Subjective Questionnaire

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?

Answer: In the case of ridge regression:- When we plot the curve between negative mean absolute error and alpha we see that as the value of alpha increase from 0 the error term decrease and the train error is showing increasing trend when value of alpha increases .when the value of alpha is 1 the test error is minimum so we decided to go with value of alpha equal to 1 for our ridge regression.

For lasso regression I have decided to keep a very small value that is 0.0001, when we increase the value of alpha the model tries to penalize more and try to make most of the coefficient value zero. Initially it came as 0.001 in negative mean absolute error and alpha.

When we double or let's say 10x the value of alpha for our ridge regression and take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set .from the graph we can see that when alpha is 10 we get more error for both test and train. Similarly when we increase the value of alpha for lasso we try to penalize more of our model and more coefficient of the variable will be reduced to zero, when we increase the value of our r2 square also decreases. The most important variable after the changes has been implemented for ridge regression are as,

Ridge Regression

For alpha = 1

Variable	Coeff
constant	0.169
GrLivArea	0.168
1stFlrSF	0.147
OverallQual	0.128
2ndFlrSF	0.098
LotArea	0.081
OverallCond	0.081

RSquare = 0.9197

For alpha = 10

Variable	Coeff
constant	0.303
OverallQual	0.093
GrLivArea	0.089
1stFlrSF	0.073
2ndFlrSF	0.069
GarageArea	0.060
OverallCond	0.051

RSquare = 0.8953

Lasso Regression

For alpha = 0.0001

Variable	Coeff
GrLivArea	0.4159
constant	0.1529
OverallQual	0.1496
OverallCond	0.0901
GarageArea	0.0676
LotArea	0.0565
MSZoning_RL	0.0521

RSquare = 0.9121

For alpha = 0.001

Variable	Coeff
GrLivArea	0.2956
constant	0.2420
OverallQual	0.2112
GarageArea	0.0838
CentralAir_Y	0.0311
OverallCond	0.0273
BsmtExposure_Gd	0.0249

RSquare = 0.8561

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?

Answer: It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and make the model interpretable.

Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which are identified by cross validation. 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 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 is the absolute value of magnitude of coefficients which are identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it makes the variables exactly equal to 0. Lasso also does variable 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.

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?

Answer: Top 5 columns in Lasso regression,

Before

Variable	Coeff
GrLivArea	0.4159
constant	0.1529
OverallQual	0.1496
OverallCond	0.0901
GarageArea	0.0676
LotArea	0.0565

After

Variable	Coeff
1stFlrSF	0.4308
constant	0.2635
2ndFlrSF	0.1965
MSZoning_RL	0.0657
MSZoning_RH	0.0651
Neighborhood_NridgHt	0.0573

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?

Answer: Ensuring that a machine learning model is robust and generalizable involves several practices and considerations that help it perform well on unseen data and various scenarios. Here are some strategies to achieve robustness and generalizability:

1. **Use Sufficient and Diverse Data:** Ensure your model is trained on a diverse and representative dataset that covers various scenarios, edge cases, and real-world variations.
2. **Data Preprocessing and Cleaning:** Handle missing values, outliers, and inconsistencies in the data. Apply appropriate preprocessing techniques such as normalization, scaling, encoding categorical variables, etc.
3. **Feature Selection and Engineering:** Select relevant features that contribute to the model's predictive power. Create new features that might capture important information from the data.
4. **Cross-Validation:** Use techniques like k-fold cross-validation to assess the model's performance on different subsets of the data. This helps in estimating how the model might perform on unseen data.
5. **Hyperparameter Tuning:** Optimize model hyperparameters using techniques like grid search or random search, ensuring the best configuration for the model.
6. **Regularization:** Apply regularization techniques like L1, L2 regularization, or dropout (for neural networks) to prevent overfitting and improve generalization.