **Regularization**: Regularization is a technique that adds to the model's loss function to **discourage** it from **overfitting**.
4. **Hyperparameter Tuning**: The choice of hyperparameters can greatly affect the performance of a model. Tuning these hyperparameters through techniques such as grid search or randomized search can improve the model's robustness and generalizability.

A robust and generalizable model is more likely to perform well on new, unseen data, and therefore be more accurate. In contrast, a model that is not robust or generalizable may perform well on the training data but poorly on new data, leading to lower accuracy and poor performance in practice. Additionally, a lack of robustness and generalizability can increase the risk of overfitting, which can further harm the model's accuracy.