Robot Arm Data Set

A. Feature Selection Methods are generally used to reduce the number of dimension by removing or minimizing the effect of Noise data. Thus, it results in generalization of model. It also helps in removing overfitting of data, but removing too many dimensions might also result in under fit.
For Robot Arm Data Set, Select K best along with F_Regression was the best method, as Robot Arm had only 8 dimension, thus each dimension needed to be judged carefully.
Select K best is one of the univariate feature selection method, here each dimension's strength gets evaluated with respect to output variable. It removes all the K highest scoring features, it determines Score based on methods such as, F_Regression for Regression techniques.
In F_Regression, cross correlation between each target value and evaluated value is calculated.

B. **Support Vector Regression**: It is a form of Support Vector Machines which support Regression by evaluating global minima with use of Epsilon loss function. It can be defined as:

$$f_{svr} = \sum_{d=1}^{D} w_d x_d = xw$$

Coefficients w can be regularized through l2 norm.

$$w^* = \arg min_w \; C \sum_{i=1}^{N} v_\epsilon(y_i - f_{svr}(x_i)) + ||w||^2$$

Where $v_\epsilon = 0, \quad if\ (y - f_{svr}) < \epsilon$
$|(y - f_{svr})| - \epsilon$, otherwise

Here, C is Regularization constant. It determines the severity of violations that will be tolerated. It is used as Hyper Parameter for SVM. If C is small, there will be narrow margins which gets rarely violated. It results in high fit classifier which has low bias and high variance. If C is larger, there will be wider margins, many support vectors. It will result in more bias and low variance.

Kernels: It is a method to change the decision boundaries from Linear form to High dimensional form, by enlarging feature space. $k(x_i, x_i') = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$

SVM function with Kernel Implementation can be written as, $f(x) = \beta_o + \sum_{i \epsilon S} \alpha_i k(x, x_i)$

Degree of polynomial can also be used as Hyper Parameter for Support Vector Regression.

Advantages: It performs well even when Number of dimensions is close to Number of Observations, or when observations are Highly Correlated.

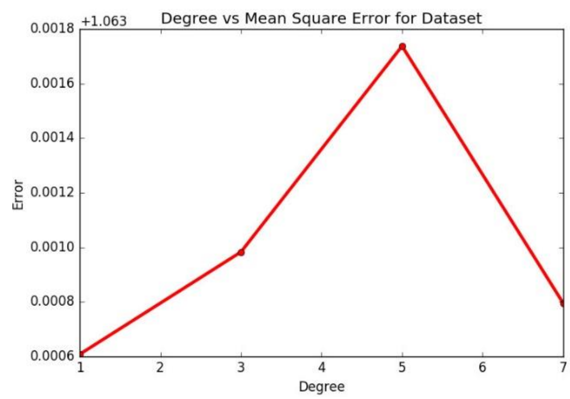
Applied Linear Regression Methods on Robot Arm Dataset such as Ridge Regression along with cross Validation, and obtained error rate of 1.28, whereas Support Vector Regression with RBF kernel is giving 1.037 error rate, thus it was best among all other Regression Methods.

C. K fold Cross Validation with Regularization Method and Degree of Polynomial as Hyper-Parameter.
Regularization Parameter (C) searched over range from **100 to 500.**
Polynomial Degree Searched over range from **1 to number of dimensions.**

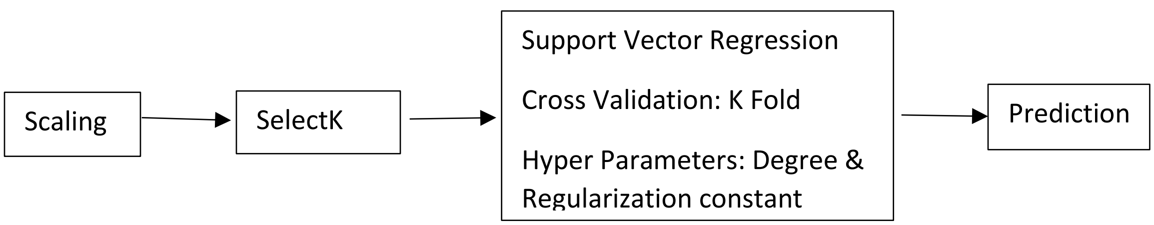
K fold Cross Validation: Cross Validation is a method to measure the accuracy of Regressor, via changing the Hyperparameter values. It is used to optimize the Hyperparameter Values. For given data set, K fold Cross Validation approach would be best, its steps are as follows:

1. Divide the given data set in to K parts.
2. For i=1,2,3…, k:
   Choose part I as Cross Validation set, and rest as Training Set.
   Train the data.
   Compute the mean square error between training and cross validation set.
   Return average of the errors for given hyperparameter value.

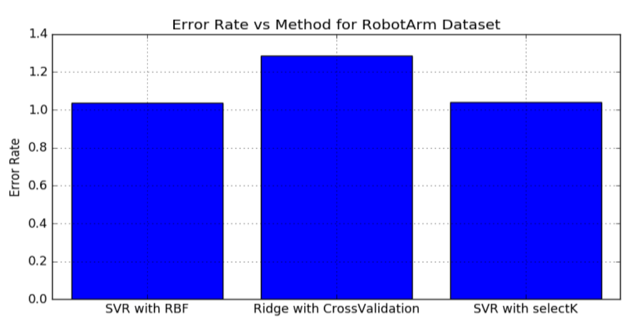
K Fold Cross Validation is one of the best approach as it considers whole dataset for cross validation as well. K Fold is better because it has low variance, as compared to other available choices such as 3 fold, 2 fold. On, the other hand Cross Validation- Cross Validation might perform better than K Fold but there will be a cost of time.



D.



Strangely on introduction of preprocessing step Error rate got increased. Scaling resulted in increment to around 1.04 whereas alone Support Vector along with Cross Validation rated as 1.037.
Normalization resulted in error rate of approximate 2.23, and introduction of SelectK as feature selection method resulted in increment of error rate to 1.04149
Thus, in final Code only original Support Vector Regression along with Cross Validation has been chosen.



Bank Queues Data Set

A. **Recursive Feature Elimination:** Given a data set with D dimensions, and we need to determine which Dimension is valuable and which is just noise, for this we use feature selection. Initially, the estimator gets trained on the first set of features which can be decided while calling the function, and weights get assigned to them. Then the features which have the small weights gets removed, the number of features to be removed at every step can also be determined during function call. Recursively, it determines weight with different set of features and eliminate till the desired number of features get selected.
Strengths: It consider all dimensions, train on them and then reject which have no significance.
Weakness: As it considers all dimensions, so It might be very slow method when it comes to very high dimensions.

B. **Ridge Regression** is a form of Linear Regression; it can be defined as:

$f(X) = \sum_{i=1}^{N} w_i * x_i + b$ $\quad$ w_i are coefficients of features x_i

w_i can be determined through finding minima of following equation:

$$w^* = \arg min_w \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i w_i)^2 + \lambda \, ||w||_2{}^2$$

Here $||w||_2{}^2$ is L2 Normalization. λ is weights penalization coefficient.
If $\lambda = 0$, this will become Linear Regression.
If $\lambda \to \infty$, All coefficients will become 0, to make sense of equation.
If $0 < \lambda < \infty$, coefficients will lie between 0 and that of simple Linear Regression.

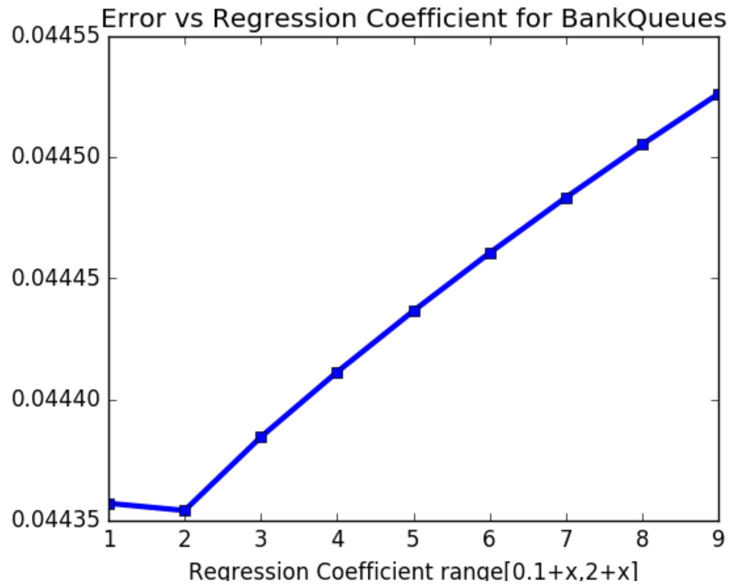
To find global Minima of weights, w^* gets differentiated with respect to weight and equated to zero. It sets Coefficients of Noise terms to minimal value, so that they play no role in equation.
Ridge Regression works better in case of Multiple Collinearity, and it can be determined using the *eigen values* of *correlation matrix*.
Here, Eigen value is low thus Ridge Regression was a good approach for this dataset.
Other Methods that has been Implemented are:

1. Support Vector Regression with Cross Validation of Regularization constant and Linear nature kernel. It resulted in error rate of 0.05183
2. Support Vector Regression with Cross Validation of Regularization constant and RBF nature. This resulted in error rate of 0.04780

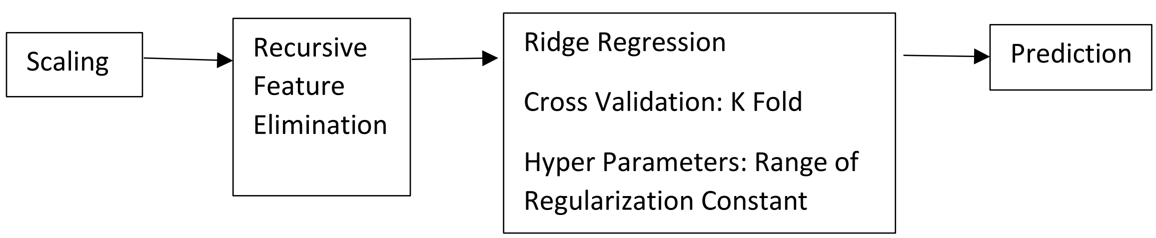
Thus RidgeCV along with Nested K Fold CV is the best method as its error rate is 0.04013

C. **RidgeCV**: Ridge Regression Function have Cross Validation form in Scikit Learn. In this it compares all degree polynomials using K Fold Cross Validation on dataset and return the one with Minimum Error.
K Fold Cross Validation: α is Regularization Constant, whose optimal range can be determined using Cross Validation.
This nested way of cross validation will help in traversing through different range of Degree Polynomials along with varying Regularization Constant.

For Bank Queues Data Set, on two-fold cross validation optimal α's range came out to be [2.1,4]



D.



In Bank Queues Data set as well, introduction of Preprocessing step, Scaling led to increment in error rate. But addition of Recursive Feature Elimination resulted in error rate of 0.04013
Other Method implementation resulted in following errors:
Recursive Feature Elimination with LassoCV: 0.04127
SelectK Feature Elimination with LassoCV: 0.04286
Thus in Final code, only Recursive Feature Elimination along with Ridge Regression Cross Validation has been chosen.