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?

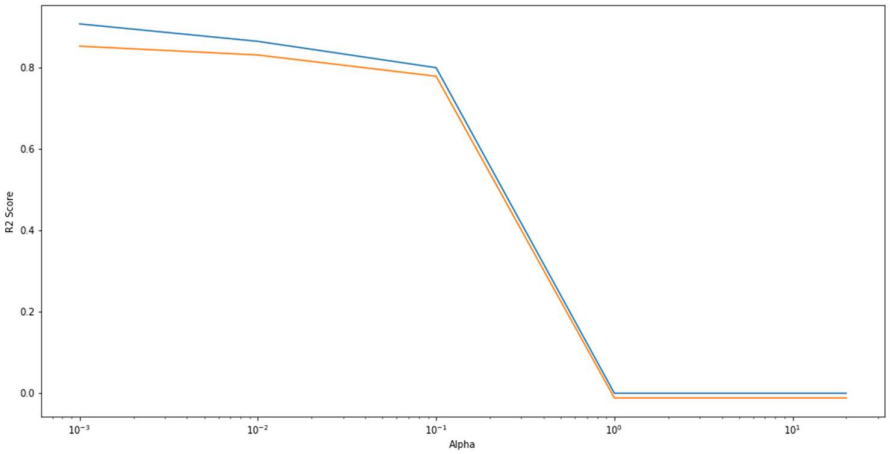
Answer:

a. What is the optimal value of alpha for ridge and lasso regression

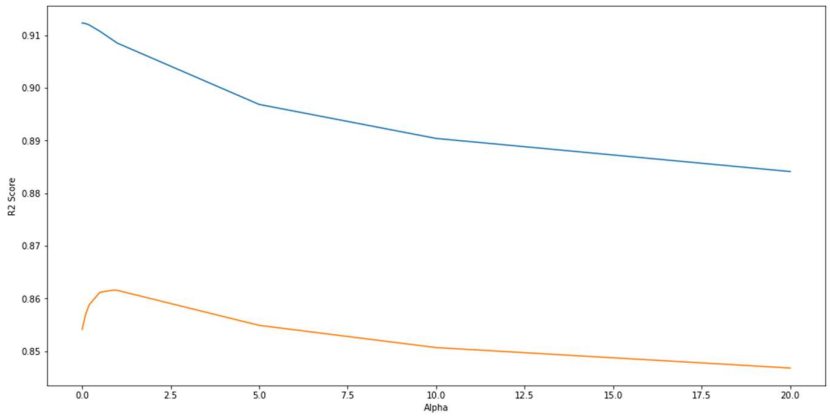
The best values are as below basis the model that is created

- Best alpha value for Lasso : {'alpha': 0.001}
- Best alpha value for Ridge : {'alpha': 0.9}

Lasso:



Ridge regression:



b. What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

R2 score will further reduce and this can be seen from the graphical plot above.

c. What will be the most important predictor variables after the change is implemented

The predictor variables would remain the same but he coefficient's are further pushed towards lesser and lesser values.

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?

Model was created in Ridge and Lasso, we can see that the r2_scores are although not very different/relative around same for both of them but it was seen that lasso will penalize more the features. This will help us in feature selection.

Thus I would consider Lasso model than Ridge model.

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?

Before dropping the significant predictor variables the model co-efficient and predictor variables are as follows:

Featuere	Coef	
49	YrSold_Old	1.365558
32	Fireplaces	0.447815
31	TotRmsAbvGrd	0.345647
33	GarageFinish	0.305298
15	BsmtFinSF2	0.280644
66	Neighborhood_Crawfor	0.271929
28	BedroomAbvGr	0.259023
2	LotShape	0.172878
27	HalfBath	0.161633
5	OverallCond	0.133923

After dropping the significant predictor variables, the model co-efficient and predictor variables are as follows

Dropped variables are

df.drop(['Fireplaces','GarageFinish','TotRmsAbvGrd','LowQualFinSF','BedroomAbvGr'],axis=1)

Featuere	Coef	
43	PoolArea	1.478633
27	HalfBath	0.429337
67	Neighborhood_Edwards	0.363583
12	BsmtFinType1	0.313600
26	FullBath	0.263799
22	LowQualFinSF	0.246891
1	LotArea	0.215448
19	CentralAir	0.210522
16	BsmtUnfSF	0.173661
44	MiscVal	0.151890

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?

- Here are some changes you can make to your model:
  - **Use a model that's resistant to outliers.**
  - The model should not impacted by the outliers: Outlier treatment is most important to get the robust model. We can detect outliers in the dataset using box plots, Z score etc. Treating the outliers will not affect mean, median etc. so that we can impute correct values to missing values. , the outlier analysis needs to be done and only those which are relevant. This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis
  - **The predicted variables should be significant.** Model significance can be determined the P-values, R2 and adjusted R2. Always a simple model can be more robust

Here are some changes you can make to your data:

1. **Data collection:** Always getting more data helps in better model building, always attempt to get as much data as possible
2. **Fix missing values and outliers:** If the data has missing values and outliers can lead to inaccurate model. Outliers can affect the mean, median that we are imputing to continuous variables
3. **Transform your data.** If your data has a very pronounced right tail, try a log transformation. User derived variables and drop the given variable if this adds more value to the data
4. **Remove the outliers.** This works if there are very few of them and its certain they're anomalies and not worth predicting
5. **Feature Selection:** Domain knowledge plays an important role in feature selection, additional techniques like data visualization also helps the selecting the features. Statistical parameters like p-Values, VIF can give us significant variables.
6. **Algorithm selection:** Choosing the right machine learning algorithm is very important to get accurate model.
7. **Cross validation:** To reduce overfitting user cross validation i.e. leave a sample on which you do not train the model & test the model on this sample before got to the final model