ASSINGMENT PART -2

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 of alpha is:

Optimal Value of alpha for Ridge

- Ridge - 3.0
- Lasso - 0.0001

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.934264	0.896795	0.899406
1	R2 Score (Test)	0.896435	0.857411	0.858905
2	RSS (Train)	9.136517	14.344308	13.981326
3	RSS (Test)	6.846235	9.425909	9.327188
4	MSE (Train)	0.097052	0.121606	0.120057
5	MSE (Test)	0.128286	0.150527	0.149737

Double the values of alpha

- Ridge - 9.0
- Lasso - 0.0002

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.893236	0.882257
1	R2Score Test	0.855929	0.850513
2	RSS Train	14.838918	16.364826
3	RSS Test	9.523880	9.881955
4	MSE Train	0.015298	0.016871
5	MSE Test	0.022894	0.023755

The Mean Squared error in case of Ridge and Lasso are:¶

- Ridge - 0.121606
- Lasso - 0.120057

The Mean Squared Error of ridge is slightly lower than that of lasso

Also, since Lasso helps in feature reduction (as the coefficient value of one of the features became 0), Lasso has a better edge over Ridge.

Hence based on Lasso, the factors that generally affect the price are the Zoning classification, Living area square feet, Overall quality, condition of the house, Foundation type of the house, Total basement area in square feet, neighbourhood etc

Therefore, the variables predicted by Lasso in the above bar chart as significant variables for predicting the price of a house.
Let's analyze the model with these alpha values
Here Lasso given the very close result of R2 score for both test and train. The most important feature after double the value of alpha is:
- MSZoning_FV
- MSZoning_RL
- GrLivArea
- OverallQual
- TotalBsmtSF
- Neighborhood_Crawfor
- Foundation_PConc
- Neighborhood_NridgHt
- SaleCondition_Normal
- GarageCars
- OverallCond
- SaleType_New

Question2

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: Based on the alpha/Lambda values I have got, Ridge regression does not zero any of the coefficient, Lasso zeroed one or two coefficients in the selected features, Lasso is better option and it also helps in the some of the feature elimination.

Optimal Value of alpha for Ridge
- Ridge - 3.0
- Lasso - 0.0001

We choose lasso regression it will give better result.

QUESTION3

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:

After removing the five most important features that we have got prior "MSZoning_FV", "GrLivArea", "MSZoning_RL", "OverallQual", "Foundation_PConc" I have got the other important features to predict the sales price with Overall condition, Lot area, shape, Condition1, IsRemodelled.

	Featuere	Coef
0	MSSubClass	11.546662
4	OverallCond	0.143045
26	IsRemodelled	0.137121
3	LotShape	0.134221
28	OldOrNewGarage	0.129265
2	LotArea	0.119874
1	LotFrontage	0.087476
36	LotConfig_FR3	0.073304
37	LotConfig_Inside	0.065800
47	BldgType_Duplex	0.062690

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 needs to be made robust and generalizable so that they are not impacted by outliers in the training data.

-The model should also be generalizable so that the test accuracy is not lesser than the training score.

-The model should be accurate for datasets other than the ones which were used during training.

Too much weightage should not give to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis.

Implications of Accuracy of a model:

1. Gain of more data allows the data to train itself, instead of depending on the weak correlations and assumption.

2. Fix missing values and outliers that can affect the mean, median value and outliers can lead to inaccurate model

3. Featuring Engineering or newly derived columns/Standardize the values that can extract the new data from the existing data.

4. Feature Selection is purely based on the domain knowledge, so that we can select important features that have good impact on the target variable. Data visualization also helps the selecting the features. Statistical parameters like p-Values, VIF can give us significant variables.

5. Applying the right algorithm to choose the right machine learning.

6.Some times more accuracy will cause overfitting, then we can use cross validation technique.