**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?

**Optimal value of alpha for ridge and lasso regression:**

**Ridge Regression Alpha:** 3.0

**Lasso Regression Alpha:** 0.0001

| | Metric | Ridge Regression - Double | Ridge Regression | Lasso Regression - Double | Lasso Regression |
|---|---|---|---|---|---|
| 0 | R2 Score (Train) | 0.935624 | 0.940562 | 0.931867 | 0.938596 |
| 1 | R2 Score (Test) | NaN | 0.922261 | NaN | 0.921105 |
| 2 | Diffrence in R2 test and train | NaN | 0.018302 | NaN | 0.017491 |
| 3 | RSS (Train) | 1.422989 | 1.313826 | 1.506023 | 1.357287 |
| 4 | RSS (Test) | NaN | 0.845913 | NaN | 0.858485 |
| 5 | MSE (Train) | 0.001478 | 0.001364 | 0.001564 | 0.001409 |
| 6 | MSE (Test) | NaN | 0.002048 | NaN | 0.002079 |
| 7 | RMSE (Train) | 0.03844 | 0.036937 | 0.039546 | 0.037542 |
| 8 | RMSE (Test) | NaN | 0.045257 | NaN | 0.045592 |

Metric

| | Ridge - Doubled | Ridge | Lasso - Doubled | Lasso |
|---|---|---|---|---|
| GrLivArea | 0.088662 | 0.102379 | 0.263735 | 0.260194 |
| TotalBsmtSF | 0.067986 | 0.074284 | 0.093509 | 0.094972 |
| LotArea | 0.053131 | 0.074790 | 0.089112 | 0.120001 |
| YearBuilt | 0.041315 | 0.059572 | 0.088701 | 0.109246 |
| 9OverallQual | 0.047535 | 0.043062 | 0.064770 | 0.046728 |
| BsmtFinSF1 | 0.061221 | 0.065760 | 0.056669 | 0.056357 |
| Neighborhood-StoneBr | 0.048380 | 0.057387 | 0.056485 | 0.065945 |
| Neighborhood-NridgHt | 0.045599 | 0.046188 | 0.053770 | 0.053160 |
| SaleType-New | 0.031712 | 0.039041 | 0.049490 | 0.044763 |
| Neighborhood-Crawfor | 0.034543 | 0.035812 | 0.045671 | 0.045551 |

Features and Coefficients

**Observation**: We can see that if we double the alpha,r2 score gets weaker and an increase in RSS and MSE values is observed.

**Observation**: Coefficients also have reduced a bit compared to the previous alpha values.Thisis happening because of the higher penalty being imposed.
**Observation** : We can see that all top 10 features remained intact in the list but there is change in the influence on the predictions.'LotArea' was dropped one places from2rdto3th,'YearBuilt'gained 1 places, 'TotalBsmtSF' moved up 2 position, 'Neighborhood-StoneBr' moved up 2 position, 'Neighborhood-NridgHt' gained 1 position, '9OverallQual' dropped 2 positions and so on.

- The most important predictor variables after the change is implemented are as follows:

| | Ridge - Doubled | Ridge | Lasso - Doubled | Lasso |
|---|---|---|---|---|
| **GrLivArea** | 0.088662 | 0.102379 | 0.263735 | 0.260194 |
| **LotArea** | 0.053131 | 0.074790 | 0.089112 | 0.120001 |
| **YearBuilt** | 0.041315 | 0.059572 | 0.088701 | 0.109246 |
| **TotalBsmtSF** | 0.067986 | 0.074284 | 0.093509 | 0.094972 |
| **Neighborhood-StoneBr** | 0.048380 | 0.057387 | 0.056485 | 0.065945 |
| **BsmtFinSF1** | 0.061221 | 0.065760 | 0.056669 | 0.056357 |
| **Neighborhood-NridgHt** | 0.045599 | 0.046188 | 0.053770 | 0.053160 |
| **9OverallQual** | 0.047535 | 0.043062 | 0.064770 | 0.046728 |
| **Neighborhood-Crawfor** | 0.034543 | 0.035812 | 0.045671 | 0.045551 |
| **SaleType-New** | 0.031712 | 0.039041 | 0.049490 | 0.044763 |

**Observation**: The most important predictors are 'GrLivArea', 'LotArea', 'YearBuilt','TotalBsmtSF', 'Neighborhood-StoneBr', 'BsmtFinSF1', 'Neighborhood-NridgHt', '9OverallQual',
'Neighborhood-Craw for' and 'SaleType-New'

**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** : As explained in our coding conclusion, the best model for our prediction is the **Lasso regression** model with alphas **0.0001**,even though all the metrics are in favor of the Ridge model. But we should take this with a pinch of salt, that if we need a simpler model with lesser parameters to consider then we should go with Lasso as it tends to reduce the number of features by making their coefficient to zero. Which will avoid overfitting and error in unknown data sets?

```
lasso_zero_betas = filter(lambda x: x == 0, lasso_model.coef_)
print("Lasso Zero Value Betas : ",len(list(lasso_zero_betas)))
ridge_zero_betas = filter(lambda x: x == 0, ridge_model.coef_)
print("Ridge Zero Value Betas : ",len(list(ridge_zero_betas)))
```

```
Lasso Zero Value Betas :  126
Ridge Zero Value Betas :  8
```

Lesser features are required by lasso regression model

| | Metric | Linear Regression with RFE | Ridge Regression | Lasso Regression |
|---|---|---|---|---|
| 0 | R2 Score (Train) | 0.796219 | 0.940562 | 0.938596 |
| 1 | R2 Score (Test) | 0.795216 | 0.922261 | 0.921105 |
| 2 | Diffrence in R2 test and train | 0.001003 | 0.018302 | 0.017491 |
| 3 | RSS (Train) | 4.504449 | 1.313826 | 1.357287 |
| 4 | RSS (Test) | 2.228344 | 0.845913 | 0.858485 |
| 5 | MSE (Train) | 0.004678 | 0.001364 | 0.001409 |
| 6 | MSE (Test) | 0.005396 | 0.002048 | 0.002079 |
| 7 | RMSE (Train) | 0.068392 | 0.036937 | 0.037542 |
| 8 | RMSE (Test) | 0.073454 | 0.045257 | 0.045592 |

Ridge and Lasso are comparably similar even though Ridge is slightly better

Q3: After building the model, you realized 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**: We will check current top 5 variables and drop them from our data set and build the model again with same alpha value for Lasso regression and Ridge regression

| | Lasso |
|---|---|
| **GrLivArea** | 0.260194 |
| **LotArea** | 0.120001 |
| **YearBuilt** | 0.109246 |
| **TotalBsmtSF** | 0.094972 |
| **Neighborhood-StoneBr** | 0.065945 |

```
Lasso_best_columns.index
```

```
Index(['GrLivArea', 'LotArea', 'YearBuilt', 'TotalBsmtSF',
       'Neighborhood-StoneBr'],
      dtype='object')
```

Oldtop5variables Now

we will check the new ones which are as follows :

| | Lasso - New |
|---|---|
| **1stFlrSF** | 0.211862 |
| **2ndFlrSF** | 0.158900 |
| **BsmtFinSF1** | 0.118434 |
| **BsmtUnfSF** | 0.068014 |
| **9OverallQual** | 0.058390 |

```
betas_new.sort_values(by=['Lasso - New'],ascending=False).head(5).index
```

```
Index(['1stFlrSF', '2ndFlrSF', 'BsmtFinSF1', 'BsmtUnfSF', '9OverallQual'], dtype='object')
```

**Observation**: The new top features for Lasso Regression are:**'1stFlrSF','2ndFlrSF', 'BsmtFinSF1', 'BsmtUnfSF' and '9OverallQual'**

**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**:

To ensure that the model is robust and generalisable we have to choose the model carefully considering a couple of parameters. We have to closely monitor **Bias** of the model with respect to the **Variance**. Also we should check the **total error** produced by the model, we should aim to minimize all these three parameters at the same time. Usually we have to give up on bias to gain on the variance and vise-versa. Also it is observed that by keeping the model simple we achieve this balance and our model becomesmorerobustandgeneralisabletounseendata.Wehavecoupleofmethodsto achieve this, such as **cross validation** and **regularization**

It is observed that the accuracy of the model **decreases** as we achieve this balance, this happens since **we give up on bias** of the model towards seen data.