Advanced Regression Subject 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 to 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 values of the hyperparameter - alpha (lambda) for Ridge and Lasso regressions are:

- Ridge: 1.0
- Lasso: 0.0001

If we double the values of alpha then, the following results would occur: -
In Ridge : The coefficients values would decrease i.e. lowered further
In Lasso: More number of coefficients for lesser significant features will turn 0.

The most important predictor variables after the change is implemented are those which are more significant. Below are the resultsfor our house-prices prediction assignment:

Ridge (after doubling the alpha, i.e. alpha=2.0):

Feature-Variable	Coefficient-Value
LotFrontage	0.288336
BsmtFullBath	0.241307
OverallQual_10	0.093276
GarageQual	0.084878
LowQualFinSF	0.081149

Lasso (after doubling the alpha, i.e. alpha=0.0002):

Feature-Variable	Coefficient-Value
BsmtFullBath	0.341981
LotFrontage	0.307560
OverallQual_10	0.100293
LowQualFinSF	0.099127
GarageQual	0.086350

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 : We have got the following optimal values of the hyperparameter - alpha (lambda) for Ridge and Lasso regressions:

- Ridge: 1.0
- Lasso: 0.0001

The r2_score for both the models are as follows:

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.903967	0.902198
1	R2 Score (Test)	0.876978	0.879896

Though the r2_scores are approximately same for both of them, but Lasso will penalize more on the dataset and can also help in feature elimination by making coefficients of lesser significant predictors equal to zero, we would choose to go with Lasso for the finalmodel.

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 : The five most important predictor variables in the Lasso model initially are:

Feature-Variable	Coefficient-Value
BsmtFullBath	0.351138
LotFrontage	0.289752
OverallQual_10	0.102606
LowQualFinSF	0.097570
GarageQual	0.083276

The top 5 significant variables in the new Lasso model, after removing the above five most important predictor variables in are as follows:

Feature-Variable	Coefficient-Value
FullBath	0.352471
LotArea	0.318032
GrLivArea	0.098058
BsmtFinSF2	0.093445
WoodDeckSF	0.091045

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 : We may take care of the following in order to generate a robust and generalisable model:

- **Use an outliers-resistant model -** The regression-based models are easily affected by outliers but Tree-based models are generally remain unaffected. We may use non-parametric tests instead of parametric ones while performing astatistical tests.
- **Evaluated based on a more robust error metric** – Evaluation based on mean absolute difference, Huber Loss etcinstead of mean squared error decreases the impact of outliers. Similarly, the median, which is a measure of central tendency as it represents 50^{th} percentile and remains unaffected by the outliers, while mean gets heavily impacted by the outliers. It doesn't have anything to do with any other values of the data set, so how does it "describe" the data set?

Here are some changes you can make to your data:

a. **Data Transformation** - If the data is right skewed, we apply a log-transformation on the column.
b. **Handle Outliers** – If we have sufficient size of data and it has few outliers, we may remove them, or for a small-size dataset we may fix them using the concepts of standard deviation or we may remove them in case those variables are not significant from business perspective and value modification might seriously impact our analysis.
c. **Data Winsorization** – We may manually cap the data at some threshold based on business use-case.