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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans: For ridge regression, when we plot the curve between negative mean absolute error and alpha we see that as the value of alpha increases from 0 the error term decreases. The training error shows an upward trend when the value of alpha increases.
When the value of alpha is 2 the test error is at its minimum, so we decided to go with a value of alpha equal to 2 for our ridge regression. the R2 values for Train and Test (0.940309632530345, 0.899753600072928) match well, indicating an optimum model (RMSE 0.11485785595060997).
For lasso regression we have decided to keep a very small value that is 0.01, when we increase the value of alpha the model tries to penalize more and try to make the most of the coefficient value 0. At 0.01, the R2 values for Train and Test (0.8843875277173815, 0.8856627748606648) match well, indicating an optimum model (RMSE 0.13093187849733087).

The most important predictor variables after the changes have been implemented for ridge regression are mentioned below:

1. MSZoning_FV
2. MSZoning_RL
3.Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

The most important variable after the changes have been implemented for lasso regression are mentioned below:

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces

8. LotArea
9. LotArea
10. LotFrontage

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: It is important to regularize coefficients and improve the prediction accuracy with the decrease in variance making the model interpretably. Ridge regression uses a tuning parameter called lambda as the penalty. Residual sum or squares should be small by using the penalty. The penalty is lambda times the sum of squares of the coefficients, hence the coefficients that have greater values get penalized. As we increase the value of lambda the variance in the model is dropped and bias remains constant. Ridge regression includes all variables in the final model, unlike Lasso Regression. Lasso regression uses a tuning parameter called lambda as the penalty. As the lambda value increases Lasso shrinks the coefficient towards zero and it makes the variables exactly equal to 0. Lasso also does feature selection. When the lambda value is small it performs simple linear regression and as the lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model. In this case, ridge regression will be preferred over lasso.

Ridge regression (lambda = 2): the R2 values for Train and Test are 0.940309632530345 and 0.899753600072928 while RMSE is 0.11485785595060997.

Lasso regression (lambda = 0.01): the R2 values for Train and Test are 0.8843875277173815 and 0.8856627748606648 while RMSE is 0.13093187849733087.

The ridge is better as the R2 is higher.

Q3. 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?

Ans: The 5 most important predictor variables that will be excluded are :

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageArea

Q4. How can you make sure that a model is robust and generalizable? 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 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 an error in the model when the model is weak to learn from the data. High bias means the model is unable to learn details in the data. The model performs poorly on training and testing data.
Variance: Variance is an error in the model when the model tries to overlearn from the data. High variance
means the model performs exceptionally well on training data as it is very well trained on this of data but performs very poorly on testing data as it was unseen data for the model.
It is important to have a balance in Bias and Variance to avoid overfitting and under-fitting of data.