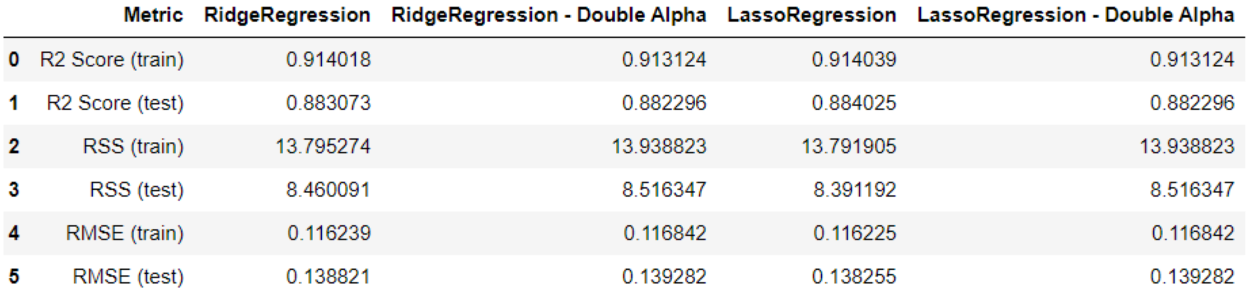
**Question-1:**

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:**

Optimal value of alpha for: Ridge = 0.5 and Lasso = 0.0001

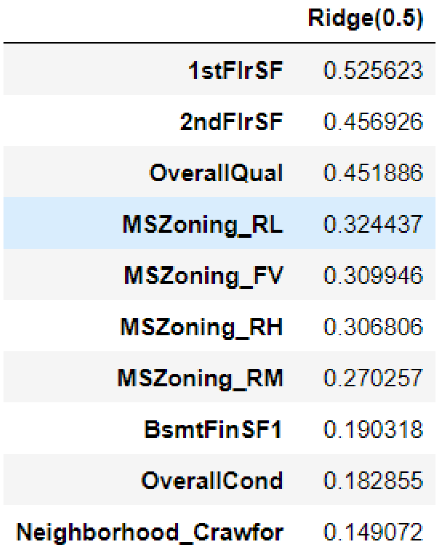
Comparison of different metrics by doubling the alpha i.e. for Ridge = 1.0 and Lasso = 0.0002



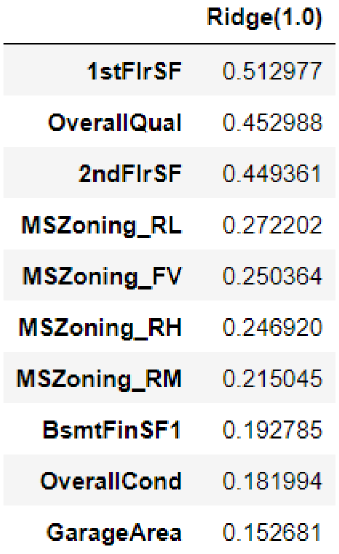
|  | Metric | RidgeRegression | RidgeRegression - Double Alpha | LassoRegression | LassoRegression - Double Alpha |
|---|---|---|---|---|---|
| 0 | R2 Score (train) | 0.914018 | 0.913124 | 0.914039 | 0.913124 |
| 1 | R2 Score (test) | 0.883073 | 0.882296 | 0.884025 | 0.882296 |
| 2 | RSS (train) | 13.795274 | 13.938823 | 13.791905 | 13.938823 |
| 3 | RSS (test) | 8.460091 | 8.516347 | 8.391192 | 8.516347 |
| 4 | RMSE (train) | 0.116239 | 0.116842 | 0.116225 | 0.116842 |
| 5 | RMSE (test) | 0.138821 | 0.139282 | 0.138255 | 0.139282 |

- We can see that that there is slight reduction in R2 score for train and test data for both Ridge and Lasso regression.
- There is very slight changes in RSS and RMSE as well for train and test data.

Comparison of Top 10 predictor variable after doubling the alpha:

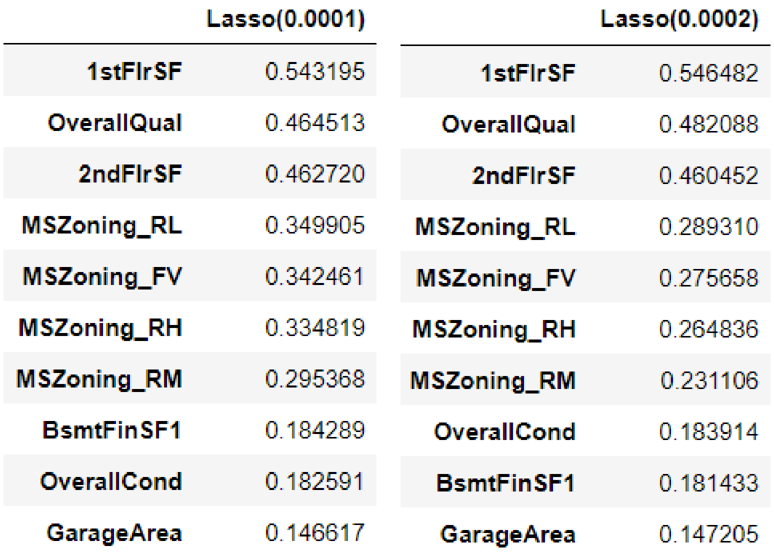


|  | Ridge(0.5) |
|---|---|
| 1stFlrSF | 0.525623 |
| 2ndFlrSF | 0.456926 |
| OverallQual | 0.451886 |
| MSZoning_RL | 0.324437 |
| MSZoning_FV | 0.309946 |
| MSZoning_RH | 0.306806 |
| MSZoning_RM | 0.270257 |
| BsmtFinSF1 | 0.190318 |
| OverallCond | 0.182855 |
| Neighborhood_Crawfor | 0.149072 |



|  | Ridge(1.0) |
|---|---|
| 1stFlrSF | 0.512977 |
| OverallQual | 0.452988 |
| 2ndFlrSF | 0.449361 |
| MSZoning_RL | 0.272202 |
| MSZoning_FV | 0.250364 |
| MSZoning_RH | 0.246920 |
| MSZoning_RM | 0.215045 |
| BsmtFinSF1 | 0.192785 |
| OverallCond | 0.181994 |
| GarageArea | 0.152681 |

- Most important predictor variable remains same as 1stFloorSF after doubling the alpha for Ridge
- Coefficients of the features slightly changing by doubling the alpha



| | Lasso(0.0001) |
|---|---|
| 1stFlrSF | 0.543195 |
| OverallQual | 0.464513 |
| 2ndFlrSF | 0.462720 |
| MSZoning_RL | 0.349905 |
| MSZoning_FV | 0.342461 |
| MSZoning_RH | 0.334819 |
| MSZoning_RM | 0.295368 |
| BsmtFinSF1 | 0.184289 |
| OverallCond | 0.182591 |
| GarageArea | 0.146617 |

| | Lasso(0.0002) |
|---|---|
| 1stFlrSF | 0.546482 |
| OverallQual | 0.482088 |
| 2ndFlrSF | 0.460452 |
| MSZoning_RL | 0.289310 |
| MSZoning_FV | 0.275658 |
| MSZoning_RH | 0.264836 |
| MSZoning_RM | 0.231106 |
| OverallCond | 0.183914 |
| BsmtFinSF1 | 0.181433 |
| GarageArea | 0.147205 |

- Most important predictor variable remains same as 1stFloorSF after doubling the alpha for Lasso
- Coefficients of the features slightly changing by doubling the alpha for Lasso

**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?

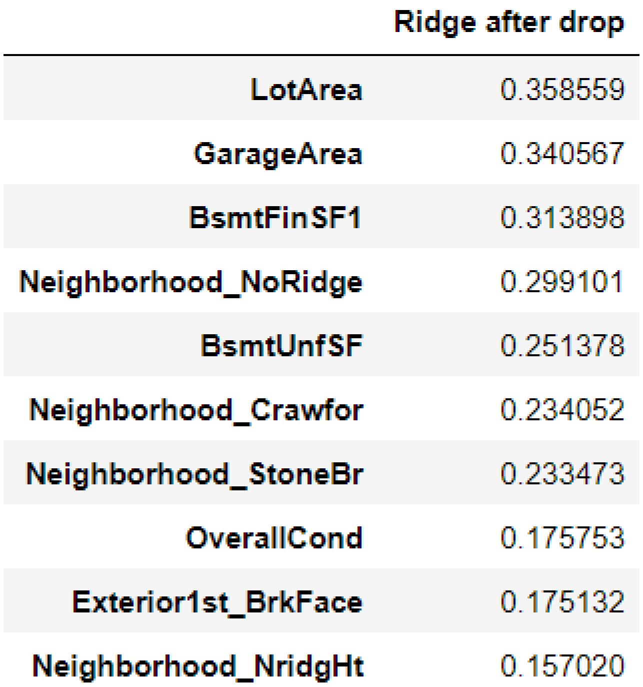
**Answer:**

We select the Lasso Regression model coefficients for its slight better score of r2 score for test data ( 0.884 over 0.883). Also Lasso regression eliminates some features without affecting the model accuracy.

**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?

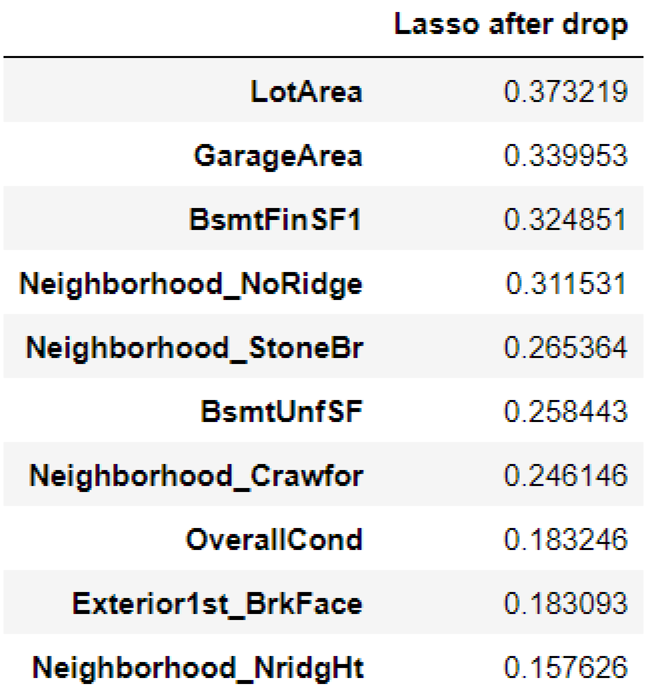
**Answer:**



| | Ridge after drop |
|---|---|
| LotArea | 0.358559 |
| GarageArea | 0.340567 |
| BsmtFinSF1 | 0.313898 |
| Neighborhood_NoRidge | 0.299101 |
| BsmtUnfSF | 0.251378 |
| Neighborhood_Crawfor | 0.234052 |
| Neighborhood_StoneBr | 0.233473 |
| OverallCond | 0.175753 |
| Exterior1st_BrkFace | 0.175132 |
| Neighborhood_NridgHt | 0.157020 |

Five most important predictor variable after dropping the top 5 features in Ridge regression are:

1. LotArea - Lot size in square feet
2. GarageArea - Size of garage in square feet
3. BsmtFinSF1 - Type 1 finished square feet
4. Neighbourhood_NoRidge - Physical locations within Ames city limits (Northridge)
5. BsmtUnfSF - Unfinished square feet of basement area



| | Lasso after drop |
|---|---|
| LotArea | 0.373219 |
| GarageArea | 0.339953 |
| BsmtFinSF1 | 0.324851 |
| Neighborhood_NoRidge | 0.311531 |
| Neighborhood_StoneBr | 0.265364 |
| BsmtUnfSF | 0.258443 |
| Neighborhood_Crawfor | 0.246146 |
| OverallCond | 0.183246 |
| Exterior1st_BrkFace | 0.183093 |
| Neighborhood_NridgHt | 0.157626 |

Five most important predictor variable after dropping the top 5 features in Lasso regression are:

1. LotArea - Lot size in square feet
2. GarageArea - Size of garage in square feet
3. BsmtFinSF1 - Type 1 finished square feet
4. Neighbourhood_NoRidge - Physical locations within Ames city limits (Northridge)
5. Neighbourhood_StoneBr - Physical locations within Ames city limits (Stone Brook)

**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:**

Model is considered to be robust if the model is not overfitting or underfitting. Linear regression can give a good r2 score for training data but it might be overfitting and hence the prediction on test data will not be accurate. To avoid overfitting, regularization is done. After regularization, the r2score of test data comes near to train data and hence we can say the prediction will be better on test data. If the model is not robust then the accuracy of the model will not be good.