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?**

   Response

   The optimal values of alpha for ridge and lasso regression are

   - Ridge Regression: 2
   - Lasso Regression: 0.00012

   The most important predictor variable for optimal value alpha in both Ridge and Lasso regressions is **GrLivArea.**

   After doubling the alpha value for both ride and lasso, the results are as follows:

   - Ridge Regression: 4
   - Lasso Regression: 0.00024

   Noticed Changes:

   - The R2 score has been reduced as the increase in APLHA leads to a more generalised model.
   - Coefficients have been changed.

   The most important top 5 predictor variables for doubled value alpha in both Ridge and Lasso regressions are **GrLivArea, TotalBsmtSF, house_age, Neighborhood_Crawfor, and GarageArea.**

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?**

   Response

   I will choose Lasso regression as it helps reduce the coefficients to zero, but that is not the case with Ridge regression. As the coefficients can become zero, there will be fewer predictor variables available in the final regression model because of the feature selection feature. Fewer predictors make the model more interpretable.

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?**

   Response

   After dropping the top 5 predictor variables, i.e., ‘GrLivArea', 'house_age', 'TotalBsmtSF', 'Neighborhood_Crawfor', 'Neighborhood_NridgHt', The R2 score on test and train data has been reduced significantly.

   Now the top 5 predictor variables are ‘GarageArea’, ‘LotArea’, ‘2ndFlrSF’, 'SaleCondition_Partial', and ‘BsmtFinSF1’.

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?

Response

Occam's razor is like a golden rule in machine learning, and it's super easy to understand.When in dilemma, choose the simpler model.

- Simple models are usually better because they work well on new, unseen data.A simpler model requires fewer data points for training
- It's easier to train a simple model because it needs less data.A simple model may make more errors in the training phase but it is bound to outperform complex model when it view new data.
- Simple models stay strong and don't change much, even if the training data has tiny tweaks.
- A simple model might mess up a bit during training, but when it sees new data, it often does better than complex models.

To make sure the model is robust and generalisable, I will do the following at a broader level:

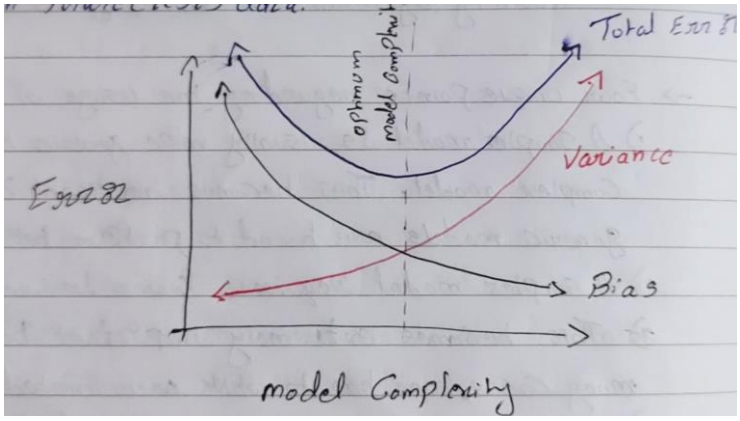
- Diverse training data is needed so that model results are not biassed.
- Clean the data really well (taking care of missing values, outliers, etc.).
- Derive new columns from the existing columns to enhance the model's performance.
- Regularisation techniques to deliberately simplify the model
- Cross-validation techniques for hyperparameter tuning.

Implication of using simpler model

Bian Varinace tradeoff

Variance: Variance in a model refers to how much its predictions differ on the same test data when the training data changes.

Bias: Bias in a model tells us how likely it is to give accurate predictions on new, unseen data.



Ideally, we aim to minimize both bias and variance because the overall error in a model is the sum of these two. So, finding the right balance between bias and variance is crucial to achieving the best performance of a model.