Advanced Regression – Questions & Answers

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 taken of alpha for Ridge and Lasso was 10 and 0.0001. From these values if we try to double the alpha values for Ridge the model will always try to apply more penalty on the curve. This will also make the model to be more generalized and simple. The most important predictor variables when the alpha is doubled in Ridge is given below.

	Features	Coefficient
0	OverallQual	0.0914
6	GrLivArea	0.0712
7	GarageCars	0.0568
9	d_BsmtQual	0.0554
1	OverallCond	0.0518
10	d_KitchenQual	0.0491
3	TotalBsmtSF	0.0468
5	2ndFlrSF	0.0459
8	OldOrNewGarage	0.0447
11	d_SaleCondition	0.0447

The R2 scores also are as follows:

```
train R2 score is 0.8870347754943094
test R2 score is 0.9029603316869086
```

Similarly for Lasso, when we doubled the alpha the R2 score decreased , however the predictor variables remained the same with reduced coeff.

	Features	Coefficient
13	MSZoning_RL	0.3162
11	MSZoning_FV	0.3136
12	MSZoning_RH	0.2742
14	MSZoning_RM	0.2165
5	GrLivArea	0.1269
42	GarageType_BuiltIn	0.1063
40	GarageType_Attchd	0.0796
0	OverallQual	0.0787
41	GarageType_Basment	0.0718
10	d_SaleCondition	0.0663

R2 Scores are given below

```
print( test R2 score is ,r2_score(y_test,y_test
```

```
train R2 score is 0.906360315064645
test R2 score is 0.8984480662627947
```

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: Selecting which regression to use depends on case to case. Given the MSE results of Ridge and Lasso, Ridge has a lower MSE (0.01241) compared to Lasso (0.01340). If its purely based on MSE scores then it would be Ridge in this case. However if interpretability is a priority and if we want to identify a subset of important features Lasso will be preferred in-spite of its slightly higher MSE value.

Other factors to consider are that Ridge includes all variables in final model unlike Lasso, Lasso always does variable selection. When lambda value is small, Lasso performs simple linear regression and as its value increases shrinking takes place. Whereas in Ridge as the lambda value increases variance in model is dropped.

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 most important or top 10 predictor variables now in Lasso are given below

	Features	Coefficient
14	MSZoning_RL	0.4013
12	MSZoning_FV	0.3938
13	MSZoning_RH	0.3700
15	MSZoning_RM	0.2982
6	GrLivArea	0.1298
46	GarageType_BuiltIn	0.1210
45	GarageType_Basment	0.0990
44	GarageType_Attchd	0.0927
0	OverallQual	0.0764
42	Foundation_Slab	0.0690

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: For a model to be effective in its objective can be achieved only if its robust and simple. The simpler the model is the more bias it has and lesser variance. Some of the techniques to be used for model robustness are:-

- Train-Test-Split
    - This is very crucial as it mimics the performance on unseen data
- Validation set
    - Apart from Train and test set we should also consider validation set. This should be used for fine-tuning

- Feature Engineering
    - Analyze the features for its usefulness
- Regularization
    - To avoid overfitting we should use regularization techniques like Ridge or Lasso
- Outlier Detection
    - Anayze and handle the outliers correctly. Outliers tend to have huge impact on the model performance

Implications of these for model accuracy are:

- Train Vs Test accracy
    - Good training does not mean accurate test. Model should perform well on the unseen data as well. Generalization plays an important role
- Overfitting Vs Underfit
    - Learning too much on the training data without considering the outliers and other feature impact results in overfitting. Regularization plays an important role in keeping overfit under control
    - Similary too much simple model will also not help as it will not understand the behavoir correctly and end up performing poorly on both train and test sets