See discussions, stats, and author profiles for this publication at: [https://www.researchgate.net/publication/235637591](https://www.researchgate.net/publication/235637591_A_comparative_analysis_of_predictive_data_mining_techniques?enrichId=rgreq-ec787b1937416ee161448bff254c016c-XXX&enrichSource=Y292ZXJQYWdlOzIzNTYzNzU5MTtBUzozMDcwMjcxNzQxOTkyOThAMTQ1MDIxMjM5MDg3Mw%3D%3D&el=1_x_2&_esc=publicationCoverPdf)

[A comparative analysis of predictive data mining techniques](https://www.researchgate.net/publication/235637591_A_comparative_analysis_of_predictive_data_mining_techniques?enrichId=rgreq-ec787b1937416ee161448bff254c016c-XXX&enrichSource=Y292ZXJQYWdlOzIzNTYzNzU5MTtBUzozMDcwMjcxNzQxOTkyOThAMTQ1MDIxMjM5MDg3Mw%3D%3D&el=1_x_3&_esc=publicationCoverPdf)

**Article** *in* International Journal of Rapid Manufacturing · January 2009

DOI: 10.1504/IJRAPIDM.2009.029380

CITATIONS READS

29 3,329

**3**

**authors**

, including:



Xueping L

[i](https://www.researchgate.net/profile/Xueping-Li-4?enrichId=rgreq-ec787b1937416ee161448bff254c016c-XXX&enrichSource=Y292ZXJQYWdlOzIzNTYzNzU5MTtBUzozMDcwMjcxNzQxOTkyOThAMTQ1MDIxMjM5MDg3Mw%3D%3D&el=1_x_5&_esc=publicationCoverPdf)

University of Tennesse

[e](https://www.researchgate.net/institution/University-of-Tennessee?enrichId=rgreq-ec787b1937416ee161448bff254c016c-XXX&enrichSource=Y292ZXJQYWdlOzIzNTYzNzU5MTtBUzozMDcwMjcxNzQxOTkyOThAMTQ1MDIxMjM5MDg3Mw%3D%3D&el=1_x_6&_esc=publicationCoverPdf)

**144**

PUBLICATIONS

**2,681**

CITATIONS

SEE PROFIL

[E](https://www.researchgate.net/profile/Xueping-Li-4?enrichId=rgreq-ec787b1937416ee161448bff254c016c-XXX&enrichSource=Y292ZXJQYWdlOzIzNTYzNzU5MTtBUzozMDcwMjcxNzQxOTkyOThAMTQ1MDIxMjM5MDg3Mw%3D%3D&el=1_x_7&_esc=publicationCoverPdf)

All content following this page was uploaded by [Xueping Li](https://www.researchgate.net/profile/Xueping-Li-4?enrichId=rgreq-ec787b1937416ee161448bff254c016c-XXX&enrichSource=Y292ZXJQYWdlOzIzNTYzNzU5MTtBUzozMDcwMjcxNzQxOTkyOThAMTQ1MDIxMjM5MDg3Mw%3D%3D&el=1_x_10&_esc=publicationCoverPdf) on 15 December 2015.

The user has requested enhancement of the downloaded file.

*Int. J. Rapid Manufacturing, Vol. 1, No. 2, 2009*

A Comparative Analysis of Predictive Data Mining Techniques

Xueping Li\*, Godswill Chukwugozie Nsofor, Laigang Song

Department of Industrial and Information Engineering,

University of Tennessee,

Knoxville, TN 37996, USA

Tel: +1-865-974-7648

Fax: +1-865-974-0588

Email: Xueping.Li@utk.edu, cnsofor@utk.edu , SongLG@utk.edu

\*Corresponding author

**Abstract:** It is non-trivial to select the appropriate prediction technique from a variety of existing techniques for a datasets, since the competitive evaluation of techniques (bagging, boosting, stacking and meta-learning) can be time consuming. This paper compares five predictive data mining techniques on four unique datasets that have a combination of the following characteristics: few predictor variables, many predictor variables, highly collinear variables, very redundant variables and the presence of outliers. Different data mining techniques, including multiple linear regression (MLR), principal component regression (PCR), ridge regression, partial least squares (PLS) and non-linear partial least squares (NLPLS), are applied to each of the datasets. The comparisons are based on different criteria: R-square, R-square adjusted, mean square error (MSE), mean absolute error (MAE), coefficient of efficiency, condition number (CN) and the number of variables of features included in the model. The advantages and disadvantages of the techniques are discussed and summarised.

**Keywords:** Data Mining, Statistical Analysis, Knowledge Discovery

**Reference** to this paper should be made as follows: Nsofor, G.., Li, X., Song, L. (2009) ‘A Comparative Analysis of Predictive Data Mining Techniques’, *IJRAPIDM*, Vol. 1, No. 2, pp. 150-172.

**Biographical notes:** Xueping Li is an is an Assistant Professor of Industrial and Information Engineering and the Director of the Intelligent Information Engineering Systems Laboratory (IIESL) at the University of Tennessee - Knoxville. He holds a Ph.D. from Arizona State University. His research areas include information assurance, scheduling, web mining, supply chain management, lean manufacturing and sensor networks. He is a member of IIE, IEEE and INFORMS.

Godswill Chukwugozie Nsofor holds a MS degree from the University of Tennessee, Knoxville. His areas of research interest include data mining, quality management, and system reliability. He has been actively involved in sponsored research and industrial projects.

1

*Int. J. Rapid Manufacturing, Vol. 1, No. 2, 2009*

Laigang Song is a Ph.D. candidate in the Department of Industrial and Information Engineering at University of Tennessee, Knoxville. His areas of research interest include web mining and web intelligence, simulation, and information assurance.

2

**1. Introduction**

In recent years, data mining has become one of the most valuable tools for manipulating data and establishing patterns of useful information for decision-making. With breakthroughs in data-collection technology and the advent of the Internet, people are deluged by data (Lyman, 2003; Han and Kamber, 2006). This tremendous growth in data and databases has spawned a pressing need for new techniques and tools that can intelligently and automatically transform data into useful information and knowledge. The needs include the automatic summarization

of data, the extraction of the “essence” of information stored, the discovery of patterns in the raw data, and ultimately the prediction of the future from prior samples, namely predictive data mining (Usama, 1996; Indurkhya, 1998; Berson, 1999; Witten and Frank, 2000). A variety of predictive data mining techniques have been developed. However, it is nontrivial to choose appropriate prediction techniques for a given dataset since the competitive evaluation of techniques (bagging, boosting, stacking, and meta-learning) can be time consuming. This paper compares five predictive data mining techniques on four unique datasets and provides a guideline to identify the best technique for a given dataset and to use it directly instead of the usual trial and error method in predictive data mining analysis.

Data mining is an analytic process in search of a consistent pattern or a systematic relationship between variables and then to validate the findings by applying the detected patterns to new datasets (Berry, 2000; Giudici, 2003; Mitra and Acharya, 2006). The roots of data mining originate in three areas: classical statistics, artificial intelligence (AI), and machine learning (Han and Kamber, 2006). Pregibon (1997) describes data mining as a blend of statistics, artificial intelligence, and database research, and states that it is not a field of interest to many until recently. Data mining can mainly be divided into two tasks (Usama, 1996): predictive tasks and descriptive tasks. Predictive data mining is the most common type and has the most application to businesses. The unexpected nugget of information can lead to new markets, new ways to reach customer and new ways of doing business (Betts, 2003).

Many frameworks have been proposed for building data mining models (Jolliffe, 1986; Chapman, 2000; Lessmann and Vob, 2009). Complex data mining projects require the coordinated efforts of various experts, stakeholders, or departments throughout the entire organization. Therefore, it is necessary to identify frameworks that can serve as blueprints for data collection, analysis, results dissemination and implementing of the organization. CRISP,

DMAIC and SEMMA are the three most common frameworks. CRISP stands for *Cross-Industrial Standard Process* for data mining, which is proposed by a European consortium of companies in the mid-1990s (Chapman, 2000). DMAIC denotes for *Define-Measure-Analyze-Improve-Control* processes, a six-sigma methodology for

eliminating defects, waste, or quality-control problems of all kinds in manufacturing, service delivery, management and other business activities (Pyzdek, 2003). SEMMA is for *Sample-Explore-Modify-Model-Assess*, which is another framework similar to six-sigma and is proposed by the SAS Institute (Jolliffe, 1986).

This paper takes a different avenue to assist the data mining modeling, especially within the predictive data mining domain. We compare five predictive data mining techniques on four unique data sets that have a combination of the following characteristics: few predictor variables, highly collinear variables, very redundant variables and the presence of outliers.

Different performance measurements or model evaluation criteria are used to identify the best technique for datasets with special characteristics. These predictive data mining techniques have wide applications in financial services, telecommunications, retail, healthcare, pharmaceuticals and many other fields.

The rest of the paper is organized as follows. Section 2 briefly introduces the data preparation techniques. Five common predictive data mining techniques are discussed in section 3. The selected datasets and criteria for model comparisons are given in section 4. In section 5, we apply the predictive data mining techniques to each dataset, compare the performance, and discuss the advantages of each technique over the others. Section 6 concludes this study.

1. **Data Preparation**

Poor data quality problems are typically encountered in the course of acquiring data and can hinder the purpose of predictive data mining. A right step taken in data acquisition and handling will help the modeler in getting reliable results and better prediction. Data acquisition and handling has so many steps and is a big topic of its own; but for the purpose of this work, only those topics relevant to predictive data mining are briefly mentioned. In particular, we will discuss data filtering and smoothing, principal component analysis (PCA), and correlation coefficient analysis (CCA).

* 1. Data Filtering and Smoothing

Sometimes during data preprocessing, there may be a need to smooth the data to get rid of

outliers and noise. This depends to a large extent, however, on the modeler‟s definition of

"noise." To smooth a dataset, filtering is used. A filter is a device that selectively passes some data values and holds some back depending on the modeler‟s restrictions (Pyle, 1999). There are several ways to filter data.

* + - Moving average: This method is used for general-purpose filtering for both high and low frequencies (Pyle, 1999; Gencay, 2002; Olaf, 2002).
    - Median filtering: This technique is usually used for time-series data sets in order to remove outliers or bad data points. It is a nonlinear filtering method and tends to preserve the features of the data (Olaf, 2002; Kassama, 1985).
    - Peak-valley mean (PVM): It takes the mean of the last peak and valley as an estimate of the underlying waveform. The peak is the value higher than the previous and next values and the valley is the value lower than the last and the next one in the series (Pyle, 1999; Olaf, 2002).
    - Normalization/standardization: This method changes the instance values in specific and clearly defined ways to expose information content (Pyle, 1999; Olaf, 2002). The measured values can be scaled to a range from -1 to +1. This method includes both the decimal and standard deviation normalization techniques. In this paper, the latter is used. It makes the data set have column means of zero and column variances of one. Every record has the equal opportunity of showing up in the model.

*MCi*  *xi* 

Column mean-centering; *MCi* has a column means of zero.

*MCi**SCi* 



Column scaling *SCi* has a column means of zero and variances of 1.

* + - Fixing missing and empty values: Most data mining algorithms can not handle the missing and empty values correctly. These values are expected to be removed before data mining process. There are many ways to fix the missing and empty values. One way is by replacing the missing values by the mean value.
  1. Principal Component Analysis (PCA)

PCA (Jolliffe, 1986) is an unsupervised parametric method that reduces and classifies the number of variables by extracting those with a higher percentage of variance in the data (called principal components, PCs) without significant loss of information. PCA transforms a set of correlated variables into a new set of uncorrelated variables and allows the analyst to use a reduced number of variables, though with some loss of information. Since the noises are normally weaker than the patterns, dimensionality reduction can eliminate much of the noise.

PCA is appropriate only in those cases that variables are measured in the same units or at least in comparable units, and the variances are roughly similar in size. In case the variables are not measured in comparable units, they should be standardized or normalized before a PCA analysis. Standardization will give all variables equal weighting and eliminate the influence of one variable over the rest. For almost all data-analysis situations, PCA can be recommended as a first step (Johnson, 1998). In the process of doing this, new variables (factors) called principal components can be formed in decreasing order of importance, so that (1) they are uncorrelated and orthogonal, (2) the first principal component accounts for as much of the variability in the data as possible, and (3) each succeeding component accounts for as much of the remaining variability as possible. The PCA is computed using singular value decomposition (SVD) (Jolliffe, 1986), which is a method that decomposes the *X* matrix into a unitary matrix *U*, and a diagonal matrix S that have the same size as *X*, and another square matrix *V* which has the size of the number of columns of *X*.

*X U S V*   *T*

*Where U* is an orthonormal (*m m* ) matrix, *S* is a diagonal (*m n* ) matrix, *n* is the rank of *X* and the diagonals are known as the singular values and decrease monotonically. When these singular values are squared, they represent the eigenvalues.

*V* = Orthonormal matrix (*n n* ) of the eigenvectors, called the loading vectors or the principal components:

*z U S* 

or

*z X V* 

where ***Z*** is an *m n*matrix called the score matrix, ***X*** is an *m n* matrix of original data, and ***V*** is an *n n* transformation matrix called the loading matrix. *m* is the dimensionality of original space, *n* is the dimensionality of the reduced PC space, and *m* is the number of observations in either space.

* 1. Correlation Coefficient Analysis (CCA)

CCA (Cohen, 2003) assesses the linear dependence between two random variables. CCA is equal to the covariance divided by the largest possible covariance and has a range from -1 to +1. A negative correlation coefficient means the relationship is an indirect one, or, as one goes up, the other tends to go down. A positive correlation coefficient shows a direct proportional relationship. The correlation coefficient can be shown with an equation of the covariance relationship:

If the covariance matrix is given by

 *x*2 *xy* cov( ,*x y*)  2 

 *xy y* 

The correlation coefficient is:

*xy*

*pxy* 

*x y*

The correlation coefficient function returns a matrix of the following form:

 1 *pxy*

*corrcoef*(*x*, *y*) *pxy* 1 

A correlation coefficient that is less than 0.3 indicates very little correlation. A correlation coefficient that is greater than 0.3 but less than 0.7 is said to be fairly correlated. A correlation coefficient that is greater than 0.7 means a strong linear relationship. The correlation coefficient of any constant signal (even with noise) with another signal is usually small. To get a good estimate of the correlation coefficient, especially for data sets with varying magnitudes, the data should first be scaled or normalized. Otherwise, it will give more importance to inputs of larger magnitude. **3. Predictive Data Mining Techniques**

3.1. Multiple Linear Regression Techniques (MLR)

The multiple linear regression model maps a group of predictor variables *x*to a response variable *y* (Berk, 1977). The equation of the mapping is in the form:

*y**w*1*x*1 *w*2*x*2 *w*3*x*3 ...*wpxp* 

where ***w***i is the coefficient of the regression. This can also be represented in a matrix formation, in which case *b* is equivalent to the intercept on the *y* axis:

*w*

*y*  *Xw**b*[*X* 1]\**b*



We can solve the above for an optimal weight matrix, ***w***i being the weight or slope. This weight matrix is optimal when the sum of squares error (SSE) is minimal. Below is an estimation of „*e*‟,

*n n*

*SSE**yi* *y*ˆ*i* 2 **y****Xw**2

*i*1

where there are *n* parameters and *y*ˆ is the prediction of *y*.

One assumption made here is that the error term is orthogonal (independent) and Gaussian (it has a mean of zero and a known variance; other assumptions have been stated before).

Collinearity causes a model to be ill-conditioned. Collinearity is a situation where the variables are correlated and the condition number (CN) is very high. The condition number serves the same purpose as variance inflation factor (VIF), tolerance or condition index (CI) (Berk, 1977; Draper 1981).

The ultimate aim of every prediction technique is to minimize the combination of error and complexity. A widely known maxim is that the simpler the model the better. Hence, a good predictive technique reduces the dimensionality of the data, reduces the prediction error, and gives a smooth regression line. Smoothing reduces the weights of the regression parameters as much as possible.

3.2. Principal Component Regression (PCR)

PCR makes usage of the principal component analysis (Jolliffe, 1986; Xie, 1997) discussed in section 2.2. PCR consists of three steps: the computation of the PCs, the selection of the PCs relevant in the prediction model, and the multiple linear regressions. The first two steps are used to take care of collinearity in the data and to reduce the dimensions of the matrix.

In principal component regression, we overcome the problems with collinear data and perform regression with a reduced set of independent and uncorrelated inputs.

3.3. Ridge Regression Modeling (RR)

The ridge regression technique shrinks the regression coefficients by imposing a penalty on their size (Trevor, 2002; Malinowski, 1977). The addition of the product of squared alpha and an identity matrix is called regularization, and alpha is the regularization parameter or ridge coefficient:

*w*(*X XT*  2 *I*)1*X yT*

This parameter α controls the trade-off between the smoothness of the solution and the fitness of the data. The ridge technique is called the smoothing technique because it is characterized by reducing the weights, in turn reducing the condition number. The ridge equation for condition number reduction is given below.

2

Without regularization coefficient "alpha", condition number = *S*max ; but with alpha,

2

*S*min

2 2 *2*  and *S2min* are maximal and minimal singular values. condition number is *S*max  , where *S max*

2 2

*S*min 

This is also very similar to the principal component regression technique in that it chooses the number of relevant PCs. The regularization parameter is related to the singular values. The optimum α value is slightly smaller than the least significant principal component that will go into the model (least significant singular value).

The regularization operation is also related to the weight by

*n*

*b* *i* \**vi* **,** *i* 1*i*  *i*

where *βi* = *uiTY*, where *ui* and *vi* are based on the singular value decomposition matrices *U* and *V*.

Small weights give a smooth solution. If *i* is greater than , then regularization has little effect on the final least-square solution. If  *i* is less than , then the corresponding term in the solution can be expressed as

*i iv* *i* \**i iv* ,

*i* *i* 

and this term approaches *0* as i tends to *0*. Making alpha (the regularization coefficient) larger helps to reduce the weight of the regression coefficients. This result is one of the benefits of ridge regression.

3.4. Partial Least Squares Modeling (PLS)

PLS involves transforming the input data (*x*) to a new variable or score (*t*) and the output data (*y*) to a new score (*u*) making them uncorrelated factors and removing collinearity between the input and output variables (Malinowski, 1977). A linear mapping (**b**) is performed between the score vectors ***t*** and ***u*** as shown inFigure 1. The score vectors are the values of the data on the loading vectors ***p*** and ***q***. Furthermore, a principle component-like analysis is done on the new scores to create loading vectors (***p*** and ***q***).

<<Insert Figure 1 around here>>

An inferential design of PLS is illustrated in Figure 1. In contrast to PCA, PLS focuses on explaining the correlation matrix between the inputs and outputs but PCA dwells on explaining the variances of the two variables. PCA is an unsupervised technique and PLS is supervised. This is because the PLS is concerned with the correlation between the input (*x*) and the output (*y*) while PCA is only concerned with the correlation between the input variables *x*.

As it can be seen in Figure 1, *b* would represent the linear mapping section between the *t* and *u* scores. The good point of PLS is that it brings out the maximum amount of covariance explained with the minimum number of components. The number of latent factors to model the regression model is chosen using the reduced eigenfactors. The eigenfactors are equivalent to the singular values or the explained variation in the PC selection and are

normally called the Malinowski‟s reduced eigenvalue. When the reduced eigenvalues are basically equal, they only account for noise.

3.5. Non Linear Partial Least Squares (NLPLS)

The NLPLS (Frank, 1990; Bakshi, 1999) is essentially the same as the PLS. It involves transforming the input data (*x*) to a new variable or score (*t*) and the y data to a new score (*u*), making them uncorrelated factors and removing collinearity between the input and output variables. An inferential design of NLPLS is shown diagrammatically in Figure 2. It is just the same as the process explained above, with the major difference being that in the linear PLS method, the inner relationships are modeled using simple linear regression while the inner relationships in the NLPLS are modeled using neural networks.

<<Insert Figure 2 around here>>

For each set of score vectors retained in the model, a Single Input Single Output (SISO) neural network is required. These SISO networks usually contain only a few neurons arranged in a two-layered architecture. The number of SISO neural networks required for a given inferential NLPLS unit is equal to the number of components retained in the model and is significantly less than the number of parameters included in the model.

1. **Datasets and Criteria for Model Comparison** 
   1. Selected Datasets

Four datasets are used in this study: the Boston Housing dataset (lib.stat.cmu.edu/datasets), the Airliner dataset, the COL dataset, and the simulated dataset (Hines, 2005). We conduct preliminary analyses on the dataset to gain insights of them. Each of these four data sets has unique properties. The Boston Housing data has thirteen input variables that are not collinear with each other and there are 506 data points. Some of its variables are categorical. These features, numbered according to the column numbers, include:

* + 1. Per capita crime rate by town (CRIM)
    2. Proportion of residential land zoned for lots over 25,000 sq. ft. (ZN).
    3. Proportion of non-retail business acres per town (INDUS).
    4. Charles River dummy variable (1 if tract bounds river; 0 otherwise) (CHAS).
    5. Nitric Oxide concentration (parts per 10 million) (NOX).
    6. Average number of rooms per dwelling (RM).
    7. Proportion of owner-occupied units built prior to 1940 (AGE).
    8. Weighted distances to five Boston employment centers (DIS).
    9. Index of accessibility to radial highways (RAD).
    10. Full value property tax per $10,000 (TAX).
    11. Pupil-teacher ratio by town (PTRATIO).
    12. 1000\*(Bk – 0.63)2 where Bk is the proportion of African-American residents by town (B).
    13. Percentage lower status of the population (LSTAT).
    14. Median value of the owner-occupied homes in $1000's (Mval).

The Airliner dataset has 19 variables (18 input variables and 1 output variable) and 836 data points. The COL dataset has only seven input variables and a response variable with 9559 data points. These variables have a nearly perfect correlation with each other and with the response variable. The simulated dataset has 44 variables and 5,000 data points. The data preprocessing on this data set help to reveal this property of the data and hence in the division of the dataset into training and test validation set, the data points are slashed into blocks of 200 before assigning the odd blocks to train set and the even blocks as test set. From these diagnoses, the Airliner dataset shows some correlation between the variables and some collinearity but not as strong as the COL dataset. The simulated dataset has a large number of input variables and most of them are not helpful to the prediction.

* 1. Criteria for Model Comparison

Many criteria can be used to evaluate the predictive abilities of the different data mining techniques. In this paper, the following criteria will be used to compare techniques: R-square, R-square adjusted, mean square error (MSE), mean absolute error (MAE), coefficient of efficiency, condition number (CN), and the number of features included in the model.

* + - R-square (R2 or R-sq) measures the percentage variability in the given data matrix accounted for by the built model (values from 0 to 1).
    - R-square adjusted (R2adj) gives a better estimation of the R2 because it is not particularly affected by outliers. While R-sq increases when a feature (input variable)

is added, R2adj only increases if the added feature has additional information added to the model. R2adj values ranged from 0 to 1.

* + - Mean square error (MSE): The MSE of the predictions is the mean of the squares of the difference between the observed dependent variable values and predicted dependent variable values by the model. It is the mean of the squared difference between the observed and the predicted values or the mean square of the residuals. MSE can reveal how good the model is in terms of its ability to predict when new sets of data are given. A low value is always desirable. Outliers can make this quantity larger than it actually is. MSE gives equivalent information as R-square adjusted

(R2adj).

* + - Mean absolute error (MAE): This measurement is the summation of the absolute values of the errors (difference between the observed and the prediction). MAE has an advantage over MSE because it takes care of over-estimation due to outliers. Using MSE, a data set that has a lot of outliers gets bloated when they are squared, and this affects the resulting numbers even when the square root is computed.
    - Modified Coefficient of Efficiency (E-mod.): This has been used in many fields of science for evaluating model performance (Legates, 1999; Nash, 1970; Willmott, 1985). According to Nash et al.(1970),the coefficient of efficiency is can be defined as



*i**n*1(*Oi*  *Xi*)2  1 *MSE* **.**

*CE*  1 *n* 2 *Variance* \_ *of* \_ *Observed i*1 *i*  *X* )

(*O*

The ratio of the mean square error to the variance of the observed data is subtracted from unity. It ranges from -1 to +1, where -1 indicates a very bad model, since the observed mean is a better predictor than the predicted variables. A value of zero would show that observed mean is as good as the predicted model.

* + - The condition number (CN)/weight of the regression coefficients: a measure that indicates the stability of the model. High condition numbers (>100) show that the problem is ill-conditioned and hence cannot give consistent or stable results.
    - The number of variables or features included in the model (N): A good predictive data mining technique accounts for most of the information available. It builds a model that gives the most possible information representative of the system being predicted with the least possible MSE. However, when more features are added, the mean square error tends to increase. The addition of more information added increases the probability of adding irrelevant information into the system. A good DM model selects the best features that will account for the most information.

1. **Results and Comparisons**

Figure 3 shows a diagram of the methodology used in this paper. The four data sets are first introduced, as well as the preliminary diagnoses done on each data set to gain an insight into their properties. The relationship check is made by plotting the inputs over the output of the raw data sets. The data is preprocessed by scaling or standardizing them (data preparation) to reduce the level of dispersion between the variables in the data set. The correlation coefficients of each of the various data sets are computed to verify more on the relationship between the input variables and the output variables. This is followed by finding the singular value decomposition of the data sets transforming them into principal components. This also will be helpful in checking the relationship between the variables in each data set.

<<Insert Figure 3 around here>>

At this stage, the data sets are divided into two equal parts, setting the odd number data points as the "training set" and the even number data points as the "test validation data set." Now the train data for each data set is used for the model building. For each train data set, a predictive data mining technique is used to build a model, and the various methods of that technique are employed. The unavailability of different but similar real-life data sets has limited this study to using only the test data set for the model validation. This is not a serious problem since this work is limited to model comparison and is not primarily concerned with the results after deployment of the model.

Finally, all the methods of all the techniques are compared. The best technique or algorithm gives the best prediction for that particular type of data set.

5.1. Boston Housing Data Analysis

5.1.1. Multiple Linear Regression Model

In the MLR model, three methods are considered: full model regression, stepwise regression, and selection of variables based on their correlation with the response variable using the correlation coefficient matrix of all the variables.

1. Full model regression, (all thirteen variables). The full model used all the response variables to predict the *ith* observed output *yi*.

*k*

*yi*  0  *j xij*  *i* , where *k* =1, 2,…, 13.

*j*1

1. Correlation based model. The correlation coefficient matrix is used to choose variables that are best correlated with the output variables.
2. Stepwise regression. The stepwise regression model built only gives results of the training data set prediction. Variables that are significantly different from zero made the model, and the same variables were used to build a multiple linear regression.

5.1.2. Principal Component Regression Model

The summary of the PCR built models are given in Table 1. It can be observed that there is controversy over which model is better, between the model with 11 PCs (=>90% variation) and

that built with 10 PCs. Using the MSE, R2adj., CN, and considering the simplicity of the model, the model with ten PCs came out over that with eleven PCs. Using MAE and modified coefficient of efficiency, however, the model built with eleven PCs looked better. Thus, of those two options, the model with ten PCs will rule over that with eleven PCs.

5.1.3. Ridge Regression (RR)

Table 1 shows the ridge regression results. The best models were those built from standardized data. Among these, the best solution came from the model built with a regularization parameter (optimal alpha value) of 4.0949. It gave a very stable result with comparatively good MSE, good modified coefficient of efficiency and a good condition number. The model with alpha value of 1 was also good but the stability of the result compared to that of alpha value of

4.0949 was not very good.

<<Insert Table 1 around here>>

5.1.4. Partial Least Squares (PLS) Model

From the Summary Table 1, it can be observed that the model with nine factors outperformed every other model in PLS. The model with all the factors (thirteen factors) was better only with R2adj and performed the same as the model with nine factors in terms of MSE. In terms of every other criterion except MAE, the model with nine factors performed better. The best model in terms of MAE was the model with three factors and its condition number was very good at 7.2. 5.1.5. Nonlinear Partial Least Squares (NLPLS) Model

The result of the NLPLS is shown in Table 1. All the measured parameters are better in the NLPLS. It is possible that NLPLS mapped also nonlinearity into the model. The MSE was

17.9547, which was the least among all the models.

5.2. COL Data Set Analysis

5.2.1. Multiple Linear Regression (MLR) Models

Three models were built, as was done earlier in the Boston Housing data set analysis. Table 2 gives the result of the MLR on the COL data. It shows that all the methods gave the same result. For a nearly perfectly correlated data set, the use of stepwise or correlated variables does not make much difference. Figure 4 shows that the prediction looks perfect, but that there is a serious problem of collinearity. The condition number is too high, and this will make the model very unstable.

<< Insert Table 7 around here>>

<<Insert Figure 4 around here>>

5.2.2. Principal Component Regression (PCR)

From Table 2, with the MSE, the full model outperformed other models but the full model had a serious problem of very high condition number. The model is therefore very unstable. The model that gave the best results with reasonable consistency was the correlation-based model. The condition number was below 100 but the MSE was high compared to the full model. The model built with variables whose scores were correlated with the output was the best in PCR on the COL data. Even with fewer PCs, it outperformed the model built with three PCs.

5.2.3. Ridge Regression on the COL Data

Table 2 shows that the ridge regression with the optimal alpha value of 9 stands out in the ridge model. The data set is highly collinear and hence is very ill-conditioned. Only an alpha value that will make a compromise with MSE while smoothing will give a fairly consistent and stable result. With an alpha value of 3.6, the solution seemed good, but the condition number was still very high (1965.6); hence, the model was very unstable because the data set was highly ill-conditioned. The first two ridge models are a proof that ridge regression performs better when the data are standardized before being analyzed. Therefore with an unstandardized data, smoothing (using regularization parameter) gives unreasonable results.

5.2.4. Partial Least Squares (PLS) on COL data

As can be seen in Table 2, the best model is the optimal Eigen value model (four factors). The solution of the optimal factors (4 factors) and that of the model built with all the factors looked almost the same in terms of the R.Sq. the R2adj, the MAE and the modified coefficient of efficiency. Their condition numbers CN were above 2,000. The model built with only two factors had a good condition number (49), and the R.Sq and R2adj were not bad, but the MSE was relatively high (57.8).

5.2.5. Non-Linear Partial Least Squares (NLPLS) on the COL Data

The results of the three NLPLS models are given in Table 2

. The model with only two factors outperformed the one with four factors. The optimal latent factors of 5 (C) gave MSE 22.7752 and MAE of 3.4819. The solution was not stable with NLPLS and therefore was unreliable. It was observed that when the data were retrained, new optimal results emerged. This was repeated many times over, and different optimal results were obtained each time.

5.3. Multiple Linear Regression on the Airliner Data

The summary of the MLR results are shown in Table 3. The full model obviously outperformed the rest in MLR. It can be observed that all the models using MLR had very high condition numbers; hence, the solutions from these models were very unstable and hence unrealistic.

5.3.1. Principal Component Regression on the Airliner Data

Table 3 summarized the results of five PCR models. In terms of MSE, the full model and the model with 13 PCs were the best in the group. The full model gave the least MSE, but the high condition number indicate highly unstable and does not give a unique solution. The model with ten PCs gave a relatively good MSE, MAE, R2adj, and E-mod, and the condition number was also very good. Hence, it is the first choice in the PCR, followed by the correlated PCs with 1 to 4 PCs.

<<Insert Table 3 around here>>

5.3.2. Ridge regression on the Airliner data

The best model, as seen in the Table 3, was the model built with an optimal alpha value of 6.65. 5.3.3. Partial Least Squares (PLS) on the Airliner data

The summary of the PLS results on the Airliner data is shown in Table 3. The model with all the factors performed poorly in terms of the condition number and simplicity of the model but was best based on every other criterion of model measurement. On the other hand, the model with three factors had good condition numbers but a poor MSE as compared to the MSE with the optimal number of factors.

5.3.4. Non-Linear Partial Least Squares on Airliner Data

The NLPLS model used the neural network training function to learn from the training data. Using latent factor14, 15 to build the model, the mean absolute error became 1.2992 and 1.4228. This is shown in Table 3. The MSEs are 4.0712 and 4.9818, but the condition numbers are high.

5.4. Simulated Dataset Analysis

5.4.1. Multiple Linear Regression on Simulated Data Set

Table 4 shows the results of the MLR on the simulated data set. The full model was the best in terms of all other measurement criteria except for the condition number. The condition numbers of the entire three models were quite high. The MSEs were not significantly different. The plots of the predicted test data outputs for the three MLR models on the simulated data showed good predictions.

<<Insert Table 4 around here>>

5.4.2. Principal Component Regression on Simulated Data Set

From Table 4, Models 3, 4, 5, and 6a were good because they had good MSE, but Model 5 had a very high condition number (1951.7). Model 3 was the best with 26 PCs (the PCs that had up to 90% of the explained information). The model had the smallest MSE and the condition number is below 100. Model 3 was better than Model 4 because it looked simpler with 26 PCs, as against 29 PCs in Model 4. The correlation based model was very close to these two and was also good because, with almost half the number of PCs of Models 3 and 4, it had a reasonable MSE and condition number.

5.4.3. Ridge Regression on the Simulated data set

Table 4 is the summary of the ridge regression on the Simulated data set. The best solution, from Table 4, was the model built with an optimal alpha value of 18.44. In this model, there was little compromise between the smoothing parameter and the bias, MSE. The condition number was reduced from 10,000 to 87. The model built with an alpha value of 3 looked very much like the model with an alpha value of zero. It can be concluded that at this alpha value, smoothing has not started. It can also be noticed that both α of 0 and 3.06 looked the same as the full model MLR and full model PCR with a difference only in the condition numbers.

5.4.4. Partial Least Squares on Simulated Data Set

From Table 4, the best solution using PLS was the model built with the optimal number of factors (8) from the iterative (generalization) method. It has the best MSE compared to the others, and the condition number was below 100.

5.4.5. Non-Linear Partial Least Squares (NLPLS) on Simulated Data Set

Table 4 is a summary of the NLPLS on the simulated data set. The best NLPLS model was the optimal factor model built with eight factors; it had a better MSE than the full model and had a good condition number.

5.5. Summary of comparisons

The comparisons of the above predictive data mining techniques are summarized in Table 5 to Table 8, which can serve as a guideline for researchers and practitioners to choose appropriate models for different datasets.

<<Insert Table 5 to Table 8 around here>>

**6. Conclusions**

This paper compares five popular predictive data mining techniques, including multiple linear regression (MLR), principal component regression (PCR), ridge regression, partial least squares (PLS) and nonlinear partial least squares (NLPLS) models, on four unique data sets: the classic Boston housing data, the COL data, the Airline data, and the simulated data. The comparisons of the models are based on different criteria: R-square, R-square adjusted, mean square error (MSE), mean absolute error (MAE), coefficient of efficiency, condition number (CN), and the number of variables of features included in the model.

PLS generally performs better than the other four techniques when building linear models. It deals with the collinearity in the COL data and gives the simplest model that makes the best predictions. The PLS also reduces the dimensionality of the data. The study shows that supervised techniques demonstrate a better predictive ability than unsupervised techniques. It can be seen that in MLR and PCR, the correlation-based models which are supervised techniques performed reasonably better than most models where variables and PCs are randomly selected to build the model. The variables that add valuable information to the prediction models are those have correlation with the output being predicted.

Ridge regression also performs very well with the ill-conditioned Airliner data. It reduced the condition number of the data matrix from 137,000,000 to just 62 with very little compromise on the MSE (bias). It also performs better than most other techniques on the COL data: the condition number is pulled down from 4,000,000 to just 383, and it is beaten only by PLS with two factors.

Some predictive data-mining problems are of the non-linear type. For very complex prediction (or forecasting) problems, non-linear algorithms or a blend of both linear and non-linear techniques are helpful. Effort should be geared towards building combined models. It is worthwhile to evaluate the strengths of the following techniques: the neural network (NN), support vector regression (SVR), the regression trees, kernel regression, kernel SVR, and so on.

**Acknowledgements**

This work was supported in part by a start-up grant from the University of Tennessee. The authors would like to thank Wesley J. Hines at the University of Tennessee for providing the datasets and his inputs to our research work.

**Reference:**

Bakshi R. Bhavik, and Utomo Utojo (1999) „A Common Framework for the Unification of neural, Chemometric, and Statistical modeling methods‟, *Analytica Chimica Acta*, 384, pp. 227-274.

Berk, Kenneth, N. (1997) „Tolerance and Condition in Regression Computations‟, *Journa of American Statistical Association*, 72: 360, December, pp. 865-866.

Berry, M. J. A., and G. S. Linoff (2003) *Mastering Data Mining*. New York: Wiley.

Berson, A., K. Thearling and J. Stephen (1999) *Building Data Mining Applications for CRM*, USA, McGraw-Hill.

Betts, M. (2003) „The Almanac: Hot Tech‟, *ComputerWorld* 52, Nov. 17.

Chapman, P. et al. (2000) „CRISP-DM 1.0: Step by Step Data Mining Guide‟, *CRISP-DM Consortium*, http://www.crisp-dm.org.

Cohen, Jacob, et al. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Draper, N., and H. Smith (1981) *Applied Regression Analysis*, 2nd edition. New York: John Wiley, pp. 307-312.

Frank, I. E. (1990) „A Nonlinear PLS Model‟, *Journal of Chemometrics and Intelligent Laboratory Systems*, 8, pp. 109-119.

Gencay, Ramazan, and F. Selcuk (2002) *An Introduction to Wavelets and other*

*Filtering Methods in Finance and Economics*. San Diego, CA: Elsevier.

Giudici, P. (2003) *Applied Data-Mining: Statistical Methods for Business and Industry*. West Sussex, England: John Wiley and Sons.

Han, J., and M. Kamber (2006) Data Mining: Concepts and Techniques, Edition 2, New York: Morgan Kaufman.

Hines, J. W. (2005), Personal Communication (*Advanced Monitoring and Diagnostic*  *Techniques,* NE 579), University of Tennessee, Knoxville.

Indurkhya Nitin (1998) Predictive Data Mining: a practical guide, New York: Morgan Kaufman.

Johnson, D.E. (1998) *Applied Multivariate Methods for Data Analysis,* Pacific Grove CA: Brooks/Cole.

Jolliffe, I.T. (1986) *Principal Component Analysis*, New York: Springer-Verlag.

Kassams Lee Yong (1985) „Generalized Median Filtering and Related Nonlinear Techniques‟, *IEEE Transactions on Signal Processing*, 33:3, pp. 672-683.

Koh, Chye Hian, and Kee Chan Low (2004) „Going Concern Prediction Using Data Mining Techniques‟, *Managerial Auditing Journal*, 19:3.

Legates, David R. and Gregory J. McCabe (1999) „Evaluating the Use of Goodness of Fit Measures in Hydrologic and Hydroclimatic Model Validation‟, *Water Resources Research,* 35, pp. 233-241.

Lessmann Stefan and Vob Stefan (2009), „reference model for customer-centric data mining with support vector machines‟, *European Journal of Operational*  *Research*, 199, 2, pp. 520-530.

Lyman, P., and Hal R. Varian (2003) „How much storage is enough?‟, *Storage*, 1:4.

Malinowski, E. R. (1977) „Determination of the Number of Factors and The Experimental Error in a Data Matrix‟, *Anal. Chem*. 49, pp. 612-617.

Mitra S. and Acharya T. (2006) Data Mining: Multimedia, Soft Computing, and Bioinformatics, Edition 1, Wiley-Interscience.

Nash, J.E. and J.V. Sutcliffe (1970) „Riverflow Forecasting through Conceptual Models: Part 1-A Discussion of Principles‟, *J. Hydrology*, 10, pp. 282-290.

Olaf, Rem, and M. Trautwein (2002) „Best Practices Report Experiences with Using the Mining Mart System‟, *Mining Mart Techreport.* No. D11.3.

Pregibon, D. (1997) „Data Mining‟, *Statistical Computing and Graphics*, pp. 7-8.

Pyle, Dorian (1999) *Data Preparation for Data-Mining.* San Francisco, Morgan Kaufmann.

Pyzdek, Thomas, (2003) *The Six Sigma Handbook, Revised and Expanded.* New York:

McGraw-Hill.

Tikhonov, A. N. (1963) „Solution to Incorrectly Formulated Problems and the Regularization Method‟, *Soviet Math* DokI 4, 1035-1038, English Translation of *DokI Akod Nauk* SSSR 151, pp. 501-504.

Trevor, H., T. Robert, and F. Jerome (2002) *The Elements of Statistical Learning*, New

York: Springer-Verlag.

Usama, M. Fayyad, et al. (1996) *Advances in Knowledge Discovery and Data Mining.* Cambridge, Mass.: MIT Press.

Usama, M. Fayyad (1996) „Data-Mining and Knowledge Discovery: Making Sense Out of Data‟, *Microsoft Research IEEE Expert,* 11:5., pp. 20-25.

Willmott, C.J., S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, J. O'Donnell, and C.M. Rowe (1985) „Statistics for the evaluation and comparison of models‟, *J. Geophy. Research* 90, pp. 8995-9005.

Witten, I. H. and Frank, E. (2000) Data mining. New York: Morgan-Kaufmann.

Xie, Y.-L, and J.H. Kaliva (1997) „Evaluation of Principal Component Selection Methods to form a Global Prediction Model by Principal Component Regression‟, *Analytica Chemica Acta,* 348:1, August. pp. 19-27.

**List of Figures**

Figure 1: Schematic diagram of the PLS Inferential Design........................................................... 31

Figure 2: Schematic diagram of the Non Linear Partial Least Squares Inferential Design. ............ 32

Figure 3: Flow chart of the methodology ........................................................................................ 33

Figure 4: Predicted test output plotted on the output test data ........................................................ 34

Input Data X

b

t

,

p

u

,

q

Figure 1: Schematic diagram of the PLS Inferential Design

Output Y

b

Input Data X

t

,

p

u

,

q

ANN

1

ANN

2

ANN

3

ANN

4

Figure 2: Schematic diagram of the Non Linear Partial Least Squares Inferential Design.

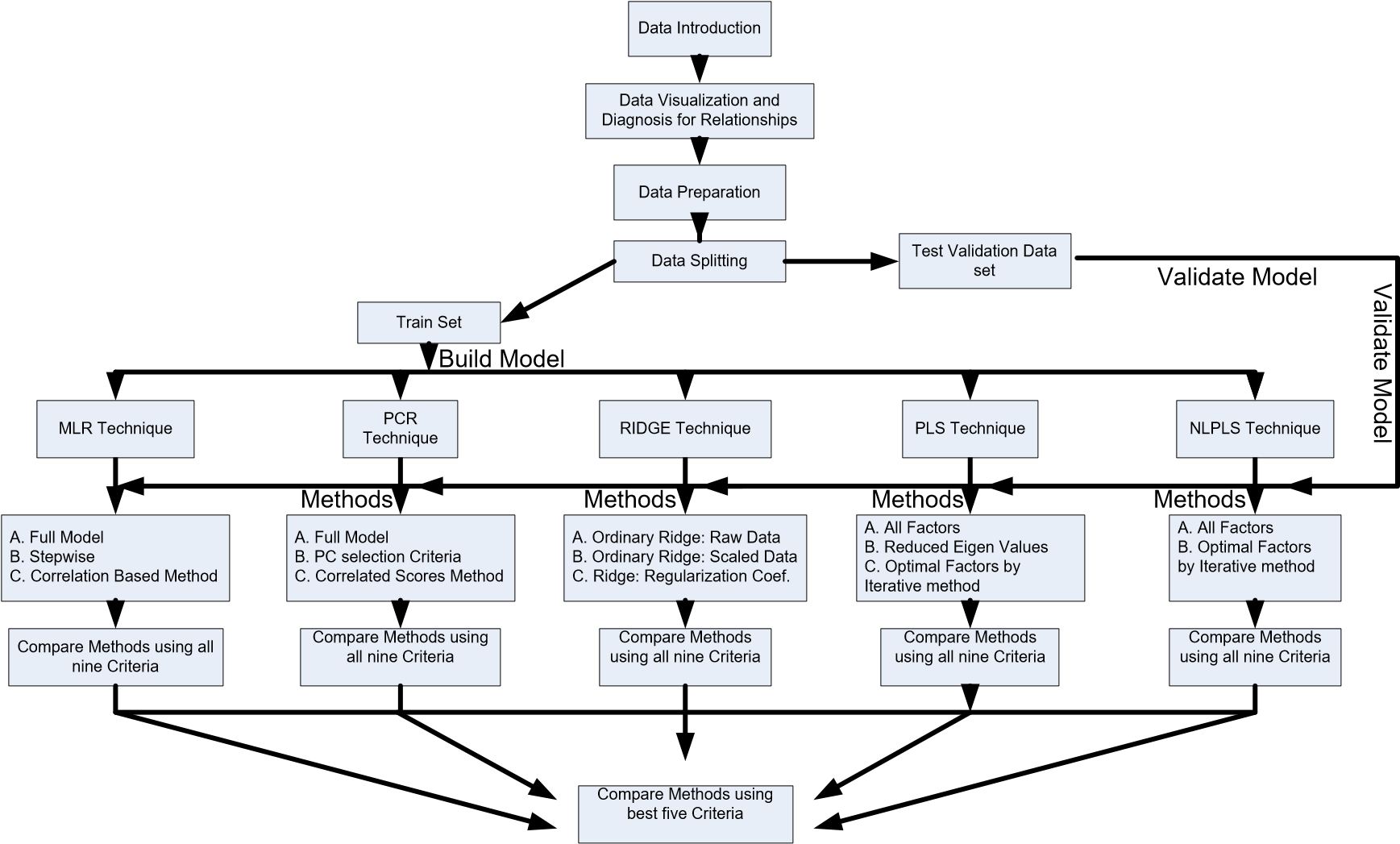


Figure 3: Flow chart of the methodology

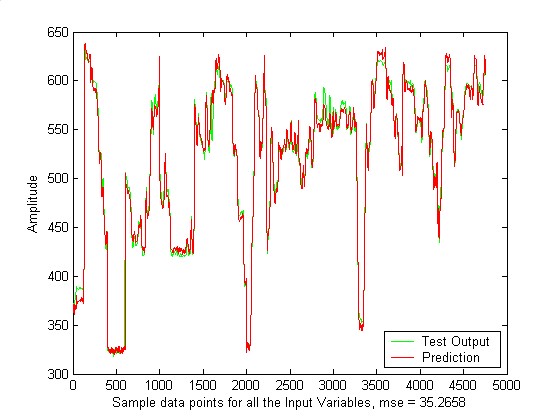


Figure 4: Predicted test output plotted on the output test data

**List of Tables**

[Table 1: Summary of the Boston housing dataset analysis 36](#_Toc81783)

[Table 2: Summary of the COL dataset analysis 37](#_Toc81784)

[Table 3: Summary of the airline dataset analysis 38](#_Toc81785)

[Table 4: Summary of the simulated dataset analysis 39](#_Toc81786)

[Table 5: Linear models compared with non-linear partial least squares 40](#_Toc81787)

[Table 6: Comparison of MLR with PCR, PLS and Ridge regression techniques 41](#_Toc81788)

[Table 7: PCR compared with PLS 42](#_Toc81789)

[Table 8: PLS/PCR compared with RR 43](#_Toc81790)

# Table 1: Summary of the Boston housing dataset analysis.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MLR** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| Full model | 0.7445 | 0.7306 | 21.1503 | 3.2500 | 0.4405 | 7.33e+7 | 13 |
| Cor.Coeff. | 0.7038 | 0.6915 | 24.5201 | 3.4430 | 0.3645 | 7.14e+7 | 11 |
| Stepwise | 0.6727 | 0.6968 | 24.5971 | 3.3989 | 0.3809 | 2.122+7 | 6 |
| **PCR** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **# of PCs** |
| Full | 0.7445 | 0.7317 | 21.1503 | 3.250 | 0.4405 | 87.5639 | 13 |
| =>90% | 0.7160 | 0.7043 | 23.5042 | 3.3517 | 0.3948 | 31.9635 | 11 |
| 2nd Knee | 0.7181 | 0.7077 | 23.3328 | 3.3714 | 0.3833 | 29.7520 | 10 |
| 1st Knee | 0.6943 | 0.6906 | 25.3053 | 3.4919 | 0.3432 | 7.1561 | 4 |
| Cor.PCs (1-3) | 0.6716 | 0.6690 | 27.1818 | 3.5908 | 0.3393 | 7.1561 | 3 |
| Cor. PCs (1-5,12) | 0.7280 | 0.7225 | 22.5133 | 3.3975 | 0.3919 | 36.3902 | 6 |
| **RR** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| Raw Data | 0.000 | 0.0000 | 340100 | 460 | 0.000 | 7.3 e+7 | 13 |
| Scaled data α=0 | 0.7160 | 0.7043 | 23.5042 | 3.3517 | 0.3948 | 31.9635 | 13 |
| Scaled data,α=1 | 0.7444 | 0.7305 | 21.1576 | 3.2429 | 0.4396 | 82.8704 | 13 |
| Scaled data, α = 4.0949 | 0.7387 | 0.7257 | 21.6261 | 3.2255 | 0.4166 | 45.1361 | 13 |
| **PLS** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **# of Factors** |
| Red.eig 1 | 0.7212 | 0.7201 | 23.0754 | 3.3019 | 0.3952 | <7 | 2 |
| Red.eig 2 | 0.7330 | 0.7309 | 22.0992 | 3.2433 | 0.4204 | 7.2 | 3 |
| Min.eig. | 0.7400 | 0.7359 | 21.5159 | 3.2928 | 0.4360 | <36 | 5 |
| Optimal | 0.7446 | 0.7307 | 21.1395 | 3.2498 | 0.4408 | <29 | 9 |
| All factors | 0.7445 | 0.7317 | 21.1503 | 3.2500 | 0.4405 | 87.5639 | 13 |
| **NLPLS** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **Factors** |
| 1st | 0.7831 | 0.7805 | 17.9547 | 2.9545 | 0.5121 | - | 4 |
| 2nd | 0.7921 | 0.7853 | 17.9547 | 2.9120 | 0.5180 | - | 9 |
| 3rd | 0.7925 | 0.7866 | 17.1734 | 2.9182 | 0.5216 | - | 8 |

# Table 2: Summary of the COL dataset analysis.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MLR** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
|  | 0.9944 | 0.994 | 35.2658 | 4.7274 | 0.9266 | 4.24E+06 | 7 |
| **PCR** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **PCs.** |
| Knee | 0.9899 | 0.9899 | 63.5893 | 6.1286 | 0.9053 | 49.1301 | 2 |
| =>90% | 0.9898 | 0.9898 | 64.4823 | 6.152 | 0.9053 | 2311.4 | 3 |
| Cor.built | 0.9899 | 0.9899 | 63.5893 | 6.1286 | 0.9053 | 49.1301 | 2 |
| <1% out | 0.9942 | 0.9942 | 36.6783 | 4.7118 | 0.927 | 2767.1 | 6 |
| All | 0.9944 | 0.9944 | 35.2658 | 4.7274 | 0.9266 | 8940.5 | 7 |
| **RR** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| Raw data α = 0 | 0 | 0 | 3.30e+9 | 10,000 | 0 | 4.2e+6 | 7 |
| scaled data α =0 | 0.9944 | 0.9944 | 35.2658 | 4.7274 | 0.9909 | 8.9e+3 | 7 |
| α = 3.6 | 0.9948 | 0.9948 | 33.0698 | 4.6334 | 0.9279 | 1965.6 | 7 |
| α = 9 | 0.9914 | 0.9914 | 54.5487 | 5.7354 | 0.9093 | 382.69 | 7 |
| **PLS** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| Red. Eig.Val | 0.9908 | 0.9908 | 57.8274 | 5.8683 | 0.9094 | 49.13 | 2 |
| Optimal Val. | 0.9946 | 0.9946 | 33.9342 | 4.651 | 0.9278 | 2311.4 | 4 |
| All factors | 0.9944 | 0.9944 | 35.2658 | 4.7274 | 0.9266 | 8940.5 | 7 |
| **NLPLS** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| 4 Lat.Factors.(A) | 0.9942 | 0.9942 | 36.8616 | 3.4720 | 0.9457 | - | 4 |
| 2 Lat.Factors (B) | 0.9958 | 0.9958 | 26.7676 | 3.3850 | 0.9471 | - | 2 |
| 5 Lat.Factors (C) | 0.9964 | 0.9964 | 22.7552 | 3.4819 | 0.9460 | - | 5 |

# Table 3: Summary of the airline dataset analysis.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MLR** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| Full model | 0.9954 | 0.9952 | 1.1840 | 0.8505 | 0.9307 | 1.37e+08 | 18 |
| Correlation | 0.9917 | 0.9915 | 2.1177 | 1.0761 | 0.9129 | 2.81+07 | 12 |
| Stepwise | 0.9895 | 0.9892 | 2.7089 | 1.2584 | 0.8986 | 1.95e+12 | 8 |
| **PCR** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **PCs** |
| Knee | 0.984 | 0.9836 | 4.0969 | 1.5998 | 0.8707 | 32.0314 | 10 |
| PCs >1% | 0.9928 | 0.9925 | 1.8508 | 1.0187 | 0.9174 | 361.892 | 13 |
| Full Model | 0.9954 | 0.9952 | 1.184 | 0.8505 | 0.9307 | 1.19E+04 | 18 |
| Cor. PC 1-3 | 0.9489 | 0.9487 | 13.051 | 2.8854 | 0.7739 | 2.7382 | 3 |
| Cor. PC 1-4 | 0.97 | 0.9698 | 7.6625 | 2.1825 | 0.8262 | 5.2562 | 4 |
| **RR** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| Raw α = 0 | 0 | 0 | 2.1e+9 | 0 | 0 | 1.37E+08 | 18 |
| Scaled, α=0 | 0.9954 | 0.9952 | 1.184 | 0.8505 | 0.9989 | 1.37E+08 | 18 |
| α = 6.6494 | 0.9888 | 0.9883 | 2.874 | 1.3361 | 0.8893 | 61.8195 | 18 |
| **PLS** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| Eig-val. | 0.9797 | 0.9796 | 5.1856 | 1.8057 | 0.8547 | 2.7382 | 3 |
| Optimal | 0.9954 | 0.9952 | 1.184 | 0.8505 | 0.9307 | 1.19e+04 | 18 |
| **NLPLS** | **R-sq** | **R-sq-Adj** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| 14 Lat. Factors | 0.9841 | 0.9836 | 4.0712 | 1.2992 | 0.8959 | 1.933e+03 | 14 |
| 15 Lat. Factors | 0.9805 | 0.9798 | 4.9818 | 1.4228 | 0.8878 | 1.933e+03 | 15 |

# Table 4: Summary of the simulated dataset analysis.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MLR** | **R-sq** | **R-sq-Adj.** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| FULL | 0.9065 | 0.9049 | 0.0604 | 0.1946 | 0.6993 | 9.885e+03 | 43 |
| COR | 0.8939 | 0.8932 | 0.0685 | 0.2063 | 0.6825 | 2.35e+03 | 17 |
| Stepwise | 0.8897 | 0.8876 | 0.0698 | 0.2081 | 0.6726 | 1.083e+03 | 9 |
| **PCR** | **R-sq** | **R-sq-Adj.** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| 1st knee (1) | 0.8102 | 0.8098 | 0.1226 | 0.2894 | 0.5079 | 12.0198 | 6 |
| 2nd knee (2) | 0.8101 | 0.8095 | 0.1226 | 0.2894 | 0.5081 | 33.5263 | 10 |
| =>90%(3) | 0.9069 | 0.906 | 0.0601 | 0.1943 | 0.6998 | 82.8167 | 26 |
| PCs >1% (4) | 0.907 | 0.906 | 0.0601 | 0.1942 | 0.6998 | 89.2839 | 29 |
| Full (5) | 0.9065 | 0.9049 | 0.0604 | 0.1946 | 0.6993 | 1951.7 | 43 |
| cor. scores (6a) | 0 | 0 | 0.6459 | 0.6817 | -59.292 | 45.7169 | 2 |
| cor. scores 6(b) | 0.8463 | 0.8455 | 0.0993 | 0.2503 | 0.591 | 62.2957 | 14 |
| **RIDGE** | **R-sq** | **R-sq-Adj.** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| Raw data α = 0 | 0.8463 | -0.381 | 0.8768 | 0.6794 | 0.4391 | 9885 | 43 |
| Scaled, α = 0 | 0.9065 | 0.9049 | 0.0604 | 0.1946 | 0.7616 | 9885.1 | 43 |
| α = 3.06 | 0.9067 | 0.9065 | 0.0603 | 0.1946 | 0.698 | 9885.1 | 43 |
| αopt = 18.44 | 0.891 | 0.8908 | 0.0704 | 0.2135 | 0.6481 | 87.4899 | 43 |
| α = 23.2857 | 0.8804 | 0.8801 | 0.0773 | 0.225 | 0.622 | 56.1658 | 43 |
| α = 26.165 | 0.874 | 0.8718 | 0.0814 | 0.2315 | 0.6071 | 44.9507 | 43 |
| **PLS** | **R-sq** | **R-sq-Adj.** | **MSE** | **MAE** | **E-mod.** | **CN** | **N** |
| Red. Eig. (a) | 0.8388 | 0.8386 | 0.1041 | 0.2639 | 0.5653 | 1.83 | 3 |
| Red. Eig. (b) | 0.9045 | 0.9044 | 0.0617 | 0.197 | 0.6942 | 4.3104 | 5 |
| Opt. factor | 0.907 | 0.9067 | 0.0601 | 0.1942 | 0.6998 | 12.2608 | 8 |
| All factors | 0.9065 | 0.9049 | 0.0604 | 0.1946 | 0.6993 | 1951.7 | 43 |
| **NLPLS** | **R-sq** | **R-sq-Adj.** | **MSE** | **MAE** | **E-mod.** | **CN** | **Factors** |
| 1st Opt.factors | 0.8928 | 0.8925 | 0.0692 | 0.208 | 0.6952 | 12.2608 | 8 |
| 2nd Opt factors | 0.8978 | 0.8969 | 0.0664 | 0.2042 | 0.6971 | 12.2608 | 9 |
| All factors | 0.8748 | 0.8726 | 0.0809 | 0.2162 | 0.6778 | - | 43 |

# Table 5: Linear models compared with non-linear partial least squares.

|  |  |  |
| --- | --- | --- |
|  | **LINEAR MODELS** | **NLPLS** |
| 1 | Models only linear relationship | Models both linear and non-linear relationship |
| 2 | Computationally less expensive | Computationally more expensive |
| 3 | Good for linear models only | Good for linear models and models containing about 20% non-linearity |
| 4 | For some collinear data, performs better | Can give unstable results (see COL data analysis) |
| 5 | Good generalization for linear models. | Cannot give good generalization for linear models |

# Table 6: Comparison of MLR with PCR, PLS and Ridge regression techniques.

|  |  |  |
| --- | --- | --- |
|  | **MLR/PLS** | **PCR, PLS, RR** |
| 1 | No standardization or scaling required | Standardization or scaling needed |
| 2 | Gives good predictions when the inputs variables are truly independent | Predicts better when input variables are not independent of each other. |
| 3 | Good when the input variables are all useful in predicting the response | Better when there need for variable reduction.  except for ridge. |
| 4 | Computationally inexpensive | Computationally expensive |
| 5 | Simpler to understand and interpret | More complex in its solutions |
| 6 | Most times results in large regression coefficients | Regression coefficients are much lesser |
| 7 | Sometimes gives unstable results | Most times gives stable results but may give a solution that is not representative of the matrix being modeled |
| 8 | Does not take care of ill-conditioned data or collinear data | Does better with ill-conditioned or collinear data |
| 9 | Does not take care of Collinear data | Removes collinearity |
| 10 | Not better when there are many redundant variables in the input. | Better for dimensionality reduction or feature selection |
| 11 | maximizes the squared correlation between projected inputs and output | PLS maximizes the covariance between projected inputs and output, PCR maximizes variance of the projected inputs, Ridges works same like OLS but uses regularization parameter to reduce the regression weights. |
| 14 | Not easy to detect presence of non-linearity in the model | Easy to detect using the scores from the PCA. |

# Table 7: PCR compared with PLS

|  |  |  |
| --- | --- | --- |
|  | **PCR** | **PLS** |
| 1 | Considers only input variables in its transformation | Considers both input and output variables in its transformation |
| 2 | Unsupervised technique | Supervised technique |
| 3 | Less complex computation | More complex computation |
| 4 | Takes care of collinear data prediction | A better prediction model for collinear data set |
| 5 | Gives good prediction model | Makes better prediction model |

# Table 8: PLS/PCR compared with RR

|  |  |  |
| --- | --- | --- |
|  | **PLS/PCR** | **RR** |
| 1 | Transforms data into orthogonal space | Does not transform data |
| 2 | Takes care of collinearity | Takes care of collinearity |
| 3 | Removes collinearity by transforming data into orthogonal space | Removes collinearity by using regularization coefficients |
| 4 | Performs well with collinear problem | Always work well with collinear problem |
| 5 | Easy to detect non-linearity in the model | Not easy to detect. |
| 6 | Results is dependent on the number of PCs, factors | Uses full variables all the time but result dependent on the regularization parameter |
| 7 | Deals with ill-conditioned regression problems by dropping PCs associated with small eigen values | Damps the minor components |
| 8 | truncates singular values when there is clear gap between two eigenvalues | works well when there is no clear gap between two eigenvalues |

[View publication stats](https://www.researchgate.net/publication/235637591)