**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:**

**Ridge Regression -**

**Optimum alpha for ridge is 50.000000**

Ridge Regression with 50

========================================

R2 Score (train): 0.9174162656466028

R2 Score (test): 0.8878676285719543

RMSE (train): 0.11481863707452225

RMSE (train): 0.13344171276419464

**Lasso Regression -**

**Optimum alpha for lasso is 0.001000**

lasso Regression with 0.001

========================================

R2 Score (train): 0.9220287606082328

R2 Score (test): 0.890957118006536

RMSE (train): 0.11156612459354591

RMSE (train): 0.1315905690055296

**Changes after doubling the doubling the alphas for both ridge and lasso regression:**

**Ridge Regression -**

**Optimum alpha for ridge is 100.000000**

ridge Regression with 100

========================================

R2 Score (train): 0.9123417904623115

R2 Score (test): 0.8858273861032065

RMSE (train): 0.11829364842992007

RMSE (train): 0.13465022274529348

* The R-Square Score drops slightly from 0.887 to 0.885 on the test data.
* The RMSE Score increases slightly from 0.1334 to 0.1346 on the test data.
* Most important predictor after change is still **1stFlrSF.**

**Lasso Regression –**

**Optimum alpha for lasso is 0.002000**

lasso Regression with 0.002

========================================

R2 Score (train): 0.9137299442532767

R2 Score (test): 0.8884981934990942

RMSE (train): 0.11735326257111758

RMSE (train): 0.13306598588927357

* The R-Square Score drops slightly from 0.890 to 0.888 on the test data.
* The RMSE Score increases slightly from 0.1315 to 0.1330 on the test data.
* Most important predictor after change is still **1stFlrSF.**

**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:**

* **Ridge Regression** model has achieved a **R2-score of 0.887** on the test data, i.e., 88.7% of the variation in the test data can be explained by the model.
* **Root Mean Squared Error** for Ridge = **0.1334** on the test data meaning the prediction made by the model could be off by 0.1334 unit.
* **Lasso Regression** model has achieved a **R2-score of 0.890** on the test data, i.e., 89.0% of the variation in the test data can be explained by the model.
* **Root Mean Squared Error** for Lasso = **0.1315** on the test data meaning the prediction made by the model could be off by 0.1315 unit.

Lasso Regression produced slightly better R2-Score on the test data than Ridge Regression. Therefore, we are choosing Lasso as the final model.

**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:**

5 most important predictor variables are:

* 1stFlrSF: First Floor square feet
* RoofMatl\_CompShg: Roof Material - Standard (Composite) Shingle
* 2ndFlrSF: 2nd Floor square feet
* OverallCond: Rates the overall condition of the house
* RoofMatl\_Tar&Grv : Roof Material - Gravel & Tar

After Dropping the 5 variables, the most important predictors will be:

* MSZoning\_RL: Identifies the general zoning classification of the sale - Residential Low Density
* MSZoning\_RM: Identifies the general zoning classification of the sale - Residential Medium Density
* MSZoning\_FV: Identifies the general zoning classification of the sale - Floating Village Residential
* GarageArea: Size of garage in square feet
* LotArea: Lot size in square feet

**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:**

The model is required to be as simple as possible. For a simpler model the accuracy might be low but it would be more robust and generalisable. This can be understood through bias-variance trade off.

Bias is when the model is unable to learn from the data provided. A high bias in the model means that model performs poorly on training and test data.

Variance is when the model over learns from the training data. High variance means that the model performs really well on training data but performs poorly on the test data.

The robustness and generalization of a model depends on the model’s ability to strike a balance between bias and variance.