Subjective Questions of 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?

Answer:

My optimal value of alpha for Ridge and Lasso are:

Ridge: 0.01

Lasso: 0.001

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.910955	0.880035
1	R2Score Test	0.891049	0.873730
2	RSS Train	1.471569	1.982547
3	RSS Test	0.718239	0.832413
4	MSE Train	0.001503	0.002025
5	MSE Test	0.001706	0.001977
6	RMSE Train	0.001503	0.002025
7	RMSE Test	0.041304	0.044466

If we double the value of alpha then,

Ridge: 0.02

Lasso: 0.002

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.910953	0.844793
1	R2 Score Test	0.891093	0.843258
2	RSS Train	1.471596	2.564970
3	RSS Test	0.717951	1.033291
4	MSE Train	0.001503	0.002620
5	MSE Test	0.001705	0.002454
6	RMSE Train	0.001503	0.002620
7	RMSE Test	0.041296	0.049542

We have got very similar values for Ridge but for lasso almost 3-4% difference in R2 Score

Important Predictors:

- OverollQual
- WoodDeckSF
- MasVnrArea
- GrLiveArea
- LotArea

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 almost 3% difference in R2 score for both Ridge and Lasso regression when the alpha value is 0.01 and 0.001.

In case of Ridge regression, 3% r2 value is high but we don't have any zero coefficient value so that we have to use all the selected features.

In the case of Lasso Regression, there are many coefficients with zero values in the selected features and it helps in some of the feature elimination. So better to use Lasso Regression.

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:

Important five features are 'GrLivArea','OverallQual','MSSubClass','LotArea','OverallCond'. I have dropped this five features and rebuilt the model.

Steps I have followed:

- Dropped the five important features.
- Done the splitting
- Followed by scaling(MinMax Scaling)
- Linear Regression and RFE
- Lasso regression model building has done (check the screenshot below)

```
r2_train_lr = r2_score(y_trainLS ,y_train_predLS )
print('Train R2 Score: ', r2_train_lr)

r2_test_lr = r2_score(y_testLS, y_test_predLS)
print('Test R2 Score: ', r2_test_lr)

rss1_lr = np.sum(np.square(y_trainLS-y_train_predLS))
print('Train RSS: ', rss1_lr)

rss2_lr = np.sum(np.square(y_testLS - y_test_predLS))
print('Test RSS: ',rss2_lr)

mse_train_lr = mean_squared_error(y_trainLS,y_train_predLS)
print('Train MSE: ',mse_train_lr)

mse_test_lr = mean_squared_error(y_testLS , y_test_predLS)
print('Test MSE: ',mse_test_lr)
```

```
Train R2 Score:  0.786461304172551
Test R2 Score:  0.7796714509847087
Train RSS:  3.528956010401741
Test RSS:  1.4524766853498163
Train MSE:  0.0036046537389190407
Test MSE:  0.0034500633856290176
```

Observations:

- R2 score has been reduced by almost 10% for both train and test data.
- RSS value has increased almost double
- MSE is some slight changes

After Remodeling, I have got the below important features "LotFrontage", "OpenPorch", "HalfBath", "ExternalCond", "IsRemodeled".

Question 5:

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:

<u>The model is robust and generalized:</u>

- Test accuracy is not much lesser than the training set then it leads to overfitting
- The model should not be impacted by the outliers. Outlier treatment is most important to get the robust model.
- Predicted variable should be significant
    - Model significance can be determined from the p-value, R2 and Adjusted R2.
    - Always simple models can be more robust

Implications of Accuracy of model:

- Fix missing values and outliers
- Feature engineering / newly derived features / standardize the value
- Feature selection:
  - Important Features have good impact on the unseen data(target data)
  - Data visualization also helps us to select the features(EDA)
- Choosing the right algorithm plays an important role to get the accurate model
- Cross-validation: Sometime high accuracy leads to overfitting, then we can use cross-validation. We will get a more accurate estimate out-of-sample accuracy.