Advanced Linear Regression- Subjective Question Answers

Question-1

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: The optimal value of alpha for **Ridge Regression is 3**, **Ridge Regression with RFE is 0.7** and **Lasso Regression is 0.0001**.

When the alpha value is doubled in case where the new alpha values for **Ridge Regression is 6**, **Ridge Regression with RFE is 1.4** and **Lasso Regression is 0.0002**. The R2 Score of the models do not change, but there is a significant change in the model coefficient values. If there is a huge variation in the alpha then the coefficients have high changes, if we double the alpha values on Ridge, we see that there is a huge change in the coefficient. The same if we double the Ridge with RFE and Lasso alpha since there is not such a huge jump in the alpha there is a very small change in the Coefficient in Ridge with RFE and Lasso. The top 20 variables are below for each of the model.

	Ridge Co-Efficient
GrLivArea	0.088821
1stFlrSF	0.075696
2ndFlrSF	0.064166
OverallQual_Excellent	0.057961
FullBath	0.054315
TotRmsAbvGrd	0.050534
GarageCars	0.049321
RoofMatl_WdShngl	0.049224
MSZoning_RL	0.042335
OverallQual_Very Good	0.039875
BsmtFullBath	0.038787
MSZoning_FV	0.038768
MSZoning_RH	0.038401
SaleType_ConLD	0.037654
Neighborhood_NoRidge	0.035947
Neighborhood_Crawfor	0.035775
LotArea	0.034821
Fireplaces	0.034639
Neighborhood_StoneBr	0.033259
Neighborhood_NridgHt	0.032716

	Ridge Co-Efficient(Double Alpha)
GrLivArea	0.071064
1stFlrSF	0.059212
2ndFlrSF	0.054236
FullBath	0.053455
TotRmsAbvGrd	0.052886
OverallQual_Excellent	0.052510
GarageCars	0.044415
OverallQual_Very Good	0.039257
Fireplaces	0.038159
BsmtFullBath	0.034838
Neighborhood_NoRidge	0.033501
Neighborhood_Crawfor	0.033083
RoofMatl_WdShngl	0.032168
GarageArea	0.030213
MSZoning_RL	0.029692
Neighborhood_NridgHt	0.029574
TotalBsmtSF	0.029045
WoodDeckSF	0.027797
Neighborhood_StoneBr	0.027719
CentralAir_Y	0.026508

	Ridge RFE Co-Efficient
OverallQual_Very Excellent	0.160601
LotArea	0.157604
OverallQual_Excellent	0.149425
OverallQual_Very Good	0.103786
MasVnrArea	0.082816
Fireplaces	0.081976
Neighborhood_NoRidge	0.067800
Exterior1st_Stone	0.064117
OverallQual_Good	0.061826
WoodDeckSF	0.053860
SaleType_ConLD	0.053218
HalfBath	0.052739
3SsnPorch	0.050846
Neighborhood_Crawfor	0.049542
Neighborhood_NridgHt	0.047833
Neighborhood_Somerst	0.044102
GarageCond_Po	0.040906
Exterior1st_BrkFace	0.039173
Neighborhood_SWISU	0.036964
Condition1_PosA	0.036327

	Lasso Co-Efficient
GrLivArea	0.379188
GarageCars	0.066876
OverallQual_Excellent	0.066320
MSZoning_RL	0.056131
MSZoning_RH	0.054028
RoofMatl_WdShngl	0.053900
MSZoning_FV	0.051376
FullBath	0.046097
OverallQual_Very Good	0.042018
BsmtFullBath	0.041144
Neighborhood_Crawfor	0.040214
MSZoning_RM	0.040114
OverallQual_Very Excellent	0.032514
Neighborhood_NridgHt	0.032152
Neighborhood_Somerst	0.032047
SaleType_ConLD	0.030998
Neighborhood_ClearCr	0.029895
LotArea	0.029212
Fireplaces	0.028816
TotRmsAbvGrd	0.027056

	Ridge RFE Co-Efficient(Double Alpha)
OverallQual_Very Excellent	0.147363
OverallQual_Excellent	0.142761
LotArea	0.122441
OverallQual_Very Good	0.101420
Fireplaces	0.084594
MasVnrArea	0.081357
Neighborhood_NoRidge	0.068785
OverallQual_Good	0.060560
WoodDeckSF	0.054956
HalfBath	0.053101
Exterior1st_Stone	0.051009
Neighborhood_NridgHt	0.050267
Neighborhood_Crawfor	0.048941
3SsnPorch	0.044683
Neighborhood_Somerst	0.044438
SaleType_ConLD	0.043195
Exterior1st_BrkFace	0.039130
GarageCond_Po	0.035646
RoofMatl_WdShngl	0.033522
Condition1_PosA	0.033209

	Lasso Co-Efficient(Double Alpha)
GrLivArea	0.338557
GarageCars	0.069525
OverallQual_Excellent	0.069410
FullBath	0.049154
OverallQual_Very Good	0.044444
BsmtFullBath	0.039097
Neighborhood_Crawfor	0.037465
TotRmsAbvGrd	0.034756
Fireplaces	0.034358
Neighborhood_Somerst	0.033773
Neighborhood_NridgHt	0.032002
Exterior1st_BrkFace	0.027158
BsmtExposure_Gd	0.027082
Neighborhood_NoRidge	0.025778
CentralAir_Y	0.025464
Neighborhood_ClearCr	0.024240
WoodDeckSF	0.022208
OverallQual_Very Excellent	0.021812
Neighborhood_StoneBr	0.019207
HalfBath	0.018469

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?

Answer: We can see here that Lasso and Ridge Regression is giving us better results. We can also see that the Regularized Models do lot better than the OLS Models in term of predicting the housing prices. Regularized models are better in generalizing the data than the OLS Models as which can be seen in the Model Building Process. We can see that:

- Linear Regression is giving us 0.96 on train and -9.73 on test, which is unable to predict.
- Ridge Regression is giving us 0.93 on train and 0.88 on test with Alpha = 3.
- Ridge Regression with RFE is giving us 0.84 on train and 0.77 on test with Alpha = 0.7.
- Lasso Regression is giving us 0.93 on train and 0.88 on test with Alpha = 0.0001.

Hence, we should choose **Lasso Regression Model as final model** since it is able **to generalize the model well and it sets the unwanted variable to 0 which are not relevant**.

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?

Answer: The current five most important predictor variables in the Lasso Regression Model are :

	Lasso Co-Efficient
GrLivArea	0.379188
GarageCars	0.066876
OverallQual_Excellent	0.066320
MSZoning_RL	0.056131
MSZoning_RH	0.054028

When a model is created excluding the above 5 predictor variables, the following are the new five most important predictor variables in the Lasso Regression:

	Lasso Co-Efficient
1stFlrSF	0.285962
2ndFlrSF	0.141967
RoofMatl_WdShngl	0.063639
GarageArea	0.058407
FullBath	0.051156

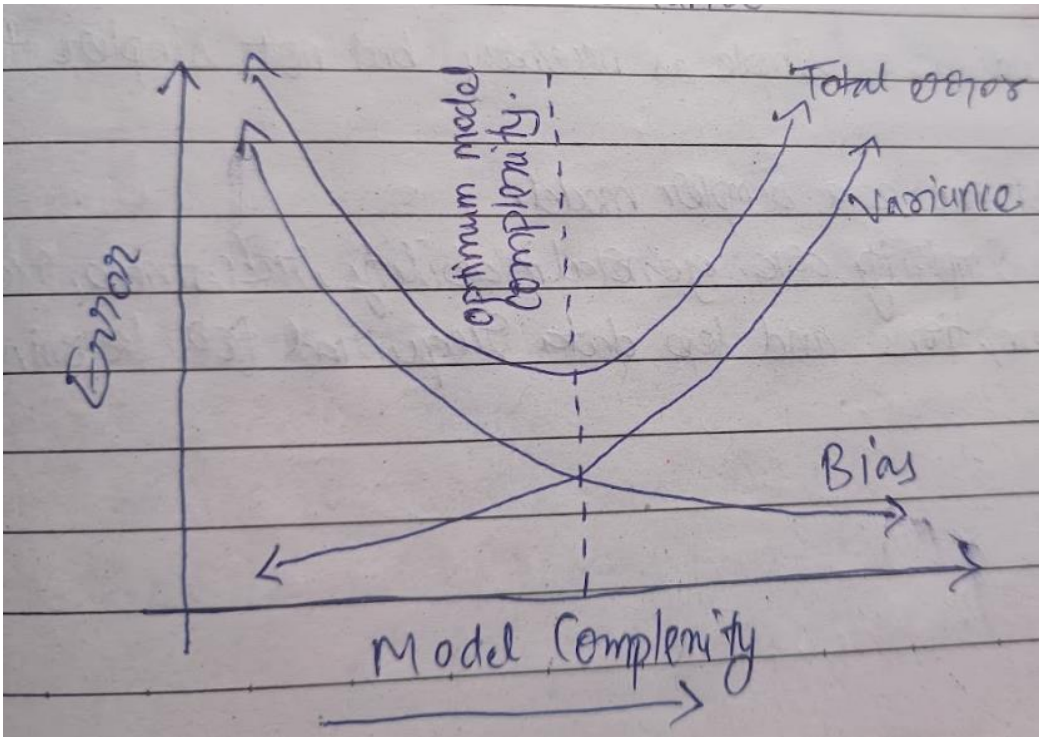
Question-4

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?

Answer: A model is robust and generlisable when its simple, by simple I don't mean that it should be really simple because a model being simpler would only have High Bias and Low Variance, and if we build more complex model then it would only result in Low Bias and High Variance, in which either of the case we have really high error rate and the model would just overfit. When you have a really complex model and you bring in some new data to the model you have to tune the model for every new data you pass it to the model, when you have a really simple model it will not be able to predict the output on very complex data. But in real life we cannot have a mode which has low variance and low bias hence we have to perform a tradeoff between the Variance and the Bias.

In Advance Regression where we deal with complex regression of nonlinear data, we have to use something called as the Regularization. Regularization helps in managing the complex model by shrinking the model coefficients towards 0 which are not needed. This is done by adding something called as the penalty coefficient in the model which puts penalty if the model goes too complex.

Whenever building model we have to always keep in mind Occam's Razor:
1. A model should be as simple as necessary but not simple that that.
2. When in doubt, choose a simpler model.
3. Advantages of simplicity are generalizability, robustness, requirement of a few assumptions and less data required for learning.



We have to make sure that we have Low Bias and Low Variance in order to have a simple model which will also have less error rate.