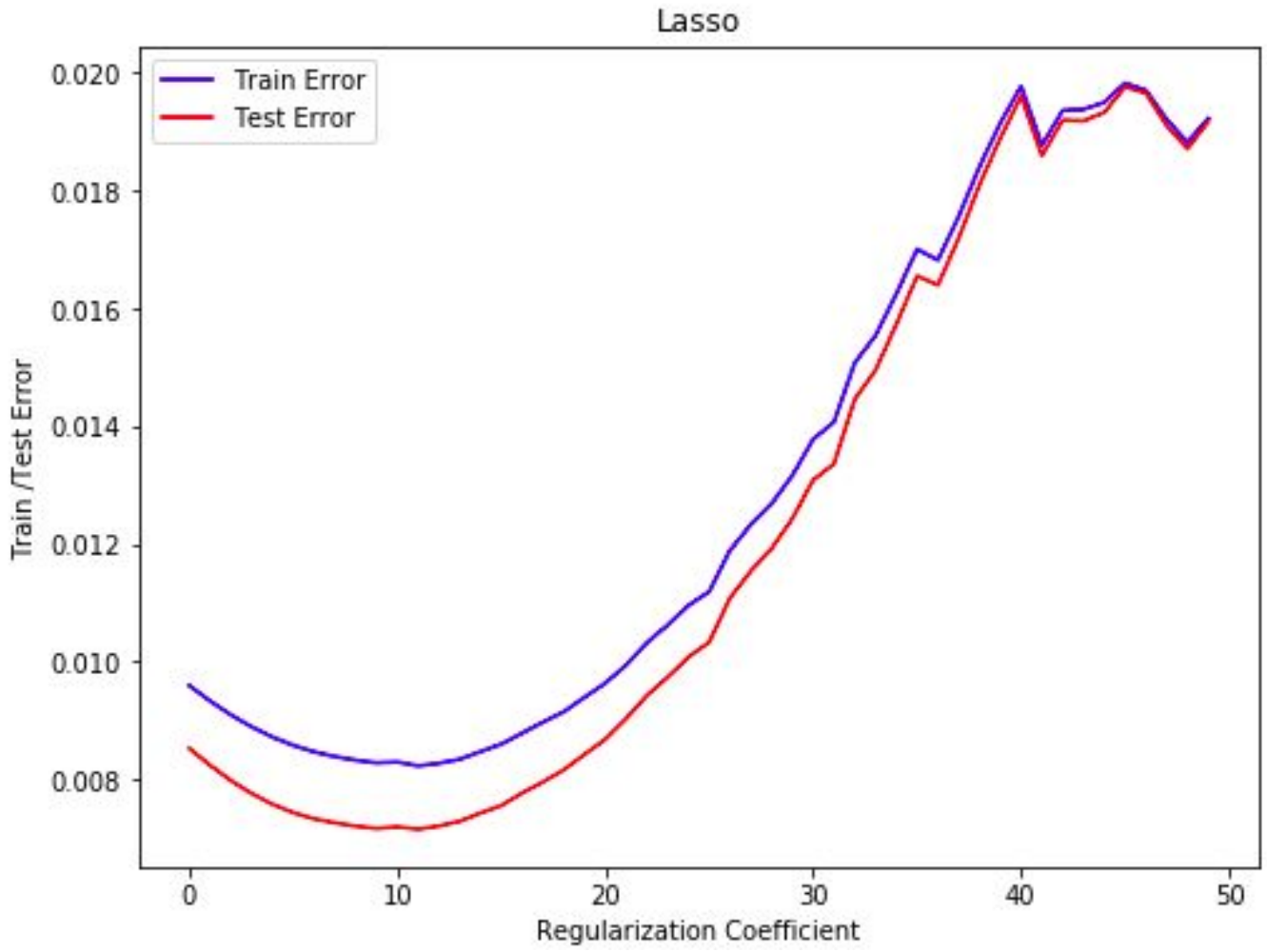
ASSIGNMENT-7 REPORT

Ques 1) Implement Lasso regression also known as L1 regularisation and plot graph between regularisation coefficient λ and error.

Answer)

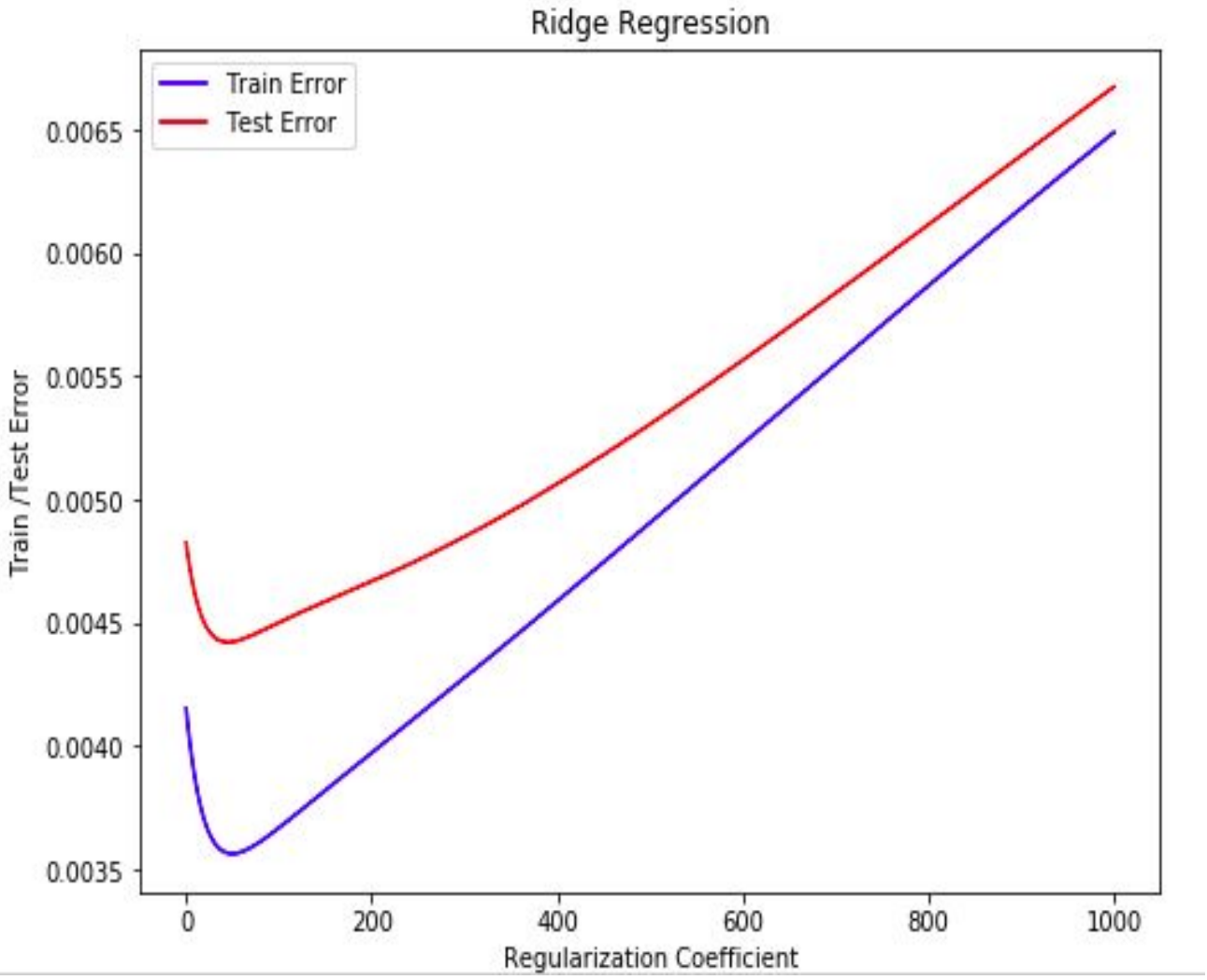
Parameters :
Learning rate : 0.01
Epochs : 400
Lambda range : 2 - 50



Ques 2) Implement Ridge regression also known as L2 regularisation and plot graph between regularisation coefficient λ and error

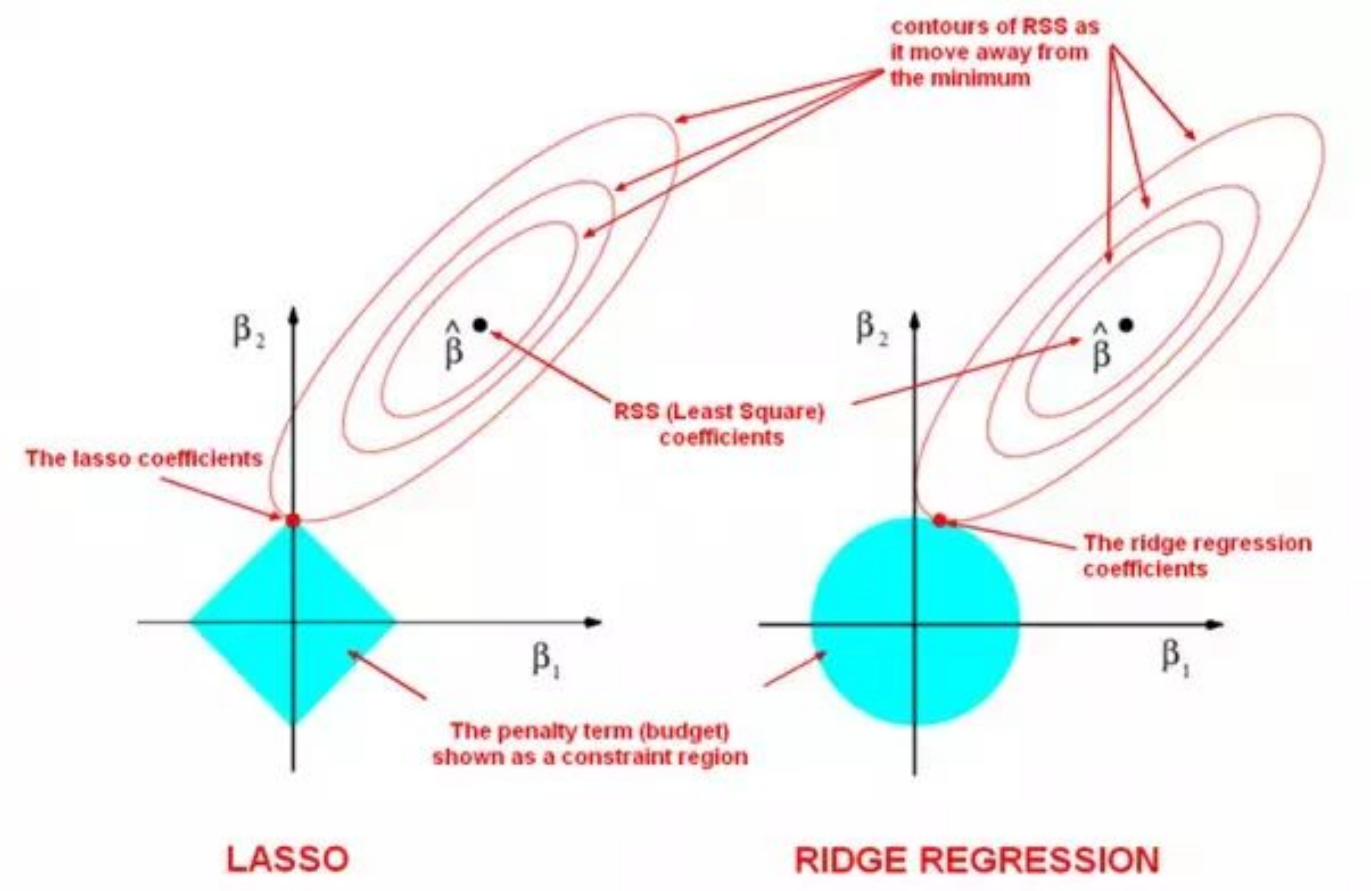
Answer)

Parameters :
Learning rate : 0.05
Epochs : 250
Lambda range : 2 - 1000

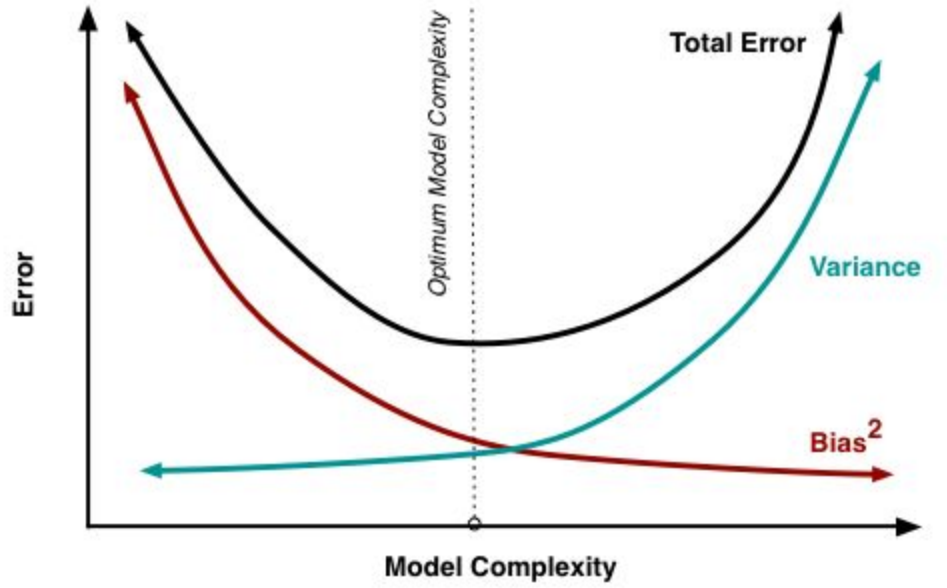


Ques 3) Analyse how the hyper-parameter λ plays a role in deciding between bias and variance.

- Linear Regression with multicollinear data has very high variance but very low bias in the model which results in overfitting. It indicates the values have a large spread from mean and among themselves.
- In such cases, the model captures the noise as a data feature.
- Similarly, if we have low variance and high bias in our model then it would result in underfitting of the data. These models are usually simple models.
- For Linear Regression with multicollinear data, a small increase in bias can result in a big decrease in the variance and which would result in a substantial decrease in Predicted error.
- One of the most common methods to avoid overfitting is by reducing the model complexity using **regularization**.
- When the model coefficients are unconstrained they can tend to have high value , which would result in very high variance in the model. In order to control the variance, we add a constraint to the model coefficients while estimating them. This constraint is nothing but the penalty parameter given by **λ (regularization coefficient).** This type of regression where we add penalty parameter λ in order to estimate the model coefficients is called **Ridge and Lasso regression**.



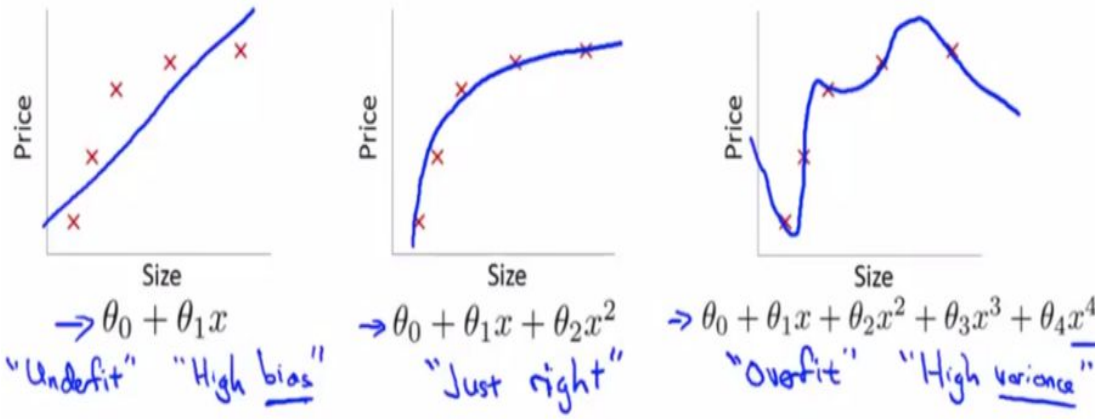
Bias Variance Tradeoff:



When we have a lot of measurements, we can be fairly confident that least squares model accurately reflects the relationships between all the features, but when we have only a few samples, say suppose 2, then we get a straight line between them which perfectly fits the training data. Now this is one of the cases of overfitting. Now for all the test samples,the sum of squared residuals will be very high. So here, in this case, we say that the line is having a **high variance.** In other words, we say that the new line overfits the training Data. The main idea behind ridge regression is to find a new line, that doesn't fit the training data very well. In other words, we introduce a small amount of **Bias** into how the new line is fit to the data. But with a small amount of bias, there is a significant drop in the variance.

So ridge regression, minimizes the **' sum of squared residuals + λ* slope² '.**

Thus we can conclude, Ridge Regression Line, which has small amount of bias due to penalty, has less variance.



1. **Ridge Regression**

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

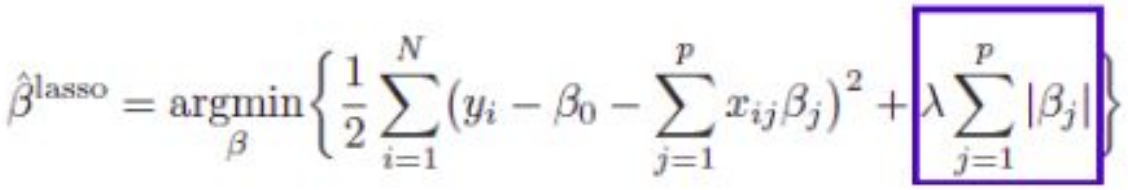
Ridge regression is an extension of Linear regression. It is a regularization method which tries to avoid overfitting of data by penalizing large coefficients. Ridge regression has an additional factor called λ (lambda) which is called the penalty factor which is added while estimating beta coefficients. This penalty factor penalizes high value of beta which in turn shrinks beta coefficients thereby reducing the mean squared error and predicted error.
So, when the value of λ is very small, it has less effect on the parameters, and hence there will be overfitting. On increasing the value of λ

When we have a lot of measurements, we can be fairly confident that least squares model accurately reflects the relationships between all the features, but when we have only a few samples, say suppose 2, then we get a straight line between them which perfectly fits the training data. Now this is one of the cases of overfitting. Now for all the test samples the sum of squared residuals will be very high. So here, in this case, we say that the line is having a high variance. In other words, we say that the new line overfits the training data. The main idea behind ridge regression is to find a new line, that doesn't fit the training data very well.

2) Lasso Regression:
The LASSO (Least Absolute Shrinkage and Selection Operator) is a regression method that involves penalizing the absolute size of the regression coefficients.

By penalizing (or equivalently constraining the sum of the absolute values of the estimates) you end up in a situation where some of the parameter estimates may be exactly zero. The larger the penalty applied, the further estimates are shrunk towards zero.

This is convenient when we want some automatic feature/variable selection, or when dealing with highly correlated predictors, where standard regression will usually have regression coefficients that are 'too large'.



Ques 4) Analyse how the two different regularisation techniques affect regression weights in terms of their values and what are the differences between the two.

Ridge Regression and Lasso Regression do similar things, but in case of lasso regression, they make our predictions of output less sensitive to the tiny training dataset.
When lambda is 0, then ridge and lasso regression, will be same as the Least Square Line. As lambda increases in value, the slope gets smaller until the slope equals to 0. The main difference between Ridge and Lasso Regression is that Ridge Regression can only shrink the slope asymptotically close to 0 while Lasso Regression can shrink the slope all the way to 0.
In case of Ridge Regression we minimize :

The sum of squared residuals + (Lambda * (sum of slope^2))

Now if there are some important parameters, while some unimportant parameters, so on increasing the Lambda value,the parameters that have more effect, shrink less, while that in case of less important parameters, the less important parameters shrink a lot near to 0, but not 0.
In case of Lasso Regression we minimize :

The sum of squared residuals + (Lambda * (sum of |slope|))

For the same case as above, the less important parameters will shrink to 0, and we are left with only the important parameters. The reason is, in this case we have to minimize the sum of |slope|, while in case of ridge regression we have to minimize the sum of **|slope^2|.**

From the above equation, it's clear that Lasso Regression can exclude useless variables from equations, so, it's little better than Ridge Regression at reducing the variance in models that contain a lot of useless variables.

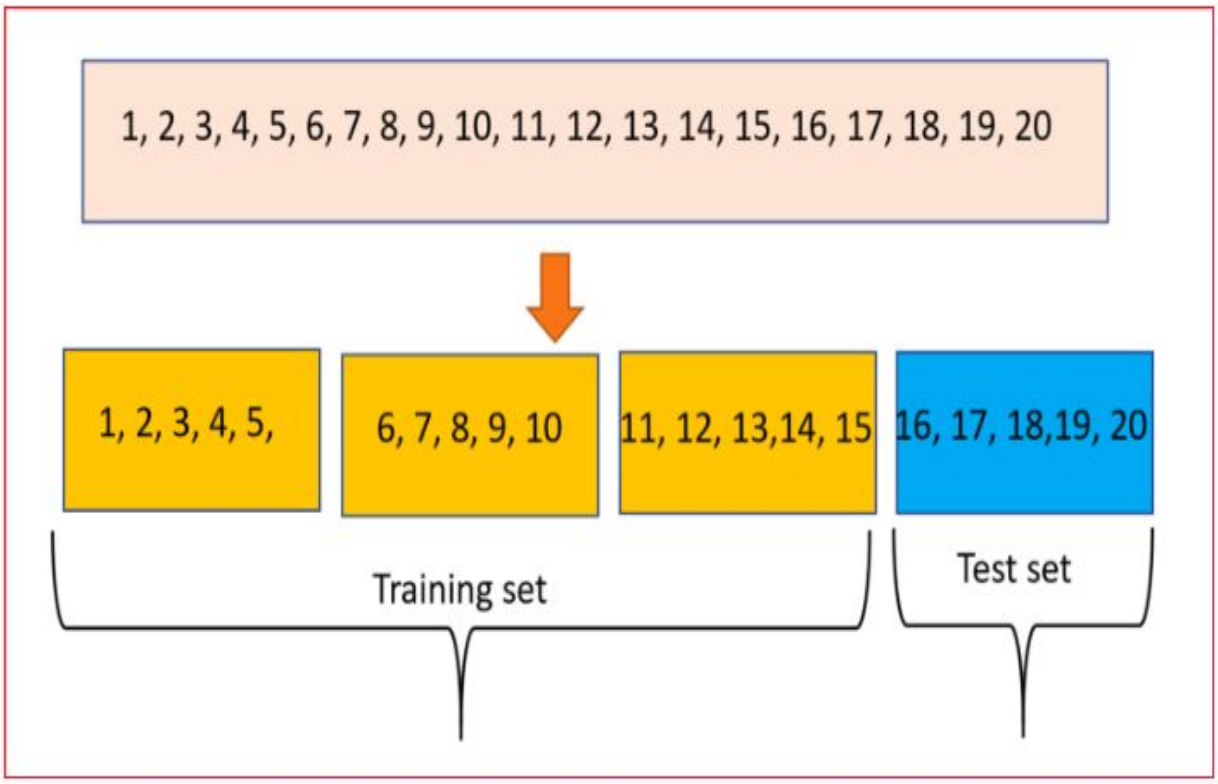
Ques 5) In this part implement regression with k-fold cross validation. Analyse how behavior changes with different values of k. Also implement a variant of this which is the leave-one-out cross validation.

K fold Cross validation

Cross-validation is a statistical technique which involves partitioning the data into subsets, training the data on a subset and use the other subset to evaluate the model's performance.

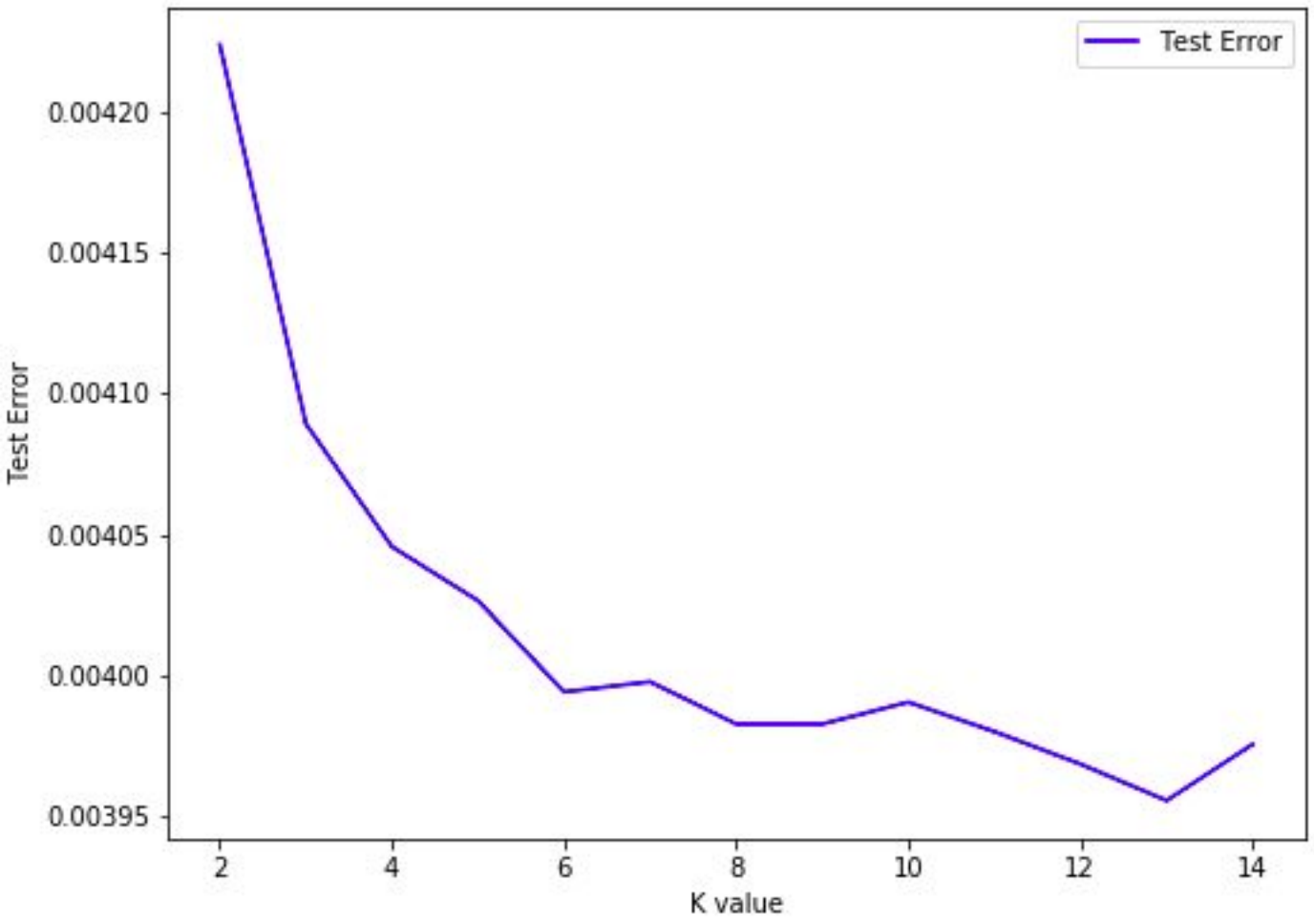
To reduce variability we perform multiple rounds of cross-validation with different subsets from the same data. We combine the validation results from these multiple rounds to come up with an estimate of the model's predictive performance.

Cross-validation will give us a more accurate estimate of a model's performance.



K Fold Splitting

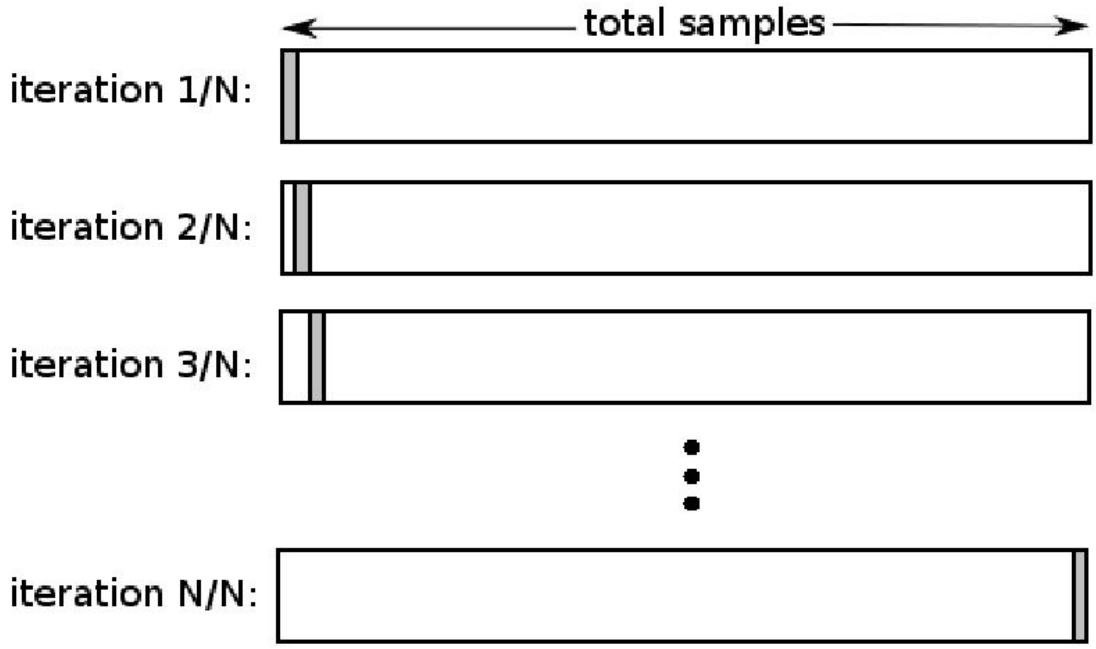
My results :



Observation : The test error rate decreases as the value of k increases.This suggests that when the training samples increase, the model learns more effective features of the data and hence performs better on the unknown data.

In general, the choice of number of folds must allow the size of each validation partition to be large enough to provide a fair estimate of the model's performance on it and at the same time K shouldn't be be too small, say 2, such that we don't have enough trained models to evaluate.

Leave one out Validation:



Leave-one-out cross validation. LOOCV uses a single observation from the original sample as the the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. This is the same as a K-fold cross-validation with K being equal to the number of observations in the original sample. Leave-one-out cross-validation is usually very expensive from a computational point of view because of the large number of times the training process is repeated.