Assignment-II

Question 1

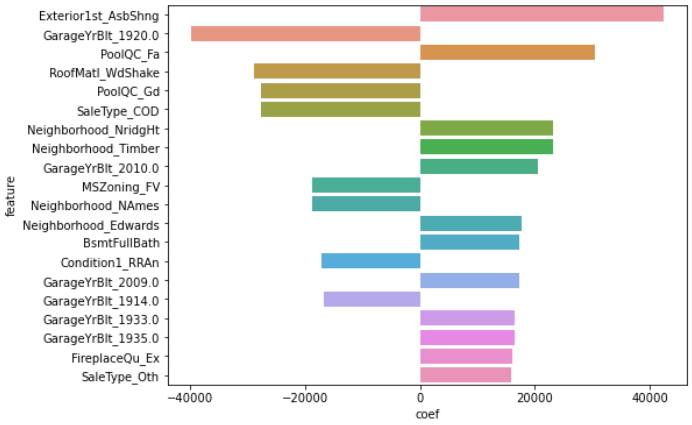
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer 1

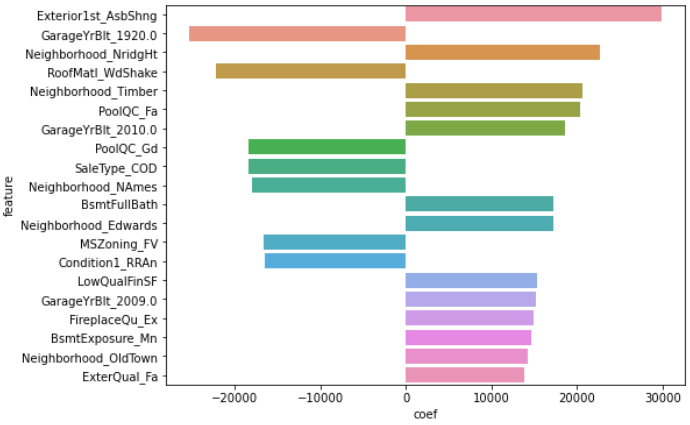
Ridge Regression:

Top Features

Optimal Lambda = 1:



(Optimal Lambda x 2) = 2:



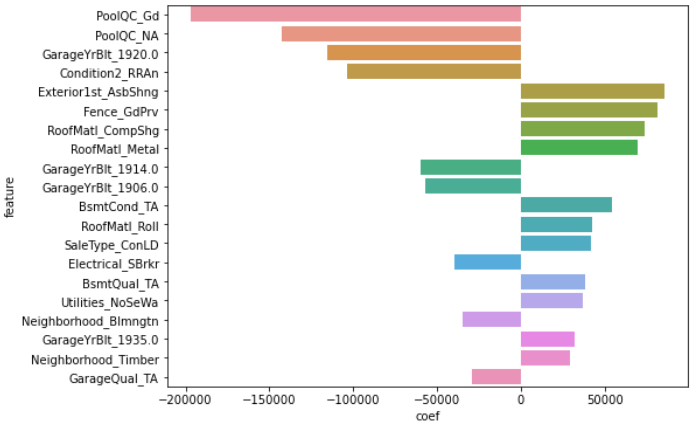
Conclusion:

- PoolQC_Fa moved from 3rd position to 6th position
- PoolQC_Gd moved from 5th position to 8th position
- Neighborhood_NridgHt was at 7th position, while it moved to 3th position

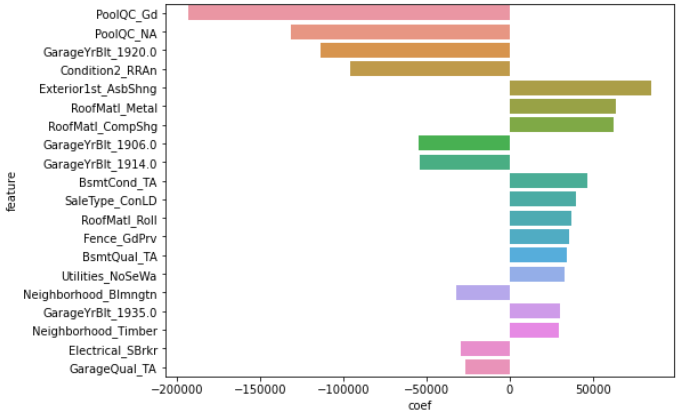
Lasso Regression:

Top Features

Optimal Lambda = 1:



(Optimal Lambda x 2) = 2:



Conclusion:

- Fence_GdPrv moved from 6th position to 13th position
- RoofMatl_Metal moved from 8th to 6th position

Final Conclusions:

- There were more changes in the top features ranking in Ridge regression when lambda was changed to 2 from 1
- Lasso was comparatively more robust, there were only some minor changes in the top features ranking, when lambda was changed to 2 from 1

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 2

Performance of Ridge Regression (optimal lambda = 1):

- Number of features with 0 coefficient: 1

	Train	Test
R2 Score	94.66%	87.76%
Root Mean Squared Error	2000	30000

Performance of Lasso Regression (optimal lambda = 1):

- Number of features with 0 coefficient: 50

	Train	Test
R2 Score	94.96%	86.26%
Root Mean Squared Error	2000	3000

Conclusion:

- Ridge regression has higher accuracy on Test data as compared to Lasso regression by approximately ~1.5%
- But, Lasso has much less features as compared to Ridge. It is because, Lasso regression also helps in feature selection, and it has found that 38 features are not important at all.

We generally prefer model with higher accuracies on Test data when we have a sensitive task. But, in this case study we are more concerned with understanding the features which influence the house prices, rather than predicted price of the houses.

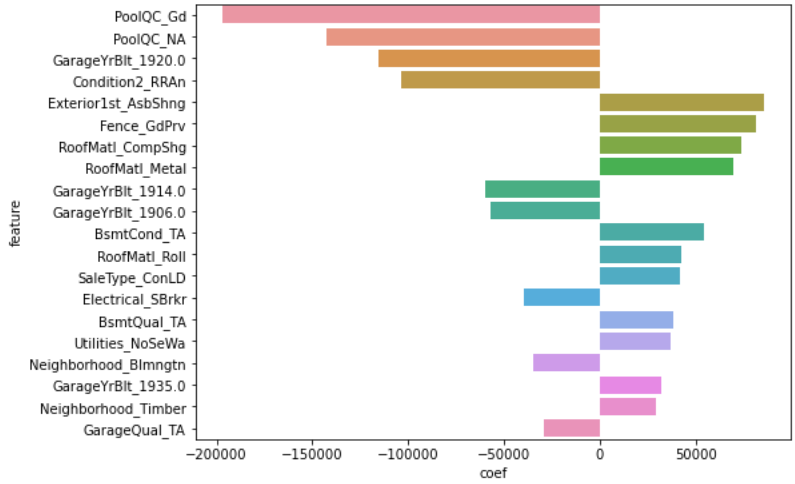
So, given the requirements we should use Lasso regression as the redundant features have been dropped and only the important ones are kept. Hence, it will be easier to understand influencing factors and explain to the clients.

Question 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?

Answer 3

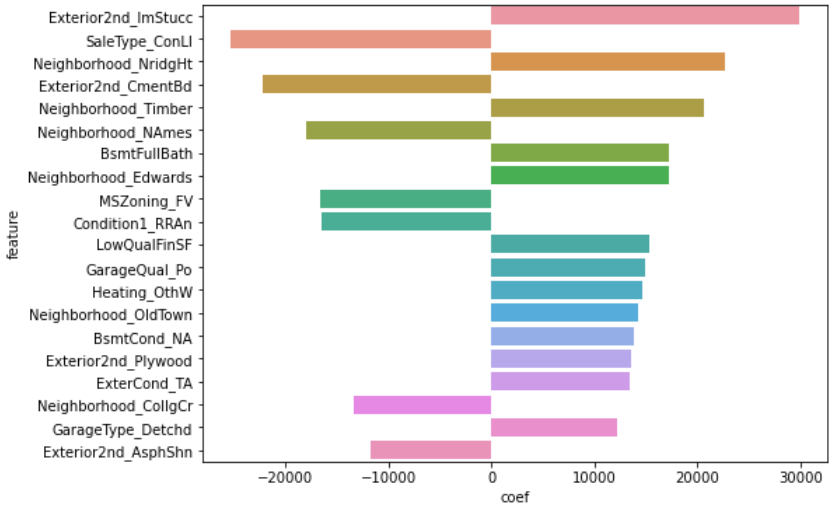
Top Features from Lasso model:



Excluding following features:

1. PoolQC
2. GarageYrBlt
3. Condition2
4. Exterior1st
5. Fence

Top 5 Features post re-training the model post exclusion of the top features:



Top 5 features:

1. Exterior2nd
2. SaleType
3. Neighborhood
4. BsmntFullBath
5. MSZoning

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?

Answer 4

To ensure model is robust and generalizable – we need to split the dataset into train, validation and test sets, the purpose of such datasets is:

- Train: model is trained on this
- Validation: this dataset is used to find optimal parameters of the model like the value of lambda
- Test: this dataset is set aside (untouched), i.e., is not used for any part of the training process and is only used to evaluate model performance.

For the modeling:

- Train consisted of 90% of the records (out of 1400)
- Test contained 10% of the records (out of 1400), 10% since the dataset is small
- Validation: here used Grid Search approach with 5 cross-validation splits to find the optimal parameters

Ridge Regression:

	Train	Test
R2 Score	94.66%	87.76%
Root Mean Squared Error	2000	30000

Lasso Regression:

	Train	Test
R2 Score	94.96%	86.26%
Root Mean Squared Error	2000	3000

Conclusion:

- Lasso regression has higher accuracy on Training data as compared to Ridge regression
- While Ridge has higher Test data as compared to Lasso regression

When both Train and Test have similar accuracy then we say model is generalized, while when Train accuracy is higher than test accuracy then model has overfitted.

In our case, both Ridge and Lasso have higher accuracies on Train as compared to Test. But, since Ridge regression has higher accuracy on Test dataset, it is a better model for taking to production.