Machine Learning II

Advanced Regression

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?

Optimal value of alpha for Ridge Regression: **3**
Optimal value of alpha for Lasso Regression: **0.0005**

Refer to the table that shows the most important predictor variables (having positive and negative influence) in case of ridge and lasso regression with the optimal value of alpha and double the value of optimal alpha. If alpha is 0, then the regularization term becomes zero. The model will try to minimize the error term. The coefficient values will increase. **As we increase the alpha value, the coefficients will become smaller.** We can observe this decrease in coefficient values of the predictor variables when we double the alpha.

In case of Lasso Regression, we can also see that the number of parameters reduced with increase in alpha. The model tends to become simpler with increase in value of alpha. But if we increase it too much then the coefficients will become zero. **The model becomes too simple with increase in the value of alpha** thereby resulting in underfitting the data.

Both Ridge and Lasso Regression simplifies the model by trying to minimize the cost function that includes the error term and regularization term. Lasso Regression is computationally intensive but the coefficient of most of the parameters becomes zero and hence we can use it for feature selection. We have applied RFE to select 120 features and then applied the Ridge and Lasso Regression. The model built with ridge regression gave us the coefficients for all 119 features, whereas the one with lasso regression gave us 94 features when using alpha = 0.0005 and 77 features when using alpha = 0.001. We can see that **as we increase alpha, the model becomes more simpler** and more and more coefficients tend to become zero.

The **r2 score** for these different models is shown below:

Ridge				Lasso			
Optimal Alpha		Double the Alpha		Optimal Alpha		Double the Alpha	
Train	Test	Train	Test	Train	Test	Train	Test
0.91	0.88	0.91	0.88	0.91	0.88	0.91	0.88

With the alpha values for ridge and lasso we see that the train r2 score decreases with increase in alpha. This is because the model become simpler and tries to generalize. However we can also see that the test r2 score is still the same.

Ridge - Optimal alpha (alpha = 3)		Ridge - Double the value of alpha (alpha = 6)	
Features with positive impact	**Features with negative impact**	**Features with positive impact**	**Features with negative impact**
Neighborhood_NridgHt (0.408) Neighborhood_Crawfor (0.397) MSZoning_RH (0.343) Neighborhood_Somerst (0.341) MSZoning_RL (0.339) Neighborhood_ClearCr (0.331) Neighborhood_StoneBr (0.309) LandContour_Low (0.287) Neighborhood_NoRidge (0.276) MSZoning_FV (0.257)	Exterior2nd_Brk Cmn (-0.081) BuildingAge (-0.103) HouseStyle_SFoyer (-0.104) Exterior2nd_Wd Shng (-0.138) SaleCondition_Abnorml (-0.149) KitchenAbvGr (-0.151) Exterior1st_BrkComm (-0.162) Neighborhood_MeadowV (-0.176) SaleType_COD (-0.21) MSSubClass_160 (-0.31)	Neighborhood_Crawfor (0.351) Neighborhood_NridgHt (0.335) Neighborhood_ClearCr (0.276) Neighborhood_Somerst (0.269) LandContour_Low (0.237) MSZoning_RL (0.225) Neighborhood_StoneBr (0.22) Neighborhood_NoRidge (0.216) MSZoning_RH (0.212) Exterior1st_BrkFace (0.212)	HouseStyle_SFoyer (-0.072) Exterior2nd_Brk Cmn (-0.08) BuildingAge (-0.105) Exterior1st_BrkComm (-0.105) Exterior2nd_Wd Shng (-0.112) KitchenAbvGr (-0.12) Neighborhood_MeadowV (-0.145) SaleCondition_Abnorml (-0.15) SaleType_COD (-0.181) MSSubClass_160 (-0.281)

Lasso - Optimal alpha (alpha = 0.0005)		Lasso - Double the value of alpha (alpha = 0.001)	
Features with positive impact	**Features with negative impact**	**Features with positive impact**	**Features with negative impact**
MSZoning_RH (0.651) MSZoning_RL (0.599) MSZoning_FV (0.505) MSZoning_RM (0.498) Neighborhood_NridgHt (0.443) Neighborhood_Crawfor (0.44) Neighborhood_Somerst (0.385) Neighborhood_StoneBr (0.364) Neighborhood_ClearCr (0.346) LandContour_Low (0.312)	HouseStyle_SFoyer (-0.053) SaleType_WD (-0.057) BuildingAge (-0.103) KitchenAbvGr (-0.121) Exterior2nd_Wd Shng (-0.136) SaleCondition_Abnorml (-0.142) SaleType_COD (-0.18) Neighborhood_MeadowV (-0.202) Exterior1st_BrkComm (-0.268) MSSubClass_160 (-0.333)	Neighborhood_Crawfor (0.407) Neighborhood_NridgHt (0.356) Neighborhood_Somerst (0.326) GrLivArea (0.326) Neighborhood_ClearCr (0.306) Neighborhood_StoneBr (0.254) LandContour_Low (0.251) MSZoning_RH (0.247) MSZoning_RL (0.222) Neighborhood_NoRidge (0.222)	GarageFinish_Unf (-0.051) SaleType_WD (-0.052) BsmtFinType1_Unf (-0.055) KitchenAbvGr (-0.076) Exterior2nd_Wd Shng (-0.084) BuildingAge (-0.112) SaleCondition_Abnorml (-0.151) Neighborhood_MeadowV (-0.156) SaleType_COD (-0.184) MSSubClass_160 (-0.316)

Note: The important predictor variables in the model are more or less the same in Ridge and Lasso Regression. MSZoning, Neighborhood, Land Contour - postive influencers (increases the price) and House Style, Building Age, KitchenAbvGr - negative influencers (reduces the price). Changes due to doubling the value of alpha in top 10 features are highlighted in yellow.

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?

We have applied RFE to reduce the number of features to 120 and then built the model using Ridge and Lasso Regularization with optimal alpha (hyper parameter).

The results are shown below:

Regularization	Optimal Alpha	r2 Score in Train	r2 Score in Test	No. of Features
Ridge (L2)	3	0.912	0.883	119
Lasso (L1)	0.0005	0.913	0.884	93

Our objectiveis to choose the simplest model. Here, both the models are giving similar r2 score, but the Lasso model has reduced the number of features to 93, thus giving a simpler model.

However, the Lasso model is computational more intense and resource consuming. The Lasso regression uses the sum of the absolute value of the coefficients as the regularization term. It does not convert into a nice invertible function and hence need to be solved iteratively. In case or ridge regularization that uses sum of the squares of the coefficients, it can converted to an invertible matrix and can thus be solved using matrix operations. This significantly lowers the computational costs associated with it. Provided the computational cost is acceptable, we can go with Lasso.

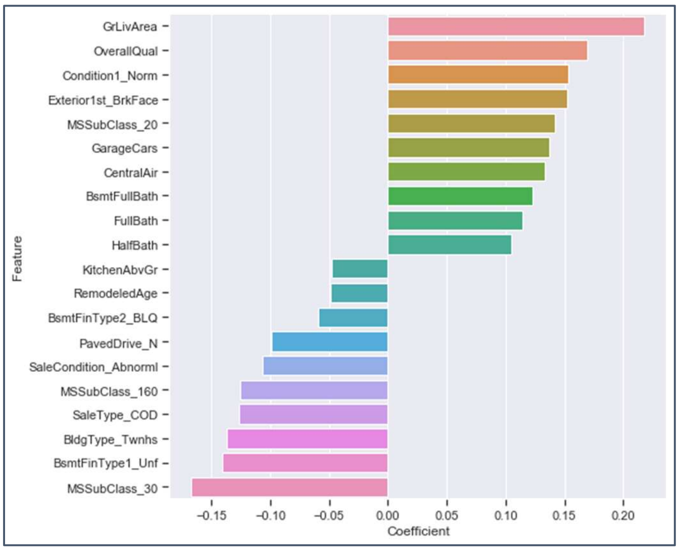
Question 3

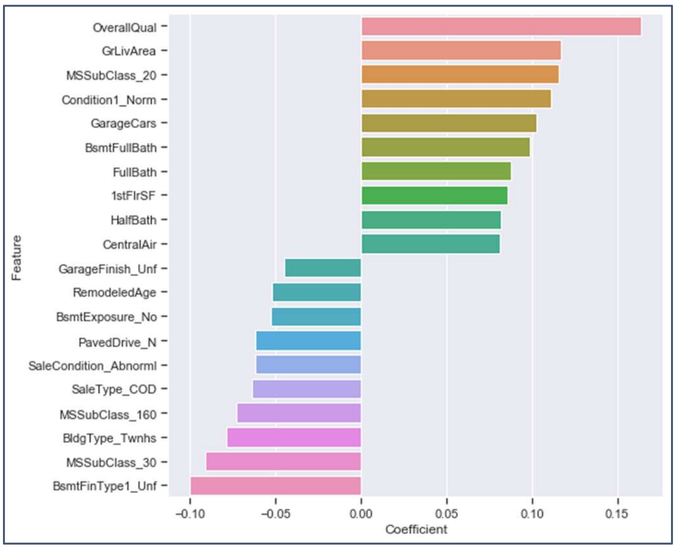
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

The five important predictor variable are: MSZoning, Neighborhood, Land Contour, Building Age, KitchenAbvGr. Dropped these columns and built the model using Ridge and Lasso Regularization. The optimal alpha for Ridge Regression is 50 and for Lasso Regression it is 0.002.

Regularization	Optimal Alpha	r2 Score in Train	r2 Score in Test	No. of Features
Ridge (L2)	50	0.893	0.878	156
Lasso (L1)	0.002	0.894	0.880	81

The new predictor variables are: **GrLiveArea, OverallQual, Condition1, GarageCars, CentralAir.** Refer to the visualizations given below.





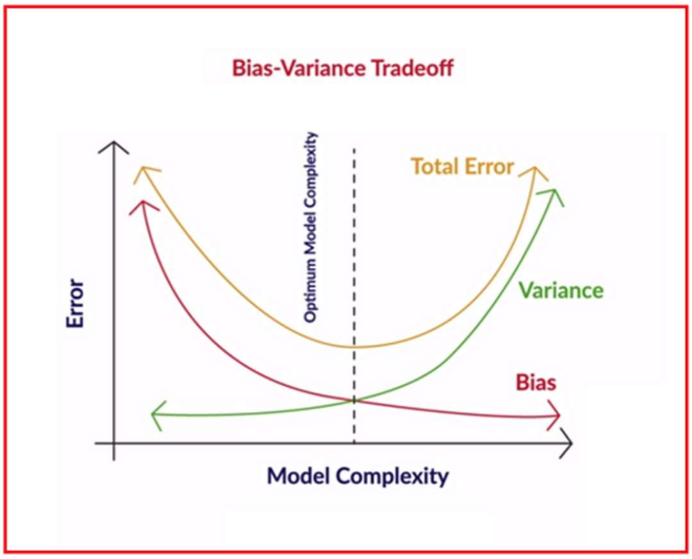
This visualization shows the top 10 variables having positive influence (increases the price) and top 10 variables having negative influence (reduces the price) as obtained from Ridge and Lasso Regression. Mostly the same features have been identified by both regularization models.

RIDGE		LASSO	
Positive Influencers	**Negative Influencers**	**Positive Influencers**	**Negative Influencers**
OverallQual (0.164) GrLivArea (0.117) MSSubClass_20 (0.116) Condition1_Norm (0.111) GarageCars (0.103) BsmtFullBath (0.099) FullBath (0.088) 1stFlrSF (0.086) HalfBath (0.082) CentralAir (0.081)	GarageFinish_Unf (-0.045) RemodeledAge (-0.052) BsmtExposure_No (-0.053) PavedDrive_N (-0.062) SaleCondition_Abnorml (-0.062) SaleType_COD (-0.064) MSSubClass_160 (-0.073) BldgType_Twnhs (-0.079) MSSubClass_30 (-0.091) BsmtFinType1_Unf (-0.1)	GrLivArea (0.218) OverallQual (0.169) Condition1_Norm (0.153) Exterior1st_BrkFace (0.152) MSSubClass_20 (0.142) GarageCars (0.137) CentralAir (0.133) BsmtFullBath (0.123) FullBath (0.114) HalfBath (0.105)	KitchenAbvGr (-0.048) RemodeledAge (-0.049) BsmtFinType2_BLQ (-0.059) PavedDrive_N (-0.099) SaleCondition_Abnorml (-0.107) MSSubClass_160 (-0.126) SaleType_COD (-0.127) BldgType_Twnhs (-0.137) BsmtFinType1_Unf (-0.141) MSSubClass_30 (-0.167)

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?

A simple model is robust and generalizable compared to a complex model. How do we know whether a model is complex or not? One way is to go by the number of parameters: lesser the parameters, simpler is the model. Or lesser the coefficient values for the parameters, simpler is the model. Simple models have low variance and high bias, whereas complex models have low bias and high variance. We have to ensure that the model is not oversimplified, which may result in underfitting. As a result the model will not perform well when we apply it on the test data. On the other hand, a complex model almost memorizes the train data, and does not generalize on the test data which it has not seen before. To make sure that the model is robust and generalizable we need to tune the hyper parameter, which, in the case of Ridge and Lasso regression, is the lambda value (or alpha) for the regularization term used for simplifying the model.



Bias quantifies how accurate the model can describe the actual task at hand i.e. how accurately it is likely to be on future (test) data. **Variance** measures how flexible the model is with respect to changes in the training data. As complexity increases, bias reduces and variance increases. The expected total error of a model is the sum of the errors in bias and the variance. We aim to find the optimal point where the total model error is the least so that we can get a model with a reasonable accuracy, but without compromising on the flexibility to generalize. In our model, we see that the r2 score on the train and test data are very close to each other, indicating that the model is able to generalize on unseen data and at the same time not over simplified.