Submitted by: Merrymel George

Advanced Regression Assignment
Subjective Questions

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?

Answer:
Optimal Alpha value for Ridge = 0.5
Optimal Alpha value for Lasso = 0.0001

As the alpha values are doubled, the R2 scores have dropped a bit for both Ridge and Lasso.

	Model	Best Alpha	R2 Train	R2 Test	RSS Train	RSS Test	RMSE Train	RMSE Test
0	Ridge	0.5000	0.9147	0.8325	12.421646	5.173278	0.107245	0.019090
1	Lasso	0.0001	0.9138	0.8332	12.548246	5.150738	0.107790	0.019006
0	Ridge	1.0000	0.9136	0.8321	12.581390	5.184278	0.107933	0.019130
1	Lasso	0.0002	0.9112	0.8321	12.925532	5.185369	0.109399	0.019134

The top 5 important predictors based on the mod value for both Ridge and Lasso models are 'LotArea', 'OverallCond', '2ndFlrSF', 'KitchenQual' and 'BsmtQual'.

Top 50 Features of Ridge -
'LotArea',
'2ndFlrSF',
'OverallCond',
'KitchenQual',
'BsmtFinType1_1',
'BsmtQual',
'BsmtUnfSF',
'OverallQual',

'GarageCond',
'BsmtFinSF1',
'Foundation_Wood',
'1stFlrSF',
'Functional_Mod',
'GarageArea',
'MSZoning_RH',
'Exterior2nd_CBlock',
'Neighborhood_MeadowV',
'BldgType_Duplex',
'HouseStyle_2.5Fin',
'RoofMatl_WdShake',
'RoofMatl_Membran',
'Built_Age',
'MSZoning_RM',
'Heating_GasW',
'MSZoning_FV',
'Neighborhood_Crawfor',
'Condition1_RRAe',
'BsmtFullBath_2',
'SaleCondition_Alloca',
'BsmtFinType2_5',
'Functional_Sev',
'Heating_GasA',
'GarageCars_4',
'BsmtFinType1_3',
'BsmtFinType1_5',
'BsmtFinType1_2',
'BsmtFinType1_4',
'Heating_Wall',
'RoofStyle_Mansard',
'BsmtFinType2_2',
'GarageType_No Garage',
'BsmtFinType1_6',
'MSZoning_RL',
'BsmtFinType2_4',
'Functional_Maj2',
'BsmtFinType2_1'

Top 50 Features of Lasso –
'LotArea',
 'OverallCond',
 '2ndFlrSF',
 'KitchenQual',
 'BsmtFinType1_1',
 'BsmtQual',
 'Foundation_Wood',
 'OverallQual',
 'GarageCond',
 'BsmtUnfSF',
 'BsmtFinSF1',
 'Functional_Mod',
 'Exterior2nd_CBlock',
 'GarageArea',
 '1stFlrSF',
 'Neighborhood_MeadowV',
 'BldgType_Duplex',
 'HouseStyle_2.5Fin',
 'RoofMatl_Membran',
 'MSZoning_RH',
 'MSZoning_FV',
 'Neighborhood_Crawfor',
 'Condition1_RRAe',
 'Heating_GasW',
 'BsmtFullBath_2',
 'BsmtFinType2_5',
 'Built_Age',
 'MSZoning_RM',
 'Heating_GasA',
 'BsmtFinType2_2',
 'BsmtFinType2_4',
 'SaleCondition_Alloca',
 'GarageCars_4',
 'Heating_Wall',
 'BsmtFinType2_1',
 'Functional_Sev',
 'BsmtFinType1_2',

'BsmtFinType1_3',
'BsmtFinType1_5',
'BsmtFinType1_4',
'GarageType_No Garage',
'Functional_Maj2',
'RoofStyle_Mansard',
'MSZoning_RL',
'BsmtFinType1_6',
'RoofMatl_WdShake'

Q2 - 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:
Since the R2 score for Lasso Model is comparatively higher than Ridge Model, the choice of model for this use case is Lasso Regression Model

	Model	Best Alpha	R2 Train	R2 Test	RSS Train	RSS Test	RMSE Train	RMSE Test
0	Ridge	0.5000	0.9147	0.8325	12.421646	5.173278	0.107245	0.019090
1	Lasso	0.0001	0.9138	0.8332	12.548246	5.150738	0.107790	0.019006

Q3 - 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 five most important predictor variables in the Lasso Model are - 'LotArea', 'OverallCond', '2ndFlrSF', 'KitchenQual' and 'BsmtQual'.

If these features are available, then the next 5 most import predictor variables of the Lasso model are - 'OverallQual', 'Functional_Mod', 'MSZoning_RL', 'GarageType_Basment' and 'MSZoning_RH'

Q4 - 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?

If there is a significant drop in testing accuracy in comparison with the training accuracy, then the model is overfitting. This might be the result of a complex model. In line with Occam's Razor theory, the model should be as simple as possible but not simpler as it will tend to underfit. Hence, the model should be simplified accordingly and optimally. To achieve this, the first and foremost step is to have a very diligent exploratory data analysis. This helps in handling the outliers, missing data and skewness. Proper cleansing, scaling and transformations are to be applied on the data. Following this, the model can be made simpler by considering the Bias-Variance trade-off. Simplifying the model can be achieved through various regularization techniques as well. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. Performing all these will make a model more robust and generalisable, which in turn improves its accuracy.