House Price Prediction – Advanced Regression – Assignment

Subjective Questions

Question 1

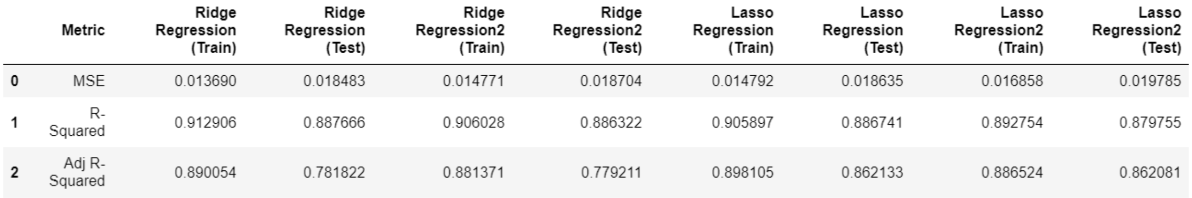
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans: 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.

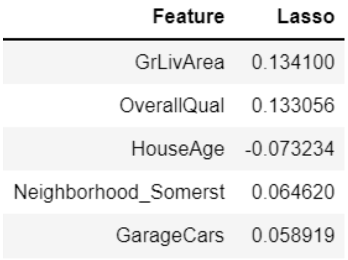


	Metric	Ridge Regression (Train)	Ridge Regression (Test)	Ridge Regression2 (Train)	Ridge Regression2 (Test)	Lasso Regression (Train)	Lasso Regression (Test)	Lasso Regression2 (Train)	Lasso Regression2 (Test)
0	MSE	0.013690	0.018483	0.014771	0.018704	0.014792	0.018635	0.016858	0.019785
1	R-Squared	0.912906	0.887666	0.906028	0.886322	0.905897	0.886741	0.892754	0.879755
2	Adj R-Squared	0.890054	0.781822	0.881371	0.779211	0.898105	0.862133	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.

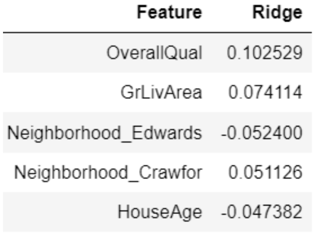
- Top 5 predictors for Lasso



Feature	Lasso
GrLivArea	0.134100
OverallQual	0.133056
HouseAge	-0.073234
Neighborhood_Somerst	0.064620
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



Feature	Ridge
OverallQual	0.102529
GrLivArea	0.074114
Neighborhood_Edwards	-0.052400
Neighborhood_Crawfor	0.051126
HouseAge	-0.047382

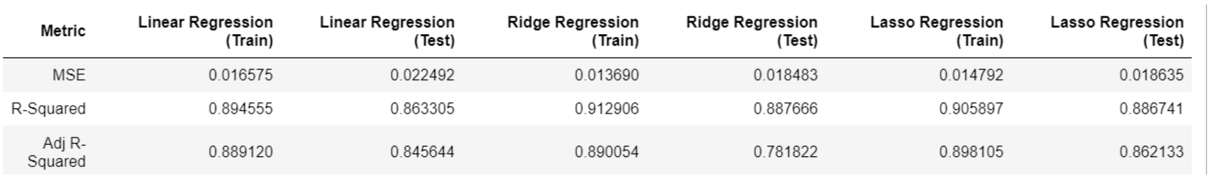
- *OverallQual: Rates the overall material and finish of the house*
- *GrLivArea: Above grade (ground) living area square feet*
- *Neighborhood_Edwards: Physical locations within Ames city limits – Edwards*
- *Neighborhood_Crawfor: Physical locations within Ames city limits – Crawford*
- *HouseAge: Age of the house [Sold Year – Construction Year]*

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: 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.



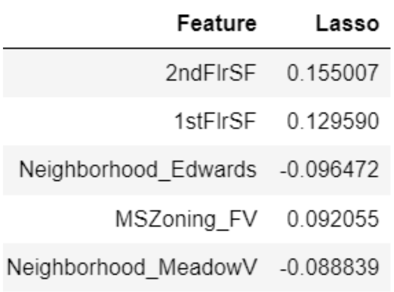
Metric	Linear Regression (Train)	Linear Regression (Test)	Ridge Regression (Train)	Ridge Regression (Test)	Lasso Regression (Train)	Lasso Regression (Test)
MSE	0.016575	0.022492	0.013690	0.018483	0.014792	0.018635
R-Squared	0.894555	0.863305	0.912906	0.887666	0.905897	0.886741
Adj R-Squared	0.889120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

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: 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:



Feature	Lasso
2ndFlrSF	0.155007
1stFlrSF	0.129590
Neighborhood_Edwards	-0.096472
MSZoning_FV	0.092055
Neighborhood_MeadowV	-0.088839

- *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*

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?

Ans: A model is robust and generalizable, when it's performance doesn't drastically degrade on un-seen data or change in the predictor variables. The test accuracy should be near to the train accuracy, without much deviation, to make sure the model is robust. Cross validation and hyper-parameter tuning will help in making the model more robust. Apply the needed transformation to the predictors and do not drop the observations, to get a more robust model, which performs well on test data. Accuracy of the model gets dropped, when it is more robust/generic. The accurate model will learn from the data and tries to overfit and performance degrades on test data, hence it is not robust.