Advanced Regression – Assignment Part II

Question 1

a. What is the optimal value of alpha for ridge and lasso regression?

The optimal alpha value in case of Ridge and Lasso is as below:

- Ridge: 0.2
- Lasso: 0.001

b. What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

Lasso:

When the value of alpha is doubled, model will be penalized more and so, the coefficient of more variables will be reduced to zero. At the same time, R2 will decrease and RMSE will increase.

The data based on House Pricing model programming exercise:

Training Data

R2 (alpha = 0.001): 0.8621306864920301
R2 (alpha = 0.002): 0.8108375037559485

Test Data

R2(alpha = 0.001): 0.8308600747150239
R2 (alpha = 0.002): 0.7745172241800786

RMSE

alpha = 0.001: 0.054265334094987634
alpha = 0.002: 0.062655052793683

No of No-zero coefficients

alpha = 0.001: 52
alpha = 0.002: 39

Ridge:

when the value of alpha is doubled, model will be penalized more and so, the coefficient of more variables will be reduced in magnitude. At the same time, R2 will decrease and RMSE will increase.

The data based on House Pricing model programming exercise:

Training Data

R2 (alpha = 0.2): 0.8621306864920301
R2 (alpha = 0.4): 0.8108375037559485

Test Data

R2(alpha = 0.2): 0.8308600747150239
R2 (alpha = 0.4): 0.7745172241800786

RMSE

alpha = 0.2: 0.054265334094987634
alpha = 0.4: 0.062655052793683

No. of No-zero coefficients

alpha = 0.2: 275
alpha = 0.4: 275

Coefficient Magnitude

	alpha = 0.2	alpha = 0.4
GrLivArea	0.196	0.176
1stFlrSF	0.184	0.165
MSZoning_RL	0.123	0.100
MSZoning_FV	0.118	0.099

c. What will be the most important predictor variables after the change is implemented?

Lasso

1. GrLivArea
2. GarageCars
3. TotRmsAbvGrd
4. YearRemodAdd
5. FireplaceQu_No Fireplace
6. OverallQual_VERY_GOOD
7. BsmtExposure_Gd
8. CentralAir_Y
9. ExterQual_TA
10. FullBath

Ridge:

11. GrLivArea
12. 1stFlrSF

13. MSZoning_RL
14. MSZoning_FV
15. RoofMatl_WdShngl
16. TotalBsmtSF
17. MSZoning_RH
18. MSZoning_RM
19. Condition2_PosA
20. 2ndFlrSF

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?

Both Ridge and Lasso are regularization techniques which are used to penalize the magnitude of coefficients of features so that the error between predictions and actual values is minimized. The key difference between those is that Lasso Regression has the ability to nullify the impact of an irrelevant feature in the data, meaning that it can reduce the coefficient of a feature to zero thus completely eliminating it and hence is better at reducing the variance when the data consists of many insignificant features. On the other hand, Ridge regression cannot reduce the coefficients to absolute zero and performs better when the data consists of features which are sure to be more useful. This is visible in case of "House Pricing" case study also. So, Lasso regression will be preferred compared to Ridge as it has capability of penalizing insignificant features by making the coefficient of those as zero and in that way, model can be easily managed without compromising accuracy.

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?

- GrLivArea
- GarageCars
- OverallQual
- TotRmsAbvGrd
- YearRemodAdd

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?

The data cleaning, data preparation and EDA analysis are very important steps before model building. The missing values, outliers, multicollinearity need to be checked and handled properly. The data set should be scaled (Minmax Salar, Standardization etc.) appropriately before being used during model building. Otherwise, model behaviour will not be appropriate. The ultimate goal should be to create a model which is simple yet robust. It should perform equally well for both the training and test data set. It should maintain good trade-off between bias (it happens when the model is tool simple such that it is unable to learn. As a result, it performs badly for both training and test data set. This is called “underfitting”) and variance (It happens when the model is too complex such that it performs well in the training data set but fails to perform equally better in test data set. This is called “overfitting”). R2 needs to be tracked on training and test data set after model building to check whether the model is overfitting or underfitting. This trade-off should be maintained by applying appropriate regularization technique (Lasso or Ridge) wherever it is required to be applied.