# Subjective Questions and Answers

## 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?

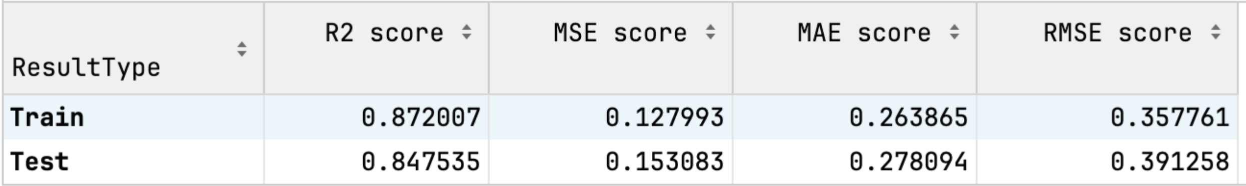
**Answer:**

The optimal values of alpha for Ridge and Lasso Regression are as below:
Ridge: alpha = 1.27
Lasso: alpha = 0.000049

### <u>For Ridge Regression</u>

**The top 10 variables before doubling alpha:**

[‘MSSubClass_OneHalfStoryUnfinished’, ‘GrLivArea’, ‘MSSubClass_OneStoryOlder’, ‘MSSubClass_OneStoryLatest’, ‘PropAgeAtSale’, ‘MSSubClass_OneHalfStoryFinished’, ‘MSSubClass_OneStoryFinishedAttic’, ‘ExterQual_encoded’, ‘BsmtFinSF1’, ‘TotalBsmtSF’].

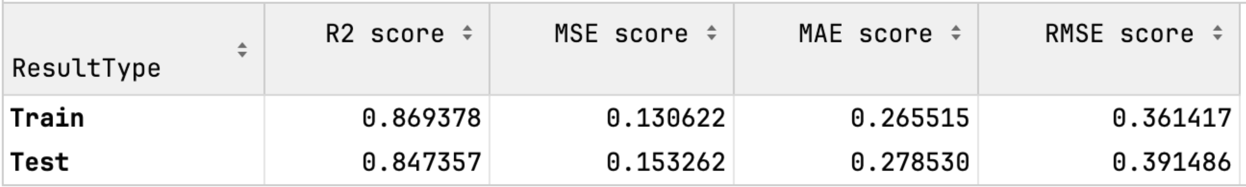


| ResultType | R2 score | MSE score | MAE score | RMSE score |
|---|---|---|---|---|
| **Train** | 0.872007 | 0.127993 | 0.263865 | 0.357761 |
| **Test** | 0.847535 | 0.153083 | 0.278094 | 0.391258 |

**After:**

[‘GrLivArea’, ‘MSSubClass_OneHalfStoryUnfinished’, ‘MSSubClass_OneStoryOlder’, ‘MSSubClass_OneStoryLatest’, ‘PropAgeAtSale’, ‘MSSubClass_OneHalfStoryFinished’, ‘MSSubClass_OneStoryFinishedAttic’, ‘ExterQual_encoded’, ‘BsmtFinSF1’, ‘TotalBsmtSF’]

R2 score decreased. MSE, MAE and RMSE increased.

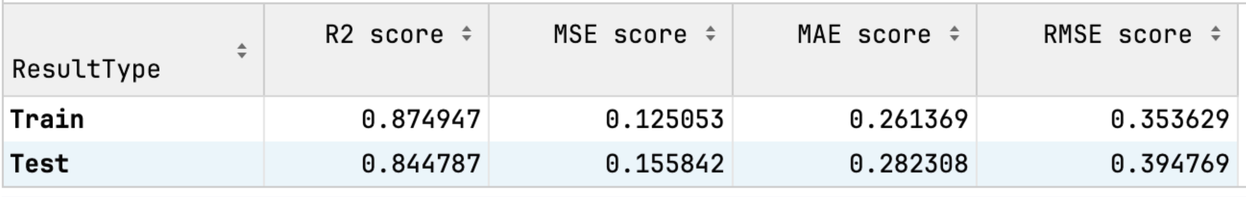


| ResultType | R2 score | MSE score | MAE score | RMSE score |
|---|---|---|---|---|
| **Train** | 0.869378 | 0.130622 | 0.265515 | 0.361417 |
| **Test** | 0.847357 | 0.153262 | 0.278530 | 0.391486 |

## For Lasso Regression:

### The top 10 variables remains before doubling alpha:

[‘MSSubClass_OneHalfStoryUnfinished’, ‘MSSubClass_OneStoryOlder’, ‘GrLivArea’, ‘MSSubClass_OneStoryLatest’, ‘MSSubClass_OneHalfStoryFinished’, ‘PropAgeAtSale’, ‘MSSubClass_OneStoryFinishedAttic’, ‘ExterQual_encoded’, ‘BsmtFinSF1’, ‘TotalBsmtSF’]

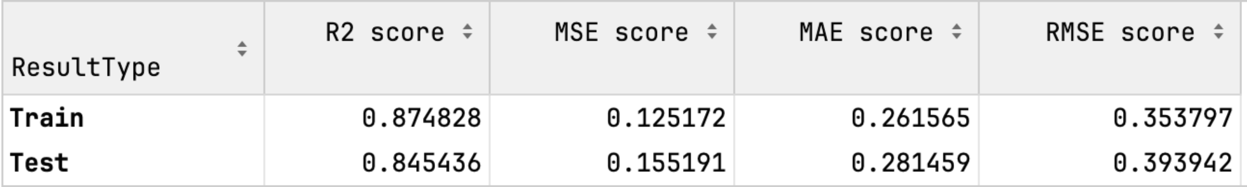


| ResultType | R2 score | MSE score | MAE score | RMSE score |
|---|---|---|---|---|
| **Train** | 0.874947 | 0.125053 | 0.261369 | 0.353629 |
| **Test** | 0.844787 | 0.155842 | 0.282308 | 0.394769 |

### After:

[‘MSSubClass_OneHalfStoryUnfinished’, ‘MSSubClass_OneStoryOlder’, ‘GrLivArea’, ‘MSSubClass_OneStoryLatest’, ‘MSSubClass_OneHalfStoryFinished’, ‘PropAgeAtSale’, ‘MSSubClass_OneStoryFinishedAttic’, ‘ExterQual_encoded’, ‘BsmtFinSF1’, ‘TotalBsmtSF’]

R2, MSE, MAE and RMSE haven't changed much.



| ResultType | R2 score | MSE score | MAE score | RMSE score |
|---|---|---|---|---|
| **Train** | 0.874828 | 0.125172 | 0.261565 | 0.353797 |
| **Test** | 0.845436 | 0.155191 | 0.281459 | 0.393942 |

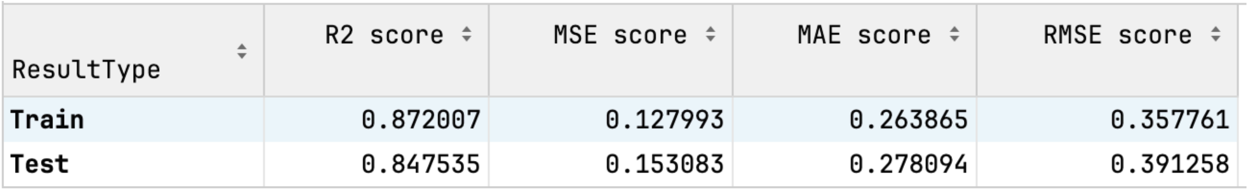
## Question 2:

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:**

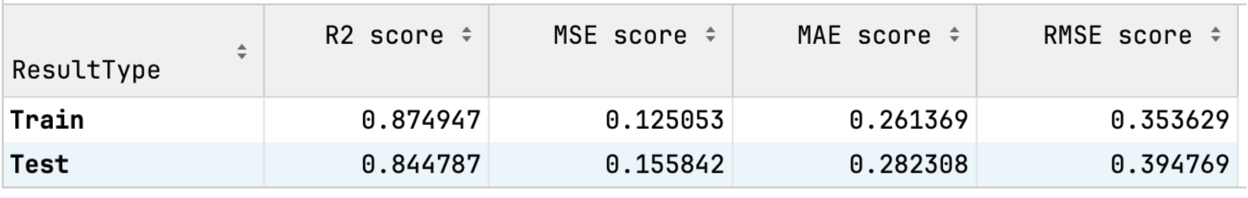
The optimal values of alpha for Ridge and Lasso Regression are as below:
Ridge: alpha = 1.27
Lasso: alpha = 0.000049

For Ridge Regression:



| ResultType | R2 score | MSE score | MAE score | RMSE score |
|---|---|---|---|---|
| **Train** | 0.872007 | 0.127993 | 0.263865 | 0.357761 |
| **Test** | 0.847535 | 0.153083 | 0.278094 | 0.391258 |

For Lasso Regression:



| ResultType | R2 score | MSE score | MAE score | RMSE score |
|---|---|---|---|---|
| **Train** | 0.874947 | 0.125053 | 0.261369 | 0.353629 |
| **Test** | 0.844787 | 0.155842 | 0.282308 | 0.394769 |

- The R2 test score on the Lasso Regression Model is slightly better than that of Ridge Regression Model. Moreover, the training accuracy is slightly reduced; hence, making the model an optimal choice as it seems to perform better on the unseen data.
- The MSE for Test set (Lasso Regression) is slightly lower than that of the Ridge Regression Model; implies Lasso Regression performs better on the unseen test data. Also, since Lasso helps in feature selection (the coefficient values of some of the insignificant predictor variables became 0), implies Lasso Regression has a better edge over Ridge Regression. Therefore, the variables predicted by Lasso can be applied in order to choose significant variables for predicting the price of a house in this analysis. Moreover, while choosing a type of regression in the real world, an analyst has to deal with the lurking and confounding dangers of outliers, non-normality of errors and overfitting especially in sparse datasets among others. Using L2 norm (Ridge) results in

exposing the analyst to such risks. Hence, use of L1 norm (Lasso) could be quite beneficial as it is quite robust to fend off such risks to a large extent, thereby resulting in better and robust regression models.

## Question 3:

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?

**Answer:**
The top 5 features of Lasso model are
['MSSubClass_OneHalfStoryUnfinished', 'MSSubClass_OneStoryOlder', 'GrLivArea', 'MSSubClass_OneStoryLatest', 'MSSubClass_OneHalfStoryFinished']

After eliminating these five predictors and building the model again, the new top 5 predictors are as below:
['Neighborhood_StoneBr', 'Foundation_Slab', 'MSSubClass_OneStoryPUD', 'MSSubClass_TwoStoryNewer', 'MSSubClass_TwoStoryOlder']

## Question 4:

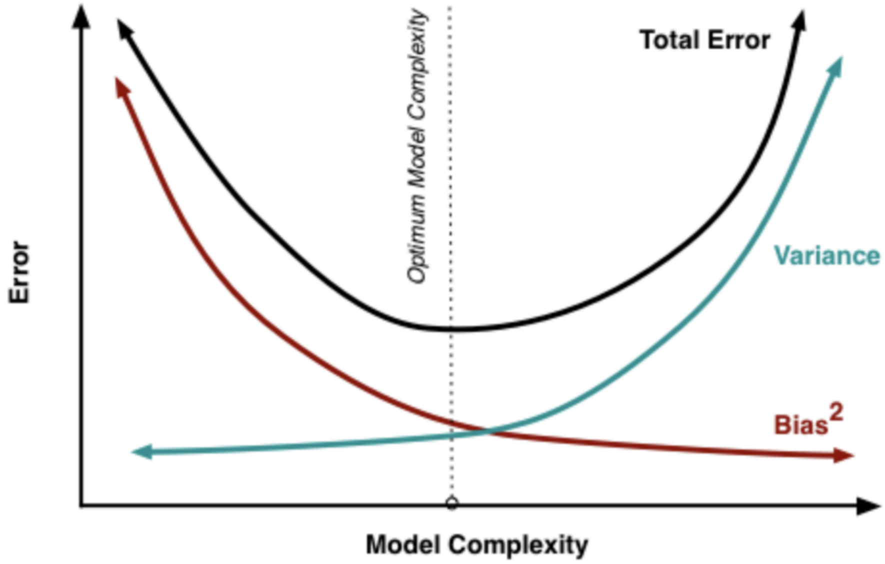
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

**Answer:**
Robustness of a model implies, either the testing error of the model is consistent with the training error, the model performs well with enough stability even after adding some noise to the dataset. Thus, the robustness (or generalizability) of a model is a measure of its successful application to data sets other than the one used for training and testing. By implementing regularization techniques, we can control the trade-off between model complexity and bias which is directly connected to the robustness of the model. Regularization helps in penalizing the coefficients for making the model

too complex; thereby allowing only the optimal amount of complexity to the model. It helps in controlling the robustness of the model by making the model optimally simpler. Therefore, in order to make the model more robust and generalizable, one needs to make sure that there is a delicate balance between keeping the model simple and not making it too naive to be of any use. Also, making a model simple leads to BiasVariance Trade-off:

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

Bias helps you quantify how accurate the model is likely to be on test data. A complex model can do an accurate job prediction provided there has to be enough training data. Models that are too naïve, for e.g., one that gives same results for all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high. Variance is the degree of changes in the model itself with respect to changes in the training data. Thus, accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph.


Optimum Model Complexity
Total Error
Variance
Bias$^2$
Error
Model Complexity