Subjective Questions

Note: Coding problem solutions are done in IPYNB file please refer to it

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?

Answer:

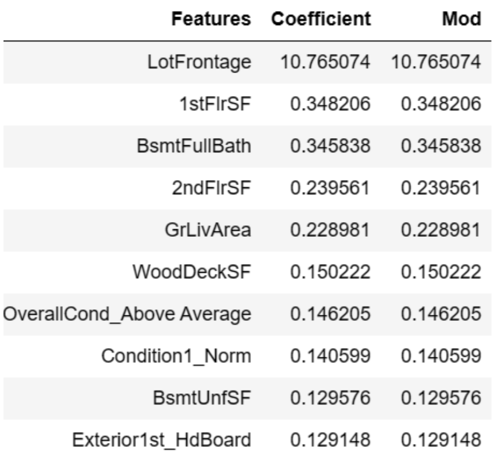
Ridge Regression - optimal value of alpha is 3.

Lasso Regression - optimal value for alpha is 0.0006.

If we choose double the value of alpha for both ridge and lasso regression, model complexity will have a greater contribution to the cost. Because the minimum cost hypothesis is selected, this means that higher λ will bias the selection toward models with lower complexity.

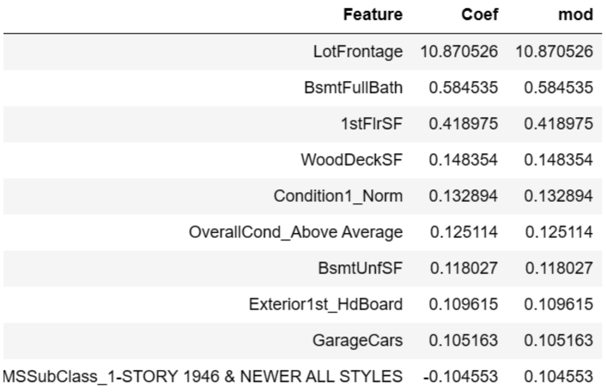
After doubling the optimal value for alpha, we saw that the coefficient values for a few features have reduced a little. The change in alpha does not seems to have significantly impacted the accuracy between the models. Below are the 10 most important predictor variables after the change.

Ridge Regression



Features	Coefficient	Mod
LotFrontage	10.765074	10.765074
1stFlrSF	0.348206	0.348206
BsmtFullBath	0.345838	0.345838
2ndFlrSF	0.239561	0.239561
GrLivArea	0.228981	0.228981
WoodDeckSF	0.150222	0.150222
OverallCond_Above Average	0.146205	0.146205
Condition1_Norm	0.140599	0.140599
BsmtUnfSF	0.129576	0.129576
Exterior1st_HdBoard	0.129148	0.129148

Lasso Regression

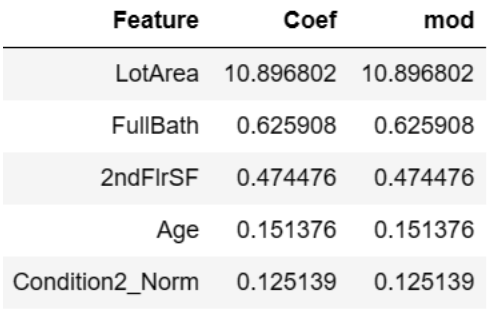


Feature	Coef	mod
LotFrontage	10.870526	10.870526
BsmtFullBath	0.584535	0.584535
1stFlrSF	0.418975	0.418975
WoodDeckSF	0.148354	0.148354
Condition1_Norm	0.132894	0.132894
OverallCond_Above Average	0.125114	0.125114
BsmtUnfSF	0.118027	0.118027
Exterior1st_HdBoard	0.109615	0.109615
GarageCars	0.105163	0.105163
MSSubClass_1-STORY 1946 & NEWER ALL STYLES	-0.104553	0.104553

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:
Lasso regression would be a better option it would help in feature elimination and the model will be more robust. Because

- In the ridge, the coefficients of the linear transformation are normal distributed and in the lasso they are Laplace distributed. In the lasso, this makes it easier for the coefficients to be zero and therefore easier to eliminate some of your input variable as not contributing to the output.
- Ridge regression can't zero out coefficients; thus, you either end up including all the coefficients in the model, or none of them. In contrast, the LASSO does both parameter shrinkage and variable selection automatically.
- Lasso regression can produce many solutions to the same problem.
- Ridge regression can only produce one solution to one problem.
- If I'm interested in identifying the most important predictors in the data, lasso regression I'll be a good choice. If I'm more concerned about reducing the variance of the model, then ridge regression might be a better option.
- If I have highly correlated predictors, lasso regression might be more effective at selecting a single predictor from each group of correlated predictors.

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:
After creating the new model below are the new top five important predictor variables based on the absolute value of their coefficients.



Feature	Coef	mod
LotArea	10.896802	10.896802
FullBath	0.625908	0.625908
2ndFlrSF	0.474476	0.474476
Age	0.151376	0.151376
Condition2_Norm	0.125139	0.125139

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:

A model can be considered generalizable when it doesn’t overfits the training data and performs equally well on the test data set as well.

A model can be considered robust if it works for broad range of input data set i.e. is does not drastically change its behavior on changing of input data. Ideally speaking accuracy should not vary much for training and test datasets