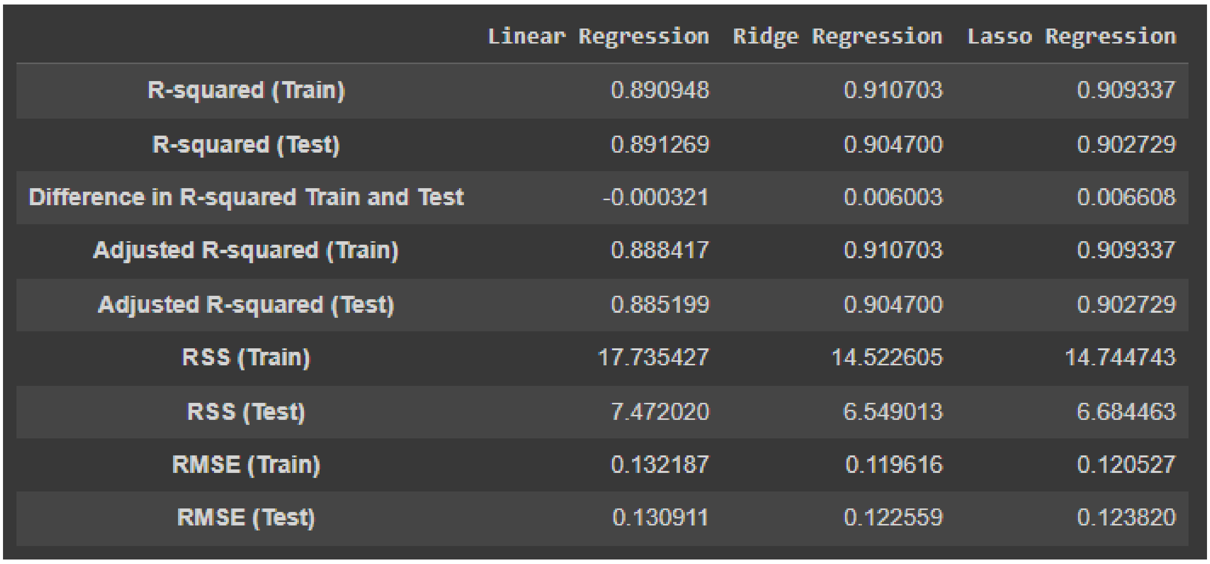
Assignment Part -II - Subjective Questions and Answers

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

I. The optimal value of ALPHA we got in case of Ridge and Lasso is:
   - ✓ Ridge - 20.0
   - ✓ Lasso - 0.001

II. The changes in the model if we choose to double the value of alpha for both ridge and lasso are:

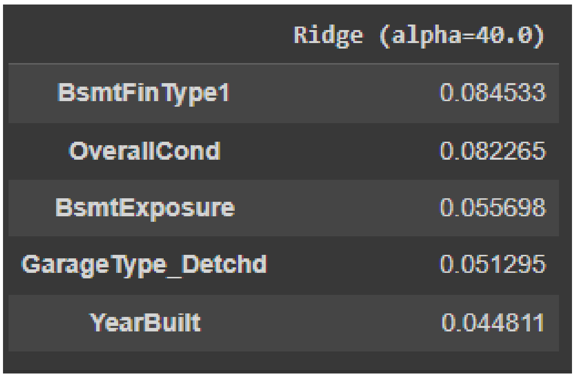


	Linear Regression	Ridge Regression	Lasso Regression
R-squared (Train)	0.890948	0.910703	0.909337
R-squared (Test)	0.891269	0.904700	0.902729
Difference in R-squared Train and Test	-0.000321	0.006003	0.006608
Adjusted R-squared (Train)	0.888417	0.910703	0.909337
Adjusted R-squared (Test)	0.885199	0.904700	0.902729
RSS (Train)	17.735427	14.522605	14.744743
RSS (Test)	7.472020	6.549013	6.684463
RMSE (Train)	0.132187	0.119616	0.120527
RMSE (Test)	0.130911	0.122559	0.123820

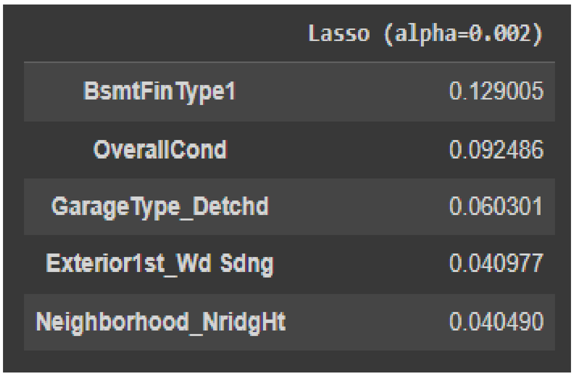
The difference of r2 between train and test is less in lasso as compared to ridge.

In addition, Lasso is advantageous in feature reduction as it can shrink the coefficient value of one of its features towards zero. This enhances model interpretability by considering the magnitude of the coefficients. Therefore, Lasso holds a superior position over Ridge.

III. The most important predictor variables (top 5) after the changes are implemented are:



	Ridge (alpha=40.0)
BsmtFinType1	0.084533
OverallCond	0.082265
BsmtExposure	0.055698
GarageType_Detchd	0.051295
YearBuilt	0.044811



	Lasso (alpha=0.002)
BsmtFinType1	0.129005
OverallCond	0.092486
GarageType_Detchd	0.060301
Exterior1st_Wd Sdng	0.040977
Neighborhood_NridgHt	0.040490

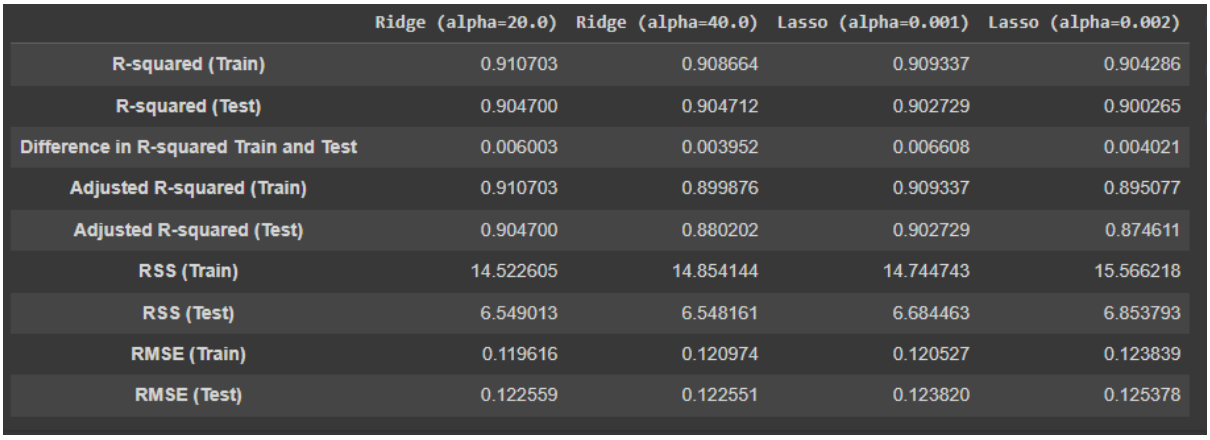
Question 2: You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer

The model we will choose to apply will depend on the use case.

- If we have too many variables and one of our primary goals is feature selection, then we will use Lasso.
- If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use Ridge Regression.

As per below matrices, the difference of r2 between train and test is less in lasso as compared to ridge and when alpha increasing, the difference is b/w R-squared is also decreasing. But RMSE values are increasing in both Ridge and Lasso.

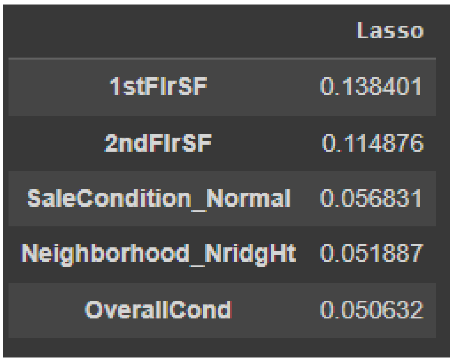


	Ridge (alpha=20.0)	Ridge (alpha=40.0)	Lasso (alpha=0.001)	Lasso (alpha=0.002)
R-squared (Train)	0.910703	0.908664	0.909337	0.904286
R-squared (Test)	0.904700	0.904712	0.902729	0.900265
Difference in R-squared Train and Test	0.006003	0.003952	0.006608	0.004021
Adjusted R-squared (Train)	0.910703	0.899876	0.909337	0.895077
Adjusted R-squared (Test)	0.904700	0.880202	0.902729	0.874611
RSS (Train)	14.522605	14.854144	14.744743	15.566218
RSS (Test)	6.549013	6.548161	6.684463	6.853793
RMSE (Train)	0.119616	0.120974	0.120527	0.123839
RMSE (Test)	0.122559	0.122551	0.123820	0.125378

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

We dropped the top 5 most important predictor variables in the lasso model and again created again model and got the below five most important predictor variables:



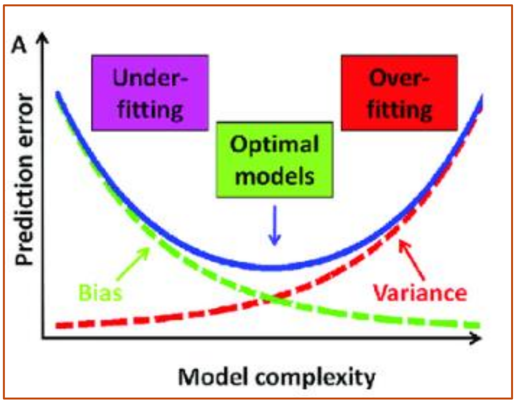
	Lasso
1stFlrSF	0.138401
2ndFlrSF	0.114876
SaleCondition_Normal	0.056831
Neighborhood_NridgHt	0.051887
OverallCond	0.050632

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?

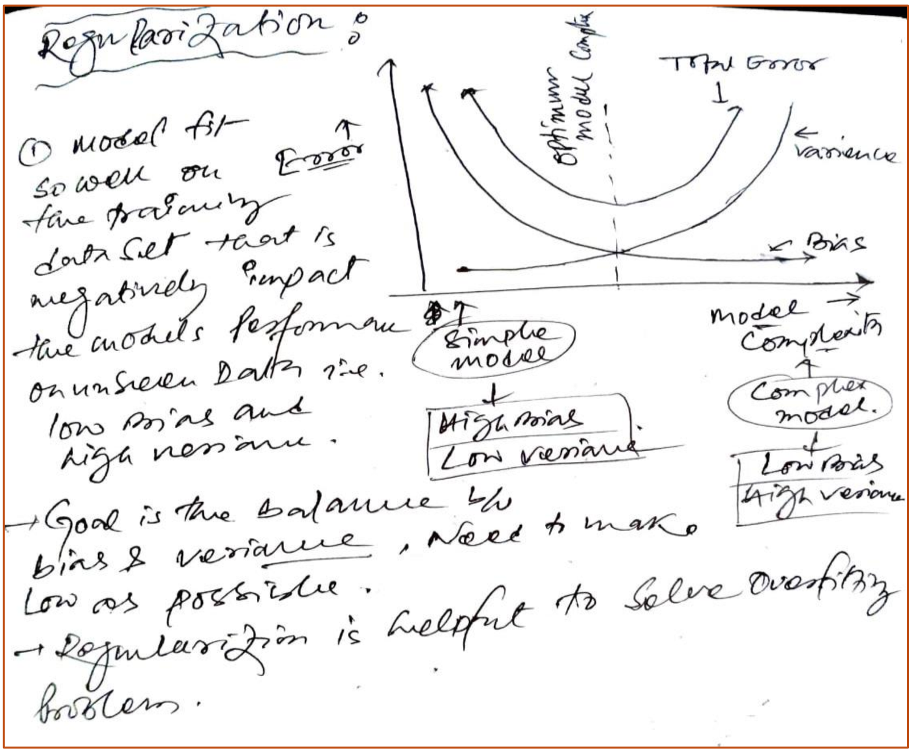
Answer

A robust model is minimally affected by variations in the data. Generalizability refers to a model's ability to adapt effectively to new, unseen data from the same distribution without overfitting. Below are few important points to ensure robust and generalizable models:

- **Model Simplicity**: A more generic model is simpler, which implies reduced complexity. Simpler models with fewer parameters are less likely to overfit to noise in the training data.
- **Bias-Variance Trade-off**: The Bias-Variance trade-off dictates that simpler models have higher bias but lower variance, making them more generalizable. Complex models exhibit low bias but high variance, risking overfitting. So, finding a balance is important for reliability.



- **Avoiding Overfitting**: Overfit models memorize training data patterns but may fail on new data. Regularization, cross-validation, and early stopping help prevent overfitting.
- **Testing on Unseen Data**: If a model performs well on new, unseen data, it is more likely to be robust and generalizable.
- **Regularization Techniques**: Techniques like Ridge Regression and Lasso, helps manage model complexity by shrinking coefficients towards zero.



Regularization :

① model fit so well on the training data set that is negatively impact the models performance on unseen Data i.e. low bias and high variance.

→ Goal is the balance b/w bias & variance, Need to make low as possible.

→ Regularization is helpful to solve overfitting problem.

Error
Optimum model Complx
Total Error
variance
Bias
model Complexity
Simple model
High bias
Low variance
Complex model.
Low bias
High variance

Implications for Model Accuracy:

- Ensuring robustness may lead to more accurate predictions on new data.
- Model simplification may cause a slight decrease in accuracy, justified for better generalizability.
- A robust model delivers consistent and reliable predictions in diverse scenarios.
- Cross-validation provides realistic estimates of model accuracy on unseen data.
- Achieving a balance in bias-variance trade-off maintains accuracy while enhancing generalizability.