ASSIGNMENT: ADVANCED LINEAR REGRESSION

V S S ANIRUDH SHARMA

ASSIGNMENT-BASED 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 TO DOUBLE THE VALUE OF ALPHA FOR BOTH RIDGE AND LASSO? WHAT WILL BE THE MOST IMPORTANT PREDICTOR VARIABLES AFTER THE CHANGE IS IMPLEMENTED?

LASSO:

Optimal alpha = 1e-6

R-squared for training = 0.8923571291216799

R-squared for testing = 0.8744220531193507

MOST IMPORTANT PREDICTORS

Feature	Coef
OverallQual	0.197478
1stFlrSF	0.162071
2ndFlrSF	0.132110
BsmtFinSF1	0.105891
LotArea	0.096079

Double the optimal alpha = 2e-6

R-squared for training = 0.8923547657554699

R-squared for testing = 0.8748183637330689

MOST IMPORTANT PREDICTORS

Feature	Coef
OverallQual	0.197452
1stFlrSF	0.162120
2ndFlrSF	0.132058
BsmtFinSF1	0.105789
LotArea	0.096036

For lasso, doubling the alpha neither impacted the R-squared scores much, nor the order of importance of variables

RIDGE

Optimal alpha = 2.0

R-squared for training = 0.8897302013515009

R-squared for testing = 0.8820836858709308

MOST IMPORTANT PREDICTORS

Feature	Coef
OverallQual	0.171786
1stFlrSF	0.153709
2ndFlrSF	0.126616
BsmtFinSF1	0.103587
YearBuilt	0.089369

Double the Optimal alpha = 4.0

R-squared for training = 0.8867760073150275

R-squared for testing = 0.8797773004012337

MOST IMPORTANT PREDICTORS

Feature	Coef
OverallQual	0.157419
1stFlrSF	0.146240
2ndFlrSF	0.122095
BsmtFinSF1	0.101779
YearBuilt	0.087674

For ridge, doubling the alpha reduced the R-squared scores, but did not affect the order of importance of variables

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?

Our main criterion of selecting a model would be R-squared scores, especially on testing data. Further consideration will be given to Lasso if it successfully selects fewer variables in the model.

Simple:

- Training Rsquared= 0.892357917249122
- Testing Rsquared = 0.8740158059391069

Lasso (λ=1e−6)

- Training Rsquared = 0.8923571291216799
- Testing Rsquared = 0.8744220531193507

Ridge (λ=2)

- Training Rsquared = 0.8897302013515009
- Testing Rsquared = 0.8820836858709308

All the 3 models have 29 variables since we used RFE and some manual techniques to select features. Lasso didn't have any zero coefficients.

Thus, we choose Ridge regression with λ=2, which shows the highest Test data R-squared.

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?

Here are the top 5 predictors of our final model

1. OverallQual
2. 1stFlrSF
3. 2ndFlrSF
4. BsmtFinSF1
5. YearBuilt

To remove them and redo the whole thing, we need to restart from RFE by removing these columns from consideration. Previously we have selected 30 columns, but now we need to increase it to compensate for the loss of crucial information. So now we will apply RFE with 40 columns.

Further, we shall remove drop one of the pair of columns correlated more than 0.85.

Thus, after optimizing for lambda and R-squared scored, our final model would be

Lasso regularized linear regression with lambda = 1e-4

We selected this because of the highest R-squared scores and since many coefficients were 0, resulting in feature selection.

R-squared for training = 0.8574128047960761

R-squared for testing = 0.8521874181894856

Now the new topmost important predictors are:

Feature	Coef
GrLivArea	0.369128
BsmtQual	0.112784
GarageArea	0.104611
Neighborhood_StoneBr (Is Neighborhood = 'StoneBr' ?)	0.091562
ExterQual	0.090832

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?

Models are not generalizable and robust when the model overfits on the training data, whereas the new unseen data can be more uncertain and varying.

Thus, making a model generalizable means reducing overfitting and simplifying it. In other words, a generalizable model is better equipped to handle the uncertainties and variations that it may encounter in the real world.

Thus, we note the following ways:

1. Using a simple model. We learned that simpler models are usually more robust than complex models. This would include reducing the number of variables, which we have done in this assignment.
2. By using regularization, we are reducing overfitting on the training data, as seen in our assignment.
3. Using a well-diversified dataset, we are making sure that the real-life new inputs wouldn't be too much of a surprise to the model.
4. By splitting data into training-validation-testing sets, we are making sure that even within all the data we have, we are not providing the information/influence of all the data we have. A model which learns on the training data, with hyperparameters tuned to show the best results on validation data, when shows good results on this unseen test data, can improve our confidence in the model's ability to deal with unseen new data.

By applying the above-mentioned techniques, or in general, when a model is made more robust and generalizable, the accuracy of such model on the data we have is usually lower than other models which are not created keeping in mind the generalizability.

This is because as discussed above, the specific models mold themselves more specific to the training and validation data, and hence score well on this dataset, compared to a robust model which doesn't fit itself too perfectly on the training data in order to accommodate for the real-life data uncertainty.