Subjective Questions

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 1

The optimal value of alpha taken for Ridge Regression is **100** while the optimal value of alpha taken for Lasso Regression is **0.004**.

Ridge Regression

alpha = 100

Train R2: 0.8823954382276953
Test R2: 0.8874056417630752
RMSE is: 0.12310960766133607

alpha = 200

Train R2: 0.8786477469819408
Test R2: 0.8854545552170129
RMSE is: 0.12524290725964823

It has been observed that as we double alpha in Ridge Regression model, the accuracy decreases slightly and error increases by fractions.

The most important predictor variables for Ridge Regression after applying changes are :-
OverallQual, GarageCars, GrLivArea, OverallCond, Total_Bathrooms, Fireplaces , TotalSF, TotRmsAbvGrd , Foundation_PConc , Neighborhood_StoneBr

Lasso Regression

alpha =0.004

Train R2: 0.8844467173196384
Test R2: 0.8892252629233568
RMSE is: 0.1211200510739345

alpha = 0.008

Train R2: 0.8826266441334811
Test R2: 0.8886661719899234
RMSE is: 0.12173135581903703

In Lasso Regression as we double the alpha, same trends we could observe. The train and test accuracy decreases slightly while mean squared error increases by fractions.

The most important predictor variables for Lasso Regression after applying changes are :-
OverallQual , GrLivArea, GarageCars, OverallCond , Total_Bathrooms, Fireplaces, Neighborhood_NridgHt, Neighborhood_StoneBr, Neighborhood_Crawfor, LotArea

Explaination:-The hyperparameter λ is a balancing factor we typically choose the λ which best captures the relative importance of error versus the model complexity. It governs the behavior of algorithm we use to build a model. When λ=0, the alphas have their least square estimate values. The actual estimates heavily depend on the training data and hence variance is high. As we increase λ, alphas start decreasing and model becomes simpler. In the limiting case of λ approaching infinity, all betas reduce to zero and model predicts a constant and has no variance.

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 2

The optimal value of lambda for ridge and lasso regression is **100 for Ridge** while **0.004 for Lasso.** I will choose these two values because as we know λ keeps a delicate balance between error and the model complexity and the optimal values taken above takes care of the fact that the model doesn't become too complex while also keeping in mind the model does perform well on training as well as on the test data. The training and test accuracy given by data is an indicator that our model is performing well while keeping the complexity low.

Also, these below two values are being chosen by GridSearchCV().best_params_ which the return the most optimum alpha parameter.

Lasso Regression

alpha =0.004

Train R2: 0.8844467173196384
Test R2: 0.8892252629233568
RMSE is: 0.1211200510739345

Ridge Regression

alpha = 100

Train R2: 0.8823954382276953

Test R2: 0.8874056417630752
RMSE is: 0.12310960766133607

By choosing the above optimal values, the model comes out to be Robust and Generalizable as most of the variance in dataset is being explained by just 29 variables (extracted using Lasso Regression) giving quite optimal accuracy.

Question 3

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?

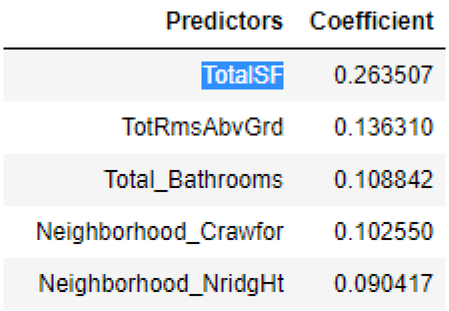
Answer

Presently the five top predictors are Overallqual, GrLivArea, GarageCars,OverallCond, Fireplaces.The accuracy with these predicators is given as:

Train R2: 0.8844467173196384
Test R2: 0.8892252629233568
RMSE is: 0.1211200510739345

If these top 5 predictors are not present in the incoming data then the five most predictors are as below:

TotalSF(Total Surface Area-Derived metrics), TotRmsAbvGrd, Total_Bathrooms(Total no of bathrooms-Derived metrics), Neighborhood_Crawfor, Neighborhood_NridgHt.



Predictors	Coefficient
TotalSF	0.263507
TotRmsAbvGrd	0.136310
Total_Bathrooms	0.108842
Neighborhood_Crawfor	0.102550
Neighborhood_NridgHt	0.090417

The accuracy reduces to:

Train R2: 0.8109049299799372
Test R2: 0.840989486676553
RMSE is: 0.17386059315765476

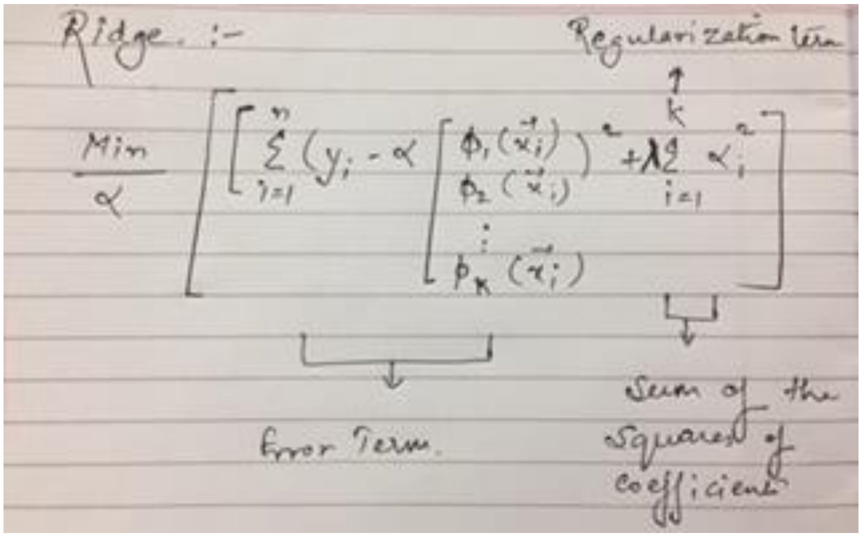
Question 4

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

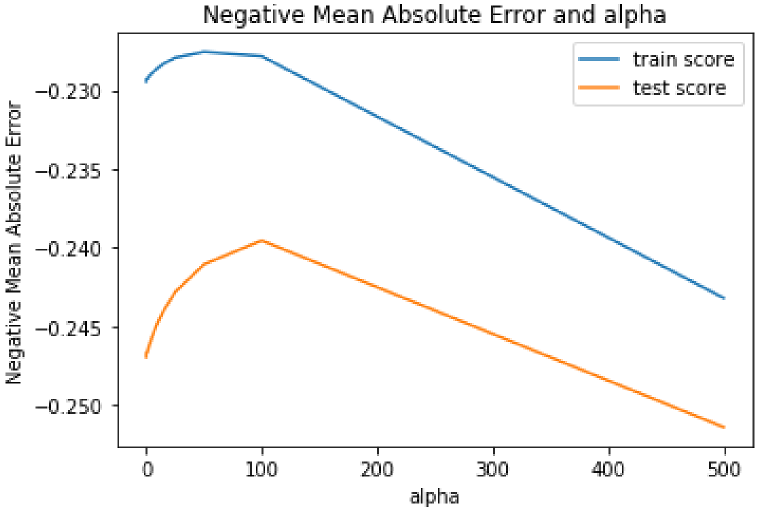
As we are already familiar, the hyperparameter λ is a balancing factor, we typically choose the λ which best captures the relative importance of error versus the model complexity. The chosen optimal values of λ which are 100 for Ridge and 0.04 for Lasso, is giving quite robust results as we have practically tried removing 5 top predictors which reduces the accuracy by 4-7% explains that the predictors we have chosen are the best parameters for explaining the variation in the dataset. **Also, we saw that by applying Lasso Regression algorithm the count of predictor variables reduces to 29**. This explains the generalizable property of our model.

Ridge Regression:



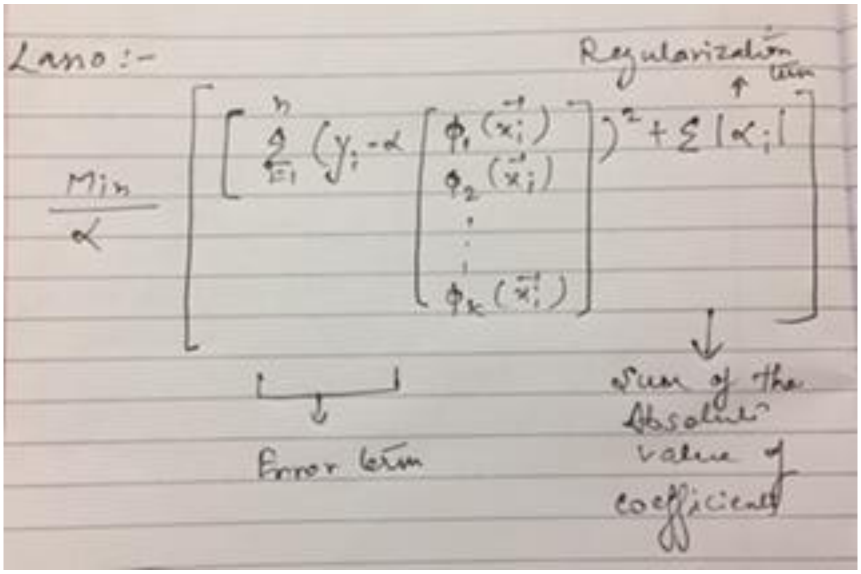
Ridge. :-

$$\min_{\alpha} \left[\underbrace{\left[\sum_{i=1}^{n} \left(y_i - \alpha \begin{bmatrix} \phi_1(\vec{x}_i) \\ \phi_2(\vec{x}_i) \\ \vdots \\ \phi_K(\vec{x}_i) \end{bmatrix} \right)^2}_{\text{Error Term.}} + \underbrace{\lambda \sum_{i=1}^{\overset{\text{Regularization term}}{\uparrow}\, k} \alpha_i^2 \right]}_{\text{Sum of the Squared of coefficient}} \right]$$



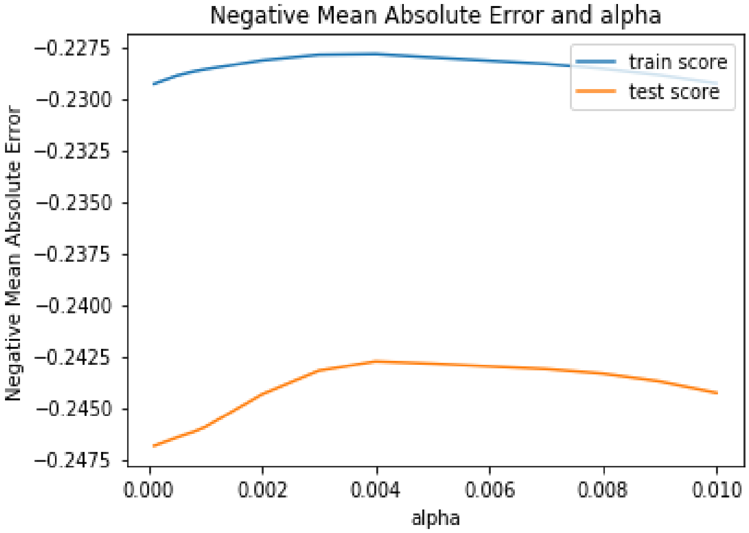
The negative mean absolute error is maximum at λ = 100.

Lasso Regression:



Lasso :-

$$\frac{\text{Min}}{\alpha} \left[\underbrace{\left[\sum_{i=1}^{n} \left(y_i - \alpha \begin{bmatrix} \phi_1(\vec{x}_i) \\ \phi_2(\vec{x}_i) \\ \vdots \\ \phi_k(\vec{x}_i) \end{bmatrix} \right)^2 \right.}_{\text{Error term}} + \overbrace{\underbrace{\sum |\alpha_i|}_{\text{Sum of the Absolute value of coefficients}}}^{\text{Regularization term}} \right]$$



Similarly for Lasso**,** the negative mean absolute error is maximum at λ = 0.004.

Hence, we can state that the model chosen is robust and generalizable as the test accuracy is almost 89% keeping the model simple with just 29 variables.