Journal of Data Mining and Knowledge Discovery
ISSN: 2229–6662 & ISSN: 2229–6670, Vol. 2, Issue 1, 2011, pp-19-29
Available online at http://www.bioinfo.in/contents.php?id=42

PREDICTIVE DATA MINING-A RELATIVE STUDY OF LINEAR TECHNIQUES

ABHISHEK TANEJA* AND CHAUHAN R.K.
*Department of Computer Applications, DIMT, Kurukshetra, taneja246@yahoo.com
Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, rkckuk@yahoo.com
*Corresponding author. E-mail: taneja246@yahoo.com

Received: May 23, 2011; Accepted: June 07, 2011

Abstract- In data mining model building and validation is done using voting, averaging, stack generalization and meta-learning. This is a very elaborate process which consumes much of the resources and wastes time. In this scenario the choice of technique depends upon the intuition of the analyst and thus jeopardizing data mining results. The aim of this study is to compare the predictive ability of four statistical data mining techniques viz., factor analysis, ridge regression, multiple linear regression (MLR), and partial least square (PLS) to prevent voting, averaging, stack generalization, meta- learning and thus saving much of our time in choosing the right technique for right kind of underlying dataset.

Keywords: Data mining, Factor Analysis, Multiple Linear Regression (MLR), Partial Least Square (PLS), Ridge Regression.

1.1 Introduction

In recent times innovation in the data compilation equipment like bar code scanners, sensors in commercial and scientific domains have led to the assortment of gigantic amount of data. This fabulous expansion in datasets has pushed the creation of proficient data mining techniques that would lead to transform these datasets into useful understanding and information. In that regard we have number of data mining techniques that would accomplish this difficult task. But all the techniques have got their own limitations and constraints. To choose right kind of technique for right kind of underlying dataset we usually resort to voting, averaging, stack generalization, and meta-learning. This seems to be a simple process but basically this is an elaborate process which requires much of time. To prevent this situation analyst normally resort to his/her intuition and thus jeopardizing predictive results. The choice of technique plays a large role in the improbability of a model. When nonlinear data are fitted to a linear model, the solution is usually biased. When linear data are fitted to a non linear model, the solution usually increases the variance**.** Hence with the influx of improved and modified prediction techniques there is the need for the analyst to know which prediction technique suits for a particular type of data set thus saving lot of time by preventing voting, averaging, stack generalization, and meta-learning.

There are many different criteria to use to evaluate a statistical data-mining model. So many, in fact it can be a bit mystifying and at times seem like a sporting event where proponents of one criterion are constantly trying to prove it is the best. Every criteria used have different story to tell, so, it is the circumstances that specifies us about the suitability of that criteria. In our study we have used many model fitness criteria, to evaluate the suitability of the model for the underlying dataset. In a given circumstances one criterion may be better than others but that will change as situations change. Normally it is recommended to use many techniques instead of one, understanding it advantages and disadvantages and then resort to the best one suitable for the underlying scenario. Scores of criterion are trivial deviation of another and a good number have residual sum of squares (RSS) in them in one manner or another. The differences may be slight but can lead to very different conclusions about the fit of a model.

There are various linear regression techniques which can be used for data mining purpose such as single equation linear regression techniques like Multiple Regression, Factor Analysis, PLS and Ridge regression and simultaneous linear regression techniques which can be only applied on more than one equation in single time like 2 SLS (two stage least square), 3 SLS (three stage least square), FIML (full information maximum likelihood), etc, but in our study we have used single equation methods. These methods require the inference of their predictor's prediction with the fulfillment of critical assumptions.

All linear regression techniques entail the specification of the regression model at first. For this purpose we have used correlation matrix of all variables of all reporting data set. For the linearity of all variables and parameters of all data set double log method has been used. Usually all linear regression models are based on two types of assumptions:

1. Non-Stochastic 2.Stochastic. Stochastic assumptions are those which are concerned with random error term in the regression model. The variable which captures influence of all omitted variable from the regression a model.

These assumptions can be comprehended as

I. E(μ_i)= 0→ No biasness

II. σ_μ^2 = constant → Homoscedasticity

III. Cov. (μ_i, μ_j)=0 → No Auto correlation

Where μ_i is random/stochastic variable.

These can be boiled down in one as

$\mu = N(0, \sigma_\mu^2)$ means random error term should be normally distributed with zero means and constant variance.

Changing variance of μ_i in the regression model may cause for heteroscedasticity for the cross-sectional data set of the study.

The association among the successive value of μ_i (1, 2,……..n) causes for spatial autocorrelation for cross-sectional data set.

The non-stochastic assumptions are those which are concerned with other part of the regression model, which is other than random error term, It can be comprehended as

$$Y_i = \underline{b_0 + b_1 X_1 + b_2 X_2 + \ldots b_n X_n} + \mu_i$$

Total *Variation*	*Explained* *Variation* *Non- stochastic* *Assumptions*	*Unexplained* *Variation* *Stochastic* *Assumptions*

For the purpose of assessment of various statistical data mining techniques we have used three unique data sets. They should be unique to have a combination of the following characteristics: few predictor variables, many predictor variables, dataset with high multi-collinearity, very redundant variables and presence of outliers. A basic assumption concerned with general linear regression model is that there is no correlation (or no multi-collinearity) between the descriptive variables. When this assumption is not satisfied, the least squares estimators have large variances and become unstable and may have a wrong sign. Therefore, we resort to biased regression methods, which stabilize the parameter estimates [1].In this study the performance of the four data mining techniques is compared on following ten parameters like mean square error (MSE), R-square, R-Square adjusted, condition number, root mean square error (RMSE), number of variables included in the prediction model, modified coefficient of efficiency, F-value, and test of normality. For models building and computing the above said ten parameters we have used various data mining tools like SPSS 17, XLstat 2009, Stata 10, Unscrambler 10.1, Statgraphics Centurion XVI and MS-Excel 2003.

1.2 Data Introduction

A basic assumption concerned with general linear regression model is that there is no correlation (or no multi-collinearity) between the explanatory variables. When this assumption is not satisfied, the least squares estimators have large variances and become unstable and may have a wrong sign. Therefore, we resort to biased regression methods, which stabilize the parameter estimates [1]. The data sets we have selected for this study have a amalgamation of the following uniqueness: a small number of explanatory attributes, several explanatory attributes, exceedingly collinear attributes, very superfluous attributes and existence of outliers.

The three data sets used in this paper viz., marketing, bank and parkinsons telemonitoring data set are taken from [2],[3], and [4] respectively.

From the foregoing, it can be observed that each of these three sets has unique properties. The marketing dataset consists of 14 demographic attributes. The dataset is a good mixture of categorical and continuous variables with a lot of missing data. This is characteristic for data mining applications.

The bank dataset is synthetically generated from a simulation of how bank-customers choose their banks. Tasks are based on predicting the fraction of bank customers who leave the bank because of full queues. Each bank has several queues, that open and close according to demand. The tellers have various affectivities, and customers may change queue, if their patience expires.

In the rej prototasks, the object is to predict the rate of rejections, i.e., the fraction of customers that are turned away from the bank because all the open tellers have full queues. This dataset consists of 32 continuous attributes and having 4500 records.

The parkinsons telemonitoring data set is made up of a array of biomedical voice dimensions from 42 people with early-stage Parkinson's infection recruited to a six-month trial of a telemonitoring device for isolated symptom succession monitoring. The recordings were routinely captured in the patient's homes. Columns in the table contain subject number, subject age, subject gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measures. Each row corresponds to one of 5,875 voice recording from these individuals. The main aim of the data is to predict the total UPDRS scores ('total_UPDRS') from the 16 voice measures. This is a multivariate dataset with 26 attributes and 5875 instances. All the attributes are either integer or real with lots of missing and outlier values.

The box plot of the three datasets (fig 1 to fig.3) shown below displays measure of dispersion between these variables, compares the mean of different variables, and also shows the outliers in three datasets. In this regard, it becomes necessary to scale these three datasets to reduce the measure of dispersion and bring all the variables of all datasets to the same unit of measure.

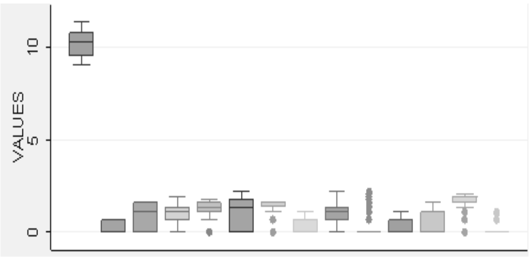


Fig. 1- Box Plot of Marketing Dataset

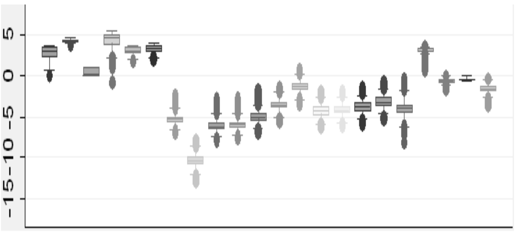


Fig. 2- Box Plot of Parkinson Dataset

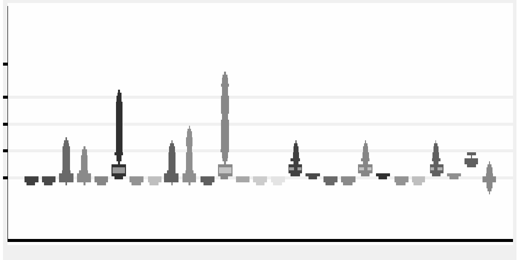


Fig. 3- Box Plot of Bank Dataset

1.3 Prediction Techniques

There are many prediction techniques (association rule analysis, neural networks, regression analysis, decision tree, etc.) but in this study only four linear regression techniques have been compared.

1.3.1 Multiple Linear Regression

Multiple linear regression model maps a group of predictors x to a response variable y [18]. The multiple linear regression is defined by the following relationship, for *i* = 1*,* 2*, n*:

$$y_i = a + b_1x_{i1} + b_2x_{i2} + \cdot \quad \cdot \quad \cdot +b_kx_{ik} + e_i$$

or, equivalently, in more compact matrix terms:

$$Y = Xb + E$$

where, for all the *n* considered observations, ***Y*** is a column vector with *n* rows containing the values of the response variable; **X** is a matrix with *n* rows and *k* + 1 columns containing for each column the values of the explanatory variables for the *n* observations, plus a column (to refer to the intercept) containing *n* values equal to 1; ***b*** is a vector with *k* + 1 rows containing all the model parameters to be estimated on the basis of the data: the intercept and the *k* slope coefficients relative to each explanatory variable. Finally ***E*** is a column vector of length *n* containing the error terms. In the bivariate case the regression model was represented by a line, now it corresponds to a (*k* + 1)-dimensional plane, called the regression plane. This plane is defined by the equation

$$\hat{y}_i = a + b_1x_{i1} + b_2x_{i2} + \cdot \quad \cdot \quad \cdot +b_kx_{ik}+\mu_i$$

Where $\hat{y}_i$ is dependent variable. $X_i{}^{'s}$ are independent variables, and μ_i is stochastic error term. We have compared three basic methods under this multiple linear regression technique. They are full method (which uses the least square approach), forward method, and stepwise approach (which used discriminant approach or all possible subsets) [5].

1.3.2 Factor Analysis

Factor analysis attempts to embody a set of exploratory attributes X_1, X_2 .... X_n in terms of a number of 'common' factors plus a factor which is unique to each variable. The common factors (sometimes called latent variables) are imaginary variables which explain why a number of variables are correlated with each other- it is because they have one or more factors *in common* [6].

Factor analysis is basically a one-sample procedure [7]. We assume a random sample $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_n$ from a homogeneous population with mean vector μ and covariance matrix $\sum$. The factor analysis model expresses each variable as a linear combination of underlying *common factors* f_1, f_2, . . . , f_m, with an accompanying error term to account for that part of the variable that is unique (not in common with the variables). For y_1, y_2, y_p in any observation vector **y**, the model is as follows:

$$\boldsymbol{y_1 - \mu_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \cdot \cdot \cdot + \lambda_{1m} f_m + \varepsilon_1}$$
$$\boldsymbol{y_2 - \mu_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + \cdot \cdot \cdot + \lambda_{2m} f_m + \varepsilon_2}$$
$$\boldsymbol{\ldots}$$
$$\boldsymbol{y_p - \mu_p = \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdot \cdot \cdot + \lambda_{pm} f_m + \varepsilon_p.}$$

Ideally, *m* should be substantially smaller than *p*; otherwise we have not achieved a parsimonious description of the variables as functions of a few underlying factors. We might regard the *f's* in equations above as random variables that engender the *y*'s. The coefficients λ_{ij} are called *loadings* and serve as weights, showing how each y_i individually depends on the *f* 's. With appropriate assumptions, λ_{ij} indicates the importance of the *j*th factor f_j to the *i*th variable y_i and can be used in interpretation of f_j. We describe or interpret f_2, for example, by examining its coefficients, λ_{12}, λ_{22}, λ_{p2}. The larger loadings relate f_2 to the corresponding *y*'s. From these *y*'s, we infer a meaning or description of f_2. After estimating the λ_{ij} 's, it is hoped they will partition the variables into groups corresponding to factors. There is superficial resemblance to the multiple linear regression, but there are fundamental differences. For example, firstly f's in above equations are unobserved, secondly equations above represents one observational vector, whereas multiple linear regression depicts all n observations.

There are a number of different varieties of factor analysis: the comparison here is limited to principal component analysis, generalized least square and maximum likelihood estimation.

1.3.3 Partial Least Square

Partial least squares (PLS) is a method for constructing predictive models when the factors are many and highly collinear. Note that the stress is on predicting the responses and not necessarily on trying to understand the underlying association between the variables. For example, PLS is not usually appropriate for selection out factors that have a negligible effect on the response. However, when prediction is the goal and there is no practical need to limit the number of measured factors, PLS can be a useful tool.

The main purpose of partial least squares regression is to build a linear model, ***Y=XB+E***, where ***Y*** is an *n* cases by *m* variables response matrix, ***X*** is an *n* cases by *p* variables predictor matrix, ***B*** is a *p* by *m* regression coefficient matrix, and ***E*** is a noise term for the model which has the same dimensions as ***Y***. Usually, the variables in ***X*** and ***Y*** are centered by subtracting their means and scaled by dividing by their standard deviations.

Partial least squares regression produce factor scores as linear combinations of the original predictor variables, so that there is no correlation between the factor score variables used in the predictive regression model. For example, suppose we have a data set with response variables ***Y*** (in matrix form) and a large number of predictor variables ***X*** (in matrix form), some of which are highly correlated. A regression using factor extraction for this type of data computes the factor score matrix ***T***=***XW*** for an appropriate weight matrix ***W***, and then considers the linear regression model ***Y***=***TQ***+***E***, where ***Q*** is a matrix of regression coefficients (loadings) for ***T***, and ***E*** is an error (noise) term. Once the loadings ***Q*** are computed, the above regression model is equivalent to ***Y***=***XB***+***E***, where ***B***=***WQ***, which can be used as a predictive regression model. Partial least squares regression produces the weight matrix ***W*** reflecting the covariance structure between the predictor and response variables.

For establishing the model, partial least squares regression produces a *p* by *c* weight matrix ***W*** for ***X*** such that ***T***=***XW***, i.e., the columns of ***W*** are weight vectors for the ***X*** columns producing the corresponding *n* by *c* factor score matrix ***T***. These weights are computed so that each of them maximizes the covariance between responses and the corresponding factor scores. Ordinary least squares procedures for the regression of ***Y*** on ***T*** are then performed to produce ***Q***, the loadings for ***Y*** (or weights for ***Y***) such that ***Y***=TQ+E. Once Q is computed, we have Y=XB+E, where B=WQ, and the prediction model is complete.

One additional matrix which is necessary for a complete description of partial least squares regression procedures is the *p* by *c* factor loading matrix ***P*** which gives a factor model ***X***=***TP***+***F***, where ***F*** is the unexplained part of the ***X*** scores.

1.3.4 Ridge Regression

Ridge Regression is a deviation of ordinary Multiple Linear Regression whose goal is to evade the problem of independent variables collinearity. It gives-up the Least Squares (LS) as a method for estimating the parameters of the model, and focuses instead of the X'X matrix. This matrix will be artificially modified so as to make its determinant appreciably different from 0. By doing so, it makes the new model parameters somewhat biased (whereas the parameters as calculated by the LS (least square) method are unbiased estimators of the true parameters since LS satisfy Gauss Markov theorem [17]). But the variances of these new parameters are smaller than that of the LS parameters and in fact, so much smaller than their Mean Square Errors (MSE) may also be smaller than that of the parameters of the LS model. This is an illustration of the fact that a biased estimator may outperform an unbiased estimator provided its variance is small enough.

Moreover, the predictions errors of the Ridge Model will also turn out to be more accurate than that of the LS regression model when independent variables exhibit near collinearity. Therefore, the idea behind of Ridge Regression is at the heart of the "bias-variance tradeoff" issue.

An extra parameter has to be introduced in the model, the "ridge parameter". Its value is assigned by the analyst, and determines how much Ridge Regression departs from LS Regression. If this value is too small, Ridge Regression cannot fight collinearity efficiently. If it is too large, the bias of the parameters become too large, and so do the parameters and predictions MSEs. There is therefore an optimal value for the ridge parameter, that theory alone cannot calculate accurately from the data only. It has therefore to be estimated by a series of trial and errors, usually resorting to cross-validation.

1.4 Literature Review

The problem of choosing a new data mining technique comes when the analyst has no knowledge of the new data set. Selection of the best technique requires the deep understanding of the data modeling technique and their advantages and disadvantages with some superficial knowledge of the underlying dataset being used for process model.

Earlier many people had done such comparisons between different data mining techniques. For example, Orsolya et.al [8], in 2005 compared Ridge, PLS, Pair-wise Correlation Method (PCM), Forward Selection (FS), and Best Subset Selection (BSS) on a quantitative structure-retention relationship (QSSR) study based on multiple linear regression on prediction of retention indices for aliphatic alcohols. They used (Mean Square Error) MSE, R^2, PRESS, and F-value for model comparison. Huang, J. et.al [9] in 2002 compared Least square Regression, Ridge and PLS in the context of the varying calibration data size using only squared prediction errors as the only model comparison criterion. Vigneau, E. et.al [10], in 1996 compared ridge, PCR and ordinary least square regression with ridge principal component, RPC (blend of ridge and PCR) on the bases of two data sets. They used PRESS and MSE as the model comparison criteria. Malthouse, C. E. et.al [11], in 2000 compared ridge with stepwise regression on direct marketing data using only MSE as model comparison criteria. Naes, T. and Irgens, C. [12] in 1985 compared MLR, ridge, (Principle Component Regression) PCR, and PLS on near infrared instrument statistical calibration using only (root mean square error) RMSE as model comparison criteria. In year 2009, Hassan, Al et.al compared ridge regression and PCR using MSE as model comparison criteria [13]. In year 2009 Noori R. et.al compared neural network and principal component regression analysis to predict the solid waste generation in Tehran. They used correlation coefficient and average absolute relative error indices for model evaluation [14]. In year 2002, Yeniay et.al compared PLS with ridge regression, OLS using PRESS and RMSE as model evaluation method[15]. In year 2005 Zurada Jozef, and Lonial Subhash compared the performance of several data mining methods for bad recovery in health care industry[16].

1.5 Methodology

For Data Mining purpose all regression techniques must satisfy all these usual assumptions and for the reliability of linear regression modeling we can use one more criteria that is criteria of desirable propertied of regression estimator/parameters. These desirable properties are known as BLUE properties i.e., Best, Linear, Unbiased and Efficient.

In our study we have used both stochastic assumptions criteria and BLUE (desirable properties) criteria to check the authenticity of regression modeling. In our study three datasets have been used for data mining to derive inherent characteristics of datasets and four linear regression techniques with their sub modeling have been used.

For the application of all linear techniques all data sets have been divided into two parts on the behalf of probability modeling. One part is training dataset on which regression modeling has been applied and another part is test validation dataset which has been chosen for getting prediction of Predictor variable on the behalf of estimates of training dataset. This step of division of data sets into parts is the rudiment of preprocessing of datasets.

Now we are going to compare our linear regression techniques with the use of above mentioned criteria.

In all linear regression techniques (used in the study) random variable has been found to satisfy all its usual assumptions. For the diagnosis of its normality two criteria have been used one is histogram of μ_i and Jarque-Bera Test. Also is all linear regression a technique of the study linearity has been conformed with the use of natural log of all response and predictor variables of all datasets.

1.6 Interpretation of all Datasets

The First criteria which we have used in our study is goodness of fit criteria (R^2) which tells that all observations lie on the fitted regression line or not. It is also called coefficient of determination.

Before we show how R^2 differs in our study. Let us consider a heuristic explanation of R^2 graphically, known as the Venn diagramed Ballentine (see fig 4).

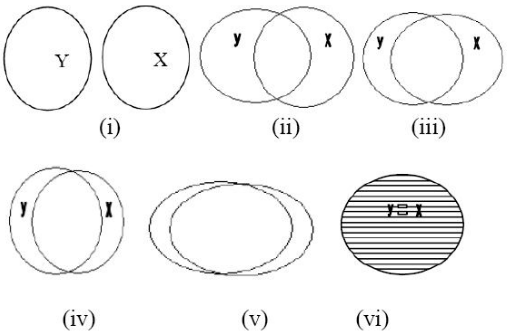


Fig 4: Venn diagramed Ballentine

The Ballative view of R^2

(i) =⟩ $R^2 = 0$ (6) $R^2 = 1$

To complete R^2 we can use if following equation:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum \hat{Y}_i^2}{\sum Y_i^2} = \frac{\sum\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2}$$

This explains percentage change in dependent variable on the basis or with respect to independent variables.

Refer to table I to table IV, the R^2 of factor analysis model was found highest on marketed dataset in comparison to MLR, PLS, and Ridge techniques which means to that factor analysis technique generates good fit regression line. The coefficient to adjusted R^2 was also found good and more than other techniques which means increasing number of variables has low effect on the good fit of the regression model. Although MLR and ridge regression techniques were found with good R^2 value in comparison to factor analysis's R^2. The gap between R^2 of these three techniques is approximately 10%.

Only those principal component are being selected which consists of relevant independent. This prior specification of the model under factor analysis makes this technique better than others.

After this technique MLR and ridge should be given weights one after another because these two techniques extracted good fit regression line with low error. However, the likely of error is more in factor analysis technique. The PLS models were found poor in case of marketing dataset regression line.

In case of Bank data set again factor analysis techniques was found with highest R^2. MLR, PLS and Ridge techniques were found here better than factor analysis.

In case of Parkinson dataset MLR, Ridge with (α = 0.0 and α = 0.25) models and PLS were found with good R^2 in comparison to factor analysis. It means factor analysis techniques was found inappropriate to get regression line for Parkinson dataset.

Dataset like Parkinson consists of less variables with more observations which is usually required to satisfy all assumption of regression model either there are non-stochastic assumption or stochastic assumption. The Parkinson dataset was found to satisfy all usually assumption of MLR, PLS and Ridge (excluding α = 0.51 model) and generated up to the mark goodness of fit. Sometimes there is a possibility to get spurious R^2 (high but not significant) in time series dataset but our study is concerned with cross-sectional data so we can discard the possibility of getting spurious R^2.

In case marketing dataset overall significance of the regression model was found high with PLS techniques. Overall all techniques were found with significant F-value set for up to the mark goodness of fit here, PLS should be considered as better techniques than others.

In case of Bank Data Set overall significant of the regression model is high with the factor analysis techniques. It means due to good extraction of regression line on R^2 the overall significant of the regression model is high. This kind of dataset is useful for the predication purposes.

In case of Parkinson dataset overall significant is high with MLR, Ridge and factor analysis in comparison to PLS techniques. The overall ANOVA (analysis of variance) has been found good with ridge modeling which means for the standardization of data set and appropriate model specification is possible with ridge regression.

Next criteria to judge the appropriate regression modeling is MSE and RMSE criteria which can be comprehended as

$$\text{MSE} = \text{Var.}\,(\hat{\theta}) + \text{Bias}\,(\hat{\theta})^2$$

Of course if the bias is zero MSE $(\hat{\theta})$ = Var. $(\hat{\theta})$

The minimum MSE criteria consist in choosing an estimator whose MSE is the least in a competing set of estimators. But notice that even if such an estimator is found, there is a trade off involved to obtain minimum variance you may have to accept some bias. Geometrically trade off between bias and variance is shown in figure 5.

In case of marketing dataset MSE and RMSE was least with PLS and Ridge models (excluding α = 0.0) which means these two techniques generate high biasness and less variance in comparison to MLR and factor analysis. Also factor analysis techniques was found poor with highest MSE and RMSE but if factor analysis is applied with asymptotic normality in the regression model then they will be likely to get less biasness for dataset like marketing.

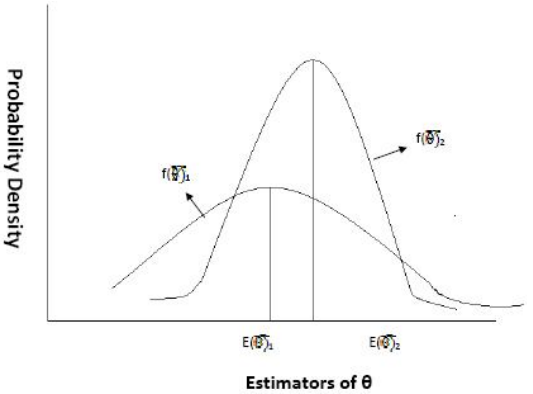


Fig. 5-Trade off between bias and variance

In case of bank dataset MSE as RMSE which are the index of less biasness and less variance were found least with factor analysis. All other models PLS, MLR and ridge were found with more MSE and RMSE, means the biasness and variance are high in there models. Dataset like bank which consist of large number of variables should use factor analysis to get least MSE as best parameters results.

In case of Parkinson dataset MLR, PLS and ridge were found with up to the mark least MSE, which signifies that the former three techniques on data set like Parkinson consist of less variables can generate efficient parameters with less error, less variance, and less biasness.

Next criteria for data mining is the condition index which is the diagnosis of multi-collinearity. Condition index can be comprehended as root of condition number which is

$$\text{K = Condition No.} = \frac{Maximum\_Eigen\_Value}{Minimum\_Eigen\_Value}$$

$$\text{Condition Index} = \sqrt{\text{K}}$$

$$\text{Condition Index} = \sqrt{\frac{Maximum\_Eigen\_Value}{Minimum\_Eigen\_Value}}$$

We have the rule of thumb. If k is between 100 and 1000 there is moderate to strong multi-collinearity and if it exceeds 1000 there is severe multi-collinearity. Alternatively, if Condition Index = $\sqrt{\text{K}}$ is between 10 and 30 there is moderate to strong multi-collinearity and if it exceeds 30 there is severe multi-collinearity.

In our study of three datasets with four techniques we have used rule of thumb to diagnose multi-collinearity.

In marketing dataset PLS and factor analysis were found with highest multi-collinearity although the impact of multi-collinearity is less in factor analysis in comparison to PLS. Another two techniques MLR and Ridge were found with her multi-collinearity. The condition index in case of all models of factor analysis and PLS is between 10 to 30 so we can say that there two techniques have the impact of multi-collinearity, which means more than one association between independent variables.

In our bank dataset again factor analysis was found with moderate multi-collinearity due to the value of condition index lies in between 10 to 30. In comparison to factor analysis all there techniques were found with low multi-collinearity which is tolerable. Only α = 0.49 model of ridge technique was found with moderate multi-collinearity except another.

Next criteria which we have used in our study are MAE which describes the predication power of the model. The model which is having good predication power should have less MAE. In marketing dataset factor analysis was found with more MAE, which means factor analysis is not fit for datasets like marketing. All other techniques were found considerably good in prediction power with less MAE except stepwise model of MLR techniques.

In our bank data set MLR and ridge were found with more MAE in comparison to PLS as factor analysis. It means MLR and ridge are poor to bank data set in context of their prediction power.

In case of Parkinson dataset MLR was found with highest MAE in comparison to all other techniques like factor analysis, PLS, and ridge which states that the perdition power of MLR for Parkinson dataset is not good.

Another prediction power measure which we have used in our study is modified coefficient of efficiency. It can be comprehended as:

$$E=1-\frac{\sum_{i=1}^{n}(O_i - X_i)^2}{\sum_{i=1}^{n}\left(O_i - \overline{X}\right)^2}=1-\frac{MSE}{Varience\_of\_Observed}$$

In Marketing dataset of our study factor analysis was found with less prediction power in comparison to PLS, MLR, and ridge regression model.

In Bank dataset MLR and ridge modeling were found with less prediction power in comparison to PLS and factor analysis with respect to modified coefficient of efficiency.

In Parkinson data set PLS, GLS of factor analysis, and full model of MLR were found with low prediction power in comparison to all other model.

The last criteria which we have used to check the satisfaction of assumption regarding to stochastic–term i.e., test of normality.

Although several tests of normality are discussed in the literature, we will consider only one i.e., chi-square goodness of fit test. The test proceeds as follows: first we run the regression, obtain the residuals $\hat{\mu}_i$; and compute sample standard deviation of $\hat{\mu}_i$.

$$\text{Var. } \mu_i = \frac{\sum(\hat{\mu}_i - \hat{\mu})^2}{n-1}$$

Then we rank the residual and put them into several groups corresponding to the number of standard deviation from zero. Ultimately we find x^2 value for checking normality χ^2 which is the measure of divergence of observed (actual frequency O_i in a class i as E_i is expected frequency in a class. If difference is small it suggests that disturbance μ_i probably come from hypothesized probability distribution. On the other hard, if the discrepancy between O and E is large we reject that disturbance that come from hypothesized probability distribution.

In our study of marketing dataset ridge model α =0.0 and factor analysis were found poor for normal distribution of its random term. So these models are not satisfying the assumption of random error term and further violate the BLUE properties of estimators.

In case of bank dataset factor analysis was found with poor normality of stochastic term in comparison to other techniques since we have to reject the NULL hypothesis in case of factor analysis for test of normality.

In Parkinson dataset MLR technique was found to violate the assumption of normality since under it the divergence between observed and predicted value of residual is large. Therefore, the probability of getting BLUE estimators is very low in case of MLR in comparison to other techniques.

1.7 Conclusion and Future Work

Overall we can suggest that linear regression modeling on randomly selected unique datasets is up-to the mark if and only if when the analysis is significant (checked through T test, F test, and R^2). The model can be used for better prediction, which means that prediction power is better with respect to satisfaction of BLUE properties of regression coefficients.

The techniques in which estimators satisfy BLUE (best, linear, unbiased, and efficient) properties of structural parameters estimates and stochastic random error term are considered better than others.

The skewness of predictors and random term in the linear regression model is creating obstacles to satisfy BLUE properties. Reducing skewness with some advance data mining tool and then comparing performance of said techniques can further enlighten us, which is an area that can be further explored.

Based on the results obtained by comparing said linear data mining techniques, one can easily generalize and answer the following queries given in the table V.

MLR and PLS techniques are simpler to understand and interpret because they do not entail high algebraic treatment. Factor analysis requires standardization to remove the effect of multi-collinearity. Same is with ridge regression, which requires up to the mark scaling until model gets efficiency. Factor analysis and ridge gives good prediction as compared to PLS and MLR when the variables are truly independent. Factor analysis is best among other but some time gives results with heteroscedasticity. MLR gives poor result that may be due the effect of non-linearity in the residual term. MLR and factor analysis give stable result since R^2 will be consistent with respect to scaling whereas PLS or ridge can be affected to estimators due to scaling.

Ridge and factor analysis are particularly suitable when multi-collinearity is there. In factor analysis up-to the mark scaling removes multi-collinearity, whereas in ridge parameter scaling is required to remove multi-collinearity. Factor analysis and PLS are suitable for ill conditioned data because factor analysis attempt to make component generalize which are having more effect on dependent variable, whereas in PLS only one variable is affected at a time. MLR and PLS are not good when redundant variables are there because they increase the variance, whereas ridge and factor analysis are robust against redundant variables are there residual is very small in both the techniques.

Factor analysis and ridge reduces the output prediction error considerably, since R^2 with low biasness is possible. MLR and PLS gives good results when all the input variables are useful due to high variance of error term. Non-linearity of the model can be easily identified through coefficient plots or plot of principle components in case of MLR, factor analysis, and ridge regression. MLR and ridge regression transforms the data into orthogonal space as by targeting principle components and removing all discrepancies respectively.

Although in our study we find ridge regression which reduces multi-collinearity by regularization of regression coefficients but in econometrics various tests have been suggested by econometricians under MLR, factor analysis, PLS, and ridge regression for regularization of regression coefficients to remove multi-collinearity.

Although, we have used the entire ten model fitness criteria's for checking their predictive abilities. Efforts should be geared to make some criteria/s that combines the advantages of two or more of these criteria's. Similarly, efforts could be geared to make a super model that incorporates features to make it fit for multiple kinds of underlying datasets.

Although the framework mentioned has been described for linear data mining techniques, yet the same framework can be extended to include non-linear techniques also.

References

[1] Al-Kassab M. (2009) *Applied Mathematical Sciences*, 3(42), 2085 – 2098.

[2] http://www-stat.stanford.edu/~tibs/ElemStatLearn/

[3]http://www.cs.toronto.edu/~delve/data/bank/desc.html

[4] http://archive.ics.uci.edu/ml/datasets.html

[5] Dash M. and Liu H. (1997) *Intelligent Data Analysis*. 1:3, 131-156.

[6] Kim Jae-on., Mueller Charles W. (1978) *Introduction to Factor Analysis-What it is and how to do it., Sage Publications, Inc*.

[7] Rencher C. Alvin (2002) *Methods of Multivariate Analysis, 2nd Edition, Wiley Interscience*.

[8] Farkas Orsolya, and Heberger Karoly (2005) *Journal of Information and Modeling,* 45(2), 339-346.

[9] Huang J. et al. (2002) *Journal of Chemometrics and Intelligent Lab. Systems*, 62(1), 25-35.

[10] Vigneau E., Devaux M. F. and Robert P. (1996) *Journal of Chemometrics*, 11(3), 239-249.

[11] Malthouse Edward C. (2000) *Journal of Interactive Marketing*, 13(1854),16-23.

[12] Naes T., Irgens C. and Martens H. (1986) *Applied Statistics*, 35(2), 195-206.

[13] Hassan Al.M.Yazid and Kassab Al.M.Mowafaq (2009) *Applied Mathematical Sciences*, 3(42), 2085-2098.

[14] Noori R., Abdoli M., Ghazizade A. and Samieifard R. (2009) *Iranian J Publ Health*, 38(1),74-84.

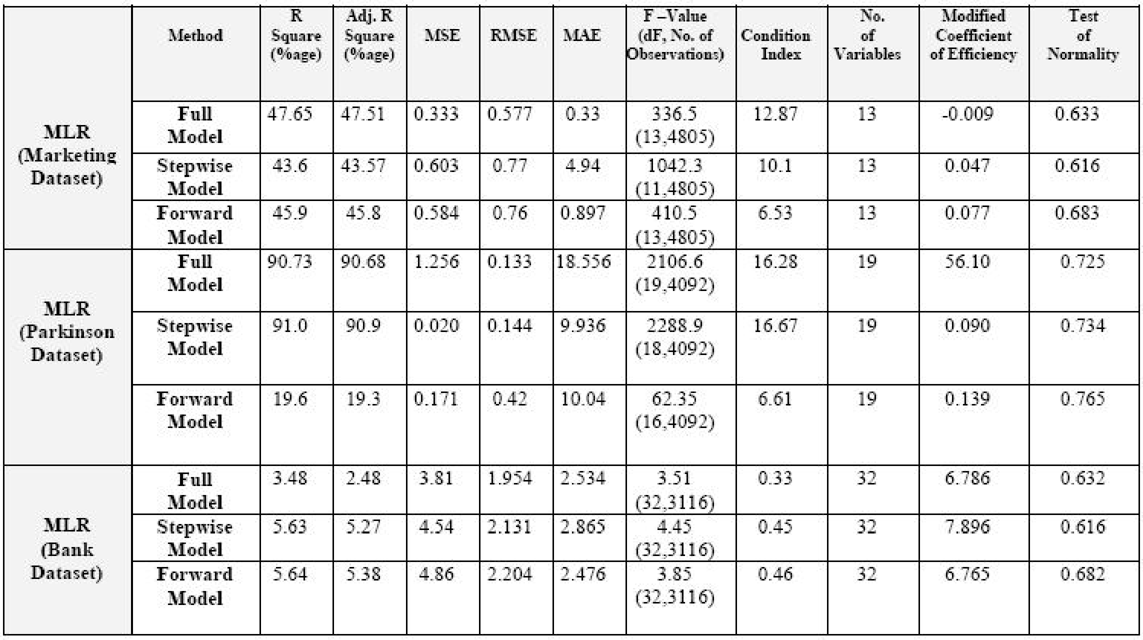
[15] Yeniay O. and Goktas A. (2002) *Journal of Mathematics and Statistics*, 31,99-111.

[16] Lonial Z. and Subhash L. (2005) *The journal of applied Business Research*, 22, 2.

[17] Gujarati N. Damodar, Sangeetha (2004) *Basic Econometrics, 4th edition, New York: McGraw Hill*, 76-79.

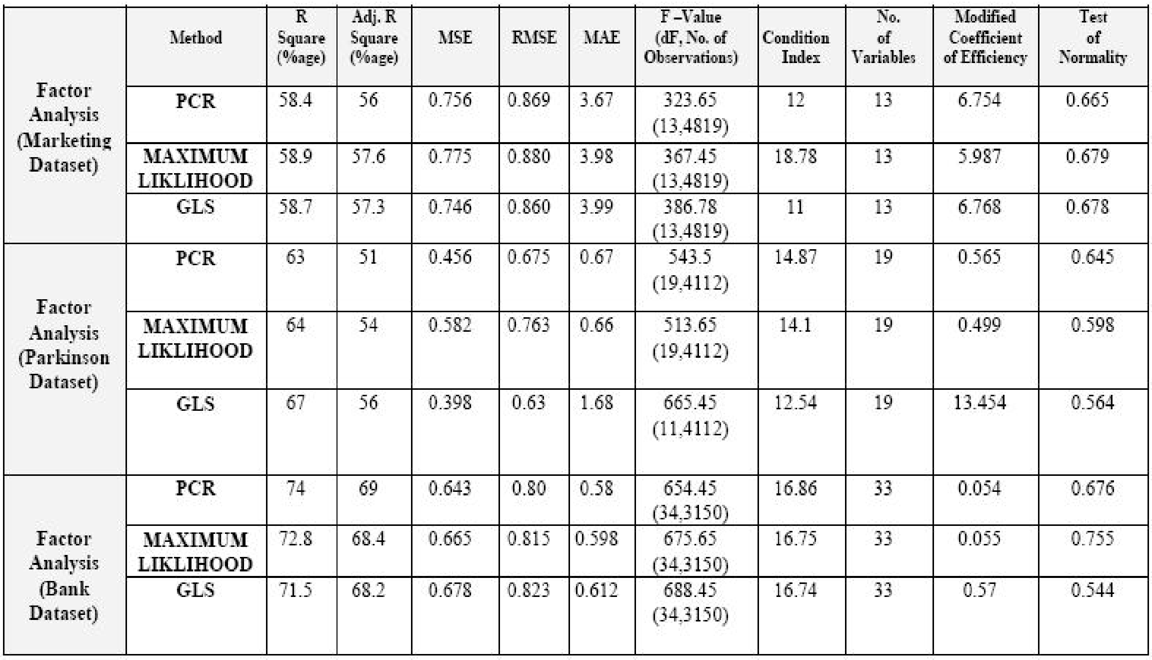
[18] Giudici Paolo (2003) *Applied Data Mining-Statistical methods for business and industry,* wiley.

Table I



	Method	R Square (%age)	Adj. R Square (%age)	MSE	RMSE	MAE	F –Value (dF, No. of Observations)	Condition Index	No. of Variables	Modified Coefficient of Efficiency	Test of Normality
MLR (Marketing Dataset)	**Full Model**	47.65	47.51	0.333	0.577	0.33	336.5 (13,4805)	12.87	13	-0.009	0.633
	Stepwise Model	43.6	43.57	0.603	0.77	4.94	1042.3 (11,4805)	10.1	13	0.047	0.616
	Forward Model	45.9	45.8	0.584	0.76	0.897	410.5 (13,4805)	6.53	13	0.077	0.683
MLR (Parkinson Dataset)	**Full Model**	90.73	90.68	1.256	0.133	18.556	2106.6 (19,4092)	16.28	19	56.10	0.725
	Stepwise Model	91.0	90.9	0.020	0.144	9.936	2288.9 (18,4092)	16.67	19	0.090	0.734
	Forward Model	19.6	19.3	0.171	0.42	10.04	62.35 (16,4092)	6.61	19	0.139	0.765
MLR (Bank Dataset)	**Full Model**	3.48	2.48	3.81	1.954	2.534	3.51 (32,3116)	0.33	32	6.786	0.632
	Stepwise Model	5.63	5.27	4.54	2.131	2.865	4.45 (32,3116)	0.45	32	7.896	0.616
	Forward Model	5.64	5.38	4.86	2.204	2.476	3.85 (32,3116)	0.46	32	6.765	0.682

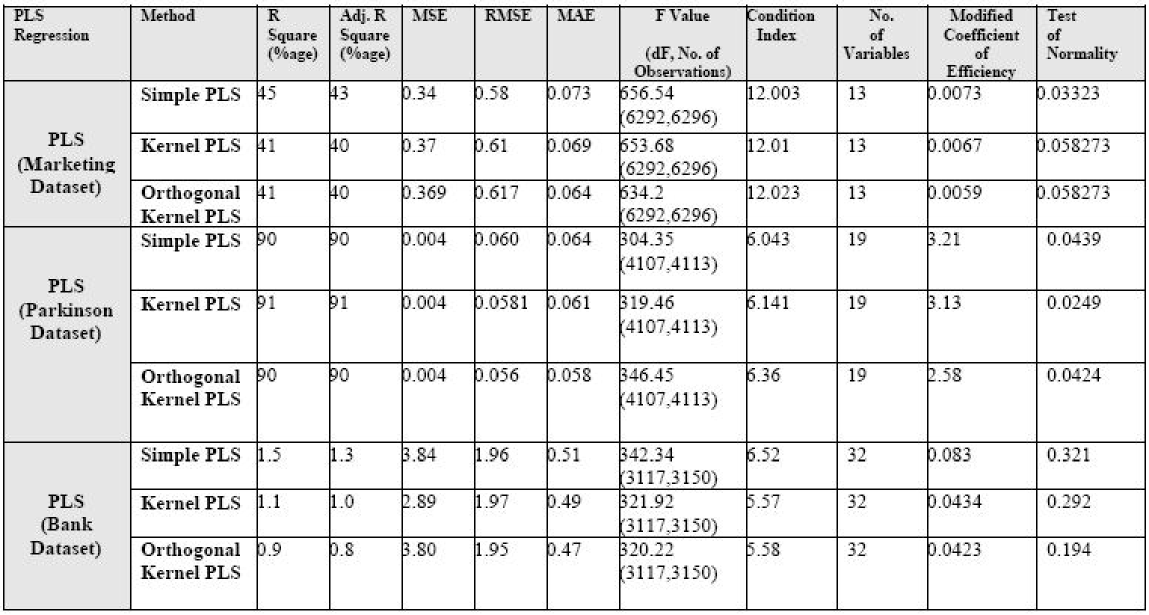
Table II



	Method	R Square (%age)	Adj. R Square (%age)	MSE	RMSE	MAE	F –Value (dF, No. of Observations)	Condition Index	No. of Variables	Modified Coefficient of Efficiency	Test of Normality
Factor Analysis (Marketing Dataset)	**PCR**	58.4	56	0.756	0.869	3.67	323.65 (13,4819)	12	13	6.754	0.665
	MAXIMUM LIKLIHOOD	58.9	57.6	0.775	0.880	3.98	367.45 (13,4819)	18.78	13	5.987	0.679
	GLS	58.7	57.3	0.746	0.860	3.99	386.78 (13,4819)	11	13	6.768	0.678
Factor Analysis (Parkinson Dataset)	**PCR**	63	51	0.456	0.675	0.67	543.5 (19,4112)	14.87	19	0.565	0.645
	MAXIMUM LIKLIHOOD	64	54	0.582	0.763	0.66	513.65 (19,4112)	14.1	19	0.499	0.598
	GLS	67	56	0.398	0.63	1.68	665.45 (11,4112)	12.54	19	13.454	0.564
Factor Analysis (Bank Dataset)	**PCR**	74	69	0.643	0.80	0.58	654.45 (34,3150)	16.86	33	0.054	0.676
	MAXIMUM LIKLIHOOD	72.8	68.4	0.665	0.815	0.598	675.65 (34,3150)	16.75	33	0.055	0.755
	GLS	71.5	68.2	0.678	0.823	0.612	688.45 (34,3150)	16.74	33	0.57	0.544

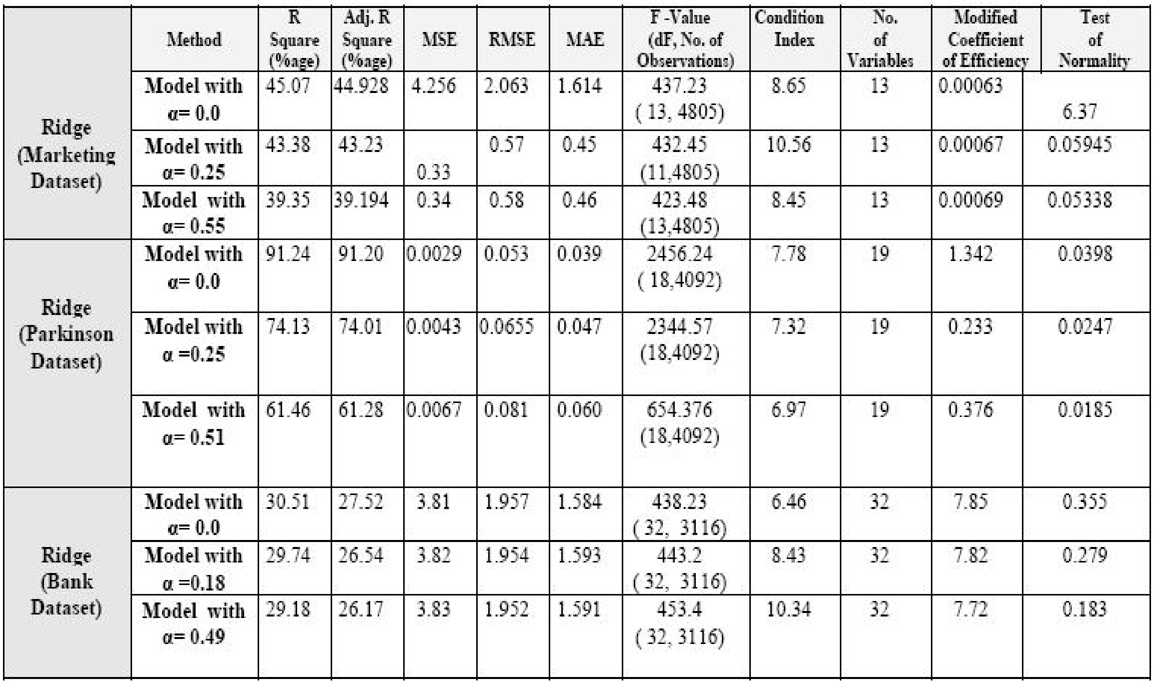
On the basis of extraction of good-fit regression line factor analysis technique was found best technique. It can also be comprehended that under factor analysis models like GLS, Maximum Likelihood and principal components all independent variables have been given weights while construction and specification of the model.

Table III



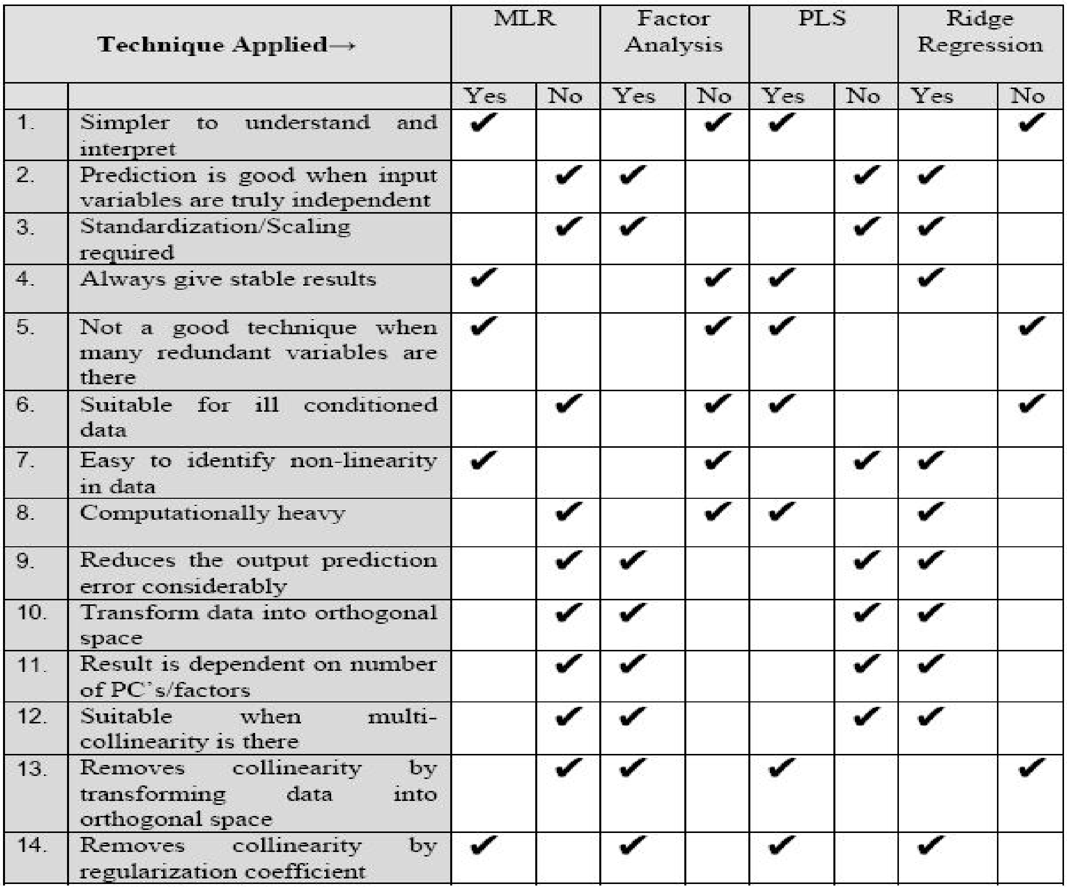
PLS Regression	Method	R Square (%age)	Adj. R Square (%age)	MSE	RMSE	MAE	F Value (dF, No. of Observations)	Condition Index	No. of Variables	Modified Coefficient of Efficiency	Test of Normality
PLS (Marketing Dataset)	**Simple PLS**	45	43	0.34	0.58	0.073	656.54 (6292,6296)	12.003	13	0.0073	0.03323
	Kernel PLS	41	40	0.37	0.61	0.069	653.68 (6292,6296)	12.01	13	0.0067	0.058273
	Orthogonal Kernel PLS	41	40	0.369	0.617	0.064	634.2 (6292,6296)	12.023	13	0.0059	0.058273
PLS (Parkinson Dataset)	**Simple PLS**	90	90	0.004	0.060	0.064	304.35 (4107,4113)	6.043	19	3.21	0.0439
	Kernel PLS	91	91	0.004	0.0581	0.061	319.46 (4107,4113)	6.141	19	3.13	0.0249
	Orthogonal Kernel PLS	90	90	0.004	0.056	0.058	346.45 (4107,4113)	6.36	19	2.58	0.0424
PLS (Bank Dataset)	**Simple PLS**	1.5	1.3	3.84	1.96	0.51	342.34 (3117,3150)	6.52	32	0.083	0.321
	Kernel PLS	1.1	1.0	2.89	1.97	0.49	321.92 (3117,3150)	5.57	32	0.0434	0.292
	Orthogonal Kernel PLS	0.9	0.8	3.80	1.95	0.47	320.22 (3117,3150)	5.58	32	0.0423	0.194

Table IV



	Method	R Square (%age)	Adj. R Square (%age)	MSE	RMSE	MAE	F -Value (dF, No. of Observations)	Condition Index	No. of Variables	Modified Coefficient of Efficiency	Test of Normality
Ridge (Marketing Dataset)	**Model with α= 0.0**	45.07	44.928	4.256	2.063	1.614	437.23 (13, 4805)	8.65	13	0.00063	6.37
	Model with α= 0.25	43.38	43.23	0.33	0.57	0.45	432.45 (11,4805)	10.56	13	0.00067	0.05945
	Model with α= 0.55	39.35	39.194	0.34	0.58	0.46	423.48 (13,4805)	8.45	13	0.00069	0.05338
Ridge (Parkinson Dataset)	**Model with α= 0.0**	91.24	91.20	0.0029	0.053	0.039	2456.24 (18,4092)	7.78	19	1.342	0.0398
	Model with α =0.25	74.13	74.01	0.0043	0.0655	0.047	2344.57 (18,4092)	7.32	19	0.233	0.0247
	Model with α= 0.51	61.46	61.28	0.0067	0.081	0.060	654.376 (18,4092)	6.97	19	0.376	0.0185
Ridge (Bank Dataset)	**Model with α= 0.0**	30.51	27.52	3.81	1.957	1.584	438.23 (32, 3116)	6.46	32	7.85	0.355
	Model with α =0.18	29.74	26.54	3.82	1.954	1.593	443.2 (32, 3116)	8.43	32	7.82	0.279
	Model with α= 0.49	29.18	26.17	3.83	1.952	1.591	453.4 (32, 3116)	10.34	32	7.72	0.183

Table V



	Technique Applied→	MLR		Factor Analysis		PLS		Ridge Regression	
		Yes	No	Yes	No	Yes	No	Yes	No
1.	Simpler to understand and interpret	✔			✔	✔			✔
2.	Prediction is good when input variables are truly independent		✔	✔			✔	✔	
3.	Standardization/Scaling required		✔	✔			✔	✔	
4.	Always give stable results	✔			✔	✔		✔	
5.	Not a good technique when many redundant variables are there	✔			✔	✔			✔
6.	Suitable for ill conditioned data		✔		✔	✔			✔
7.	Easy to identify non-linearity in data	✔			✔		✔	✔	
8.	Computationally heavy		✔		✔	✔		✔	
9.	Reduces the output prediction error considerably		✔	✔			✔	✔	
10.	Transform data into orthogonal space		✔	✔			✔	✔	
11.	Result is dependent on number of PC's/factors		✔	✔			✔	✔	
12.	Suitable when multi-collinearity is there		✔	✔			✔	✔	
13.	Removes collinearity by transforming data into orthogonal space		✔	✔		✔			✔
14.	Removes collinearity by regularization coefficient	✔		✔		✔		✔	