Advanced Linear regression assignment

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.

Optimum values of alpha for Ridge : 2.89
Optimum values of alpha for lasso : 0.0005

Effect of doubling alpha value

Ridge :

On doubling alpha value, it was observed that the R squared score reduced by 0.006 points. While the top 10 columns did not change, the order of some columns changed. Namely MSSubClass_160 and BldgType_Duplex exchanged places.

Error metrics when alpha is 2.89

```
Error metrics for alpha

r2_score train_set: 0.8338679757870402
r2_score test_set: 0.8228340874087039
======================================================================
mean_squared_error train_set: 0.02612686278759779
mean_squared_error test_set: 0.029150899902563112
======================================================================
mean_absolute_error train_set: 0.10861094932999536
mean_absolute_error test_set: 0.11965185016789956
======================================================================
rms_error train_set: 0.16163806107349157
rms_error test_set: 0.17073634616730882
```

Error metrics when alpha is doubled

```
Error metrics when alpha doubled

r2_score train_set: 0.8327662877462447
r2_score test_set: 0.822380073830256
======================================================================
mean_squared_error train_set: 0.02630012048678591
mean_squared_error test_set: 0.029225603349666215
======================================================================
mean_absolute_error train_set: 0.10883362303726751
mean_absolute_error test_set: 0.119411893687819
======================================================================
rms_error train_set: 0.16217311887851793
rms_error test_set: 0.17095497462684792
```

Lasso :

On doubling alpha value, it was observed that the R squared score reduced by 0.006 points. The list of top 10 important features changes by only one feature, BsmtEXposure_Gd becomes more important than Garage_type_other

Error metrics when alpha is 0.0005

```
r2_score train_set: 0.8337616518130264
r2_score test_set: 0.8233476952272103
======================================================================
mean_squared_error train_set: 0.026143583897770575
mean_squared_error test_set: 0.029066390812256367
======================================================================
mean_absolute_error train_set: 0.10842407113886196
mean_absolute_error test_set: 0.11932241868238054
======================================================================
rms_error train_set: 0.1616897767262067
rms_error test_set: 0.17048868235826203
```

Error metrics when alpha is doubled

```
r2_score train_set: 0.8320417004859508
r2_score test_set: 0.8228996594339302
======================================================================
mean_squared_error train_set: 0.026414073182041534
mean_squared_error test_set: 0.02914011067389141
======================================================================
mean_absolute_error train_set: 0.10868811760171351
mean_absolute_error test_set: 0.11884719566779683
======================================================================
rms_error train_set: 0.16252406954676446
rms_error test_set: 0.1707047470748585
```

The top 10 features after changes implemented

Ridge, alpha = 2*2.89

BsmtFinType1_other	-0.190503
BldgType_Duplex	-0.180345
Neighborhood_Crawfor	0.179395
MSSubClass_160	-0.174116
Neighborhood_NridgHt	0.169640
1stFlrSF	0.161205
MSZoning_other	-0.156313
2ndFlrSF	0.140321
Neighborhood_Somerst	0.127074
GarageType_other	-0.122012

Lasso, alpha =0.0005*2

BsmtFinType1_other	-0.192040
Neighborhood_Crawfor	0.182368
BldgType_Duplex	-0.180556
MSSubClass_160	-0.173659
Neighborhood_NridgHt	0.167982
1stFlrSF	0.162207
MSZoning_other	-0.149064
2ndFlrSF	0.140600
Neighborhood_Somerst	0.120575
BsmtExposure_Gd	0.116130

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:
A lasso regression model with alpha being 0.0005 gave us the highest R squared score and also the lowest rms error on the test set.

Comparing the two models on test dataset

R squared score for ridge model : 0.82238
R squared score for lasso model : 0.82234

RMS error for ridge model : 0.17095
RMS error for ridge model : 0.17048

It can be seen that the R score is higher for the lasso model and the rms error is also lower on the test set. As the metrics are on a test set, data the model hasn't learnt on, it can be safe to say that for new data the lasso model will perform better.

While selecting a model in real life, one might also have to consider the presence of outliers, non-normality of errors and overfitting. In such cases using lasso could be beneficial to fend off such errors to a large extent resulting in a robust model.

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?
Answer:

Top 5 features are
1. 'BsmtFinType1_other',
2. 'Neighborhood_Crawfor',
3. 'MSSubClass_160',
4. 'BldgType_Duplex',
5. 'Neighborhood_NridgHt'

On dropping these variables and training a model, we observe that the model performance reduces drastically. The new model metrics are
R squared score for test data : 0.7836
RMS error on test data : 0.18886

The new top 5 features are
1. '1stFlrSF',
2. 'MSZoning_RM',
3. 'MSZoning_other',
4. '2ndFlrSF',
5. 'Neighborhood_Somerst'

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 model is considered as robust and generalizable when it provides similar performance metrics when it works on new test data. Data that may consist of

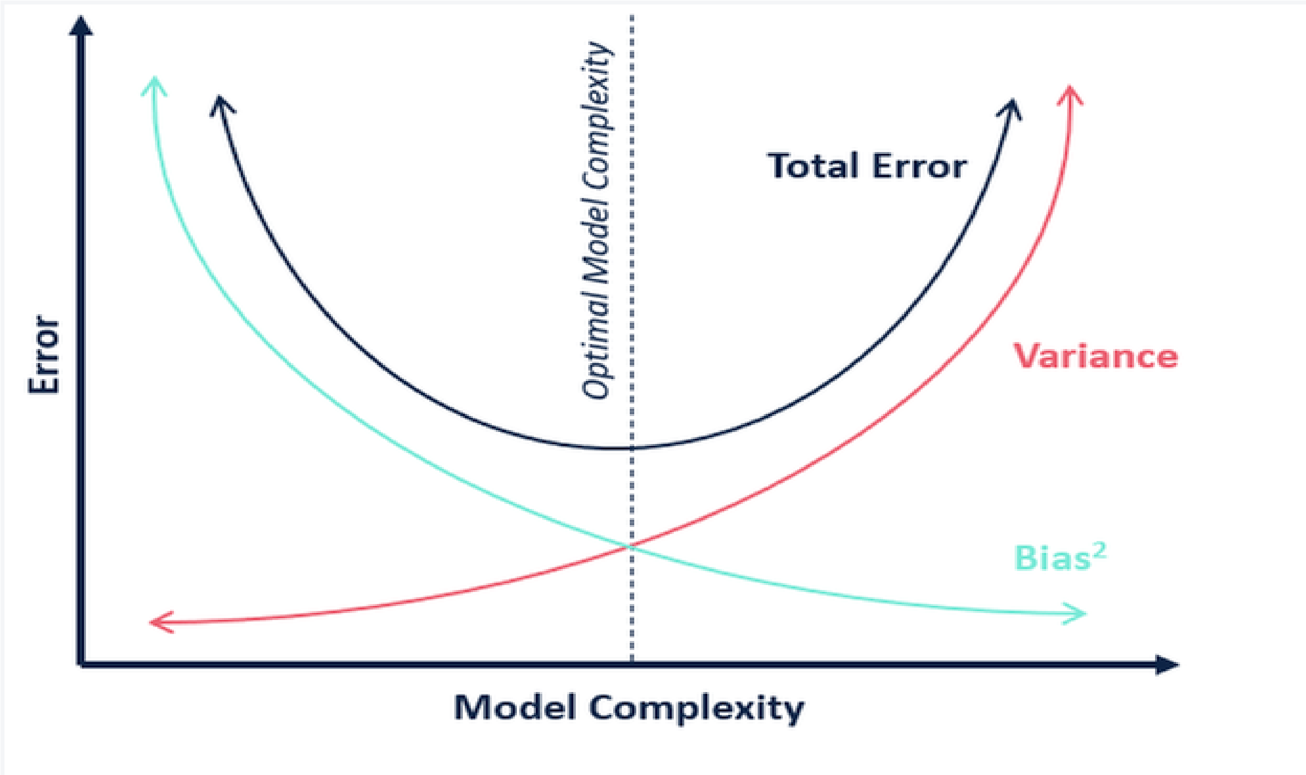
significant outliers and noise. In other words, If our model were to perform in a similar way on new data with noise, then we consider it to be robust.

Regularization can be used to make a model simple and robust. It helps in striking a balance between keeping the model simple and making the model accurate on the training data. For regression tasks, it involves adding a penalty term for overfitting terms.

Regularization will help us select the best performing and yet simple model. A model which does not overfit and neither does it underfit. This is known as bias-Variance tradeoff.

- A complex model will have to change its parameter for every new data it learns from, even if its noise, it changes the parameter values
- A very simple model will not change its parameter even if it is trained on the correct data, leading to a model that in the end does not learn anything.
- The right model will be able to identify noise and change its parameter only if the data is correct.

Bias quantifies how accurate the model is likely on test data. Variance refers to the degree of changes in the model itself when training data changes.



Thus, the accuracy of the model can be maintained by striking a balance between bias and variance, The valley in the graph shows the optimum model.