Assignment 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?

Optimal value for ridge is 10.

```
In [80]: #final ridge model
         alpha = 10
         ridge = Ridge(alpha=alpha)

         ridge.fit(X_train, y_train)
         ridge.coef_
```

```
Out[80]: array([-2.13588570e-02,  1.89916661e-02,  7.32789555e-03,  7.96272275e-02,
                 4.07794642e-02,  3.72503132e-02,  2.05613215e-02, -2.78746650e-03,
                -1.30338934e-03,  7.00573270e-03,  4.35792500e-03,  5.54556617e-03,
                 1.48198007e-02,  3.71802844e-02,  4.20641346e-02,  8.38742474e-04,
                 6.23081651e-02,  2.96129113e-02,  2.70982155e-03,  1.87339500e-02,
                 1.07449570e-02,  1.20912671e-02, -1.36072850e-02,  1.55763992e-02,
                 2.26754481e-03, -5.94799196e-05,  4.02868362e-02,  4.26535353e-03,
                 1.44146209e-03,  1.36106893e-02, -2.04566096e-03,  8.78577136e-03,
                 7.21093263e-03,  1.06190103e-02, -1.48410955e-02, -1.45446773e-06,
                -1.10808590e-03, -6.42496774e-03,  4.82665080e-02,  3.90987203e-02,
                 6.20215605e-02,  1.56874391e-02,  2.13169821e-03,  2.59754430e-02,
                 2.43132363e-02, -5.00429742e-02,  4.82147056e-03,  3.82128670e-02,
                 3.57786613e-02,  4.68948900e-02, -1.30779655e-02,  3.69458991e-02,
                -3.54872267e-02, -7.81303691e-03, -7.13321112e-03,  2.74817043e-02,
                -6.57361354e-03, -4.05279914e-03, -2.37900445e-02,  1.68568833e-02,
                 5.11106484e-02, -1.48457912e-02,  1.03581702e-01, -7.88181138e-02,
                -2.85607776e-02, -6.06433919e-02, -4.79013337e-02, -2.79855433e-02,
                -9.83064176e-03, -1.10216221e-02, -1.51395258e-02,  4.35567573e-02,
                 8.10376872e-02, -2.43051639e-02,  1.92119395e-02, -2.90898522e-02,
```

```
In [81]: #lets predict the R-squared value
         y_train_pred = ridge.predict(X_train)
         print(metrics.r2_score(y_true=y_train, y_pred=y_train_pred))
```

```
0.9220052627340902
```

```
In [82]: # Prediction on test set
         y_test_pred = ridge.predict(X_test)
         print(metrics.r2_score(y_true=y_test, y_pred=y_test_pred))
```

```
0.8855289803702383
```

Optimal value for Lasso is 0.0001

```
                return_train_score=True, scoring='neg_mean_absolute_error',
                verbose=1)
```

```
]: cv_results_l = pd.DataFrame(lasso_cv.cv_results_)
```

```
]: print(lasso_cv.best_params_)
   print(lasso_cv.best_score_)
```

```
{'alpha': 0.0001}
```

once we double it is 20 for ridge and 0.0002 for lasso

Most important predictor variable for ridge and lasso before and after doubling the alpha is same .

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?

Lasso regression works better and can be used for feature selection as well.

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?

Lot area

Lot fontage

yearbuilt

GrLivArea

Garage cars

Screenshot before and after are attached below

before:

```
LotFrontage 0.0525685950804696
LotArea 0.07500549205434695
YearBuilt 0.11370238107620331
YearRemodAdd 0.11702715517501418
MasVnrArea 0.08095677123082122
BsmtFinSF1 0.010905283564455932
BsmtFinSF2 -0.01884506578608505
BsmtUnfSF 0.0
TotalBsmtSF 0.11668097050857894
1stFlrSF 0.0
2ndFlrSF 0.010301614979598734
LowQualFinSF -0.02999369865544564
GrLivArea 0.3716999733138868
BsmtFullBath 0.07961391719918515
BsmtHalfBath 0.01762115205847727
FullBath 0.049298047708201266
HalfBath -0.038518377119787804
BedroomAbvGr -0.11182706905754818
KitchenAbvGr -0.10676719016940561
TotRmsAbvGrd 0.07489213341563498
Fireplaces 0.06559380620303708
GarageYrBlt 0.042958771484868596
GarageCars 0.14258354531218703
GarageArea -0.015649489814901673
WoodDeckSF 0.03532516641323631
OpenPorchSF -0.01165721947677007
EnclosedPorch 0.018307047068372075
3SsnPorch 0.0032986315605546012
ScreenPorch 0.028133550625998288
PoolArea -0.04120917813481295
MiscVal 0.0
MoSold 0.010869262147546006
YrSold -0.00850379757063662
```

After

```
LotFrontage0.06095999524470457
LotArea0.06853067763972061
YearRemodAdd0.14475985053715698
MasVnrArea0.09310730966920644
BsmtFinSF10.14343246339097837
BsmtFinSF20.028003004355312826
BsmtUnfSF0.1356308150068319
1stFlrSF0.2414659277347859
2ndFlrSF0.2790257362182471
LowQualFinSF-0.011788374193507948
BsmtFullBath0.08846144389285251
BsmtHalfBath0.021160718095745115
FullBath0.10005998236496688
HalfBath0.025503176593456245
BedroomAbvGr-0.11655676404090513
KitchenAbvGr-0.11343159058441558
TotRmsAbvGrd0.07441763698512234
Fireplaces0.08070853289417666
GarageYrBlt0.07708990915788129
GarageArea0.11983728729729255
WoodDeckSF0.034626823197238386
OpenPorchSF-0.02468439552168041
EnclosedPorch-0.0031419199589251393
3SsnPorch0.0016620370432439515
ScreenPorch0.024288485153988796
PoolArea-0.0451779948560824
MiscVal-0.0012813191468553499
MoSold0.011806664207458215
YrSold-0.014503578274094276
```

MasvnrArea,1stflrsf0,fullbath0garagearea

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 :

The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training.
Bias-variance tradeoff - If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.To make sure that a model is robust and generalisable, we must ensure that our model is resistant to outliers and use more robust error metrics.

▪ To take care of the existing outliers in the data, we use various techniques to remove them
o Capping the values at a certain threshold
o Removingtheoutliersmanually
o Transformingcertainvalues(exp,logetc

If a model is too complex, it will have low bias and high variance. But as it has overfitted the training data, model will give high accuracy on training dataset, but is more likely to perform poorly in unseen test dataset.

If a model is too simple, it will have high bias and low variance. As it is too simple, it will fail to identify the underlying patterns in the data, and as a result it will high a low training score as well as test score.

If we take the point in the bias variance trade off graph, where both intersect each other, that point will give perfect balance between bias-variance. It will ensure that model does not overfit while still having good variance.