Assignment II

Assignment-based Subjective Question

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:

Using Ridge Regression, we found optimal value of alpha is 2.0
Using Lasso Regression, we found optimal value of alpha is 50

Changes after doubling the value of alpha:

Note: Please check the Vipin_Mishra.ipydb file for calculation

Ridge:

Values when alpha=2
```
r2_train_lr 0.9278166617127945
r2_test_lr 0.9165003989976092
rss1_lr 378787958828.7213
rss2_lr 192074847325.443
mse_train_lr 424174645.9448167
mse_test_lr 435543871.4862653
```

Values when alpha=4
```
r2_train_lr  0.9228351639914778
r2_test_lr 0.9182971026491835
rss1_lr 404928774680.98474
rss2_lr 187941874527.70062
mse_train_lr 453447676.0145406
mse_test_lr 426172051.0832214
```

Lasso:
Values when alpha=50
```
r2_train_lr 0.9257498019725721
r2_test_lr 0.9179618770441361
rss1_lr 389633974000.10474
rss2_lr 188712996858.0603
mse_train_lr 436320239.6417746
mse_test_lr 427920627.7960551
```

Values when alpha=100

```
r2_train_lr 0.920088431663279
r2_test_lr 0.9200972835191552
rss1_lr 419342476744.85406
rss2_lr 183800903054.69824
mse_train_lr 469588439.80386794
mse_test_lr 416782093.0945538
```

Below are the most important predictor variables after the change is implemented

- LotArea
- OverallQual
- OverallCond
- YearBuilt
- MasVnrArea
- BsmtFinS
- TotalBsmtSF
- TotRmsAbvGrd
- GarageCars
- Neighborhood_StoneBr
- Street_Pave

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:

Generally, Lasso should perform better in situations where only a few among all the predictors that are used to build our model have a significant influence on the response variable. So, feature selection, which removes the unrelated variables, should help. But Ridge should do better when all the variables have almost the same influence on the response variable. It is not the case that one of the techniques always performs better than the other – the choice would depend upon the data that is used for modelling.

In our case, we can see from below calculation:

Ridge:

Values when alpha=2

```
r2_train_lr 0.9278166617127945
r2_test_lr 0.9165003989976092
rss1_lr 378787958828.7213
rss2_lr 192074847325.443
mse_train_lr 424174645.9448167
mse_test_lr 435543871.4862653
```

Lasso:
Values when alpha=50
```
r2_train_lr 0.9257498019725721
r2_test_lr 0.9179618770441361
rss1_lr 389633974000.10474
rss2_lr 188712996858.0603
mse_train_lr 436320239.6417746
mse_test_lr 427920627.7960551
```

R2 score is used to evaluate the performance of a linear regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable(s).

In other words, r R2 score shows how well the data fit the regression model

From above calculation we can say,
r2_test_lr is slightly higher for Lasso in comparison to Ridge, so we can conclude that Lasso will be better choice.

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:

Note: Please refer Q3 calculation in Vipin_Mishra.ipydb file.

Based on the calculation, below are the table.

	LassoNew
OverallCond	34258.150338
MasVnrArea	32959.005279
BsmtFinSF2	-0.000000
BsmtUnfSF	-32050.551609
1stFlrSF	101209.986339
2ndFlrSF	0.000000
LowQualFinSF	0.000000
GrLivArea	173246.762569

	LassoNew
FullBath	14386.655870
BedroomAbvGr	-32198.995819
KitchenAbvGr	0.000000
TotRmsAbvGrd	10816.907093
GarageCars	42453.931855
OpenPorchSF	19095.128325
ScreenPorch	9736.687934
MSZoning_FV	22351.157049
MSZoning_RH	16779.841566
MSZoning_RL	21244.883071
MSZoning_RM	8108.682992
Street_Pave	62126.832251

We can see below are the five most important predictor variables now.

1. **1stFlrSF**
2. **GrLivArea**
3. **Street_Pave**
4. **Neighborhood_StoneBr**
5. **GarageCars**

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:

Model should not overfitted, When the model memorizes and fits too closely to the training set, the model becomes overfitted, and it is unable to generalize well to new data. If a model cannot generalize well to new data, then it will not be able to perform the classification or prediction tasks that it was intended for. Ideal and robust model should work similarly on test data in comparison to training data.

Accuracy of any model can be increased by following below points.

- Treat missing values
- Treat outlier values

- Feature scaling
- Feature selection
- Model optimization
- Cross validation