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: In the case of ridge regression, we can observe that the error term decreases as the value of alpha increases from 0 and that the train error is trending upward as the value of alpha increases when we plot the curve between negative mean absolute error and alpha.We chose to use a value of alpha equal to two for our ridge regression since the test error is at its lowest when alpha is at 2.
I have chosen to maintain a very modest value for the lasso regression, which is 0.01; as alpha increases, the model attempts to penalize more and attempt to make the majority of the coefficient values zero. It started out with a negative mean absolute error and alpha of 0.4.
The model will place more penalty on the curve and attempt to become more generalized, which is making the model more simpler and not thinking to fit every data point of the data set, when we double the value of alpha for our ridge regression no. We will take the value of alpha equal to 10.The graph shows that there is an increase in error for both test and train when alpha equals 10.
Similar to how we strive to penalise more of our model as we increase the alpha for lasso, more coefficients of the variable will be lowered to zero, and our r2 square value similarly declines as we increase.
Following the implementation of the adjustments for ridge regression, the following variables are the most crucial ones:
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
Following the implementation of the adjustments for lasso regression, the following variables are the most crucial ones:
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 also with the decrease in variance, and making the model interpretability.

- Ridge uses L2 regularization, which adds the squared magnitude of coefficients to the cost function. This tends to shrink all coefficients toward zero, but it rarely sets them exactly to zero.
- Lasso uses L1 regularization, which adds the absolute magnitude of coefficients to the cost function. Lasso has a feature selection property, meaning it can set some coefficients to exactly zero, effectively eliminating certain features.

The choice between Ridge and Lasso regression depends on the specific characteristics of your data and your modeling goals. It's not uncommon to try both methods and compare their performance to determine which one is more suitable for your task.

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?
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 simpler the model, the more resilient and generalizable it will be, even if its accuracy will decline. The Bias-Variance trade-off can also be used to understand it. There is more bias but less variation and greater generalizability in simpler models. In terms of accuracy, it implies that a reliable and generalizable model will function similarly on training and test data, meaning that there will be minimal differences in accuracy between the two sets of data.
Bias: When a model is unable to adequately learn from the data, it exhibits mistake. High bias indicates that the model cannot pick up on subtleties in the input. Model's performance on testing and training data is low.

Variance: Variance is the result of a model's attempt to learn too much from the data. High variance indicates that the model performs very well on training data since it has been trained on this type of data, but poorly on testing data because the testing data was not familiar to the model.
Maintaining equilibrium between Bias and Variance is crucial in order to prevent both overfitting and underfitting of data.