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?

- The Optimal value of alpha for ridge = 2 and for lasso = 0.0001. With these alphas ,R2 of the model = 0.83.
- Doubling the alpha values using Ridge and Lasso , the optimal value R2 around 0.82 but there is slight change in co-efficient values. These values are derived in assignment submission (code from jupyter notebook).
- Please find the co-efficients are listed below in table between Normal & Doubled alpha for both Ridge & Lasso.

Ridge Co-Efficient	**Ridge Doubled Alpha Co-Efficient**
Total_sqr_footage 0.459042	Total_sqr_footage 0.405689
GarageArea 0.215074	GarageArea 0.204762
TotRmsAbvGrd 0.155003	TotRmsAbvGrd 0.164504
LotArea 0.125733	LotArea 0.113824
OverallCond 0.101723	OverallCond 0.096920
SaleType_CWD 0.098161	SaleType_CWD0.080306
LotFrontage0.088044	LotFrontage 0.078705
HouseStyle_2.5Unf 0.075202	Total_porch_sf 0.073541
Total_porch_sf 0.072676	CentralAir_Y 0.072929
RoofMatl_WdShngl 0.072215	HouseStyle_2.5Unf 0.071178
CentralAir_Y 0.069394	RoofMatl_WdShngl 0.065830
SaleType_Con 0.062119	LandContour_HLS 0.055128
LandContour_HLS 0.060496	KitchenQual_Ex0.051273
Condition2_Norm 0.052773	SaleType_Con 0.042650
Condition2_PosA 0.051670	BsmtQual_Ex 0.039141
KitchenQual_Ex 0.049366	Condition2_Norm 0.038724
HouseStyle_1.5Unf 0.040599	MSSubClass_70 0.037222
MSSubClass_70 0.038129	PavedDrive_Y 0.036091
BsmtQual_Ex 0.036923	Neighborhood_Veenker 0.034917
Neighborhood_Veenker 0.036281	Condition2_PosA 0.033983

LASSO

Lasso Co-Efficient	**Lasso Doubled Alpha Co-Efficient**
Total_sqr_footage 0.543854	Total_sqr_footage 0.537830
GarageArea 0.217039	GarageArea 0.208540
TotRmsAbvGrd 0.137567	TotRmsAbvGrd 0.149942
LotArea 0.111823	OverallCond 0.085500
OverallCond 0.096323	CentralAir_Y 0.075181
CentralAir_Y 0.072815	Total_porch_sf 0.069284
Total_porch_sf 0.071217	LotArea 0.064586
HouseStyle_2.5Unf 0.056901	KitchenQual_Ex0.045076
SaleType_CWD0.051616	BsmtQual_Ex 0.039116
LandContour_HLS 0.047457	SaleCondition_Partial 0.034450
KitchenQual_Ex0.044304	LandContour_HLS 0.031449
LotFrontage 0.040095	HouseStyle_2.5Unf 0.030696
BsmtQual_Ex 0.037305	MSSubClass_70 0.029321
SaleCondition_Partial 0.034629	PavedDrive_Y 0.025527
MSSubClass_70 0.034425	ExterQual_Ex 0.024025
PavedDrive_Y 0.027893	Condition1_Norm 0.023043
Condition1_Norm 0.026634	BsmtCond_TA 0.021770
RoofMatl_WdShngl 0.025823	OpenPorchSF 0.019072

Alley_Pave 0.025242 OpenPorchSF 0.024802	Alley_Pave 0.017456 MasVnrType_Stone 0.015927

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

Based on the derived facts on optimum lambda value in both regression model.

Ridge Regression Model values	Lasso Regression Model values
Lambda = 1	Lambda = 0.0002
Mean Squared Error – 0.0067	Mean Squared Error – 0.0068
R2 value = 0.83	R2 value = 0.83

Based on value statistics are almost same between Ridge & Lasso , however Lasso helps in feature reduction (as coefficients of them are zero).
So I will choose Lasso Regression model for this prediction assignment for final 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

The most important predictor variables in Lasso model (after doubling alpha value).

- Total_sqr_footage 0.537830
- GarageArea 0.208540
- TotRmsAbvGrd 0.149942
- OverallCond 0.085500
- CentralAir_Y 0.075181

After removing the top five predictor above, built another Lasso model where R2 for this model = 0.72 and MSE = 0.0116

New Top five predictors are listed below.

LotArea	0.305439
LotFrontage	0.253582
Total_porch_sf	0.143257
BsmtFullBath	0.104663
HouseStyle_2.5Unf	0.102524

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?

As per **Occam's Razor**

- A model should be as simple as necessary but not simpler than that.
- When in doubt, choose a simpler model.
- Advantages of simplicity are generalisability, robustness, requirement of a few assumptions and less data required for learning

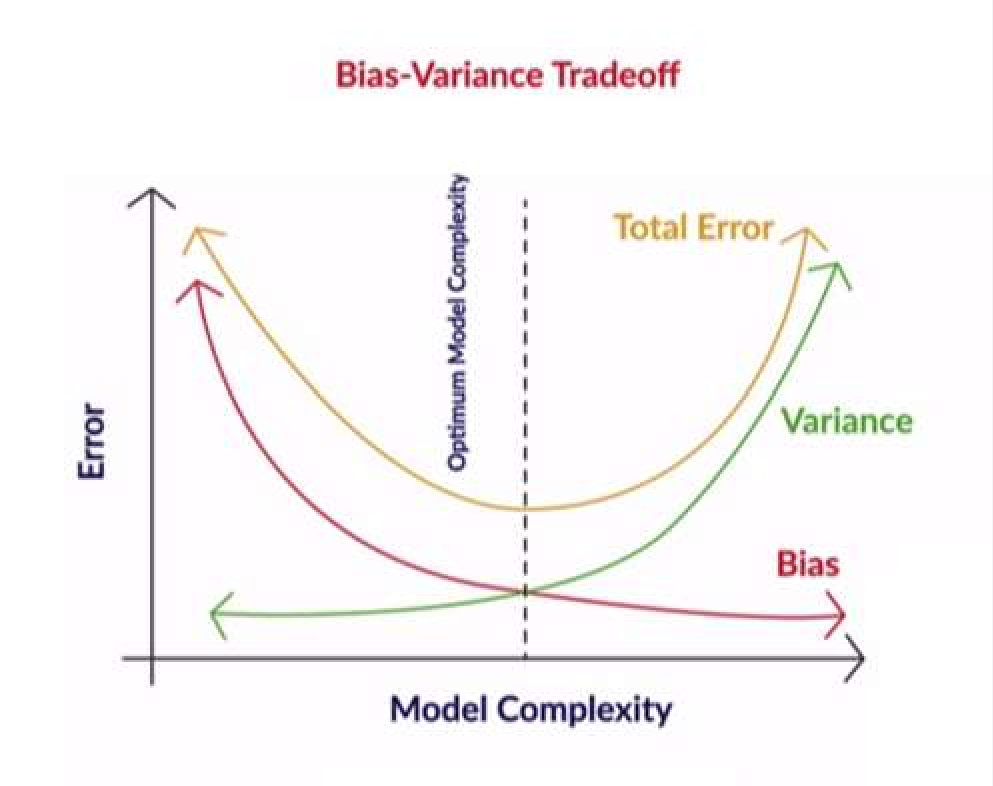
Bias-Variance Tradeoff

- Bias measures how accurately a model can describe the actual task at hand.
- Variance measures how flexible the model is with respect to changes in the training data.
- As complexity increases, bias reduces and variance increases, and we aim to find the optimal point where the total model error is the least.

Regularization

- Regularization helps model perform well with unseen data while identifying necessary underlying patterns in it. By adding a penalty term to the cost function used by OLS.
- Ridge and Lasso regression methods, which both allow some bias to get a significant decrease in variance, thereby pushing the model coefficients towards 0.

- In Lasso, some of these coefficients become 0, thus resulting in model selection and, hence, easier interpretation, particularly when the number of coefficients is very large.
- Ideally, we want to reduce both bias and variance because the expected total error of a model is the sum of the errors in bias and variance, as shown in the figure given below.


Bias-Variance Tradeoff
Optimum Model Complexity
Total Error
Error
Variance
Bias
Model Complexity