Advance Regression Assignment

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:

Optimal value of alpha (λ) for Ridge and Lasso are given below.

Regression Type	Optimal Value λ
Ridge	**35**
Laso	**0.002**

What if we double the alpha:

Change in Model (both Ridge and Lasso):

		R2 Score				
Model	**λ**	**Train**	**Test**	**RSS**	**MSE**	**RMSE**
Ridge with RFE (20 Vars)	**35**	81.5%	83.8%	213	0.11	0.33
Ridge with RFE (20 Vars)	35 *2	80.0%	83.4%	210	0.11	0.34
Lasso with RFE (20 Vars)	**0.002**	82.5%	82.6%	156	0.12	0.34
Lasso with RFE (20 Vars)	0.002 * 2	82.0%	82.7%	147	0.12	0.34

Change in Predictor Variable in descending order (* Most Significant at the top) for both Ridge and Lasso:

Ridge (Lambda=35)	Ridge (Lambda=35*2)	Lasso (Lambda =0.002)	Lasso (Lambda =0.002 *2)
Estimates	Estimates	Estimates	Estimates
GarageCars_3	GrLivArea	KitchenQual_Fa	GarageCars_3
GrLivArea	GarageCars_3	KitchenQual_TA	GrLivArea
OverallQual_Others	OverallQual_Others	GarageCars_3	KitchenQual_Fa
FullBath_3	ExterQual_TA	GrLivArea	OverallQual_Others
Fireplaces_2	BsmtFinType1_GLQ	OverallQual_Others	KitchenQual_TA
SaleCondition_Partial	TotalBsmtSF	KitchenQual_Gd	BsmtExposure_Gd
CentralAir_Y	SaleCondition_Partial	FullBath_3	Fireplaces_2
ExterQual_TA	BsmtExposure_Gd	Fireplaces_2	CentralAir_Y
BsmtExposure_Gd	Fireplaces_2	BsmtExposure_Gd	KitchenQual_Gd
BsmtFinType1_GLQ	FireplaceQu_No Fireplace	LotConfig_CulDSac	ExterQual_TA
Functional_Typ	FullBath_3	CentralAir_Y	FullBath_3
Condition1_Others	CentralAir_Y	ExterQual_TA	LotConfig_CulDSac
LotConfig_CulDSac	BsmtQual_TA	Functional_Typ	SaleCondition_Partial
OverallQual_8	Condition1_Others	Neighborhood_Edwards	Functional_Typ
Neighborhood_Edwards	OverallQual_8	SaleCondition_Partial	OverallQual_8
BldgType_Others	BldgType_Others	OverallQual_8	BsmtFinType1_Unf
TotalBsmtSF	KitchenQual_TA	BsmtFinType1_Unf	Neighborhood_Edwards
FireplaceQu_No Fireplace	Functional_Typ	Foundation_PConc	Foundation_PConc
BsmtQual_TA	Neighborhood_Edwards	Condition1_Others	Condition1_Others
KitchenQual_TA	LotConfig_CulDSac	BldgType_Others	BldgType_Others

Some of the other effects on model if we increase the value of Lambda:

If we double the value of alpha then:

a) Complexity of the model will also increase
b) Bias will become **High** and Variance will become **Low**
c) Model start underfitting

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:

For this assignment, we will choose Lasso Regression.

As the number of features are high and apart from regularization Lasso is also used for feature selecting.

As Lasso convert beta estimates to zero for those variables which are less useful for the model.

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:

If we need to remove five most important predictor (top 5 predictor),

Lasso (Lambda =0.002)	
Feature	Estimates
KitchenQual_Fa	-0.58731
KitchenQual_TA	-0.49491
GarageCars_3	0.463377
GrLivArea	0.449296
OverallQual_Others	0.417889

If the above top 5 features are unavailable in the data, then the new top features would be:

If due to data unavailability, my new five most predictor would be:

Lasso (Lambda =0.002)	
Features	Estimates
FullBath_3	0.652903
TotRmsAbvGrd_Others	0.335327
BsmtQual_TA	-0.325666
Fireplaces_2	0.314684
ExterQual_TA	-0.308238

Top 5 features for Lasso Reg. post removing previous top features.

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:

The final model is robust and generalisable because due to below reason:

a) We have used Regularization with optimal parameter to avoid overfitting.
b) Lasso Regularization also helped us in feature selection and the model build with less features this will again reduce the complexity of the model and thus robust to unseen data.
c) We have also compared R2 score for Train and Test data and both the results are not varying much, so the model should also work good for new/unseen data.
d) We have also validated L.R model assumption and both residual assumptions are almost satisfied.
    - In the Residual Vs Predicted plot the residuals are randomly spread around 0-axis i.e. seems to be constant and not showing any pattern i.e. there is no sign of heteroscedasticity in the model
    - Residual almost follow Normal Distribution.