Name: Shu-Ren Chang

Assignment- Advanced Regression - Part-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 double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

The optimal value of alpha for ridge and lasso regression
Ridge: 2.0
Lasso: 0.0001

What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

For Ridge one, the value of mean squared error increases, but the values of r2 stay the same for train and test ones. For Lasso one, the value of mean squared error increases very slightly, but the values of r2 decrease slightly for train and test ones.

For the Ridge one, below are the most 10 important predictor variables after the change is implemented:

```
                      Coefficient
OverallQual              0.119289
Total_sqr_footage        0.100625
GrLivArea                0.097832
OverallCond              0.078720
Neighborhood_StoneBr     0.071485
LotArea                  0.060507
TotalBsmtSF              0.053101
Fireplaces               0.045202
YearBuilt                0.042966
Neighborhood_Crawfor     0.038849
```

For the Lasso one, below are the most 10 important predictor variables after the change is implemented:

```
                       Coefficient
OverallQual               0.198295
GrLivArea                 0.158693
OverallCond               0.112671
YearBuilt                 0.099392
Total_sqr_footage         0.090326
```

```
Neighborhood_StoneBr      0.083571
TotalBsmtSF               0.073047
LotArea                   0.068038
SaleCondition_Partial     0.053133
GarageArea                0.045747
```

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:

Below are the optimal values of lambda, mean squared error, and r2 for ridge and lasso regression.
I think the lasso one will be the best one based on the findings below:

1. The Mean Squared Error of Lasso is much lower than the one of Ridge.
2. Lasso helps in reducing features. Also, the coefficient values of the lasso are shrunk toward to 0 that helps increase model interpretation if we look at the magnitude of the coefficients.
3. In overall, the lasso regression does perform better than ridge one.

The optimal value of alpha for ridge and lasso regression
Ridge: 2.0
Lasso: 0.0001

The Mean Squared error for Ridge and Lasso regressions are:

Ridge: 0.00342
Lasso: 0.00280

The r2 values from Ridge and Lasso regressions are:
Ridge - Train = 0.929, Test = 0.879
Lasso - Train = 0.920, Test = 0.901

Name: Shu-Ren Chang

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:

The following top five most important predictor variables in the lasso model are presented below if the original most important five ones are no longer available:

1. TotalBsmtSF
2. Neighborhood_StoneBr
3. TotRmsAbvGrd
4. LotArea
5. GarageArea

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:

In order to make sure the model we built is robust and generalizable. The following steps can be taken:

1. Data splitting for training and validation: by splitting two data sets, the model can be validated to ensure the model is robust and the results can be much generalizable.
2. Feature selection. Feature selection can help to remove features that are not making significant contributions to models. It is the important process to select the most relevant features for the model.
3. Outlier detection: The outliers can deviate model predictability. Thus, it’s very important to detect and remove outliers from the data before building up a good model for making robust predictions.
4. Data preprocessing: Data cleaning and coding are very important procedures to ensure data can be used to fit models and make build good modes for better predictions.
5. Regularization. Regularization is a technique used to reduce errors, help improve model predictability, and avoid overfitting during data training process.

Name: Shu-Ren Chang

The robust and generalizability also mean accuracy, because when your models are robust and generalizable, your models are accurate that can be generalize to more situations in term of predictions. When a model has high variance with low bias, model overfitting can occur. In an overfitted model, we could get high accuracy on training data (Seen Data), but we might get very low accuracy on test data (Unseen Data). It also means that there is a huge difference on accuracy between train and test data that will result in failing to generalize from one model to another.