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:

I have got the below optimal values for alpha

- Ridge - 0.3
- Lasso - 0.0001

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.906714	0.905679
1	R2Score Test	0.870229	0.871224
2	RSS Train	12.965627	13.109551
3	RSS Test	8.578568	8.512817
4	MSE Train	0.013367	0.013515
5	MSE Test	0.020622	0.020464

When double the value of alpha

- Ridge - 0.6
- Lasso - 0.0002

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.906107	0.888308
1	R2Score Test	0.870309	0.860208
2	RSS Train	13.050078	15.523851
3	RSS Test	8.573325	9.241049
4	MSE Train	0.013454	0.016004
5	MSE Test	0.020609	0.022214

I have got very similar values for the first time for both Lasso and Ridge.

Very slight difference in the R2 score, RSS for Ridge regression when we double the value of alpha.

A noticeable difference is there in the R2 score test and train, RSS in Lasso Regression when we double the value of alpha.

The important variables after this change is implemented are:

- MSZoning_FV
- MSZoning_RL
- GrLivArea

- OverallQual
- TotalBsmtSF
- Neighborhood_Crawfor
- Foundation_PConc
- Neighborhood_NridgHt
- SaleCondition_Normal
- GarageCars

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:

I have got similar values for both Ridge and Lasso for the alpha values 0.3 and 0.0001.

Based on the alpha/Lambda values I have got, Ridge regression does not zero any of the co efficient,

Lasso zeroed one or two coefficients in the selected features, Lasso is better option and it also helps in some of the feature elimination. So I would choose 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?

Answer:

I have excluded the five most important variable I have got prior. Those are MSZoning_FV, GrLivArea, MSZoning_RL, OverallQual, Foundation_PConc. I have created a new model after removing these columns code is mentioned the Python notebook. After the Lass Regression I have got the other important predictors are

Overall condition,

Lot area,

Lot shape,

Condition1,

IsRemodeled.

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:

The mode is robust and generalizable when

1. Test accuracy is not much lesser than the training score
2. 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
3. **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

Implications of Accuracy of a model:

1. Gain the more data as much you can:
   Having more data allows the data to train itself, instead of depending on the weak correlations and assumption, it is good to have more data.
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
   You can get the outlier values using a boxplot, treating the outliers in the data will make our mode more accurate.
3. Featuring Engineering or newly derived columns/Standardize the values:
   We can extract the new data from the existing data ex: from DOB we can get the Age of the person, after extracting the new data required we can drop the existing features.
   Scaling the values: ex: one value is in meters, the other is Kilo meters, it is important to scale these feature into one standardized unit.
   If we did this we can get accurate model.
4. Feature Selection:
   It 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
   Choosing the right machine learning algorithm is very important to get accurate model. This will comes with experience
6. Cross validation:
   Some times more accuracy will cause overfitting, then we can use cross validation technique, 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.