Problem Statement - Part II

Q1. 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?

Ans :

The optrimal values for Ridge and Lasso regression as seen above are :

1. Ridge regression : 10
2. Lasso regression : 0.001

For ridge regression, we see that the negative mean absolute error decreases as alpha increases from 0, while the train error increases as alpha increases, so we choose alpha as 10 for our ridge regression.

For lasso regression, the model does not penalize much and keeps most of the coefficients with a very small value of 0.001. The negative mean absolute error was 0.8 for this alpha value. For ridge regression, I have used a higher value of alpha, 10, which makes the model simpler and more general by penalizing more. The graph shows that this alpha value increases the errors for both test and train. For lasso regression, increasing alpha also simplifies and generalizes the model by penalizing more and reducing more coefficients to zero, which reduces the r2 score.

The most important variable after the changes has been implemented for ridge regression are as follows:-

```
Neighborhood_Crawfor    1.304172
SaleCondition_Normal    1.243996
OverallQual             1.200987
GrLivArea               1.185052
SaleType_New            1.179799
Neighborhood_StoneBr    1.173128
MSZoning_RL             1.151207
Foundation_PConc        1.150788
OverallCond             1.148480
MSZoning_FV             1.148339
Name: Ridge, dtype: float64
```

The most important variable after the changes has been implemented for lasso regression are as follows:-

```
Neighborhood_Crawfor    1.412979
SaleType_New            1.365401
GrLivArea               1.334626
SaleCondition_Normal    1.290601
MSZoning_FV             1.264530
Neighborhood_StoneBr    1.239986
OverallQual             1.203691
MSZoning_RL             1.180637
Exterior1st_BrkFace     1.166938
OverallCond             1.152987
Name: Lasso, dtype: float64
```

Q2. 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?

Ans: The accuracy and interpretability of the model can be improved by regularization, which reduces the variance and shrinks the coefficients. There are two regularization methods: ridge regression and lasso regression. They both use a parameter called lambda to adjust the penalty. Ridge regression penalizes the square of the coefficients, while lasso regression penalizes the absolute value of the coefficients. Cross validation is used to determine the best value of lambda. Ridge regression does not remove any variables from the final model, but lasso regression can make some variables irrelevant by setting their coefficients to zero. As lambda increases, ridge regression lowers the variance and maintains the bias, while lasso regression selects variables and makes the model simpler.

	Model	Alpha	MAE	MSE	RMSE	R2 Score	RMSE - Cross-Validation
1	Ridge	10	2.109300e-01	8.838000e-02	2.972900e-01	9.144400e-01	0.115232
2	Lasso	0.001	2.121600e-01	8.929000e-02	2.988100e-01	9.135600e-01	0.116269
0	LinearRegression	NA	3.566540e+07	2.345172e+17	4.842697e+08	-2.270392e+17	0.126760

Based on above table, clearly Lasso regression will produce a simple model and yield better results on unseen data.

Q3. 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?

Ans:

Those 5 most important predictor variables are :-

1. GrLivArea

2. OverallQual
3. OverallCond
4. SaleType
5. Neighborhood

Q4. 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?

Ans: The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.
It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.