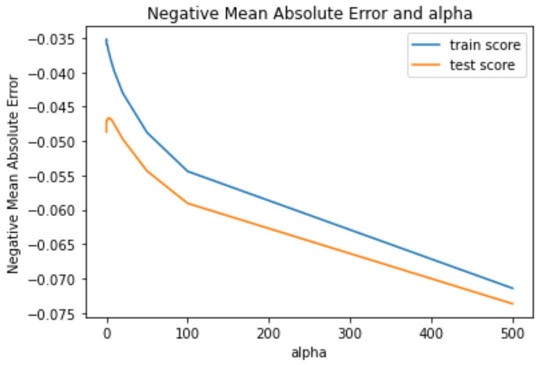
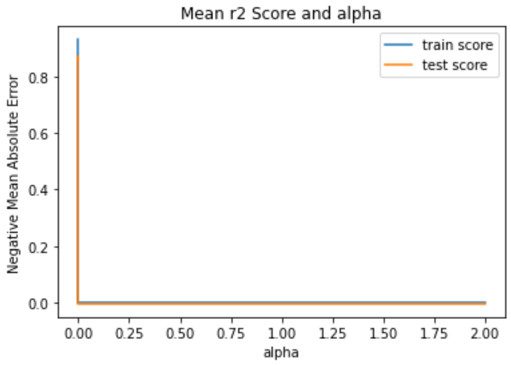
Advanced Regression Assignment

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 are the values for which the error term (MAE or Mean R2 Score) of the estimator is minimum.
On plotting the root mean squared error against a continuous set of lambda values, optimal value can be found out by the elbow curve.
As the value of alpha increases, the model complexity reduces.
Though higher values of alpha reduce overfitting, significantly high values can cause underfitting as well.

Ridge Regression	Lasso Regression
	

In above image, for Ridge Regression, plot between **negative mean absolute error (negative MAE)** and **alpha** is provided. As the value of alpha increase, **negative MAE** increases and then decreases on test score.
When the value of alpha is 3, negative MAE is maximum, or MAE is minimum.

Similarly, for lasso regression, plot between **Mean R2 Score** and **alpha** is provided.
When value of alpha increases, the model try to penalize more and try to make most of the coefficient value zero.
When the value of alpha is 0.001, Mean R2 Score is minimum.

On doubling the value of alpha for ridge, more penalty is applied on the curve and model becomes more generalized and simpler. It does not fit every data of the data set.
Similarly, on doubling the value of alpha for lasso, more penalty is applied on the curve and more coefficient of the variables are reduced to zero.

The most important variable after the changes has been implemented for ridge and lasso regression

Ridge Regression	Lasso Regression
1stFlrSF	GrLivArea
TotalBsmtSF	TotalBsmtSF
OverallQual_8	OverallQual_8
2ndFlrSF	KitchenQual_5
Neighborhood_Somerst	Neighborhood_Crawfor
ExterQual_5	Neighborhood_Somerst
Fireplaces_2	BsmtQual_5
GarageCars_4	GarageArea
GarageCars_3	BsmtFinSF1
GarageFinish_3	LotArea

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?

Normal regression gives you unbiased regression coefficients (maximum likelihood estimates).
Mean Square Error (MSE) = Variance + Bias2 + Irreducible error
The basic idea of both ridge and lasso regression is to introduce a little bias so that the variance can be substantially reduced, which leads to a lower overall MSE.

Ridge and lasso regression allow you to regularize ("shrink") coefficients. This means that the estimated coefficients are pushed towards 0, to make them work better on new data-sets ("optimized for prediction"). This allows usage of complex models and avoid over-fitting at the same time.
For both ridge and lasso, we need to set the tuning parameter that defines how aggressive regularization is performed. Tuning-parameters are usually chosen by cross-validation.

For Ridge regression the tuning-parameter is often called "alpha", α or "L2"; it simply defines regularization strength. The 2-norm of a vector is the square root of the sum of the squared values in your vector.
Minimization objective = RSS + $\lambda\Sigma\beta j^2$
As we increase the value of lambda the variance in model is dropped and bias increases. Ridge regression includes all variables in final model.

For LASSO the meta-parameter is often called "lambda", λ, or "L1". The l1-norm of a vector is the sum of the absolute values in that vector.
Minimization objective = RSS + $\lambda\Sigma|\beta j|$
As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection.

We will apply Lasso regression as it is helping us with variable selection.

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?

After dropping the top 5 variables, train score and test score are dropped.

Original Lasso Regression	Lasso Regression post dropping top 5 variables
train score: 92.2%	train score: 90.84%
test score: 89.86%.	test score: 87.9%.
TotalBsmtSF, BsmtFullBath_3, KitchenQual_5, OverallQual_8, Neighborhood_Crawfor, Neighborhood_Somerst, LotArea, 2ndFlrSF, GarageArea, GarageCars_4 Top 5 Variables	GrLivArea, BsmtQual_5, BsmtFinSF1, GarageArea, ExterQual_5, ExterQual_4, 1stFlrSF, LotArea, Neighborhood_Somerst, MSZoning_RL Top 5 Variables

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?

A model is robust if its output dependent variable is consistently accurate even if one or more of the input independent variables are drastically changed.
It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is error when model is weak to learn from data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.
Variance: Variance is error when model tries to over learn from data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.

One of the common reasons why the training accuracy is much higher than the test accuracy because the model is overfitting. This means that the model accuracy is high with respect to the dataset used while testing train accuracy but it drops significantly when a new dataset is used. This means that the model will not be useful for prediction.
It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.

In order to make the model robust and generalisable, below mentioned steps can be performed

1. **Cross Validation**
   Partitioning the dataset is a good way to check if the model fits datasets which were not used during testing. Cross-validation calculates the accuracy of the model by separating the data into two different populations. This method results in a less biased model compared to other methods since every observation has the chance of appearing in both train and test sets and hence, high accuracy.
2. **Treat missing and Outlier values**
   The unwanted presence of missing and outlier values in the training data often leads to a biased model. It leads to inaccurate predictions. This is because we don't analyse the behavior and relationship with other variables correctly.
   The outlier detection and removal method reduced the variance of the training data.
   It makes the model more generalizable and hence, improves accuracy over train and test data.
3. **Feature transformation and Feature Creation**
   New features may have a higher ability to explain the variance in the training data. Thus, giving improved model accuracy.
4. **Feature Selection**
   Ensure that all the variables which are important are added to the model and ensure the number of independent parameters (which also means the number of coefficients in a polynomial) are much lesser than the number of data points. We can use domain knowledge and PCA to achieve that.
5. **Algorithm Tuning**
   The objective of parameter tuning is to find the optimum value for each parameter to improve the accuracy of the model.