Problem Statement - Part II

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?

Ans 1: The optimal value of alpha for ridge and lasso regression is 10 and 0.001 respectively. The following changes in the model are seen when we choose double the value of alpha:

- In ridge regression metrics: R2 score of training set lowered from 0.94 to 0.93, R2 score of the test set remains unchanged at 0.93
- In lasso regression metrics: R2 score of training set decreased from 0.92 to 0.91 and test set decreased from 0.93 to 0.91

The following variables are the most important after the change is implemented:

- GrLivArea
- OverallQual_8
- OverallQual_9
- Functional_Typ
- Neighborhood_Crawfor
- Exterior1st_BrkFace
- TotalBsmtSF
- CentralAir_Y

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?

Ans 2. This will depend on the use case and our objective.

- In case we have too many variables and one of our primary objectives is feature selection, then we will use **Lasso regression**.
- In case we don't want to get too large coefficients and reduction of coefficient magnitude is one of our main motives, then we can use **Ridge 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?

Ans 3. After dropping the top 5 lasso predictors: Top 5 Lasso predictors namely, OverallQual_9, GrLivArea, OverallQual_8,Neighborhood_Crawfor and Exterior1st_BrkFace.

We compute the following new top 5 predictors: -

2ndFlrSF

Functional_Typ

1stFlrSF

MSSubClass_70

Neighborhood_Somerst

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?

Ans 4. A model is said to be **robust** when its performance does not get affected much with the variation in the data.

A **generalizable** model is the one which performs well on the unseen data and not just the data it was trained on.

- To make sure a model is robust and generalizable, we have to **take care it doesn't overfit and underfit**.
- **Overfitting**: When a model learns the training data too well, including its noise and outliers, it performs poorly on new data. This happens because the model is too complex.
- **Underfitting:** When a model is too simple, it fails to capture the underlying patterns in the training data and performs poorly on new data.
- In other words, the model should not be too complex or too simple in order to be robust and generalizable.
- From an **accuracy perspective**, an overly complex model will show very high accuracy on the training data. Therefore, to make our model more robust and generalizable, we need to reduce variance, which will introduce some bias. Introducing bias means that accuracy on the training data will decrease.
- Overall, we need to strike a balance between model accuracy and complexity. This can be achieved using regularization techniques like Ridge Regression and Lasso.
- Hence, **Bias-variance trade-off** is an important concept to ensure that the model is robust & generalisable. Properly balancing bias and variance helps create models that are accurate and robust, capable of making reliable predictions on new, unseen data.