**Problem Statement - Part II**

**Q1.** 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 changes are implemented?

**Ans:** In my final models, the optimal alpha values are 50 for Ridge and 10 for Lasso regression.

- Doubling these values does not alter the model performance significantly in either case.

- Following the adjustment, the important predictor variables are as follows:

**For Ridge Regression:**

* Neighborhood\_StoneBr
* GarageArea
* Neighborhood\_NridgHt
* TotalBsmtSF
* GrLivArea
* KitchenQual
* Neighborhood\_Names
* Neighborhood\_Edwards
* BldgType\_TwnhsE
* GarageFinish

**For Lasso Regression:**

* TotalBsmtSF
* SaleType\_New
* MSZoning\_RM
* GarageType\_Attchd
* GrLivArea
* Neighborhood\_NAmes
* Neighborhood\_OldTown
* KitchenQual
* SaleCondition\_Partial
* RoofStyle\_Gable

**Q2.** 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:** Despite both Ridge and Lasso models demonstrating closely similar performances, with the Ridge model edging slightly ahead (1% better) on the test dataset, we have deliberately opted to employ the Lasso model as our final choice. The key reasoning behind this decision lies in Lasso's capacity for feature elimination. Given our dataset's complexity encompassing over 130+ columns, the feature elimination attribute of Lasso emerges as particularly advantageous. It aids in identifying the most influential predictor variables, potentially streamlining the model and enhancing its interpretability.

Consequently, our final model selection is Lasso Regression, yielding an R2 score of 88 on the train dataset and 84 on the test dataset. These R2 scores signify that the Lasso model effectively captures 88% of the variance in the training dataset and 84% in the test dataset, indicating a solid fit. This deliberate preference for Lasso, despite its slightly lower test performance compared to Ridge, underscores the value placed on feature elimination capabilities and potential model simplification in handling our extensive dataset.

**Q3.** 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:** After removing the five most important predictor variables ('GrLivArea', 'GarageType\_Attchd', 'MSZoning\_RM', 'SaleType\_New', 'TotalBsmtSF') and rebuilding the model, the subsequent five most influential predictor variables are as follows:

* MasVnrArea
* Neighborhood\_StoneBr
* Neighborhood\_NridgHt
* Fireplaces
* GarageArea

**Q4.** 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?

**Ans:** Ensuring the model avoids overfitting and maintains simplicity is essential for enhancing its robustness and generalizability. While attempting to overfit the model might temporarily boost accuracy, it compromises its generalizability. A generalized model exhibits commendable accuracy on both the training and testing datasets, thereby fortifying its robustness. Therefore, prioritizing a model that performs well on unseen data helps guarantee its reliability and adaptability to new scenarios, contributing to overall robustness and practical applicability.