## 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 :**

The optimal value of alpha for ridge and lasso regression

1. Ridge Alpha = 0.3
2. lasso Alpha = 0.0001

| Metric | Linear Regression | Ridge Regression | Lasso Regression |
|---|---|---|---|
| R2 Score Train | 0.840192 | 0.839569 | 0.838861 |
| R2 Score Test | 0.770704 | 0.788097 | 0.792077 |
| RSS Train | 5.687107 | 5.709275 | 5.734475 |
| RSS Test | 3.874647 | 3.580751 | 3.513482 |
| MSE Train | 0.005809 | 0.005832 | 0.005857 |
| MSE Test | 0.009225 | 0.008526 | 0.008365 |
| RMSE Train | 0.076217 | 0.076366 | 0.076534 |
| RMSE Test | 0.096049 | 0.092334 | 0.091463 |

[106]:

| | Linear | Ridge | Lasso |
|---|---|---|---|
| MSSubClass | -0.081507 | -0.077529 | -0.078457 |
| MSZoning | -0.056144 | -0.056533 | -0.051615 |
| LotFrontage | -0.147273 | -0.124712 | -0.098191 |
| LotArea | 0.189829 | 0.168074 | 0.136816 |
| Street | 0.157030 | 0.147080 | 0.123167 |
| Utilities | -0.148644 | -0.112552 | -0.042532 |
| LandSlope | 0.050361 | 0.053391 | 0.051050 |
| Condition2 | -0.150373 | -0.123803 | -0.084784 |
| OverallQual | 0.352826 | 0.356576 | 0.360227 |
| OverallCond | 0.142210 | 0.141200 | 0.137280 |
| YearBuilt | 0.175854 | 0.171377 | 0.173571 |
| ExterQual | -0.056136 | -0.058145 | -0.056056 |
| BsmtQual | -0.061259 | -0.062336 | -0.062124 |
| BsmtFinSF2 | 0.054453 | 0.052533 | 0.045578 |
| TotalBsmtSF | 0.059232 | 0.060484 | 0.030649 |
| GrLivArea | 0.716484 | 0.663262 | 0.694659 |
| BsmtFullBath | 0.106874 | 0.103575 | 0.107183 |
| KitchenQual | -0.063108 | -0.063162 | -0.061334 |
| Functional | 0.067929 | 0.064413 | 0.061593 |
| Fireplaces | 0.073346 | 0.079607 | 0.077672 |
| GarageCars | 0.078568 | 0.087237 | 0.085216 |
| GarageArea | 0.079795 | 0.076577 | 0.070600 |
| ScreenPorch | 0.086382 | 0.080720 | 0.076026 |
| PoolArea | -0.437712 | -0.377045 | -0.382027 |
| MiscVal | 0.076884 | 0.062885 | 0.038098 |

These are the metrics and coefficients calculated for the optimal alpha value.

When we choose to double the value of alpha i.e., ridge alpha = 0.6 and lasso alpha = 0.0002, we get

| Metric | Linear Regression | Ridge Regression | Lasso Regression |
|---|---|---|---|
| R2 Score Train | 0.840192 | 0.838185 | 0.835515 |
| R2 Score Test | 0.770704 | 0.798683 | 0.804947 |
| RSS Train | 5.687107 | 5.758531 | 5.853544 |
| RSS Test | 3.874647 | 3.401854 | 3.296012 |
| MSE Train | 0.005809 | 0.005882 | 0.005979 |
| MSE Test | 0.009225 | 0.008100 | 0.007848 |
| RMSE Train | 0.076217 | 0.076695 | 0.077325 |
| RMSE Test | 0.096049 | 0.089998 | 0.088587 |

| | Linear | Ridge | Lasso |
|---|---|---|---|
| MSSubClass | -0.081507 | -0.074148 | -0.075575 |
| MSZoning | -0.056144 | -0.056545 | -0.047020 |
| LotFrontage | -0.147273 | -0.106794 | -0.049531 |
| LotArea | 0.189829 | 0.152256 | 0.085226 |
| Street | 0.157030 | 0.138305 | 0.089340 |
| Utilities | -0.148644 | -0.090463 | -0.000000 |
| LandSlope | 0.050361 | 0.055156 | 0.051346 |
| Condition2 | -0.150373 | -0.104990 | -0.019614 |
| OverallQual | 0.352826 | 0.358132 | 0.367837 |
| OverallCond | 0.142210 | 0.139940 | 0.132092 |
| YearBuilt | 0.175854 | 0.167793 | 0.171360 |
| ExterQual | -0.056136 | -0.059953 | -0.055656 |
| BsmtQual | -0.061259 | -0.063337 | -0.062834 |
| BsmtFinSF2 | 0.054453 | 0.051115 | 0.037741 |
| TotalBsmtSF | 0.059232 | 0.061675 | 0.000827 |
| GrLivArea | 0.716484 | 0.619093 | 0.672318 |
| BsmtFullBath | 0.106874 | 0.100619 | 0.107135 |
| KitchenQual | -0.063108 | -0.063482 | -0.059983 |
| Functional | 0.067929 | 0.061689 | 0.055414 |
| Fireplaces | 0.073346 | 0.084902 | 0.082067 |
| GarageCars | 0.078568 | 0.093453 | 0.091577 |
| GarageArea | 0.079795 | 0.075407 | 0.062115 |
| ScreenPorch | 0.086382 | 0.076496 | 0.067950 |
| PoolArea | -0.437712 | -0.330605 | -0.327112 |
| MiscVal | 0.076884 | 0.052450 | 0.000000 |

It makes some of the coefficients to zero which leads to automatic feature selection in lasso regression. By comparing the metrics, R2 score on training data has decreased but it has increased on testing data for both Ridge Regression and Lasso Regression.

Most important variables after the change are,

1. GrLivArea
2. OverallQual
3. OverallCond
4. YearBuilt
5. LotArea
6. Street
7. Fireplaces
8. GarageCars
9. GarageArea
10. ScreenPorch
11. PoolArea

## 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:**

I will choose Lasso regression to solve this problem because it has optimal test and train r2 scores comparing the other regression models.

## 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:**

After Dropping the five most important predictor, the metrics and the coefficients are

| Metric | Linear Regression | Ridge Regression | Lasso Regression |
|---|---|---|---|
| R2 Score Train | 0.713411 | 0.713209 | 0.712683 |
| R2 Score Test | 0.716183 | 0.715748 | 0.716182 |
| RSS Train | 10.198857 | 10.206057 | 10.224795 |
| RSS Test | 4.795945 | 4.803298 | 4.795960 |
| MSE Train | 0.010418 | 0.010425 | 0.010444 |
| MSE Test | 0.011419 | 0.011436 | 0.011419 |
| RMSE Train | 0.102067 | 0.102103 | 0.102196 |
| RMSE Test | 0.106859 | 0.106941 | 0.106859 |

| | Linear | Ridge | Lasso |
|---|---|---|---|
| **MSSubClass** | -0.010338 | -0.012027 | -0.010965 |
| **MSZoning** | -0.080094 | -0.077587 | -0.076333 |
| **LotFrontage** | 0.071193 | 0.073426 | 0.056436 |
| **Street** | 0.261458 | 0.237881 | 0.225848 |
| **Utilities** | -0.146674 | -0.112876 | -0.042957 |
| **LandSlope** | 0.063538 | 0.061399 | 0.057447 |
| **Condition2** | -0.118733 | -0.099256 | -0.059749 |
| **OverallCond** | 0.233673 | 0.227841 | 0.226079 |
| **YearBuilt** | 0.191974 | 0.191574 | 0.190097 |
| **ExterQual** | -0.140531 | -0.141212 | -0.139582 |
| **BsmtQual** | -0.085302 | -0.087726 | -0.087025 |
| **BsmtFinSF2** | 0.045013 | 0.045572 | 0.036461 |
| **TotalBsmtSF** | 0.386172 | 0.353358 | 0.371395 |
| **BsmtFullBath** | 0.041839 | 0.044492 | 0.041735 |
| **KitchenQual** | -0.098974 | -0.099471 | -0.098581 |
| **Functional** | 0.052458 | 0.052903 | 0.049611 |
| **Fireplaces** | 0.201575 | 0.203159 | 0.203695 |
| **GarageCars** | 0.276910 | 0.275865 | 0.276782 |
| **ScreenPorch** | 0.069256 | 0.068354 | 0.060201 |
| **MiscVal** | 0.002371 | -0.000118 | 0.000000 |

Now the top five predictors are

1. TotalBsmtSF
2. GarageCars
3. OverallCond
4. Street
5. FirePlaces

## 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:**

Develop a robust, generalized model by preprocessing data to handle outliers, retaining relevant ones, and removing irrelevant ones. Train the model with balanced training and test accuracies, employing regularization techniques to prevent overfitting. Evaluate its performance on unseen datasets to ensure trustworthiness for predictive analysis.