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?

a. The Optimal values are:
   Optimal value of alpha for Ridge {'alpha': 50}
   Optimal value of alpha for Lasso {'alpha': 0.01}

b. Metrics when alpha for Ridge = 50 and Lasso = 0.01

Metric	Ridge	Lasso
R2 Square(Train)	0.933405	0.924747
R2 Square(Test)	0.907930	0.914612

Metrics when we double the value of alpha for both Ridge = 100 and Lasso=0.02

Metric	Ridge	Lasso
R2 Square(Train)	0.930185	0.913622
R2 Square(Test)	0.909458	0.908934

c. The important 5 predictor variables after the change is implemented are:

Lasso

```
GrLivArea               0.337137
OverallQual             0.199561
BsmtFinSF1              0.085703
Neighborhood_NridgHt    0.078600
TotalBsmtSF             0.078486
```

Ridge

```
OverallQual             0.132774
GrLivArea               0.123681
Neighborhood_NridgHt    0.083271
2ndFlrSF                0.082002
TotalBsmtSF             0.078835
```

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?

Metric	Ridge	Lasso
R2 Square(Train)	0.933405	0.924747
R2 Square(Test)	0.907930	0.914612
Mean square error (Train)	0.066595	0.075253
Mean square error (Test)	0.099136	0.091941

1. Looking at the metrics, we can choose Lasso as the R2 value for test is greater than Ridge.
2. Also Lasso has many co-efficient set to 0 which in turn will help in feature elimination and keeps the model fairly simple compared to Ridge
3. Hence we will choose to apply Lasso.

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?

The top 5 predictor variables for Lasso are:

```
GrLivArea                0.327577
OverallQual              0.186756
Neighborhood_NridgHt     0.095961
BsmtFinSF1               0.079385
Neighborhood_Crawfor     0.076614
```

If we drop these top 5 predictor variables and re-run the Lasso model, then the new top 5 predictor variables will be:

```
2ndFlrSF       0.300650
1stFlrSF       0.200534
TotalBsmtSF    0.176360
ExterQual      0.105744
OverallCond    0.086057
```

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?

There are mainly 2 things that need to be taken care to make a model robust and generalisable.

1. Low bias - A model with high bias pays very little attention to the training data and oversimplifies the model leading to high error on both training and test data.

2. Low variance - A model with high variance pays lot of attention to the training data and hence doesn't generalize on the data leading to high error rates on test data.

Hence in order to make a model robust and generalisable we should have a trade-off and choose a model with low bias and low variance.

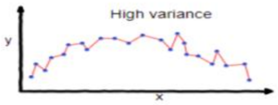
As per Occam Razor's rule, we should always choose a simple model because

a. They are generic
b. Requires less training samples
c. Are robust as complex models result in overfitting.

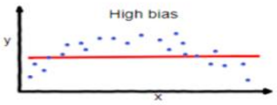
Regularization is a technique to keep the models simple by adding constraints to the cost function e.g. Ridge and Lasso.

Implications on the accuracy of the model.

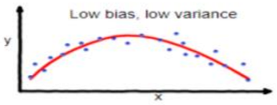
- A model with low bias will keep the model simple but usually will have high variance leading to high error rates on test data and hence not robust.
- A model with low variance will usually have high bias making the model complex.
- In order to have a balance we need to have a trade-off between bias and variance which will lead to increase in errors and hence reduction in the accuracy. So this is a trade-off that one has to make to have reasonably robust and generalisable models.



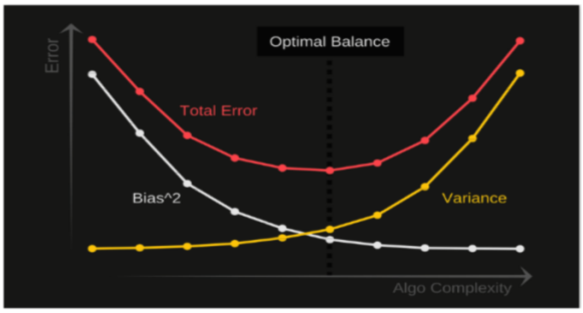
overfitting



underfitting



Good balance


Error
Optimal Balance
Total Error
Bias^2
Variance
Algo Complexity