Problem Statement - Part II

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?

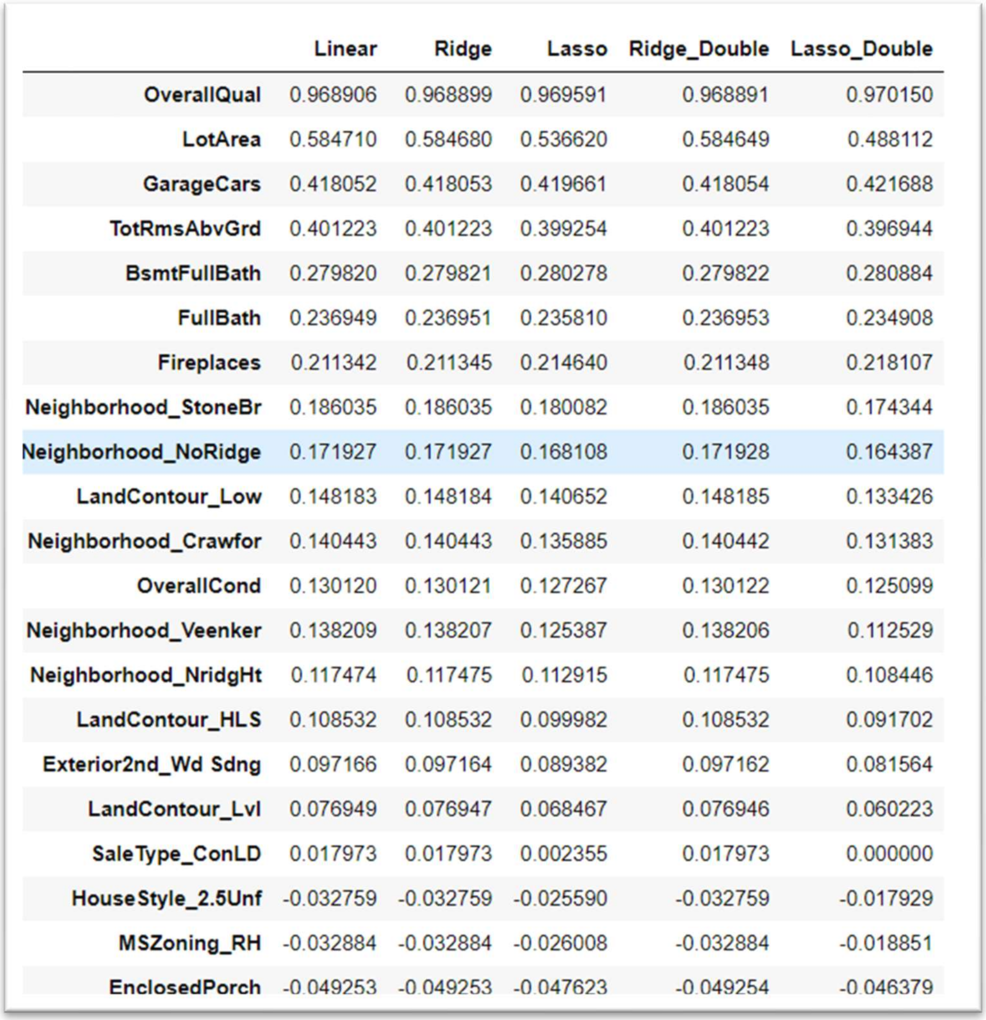
Answer:

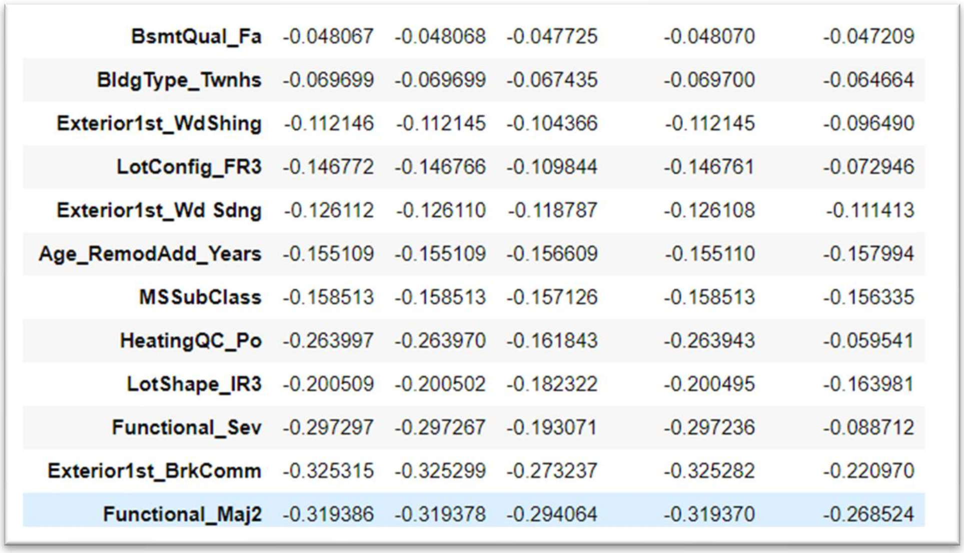
The optimal value for Ridge Regression is : 0.0001
The optimal value for Lasso Regression is : 0.0001

After we double the value of alpha both ridge and lasso values is

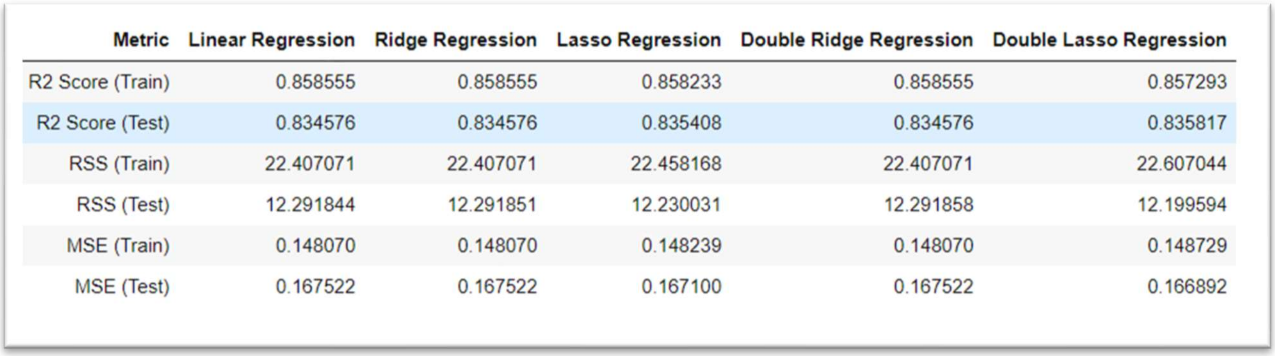
Doubled alpha values of Ridge is 0.0002 and Lasso is 0.0002

For details problem is solved in notebook.





	Linear	Ridge	Lasso	Ridge_Double	Lasso_Double
OverallQual	0.968906	0.968899	0.969591	0.968891	0.970150
LotArea	0.584710	0.584680	0.536620	0.584649	0.488112
GarageCars	0.418052	0.418053	0.419661	0.418054	0.421688
TotRmsAbvGrd	0.401223	0.401223	0.399254	0.401223	0.396944
BsmtFullBath	0.279820	0.279821	0.280278	0.279822	0.280884
FullBath	0.236949	0.236951	0.235810	0.236953	0.234908
Fireplaces	0.211342	0.211345	0.214640	0.211348	0.218107
Neighborhood_StoneBr	0.186035	0.186035	0.180082	0.186035	0.174344
Neighborhood_NoRidge	0.171927	0.171927	0.168108	0.171928	0.164387
LandContour_Low	0.148183	0.148184	0.140652	0.148185	0.133426
Neighborhood_Crawfor	0.140443	0.140443	0.135885	0.140442	0.131383
OverallCond	0.130120	0.130121	0.127267	0.130122	0.125099
Neighborhood_Veenker	0.138209	0.138207	0.125387	0.138206	0.112529
Neighborhood_NridgHt	0.117474	0.117475	0.112915	0.117475	0.108446
LandContour_HLS	0.108532	0.108532	0.099982	0.108532	0.091702
Exterior2nd_Wd Sdng	0.097166	0.097164	0.089382	0.097162	0.081564
LandContour_Lvl	0.076949	0.076947	0.068467	0.076946	0.060223
SaleType_ConLD	0.017973	0.017973	0.002355	0.017973	0.000000
HouseStyle_2.5Unf	-0.032759	-0.032759	-0.025590	-0.032759	-0.017929
MSZoning_RH	-0.032884	-0.032884	-0.026008	-0.032884	-0.018851
EnclosedPorch	-0.049253	-0.049253	-0.047623	-0.049254	-0.046379
BsmtQual_Fa	-0.048067	-0.048068	-0.047725	-0.048070	-0.047209
BldgType_Twnhs	-0.069699	-0.069699	-0.067435	-0.069700	-0.064664
Exterior1st_WdShing	-0.112146	-0.112145	-0.104366	-0.112145	-0.096490
LotConfig_FR3	-0.146772	-0.146766	-0.109844	-0.146761	-0.072946
Exterior1st_Wd Sdng	-0.126112	-0.126110	-0.118787	-0.126108	-0.111413
Age_RemodAdd_Years	-0.155109	-0.155109	-0.156609	-0.155110	-0.157994
MSSubClass	-0.158513	-0.158513	-0.157126	-0.158513	-0.156335
HeatingQC_Po	-0.263997	-0.263970	-0.161843	-0.263943	-0.059541
LotShape_IR3	-0.200509	-0.200502	-0.182322	-0.200495	-0.163981
Functional_Sev	-0.297297	-0.297267	-0.193071	-0.297236	-0.088712
Exterior1st_BrkComm	-0.325315	-0.325299	-0.273237	-0.325282	-0.220970
Functional_Maj2	-0.319386	-0.319378	-0.294064	-0.319370	-0.268524



Metric	Linear Regression	Ridge Regression	Lasso Regression	Double Ridge Regression	Double Lasso Regression
R2 Score (Train)	0.858555	0.858555	0.858233	0.858555	0.857293
R2 Score (Test)	0.834576	0.834576	0.835408	0.834576	0.835817
RSS (Train)	22.407071	22.407071	22.458168	22.407071	22.607044
RSS (Test)	12.291844	12.291851	12.230031	12.291858	12.199594
MSE (Train)	0.148070	0.148070	0.148239	0.148070	0.148729
MSE (Test)	0.167522	0.167522	0.167100	0.167522	0.166892

after double the values there is no significant changes in both metrics and features.

very minor variations but overall, it is similar

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:

R2 scores and MSE and RSS for Lasso Regression are better than Ridge Regression in this model.

In Lasso, Feature which have zero coefficient value can be removed from the model

Model complexity also reduce as we can remove feature with zero coefficients

Hence Lasso can be preferred over Ridge regression model

Question 3

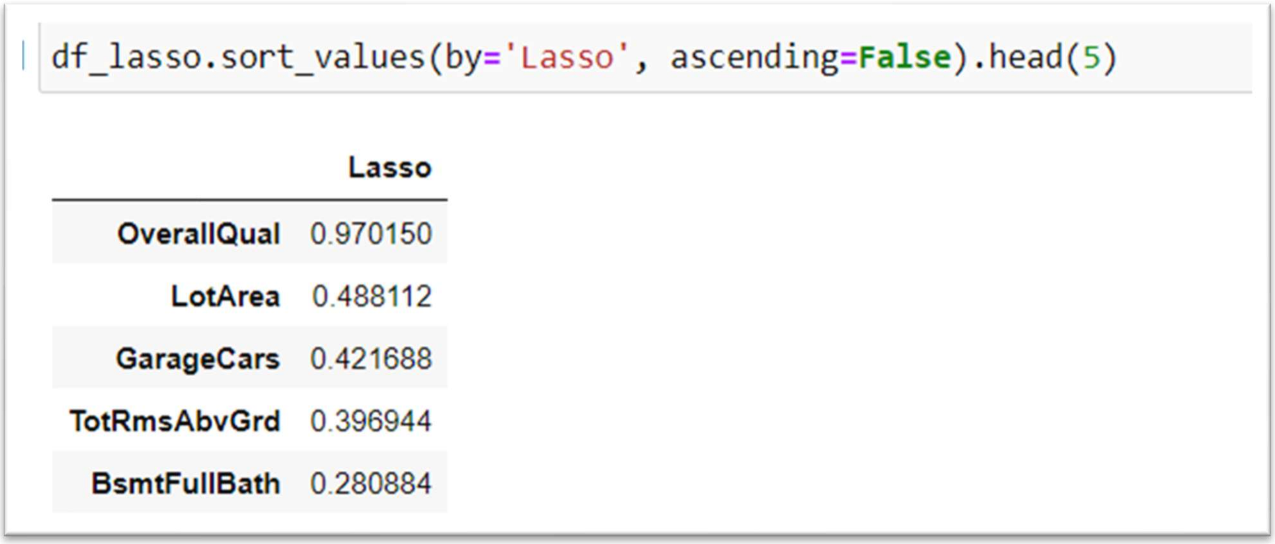
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:

Top five feature in Lasso Model are:

```
['OverallQual', 'LotArea', 'GarageCars', 'TotRmsAbvGrd', 'BsmtFullBath']
```

After removal of top 5



```
df_lasso.sort_values(by='Lasso', ascending=False).head(5)
```

	Lasso
OverallQual	0.970150
LotArea	0.488112
GarageCars	0.421688
TotRmsAbvGrd	0.396944
BsmtFullBath	0.280884

Question 4

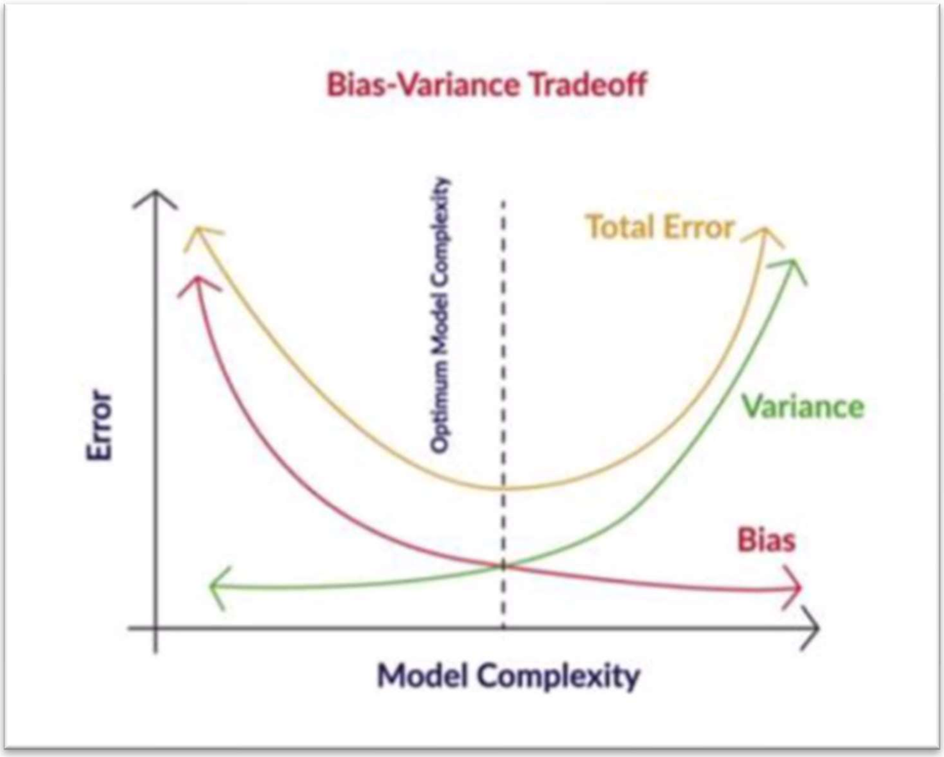
How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer:

A model with testing error and training error has enough stability even after some noise, This means that an unprecedented change in one or more features does not significantly alter the value of the predicted variable

We can control the tradeoff between model complexity and bias, regularization helps coefficient for making the model too complex, so to make the model more robust, there should be balance between keeping the model simple, making simple model lead to bias variance trade off.

Accuracy of model can be maintained by keeping the balance between bias and variance and minimize the total error.


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