Answers to the Subjective Questions

Codes necessary for answering the subjective questions have been included in the Jupyter notebook (ipynb file).

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:

Optimal value of alpha for ridge regression = 0.4
Optimal value of alpha for lasso regression = 0.0001

Cost of Ridge & Lasso regression models before doubling optimal alpha:

	Cost	Ridge Regression (train data)	Ridge Regression (test data)
0	RSS	9.774191	7.571516
1	RMSE	0.105450	0.141717
2	R2-score	0.915978	0.858661

	Cost	Lasso Regression (train data)	Lasso Regression (test data)
0	RSS	9.791744	7.531085
1	RMSE	0.105544	0.141338
2	R2-score	0.915827	0.859416

Cost of Ridge & Lasso regression models after doubling optimal alpha:

	Cost	Ridge Regression (train data)	Ridge Regression (test data)
0	RSS	9.960991	7.639197
1	RMSE	0.106453	0.142349
2	R2-score	0.914372	0.857397

	Cost	Lasso Regression (train data)	Lasso Regression (test data)
0	RSS	10.119316	7.609559
1	RMSE	0.107295	0.142072
2	R2-score	0.913011	0.857951

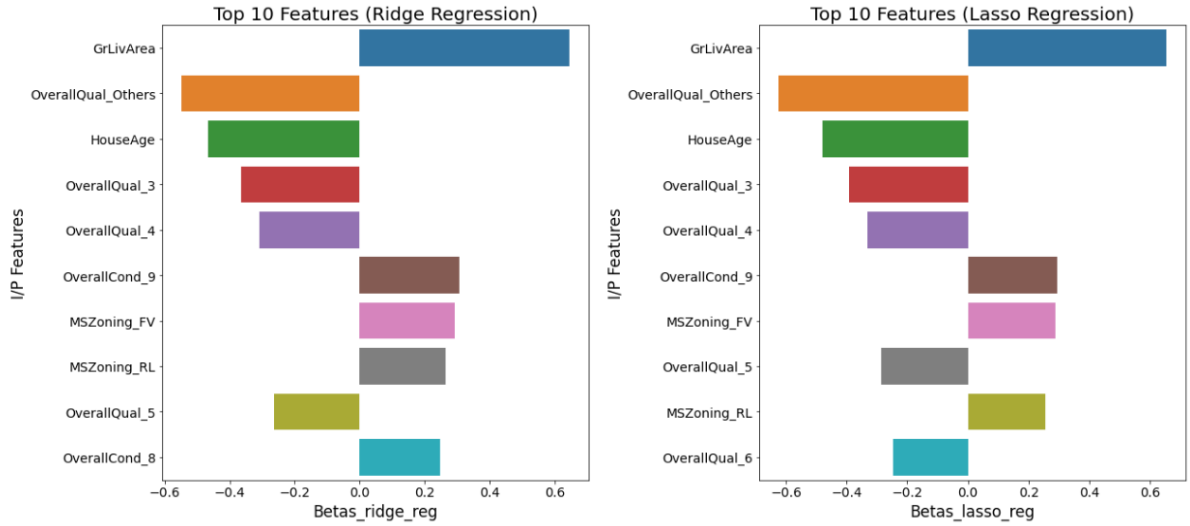
There isn't much difference in the model performance, before & after doubling the best alpha, for both Ridge & Lasso regressions. R2-score continues to be around 92% for train & 85% for test data.

Most important predictor variables (top 10) for Ridge regression model before doubling optimal alpha:

1. GrLivArea → living area above ground
2. OverallQual_Others → whether overall material and finish of the house is very poor (or) poor
3. HouseAge → age of the house as on the year 2022
4. OverallQual_3 → whether overall material and finish of the house is fair
5. OverallQual_4 → whether overall material and finish of the house is below average
6. OverallCond_9 → whether overall condition of the house is excellent
7. MSZoning_FV → whether zone classified as “floating village residential”
8. MSZoning_RL → whether zone classified as “residential low density”
9. OverallQual_5 → whether overall material and finish of the house is average
10. OverallCond_8 → whether overall condition of the house is very good

Most important predictor variables (top 10) for Lasso regression model before doubling optimal alpha:

1. GrLivArea → living area above ground
2. OverallQual_Others → whether overall material and finish of the house is very poor (or) poor
3. HouseAge → age of the house as on the year 2022
4. OverallQual_3 → whether overall material and finish of the house is fair
5. OverallQual_4 → whether overall material and finish of the house is below average
6. OverallCond_9 → whether overall condition of the house is excellent
7. MSZoning_FV → whether zone classified as “floating village residential”
8. OverallQual_5 → whether overall material and finish of the house is average
9. MSZoning_RL → whether zone classified as “residential low density”
10. OverallQual_6 → whether overall material and finish of the house is above average

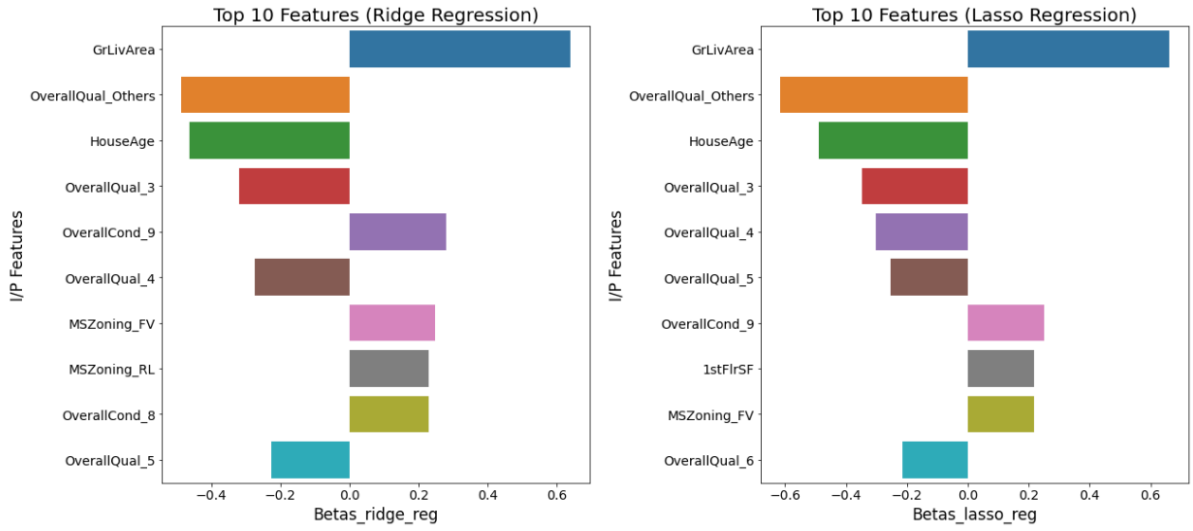


Most important predictor variables (top 10) for Ridge regression model after doubling optimal alpha:

1. GrLivArea → living area above ground
2. OverallQual_Others → whether overall material and finish of the house is very poor (or) poor
3. HouseAge → age of the house as on the year 2022
4. OverallQual_3 → whether overall material and finish of the house is fair
5. OverallCond_9 → whether overall condition of the house is excellent
6. OverallQual_4 → whether overall material and finish of the house is below average
7. MSZoning_FV → whether zone classified as “floating village residential”
8. MSZoning_RL → whether zone classified as “residential low density”
9. OverallCond_8 → whether overall condition of the house is very good
10. OverallQual_5 → whether overall material and finish of the house is average

Most important predictor variables (top 10) for Lasso regression model before doubling optimal alpha:

1. GrLivArea → living area above ground
2. OverallQual_Others → whether overall material and finish of the house is very poor (or) poor
3. HouseAge → age of the house as on the year 2022
4. OverallQual_3 → whether overall material and finish of the house is fair
5. OverallQual_4 → whether overall material and finish of the house is below average
6. OverallQual_5 → whether overall material and finish of the house is average
7. OverallCond_9 → whether overall condition of the house is excellent
8. 1stFlrSF → area of first floor
9. MSZoning_FV → whether zone classified as “floating village residential”
10. OverallQual_6 → whether overall material and finish of the house is above average



There isn’t much difference in the order of the most important predictor variables, before & after doubling the best alpha, for both Ridge & Lasso regressions. There are a few re-arrangements of existing variables and addition of new ones, with no changes in its magnitude.

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:

R2-score for train data is marginally higher in the Ridge regression whereas, R2-score for test data is marginally higher in the Lasso regression. Optimal value of lambda for lasso regression can be applied in our model, as we need a better performance on the unseen data, also perform feature selection.

Cost of Ridge & Lasso regression models with their corresponding optimal alpha's:

	Cost	Ridge Regression (train data)	Ridge Regression (test data)
0	RSS	9.774191	7.571516
1	RMSE	0.105450	0.141717
2	R2-score	0.915978	0.858661

	Cost	Lasso Regression (train data)	Lasso Regression (test data)
0	RSS	9.791744	7.531085
1	RMSE	0.105544	0.141338
2	R2-score	0.915827	0.859416

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:

Five most important predictor variables in the new Lasso regression model after excluding its previous five most important predictor variables are:

1. MSZoning_FV → whether zone classified as "floating village residential"
2. BsmtCond_Po → whether general condition of the basement is poor
3. MSZoning_RL → whether zone classified as "residential low density"
4. BsmtQual_Fa → whether general condition of the basement is poor
5. 1stFlrSF → area of first floor

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 high variance model has low bias and, vice-versa. A high variance model fits all the data points in the train data, not being able to fit sufficient data points in the test data, due to higher degree polynomial. This model will have very low to nil bias if the test data points are positioned similar to the train data. A high bias model makes many assumptions on the test data, as it didn't have sufficient data points in its training set. This model will have very low to nil variance, as any changes in the test data will not matter much, as it is any how making a lot of assumptions. A model can be made robust and generalisable by a trade-off between its variance & bias.

Regularization discourages any variance in the model, at the cost of a marginal increase in its bias. Robustness & generalizability of a model is its ability to perform well on the unseen data, even for a significant change in its input data. Regularization controls the model fit using hyperparameter tuning, thus avoiding the model to overfit as well as underfit. Minimization of error in the test data needs to be prioritized over that in the train data. So that, even if our model does not remember almost all the data points accurately, it is still able to make predictions precisely, without recalling everything.

Model's accuracy increases with a decrease in its sensitivity until an optimal threshold is reached, after which its accuracy decreases with a decrease in its sensitivity. Model's accuracy increases with an increase in its specificity until an optimal threshold is reached, after which its accuracy decreases with an increase in its specificity. A sensitive model tries to predict all the features, even if the chances of it being correct is low whereas, a high specificity model predicts only those features that have high chances of being correct. A regularized regression model reduces the model sensitivity significantly, so that it is generalizable, without affecting its specificity, so that it is robust as well.