HOUSING SALE PREDICTION ASSESSMENT

By

Varun Shenoy

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?

1. Optimum Value got from the built model

Lambda values (Alpha) for Ridge and Lasso regression are:

- Lambda for Ridge: 5
- Lambda for Lasso: 0.0001

2. Doubling the Lambda value to 10 and 0.0002

Ridge Lambda 5 (important variables)	Lasso Lambda 0.0001 (important variables)
Test R2 Score - 0.892	Test R2 Score- 0.893
OverallQual_Excellent 0.143510	MSZoning_RH 0.237041
Neighborhood_StoneBr 0.133981	MSZoning_RL 0.212952
Neighborhood_Crawfor 0.104176	MSZoning_FV 0.198215
OverallQual_Very Good 0.092749	OverallQual_Excellent 0.189126
Neighborhood_NridgHt 0.083310	Neighborhood_StoneBr 0.179699

Ridge Lambda 10 (important variables)	Lasso Lambda 0.0002 (important variables)
Test R2 Score - 0.891	Test R2 Score- 0.892
OverallQual_Excellent 0.118201	OverallQual_Excellent 0.187028
Neighborhood_StoneBr 0.102800	Neighborhood_StoneBr 0.168095
Neighborhood_Crawfor 0.091551	Neighborhood_Crawfor 0.114941
OverallQual_Very Good 0.082868	MSZoning_RH 0.105237
Neighborhood_NridgHt 0.068983	OverallQual_Very Good 0.104521

- The test R2 scores are slightly lower when the Alpha is doubled
- The Lasso regression top variables are now similar to ridge regression.
- Coefficients of ridge regression doesn't change on doubling of the alpha coefficients.

Q2. YOU HAVE2 DETERMINED THE OPTIMAL VALUE OF LAMBDA FOR RIDGE AND LASSO REGRESSION DURING THE ASSIGNMENT. NOW, WHICH ONE WILL YOU CHOOSE TO APPLY AND WHY?

- I choose to implement and interpret with lasso regression for 2 reasons
  1. In this case it has better score over Ridge (slightly)
  2. It selects features on its own without affecting the model accuracy hence simplifying the model.

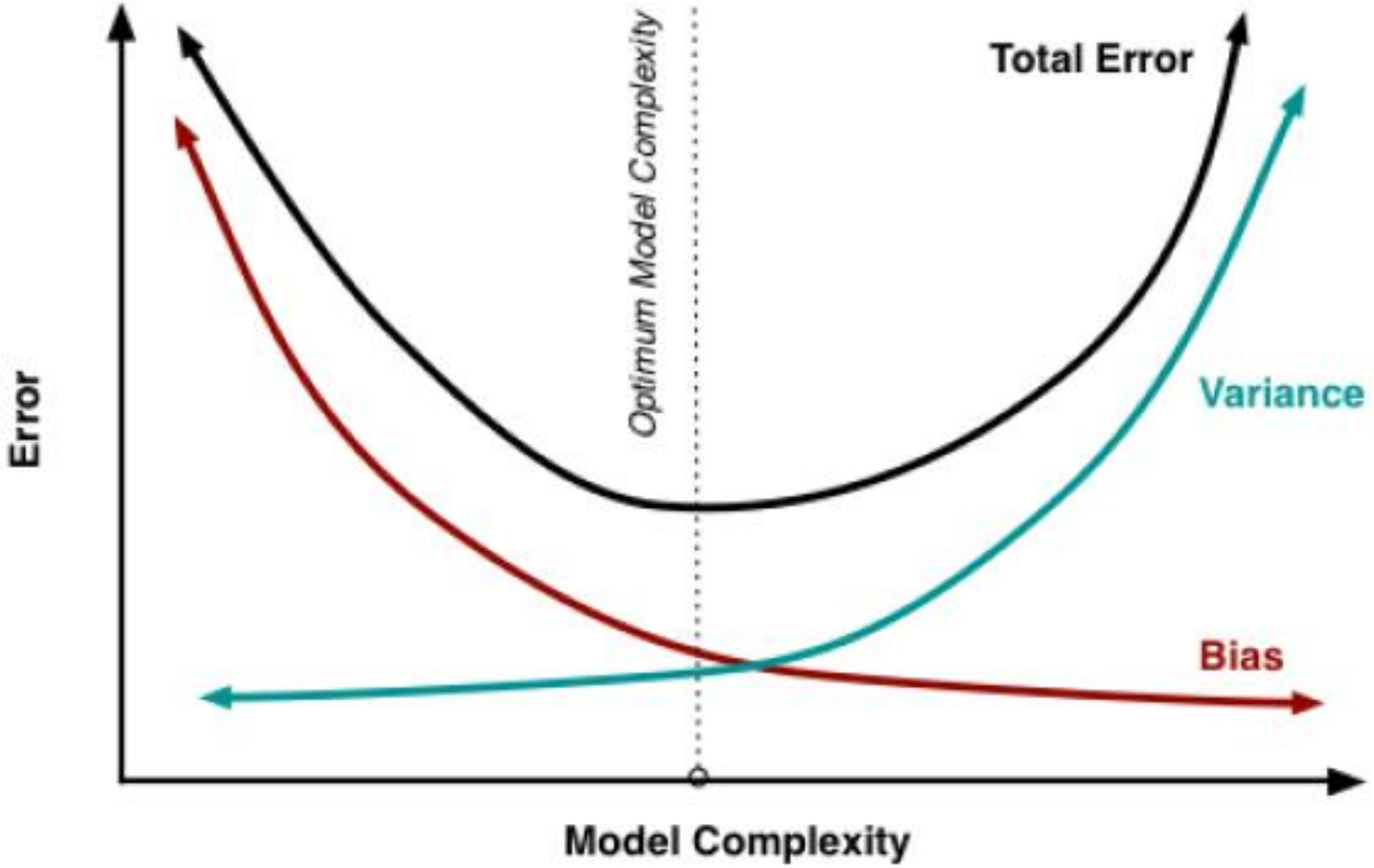
Q3. AFTER BUILDING THE MODEL, YOU REALIZED THAT THE FIVE MOST IMPORTANT PREDICTOR VARIABLES IN THE LASSO MODEL ARE NOT AVAILABLE IN THE INCOMING DATA. YOU WILL NOW HAVE TO CREATE ANOTHER MODEL EXCLUDING THE FIVE MOST IMPORTANT PREDICTOR VARIABLES. WHICH ARE THE FIVE MOST IMPORTANT PREDICTOR VARIABLES NOW?

Lasso Lambda 0.0001 (Before deleting important variables)	Lasso Lambda 0.0001 (After deleting important variables)
Test R2 Score - 0.892	Test R2 Score- 0.883
OverallQual_Excellent 0.143510	Neighborhood_Crawfor 0.097808
Neighborhood_StoneBr 0.133981	GrLivArea 0.087863
Neighborhood_Crawfor 0.104176	Neighborhood_NridgHt 0.085094
OverallQual_Very Good 0.092749	Exterior1st_BrkFace 0.084292
Neighborhood_NridgHt 0.083310	SaleCondition_Partial 0.079977

- Tip in Test R2 score is seen after removing the top 5 important variables (although slight dip)
- The Lasso regression alpha was kept 0.0001.
- Comparison table given. Refer ipynb for code
- The coefficient values too are comparitively less

Q4. How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why??

- In general overfitting is a great way to check for model. i.e how will the model perform on unseen data
- A model but be simple but not very simple that it underfits the data and not too complex that it overfits (memorize the training set) . It must be of the right balance
- Simple models have high bias and low variance, complex models on the other hand has low bias and high variance.
- Regularizing the model manages the complexity by bringing the variable coefficients close to 0. Complex models are penalized and hence the R2 score is maintained by compromising Bias to reach optimum position and minimum total error.
- Bias-Variance trade off graph is shown in next page highlighting the point of optimum complexity
- If the model is at optimum complexity point or close to it the model will be robust and generalizable.


Optimum Model Complexity
Total Error
Variance
Bias
Error
Model Complexity