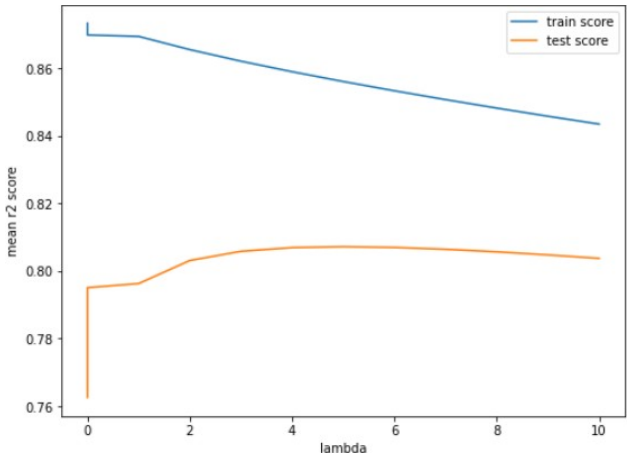
Question 1

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?

As a general trend as the Lambda value increases, the r2 score decreases. which indicates that the error increases. Because the model tends to underfit and more generalized.

Thus ,by increasing (doubling or more) the Lambda value the Variance decreases and Bias increases, i.e the model becomes more simple and thus leads to a underfitting the underlying data.

Ridge Regression:



Test score with the very low value of Lambda, the error is high as we can see the r2 value decreased. But the error for the train set is low. It means that the model is becomes overfitted with very low value of Lambda.

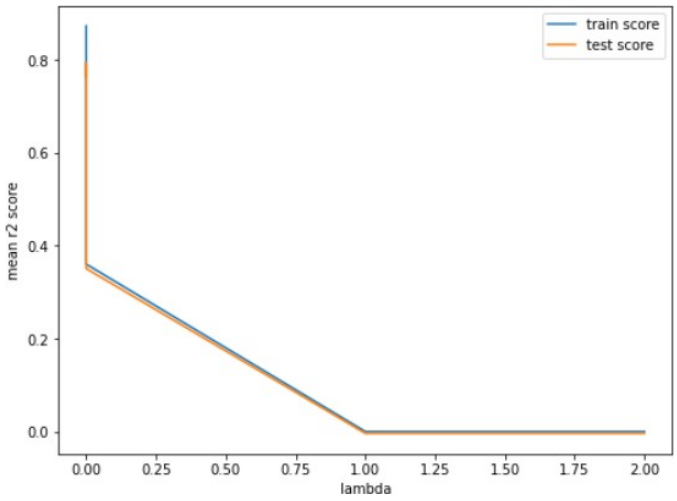
With the increasing value of Lambda(doubling or more), the error started decreasing more and it reached to a peak at lambda=4. Here, the error is least and accuracy (r2 score) is the highest.

After Lambda=4, the r2 score started decreasing as the Lambda is increasing. Hence, the model accuracy started dipping.

We need to pick the value of alpha for which the test score peaks up. In this case in Lambda=4, the error is least in the test set and hence the accuracy is around 80.5%.

So, the optimum Lambda will be 4, for which we will have a right balance between the error and the generalization of the model for creating a simpler model.

LASSO Regression:



From the above graph we can analyse that with very lower value of alpha (~ 0.01 - 0.02), the accuracy of the train and test set is the highest.

Train Score:

As the alpha (lambda) increases, the r2 score decreases. That means the error increases at a large pace. Because the model tends to get underfit and more generalised. At 0.002 (close to 0) the train set accuracy is highest(Approx 85%).

Test Score:

At alpha = 0.002 the test accuracy is highest (Approx 80%). After alpha=0.002, the r2 score starts to decrease as the alpha is increasing. Hence, the model accuracy started to dip. To pick the optimum alpha / lambda value, In this case at alpha=0.002, the error is least in the test set and hence the accuracy is around 80%. We have train set accuracy 73.77% and test set accuracy is 69.77% from R2 score.

So, the optimum alpha will be 0.002, for which we will have a right balance between the error and the generalisation of the model for creating a good model.

Lasso regression with optimal alpha / lambda: 0.002

R2 score of Ridge regression is slightly higher than Lasso. So I would choose Ridge regression for this problem.

The most important predictor variables after the changes are :

Params	Coef
TotRmsAbvGrd	0.103
Neighborhood_StoneBr	0.075
FullBath	0.075
GarageArea	0.070
Neighborhood_NoRidge	0.070
ExterQual	0.061
BsmtQual	0.061
Neighborhood_NridgHt	0.057
OverallCond	0.052
MasVnrArea	0.051
TotalBsmtSF	0.045

RIDGE regression

Params	Coef
ExterQual	0.113
TotRmsAbvGrd	0.073
GarageArea	0.069
FullBath	0.053
FireplaceQu	0.051
Neighborhood_NoRidge	0.040
BsmtExposure	0.039
BsmtQual	0.032
Neighborhood_NridgHt	0.031
GarageFinish	0.025
BsmtFinType1	0.023

Lasso Regression

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?

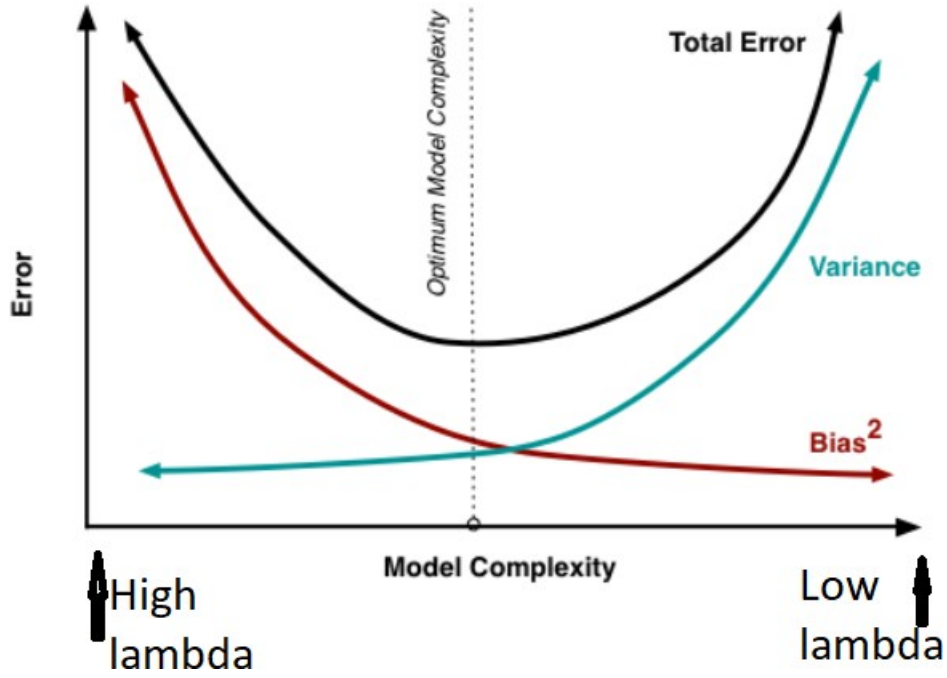


Figure 1

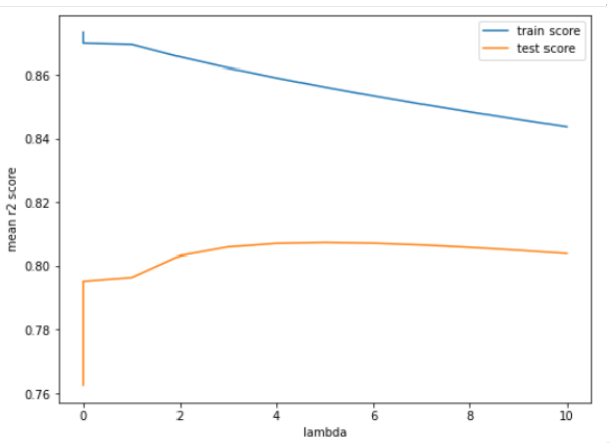


Fig2: RIDGE regression

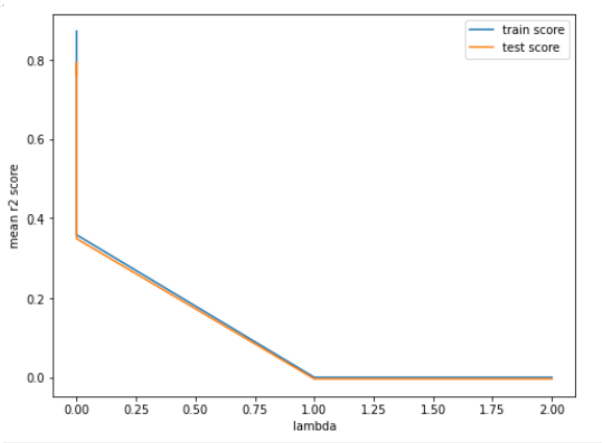


Fig3: LASSO regression

From Figure 1 it is very clear that :

When choosing a lambda value, the goal is to strike the right balance between simplicity and training-data fit:

- If the lambda value is too high, our model will be simple, and we run the risk of underfitting the given data. The model won't learn enough about the training data to make useful predictions.
- If your lambda value is too low, our model will be more complex, and we run the risk of overfitting the given data. The model will learn too much about the particularities of the

training data, and won't be able to generalize to new data. Hence the test accuracy becomes unpredictable

From Fig2 :
With the increasing value of Lambda(doubling or more), the error started decreasing more and it reached to a peak at lambda=4. Here, the error is least and accuracy (r2 score) is the highest.

We need to pick the value of alpha for which the test score peaks up. In this case in Lambda=4, the error is least in the test set and hence the accuracy is around 80.5%.

For Ridge regression I will choose lambda (alpha) =4

From Fig3:

At 0.002 (close to 0) the train set accury is highest(Approx 85%).

At alpha = 0.002 the test accuracy is highest (Approx 80%). After alpha=0.002, the r2 score starts to decrease as the alpha is increasing. Hence, the model accuracy started to dip. To pick the optimum alpha / lambda value, In this case at alpha=0.002, the error is least in the test set and hence the accuracy is around 80%.

For Lasso regression I will choose lambda (alpha) = 0.02.

R2 score of Ridge regression is much higher than Lasso and also for slightly varying values of lambda, the model R2 score for Ridge regression decreases slowly, whereas in case of Lasso the model score drops drastically. So, I would choose Ridge regression for this problem.

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?

The top 5 predictors during initial lasso regression.

ExterQual	0.113
TotRmsAbvGrd	0.073
GarageArea	0.069
FullBath	0.053
FireplaceQu	0.051

Top 5 predictors after removing the above predictors are:

Params	Coef
Fireplaces	0.080
BsmtQual	0.073
constant	0.063
Neighborhood_NoRidge	0.063
Neighborhood_NridgHt	0.063

The R2score of training and testing data has decreased considerably. The overall model fit has reduced after removing the top 5 predictor variables.

Question 4

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

A model is considered to be robust if its output and forecasts are consistently accurate even if one or more of the input variables or assumptions are drastically changed due to unforeseen circumstances. It is therefore a measure of its successful application to data sets other than the one used for training and testing.

Models can be categorized into the below:

- Underfit Model. A model that fails to sufficiently learn the problem and performs poorly on a training dataset and does not perform well on a holdout sample.
- Overfit Model. A model that learns the training dataset too well, performing well on the training dataset but does not perform well on a hold out sample.
- Good Fit Model. A model that suitably learns the training dataset and generalizes well to the old out dataset.

1. A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalizable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training.
2. Too much weightage should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model.
3. Normalizing the data before using for training the model helps in rightsizing the predictor variable co-efficient. This would help standardize the predictions made by the model. If the model is not robust , it cannot be trusted for predictive analysis.
4. Regularization by Ridge or Lasso methods should be performed so unnecessary co-efficient are not considered for modelling.

5. Multi order polynomials (order >2) should be avoided so that the model built is simple and is generalizable.

The above methods ensure the model built is usually Robust and Generalizable.