EFFECTIVENESS EVALUATION OF REGRESSION MODELS FOR PREDICTIVE DATA-MINING

Dr. Chatur P. N.*

Khobragade Anish R.

Asudani Deepak S.

Abstract

In predictive data mining, the choice of technique to use in analysing a data set depends on the understanding of the analyst. In most cases, a lot of time is wasted in trying every single prediction technique in a bid to find the best solution that fits the analyst's needs. Hence, with the advent of improved and modified prediction techniques, there is a need for an analyst to know which tool performs best for a particular type of data set.

This paper studies and proposed a model to evaluate the effectiveness of three data-mining regression techniques for prediction on different and unique data sets. Data-mining techniques, including Multiple Linear Regression MLR, based on the ordinary least-square approach; Principal Component Regression (PCR), an unsupervised technique based on the principal component analysis; Partial Least Squares (PLS), a supervised technique, were applied to each of the data sets. Five criteria used to evaluate the effectiveness of each technique, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Efficiency (COE), Coefficient of determination (R^2) and Number of features or variables used. It also covers the advantages of each technique over the other.

Keywords- *PDM, MLR, PCR, PLS, RMSE, MAE, COE, R^2*

* G.C.O.E. Amravati, Amravati, India

I. INTRODUCTION

In recent years, data-mining has become one of the most valuable tools for extracting and manipulating data and for establishing patterns in order to produce useful information for decision-making. The failures of structures, metals, or materials (e.g. buildings, oil, water or sewage pipes) in an environment are often either a result of ignorance or the inability of people to take note of past problems or study the patterns of past incidents in order to make informed decisions that can forestall future occurrences. Nearly all areas of life activities demonstrate a similar pattern. Whether the activity is finance, banking, marketing, retail sales, production, population study, employment, human migration, health sector, monitoring of human or machines, science or education, all have ways to record known information but are handicapped by not having the right tools to use this known information to tackle the uncertainties of the future [1].

These needs include the automatic summarization of data, the extraction of the "essence" of information stored, and the discovery of patterns in the raw data. These can be achieved through data analyses, which involve simple queries, simple string matching, or mechanisms for displaying data. Such data-analysis techniques involve data extraction, transformation, organization, grouping, and analysis to see patterns in order to make predictions [2]. From a statistical perspective of data mining is viewed as computer automated exploratory data analytical system for large sets of data.

Knowledge Discovery in Databases (KDD) is an umbrella name for all those methods that aim to discover relationships and regularity among the observed data. Data mining is the process of selection, exploration, and modelling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database [3]. Predictive data mining (PDM) works the same way as doe's a human handling data analysis for a small data set; however, PDM can be used for a large data set without the constraints that a human analyst has [4]. PDM "learns" from past experience and applies this knowledge to present or future situations.

Data Modelling can involve either data classification or prediction. As given in following figure 1 Data-mining steps [5], the classification methods include deviation detection, database segmentation, clustering (and so on). The predictive methods include (a) mathematical operation

solutions such as linear scoring, nonlinear scoring (neural nets), and advanced statistical methods like the multiple adaptive regression. (b) distance solutions, which involve the nearest-neighbour approach; (c) logic solutions, which involve decision trees and decision rules[6].

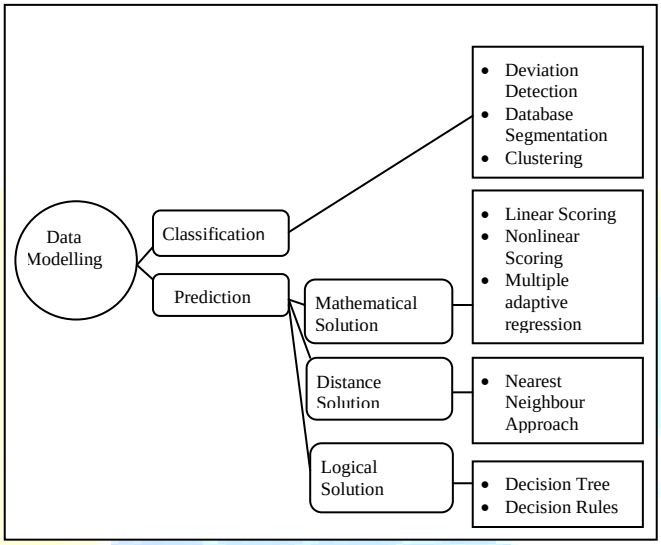

Figure 1. Data modelling steps in Data-mining

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables [7]. In data mining independent variables are attributes already known and response variables are what we want to predict.

Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line ($y = m x + b$) [8] and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation.

II. MATERIAL AND METHODS

The relationship between the response variable and the explanatory variables on the basis of features of the scatter plot. In each case, there has been some sort of functional relationship (either straight-line or a curved) between the variables, and we have made some notion about the

amount of scatter in the plot. The general regression model takes both of these features into account. The most common function is a straight line [9]

$$Y = f(X , \beta) + \varepsilon$$

It shows that regression is the process of estimating the value of a continuous target (Y) as a function (f) of one or more predictors ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), a set of parameters (β_1 , β_2 , ..., β_n), and a measure of error (ε).

- Here β is regression parameter.
- Such models are called linear regression models.
- Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 x_1 + \varepsilon$$

- Y = dependent variable, outcome variable, response variable, explained variable, predicted variable, regress
- x = independent variable, explanatory variable, control variable, predictor variable, regressor
- ε = error term, disterbuns

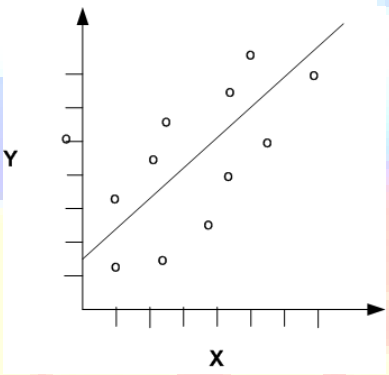

Figure 2. Linear relationship between *x* and *y*

In a linear regression scenario with a single predictor ($y = \beta_0 + \beta_1 x$), the regression parameters (called coefficients) are:

- The **slope** of the line (β_1) — the angle between a data point and the regression line and
- The ***y* intercept** ($\beta 0$) — the point where *x* crosses the *y* axis ($x = 0$)

Linear regression is arguably the most popular modeling approach across every field in the statistical sciences [10].

- Very robust technique
- Linear regression also provides a basis for more advanced empirical methods.
- Transparent and relatively easy to understand technique
- Useful for both descriptive and structural analysis

A. Multiple Linear Regression Model

The general idea of a simple linear regression model is that the response variable Y_i is a straight-line function of a single explanatory variable x_i. In this module, we extend the concept of simple linear regression models to multiple linear regression models by allowing the response variable to be a function of 'p' explanatory variables $x_i, 1, \ldots, x_i, p$. This relationship is straight-line and in its basic form [11] it can be written as

$$Y_i = \beta_0 + \beta_1 x_i, 1 + \beta_2 x_i, 2 + \cdots + \beta_p x_i, p + \varepsilon_i$$

$$Y_i = \beta_0 + \Sigma^p_{j=1} \beta_j \, X_{ij} + \varepsilon_i$$

Input Matrix: X of dimension N × (p + 1) and Y of N

$$X = \begin{bmatrix} 1 x_1, 1 & x_1, 2 \cdots & x_1, p \\ \vdots & \vdots & \vdots \\ 1 x_N, 1 & x_N, 2 \cdots & x_N, p \end{bmatrix} \quad , \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$B = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \quad , \quad \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

Estimation of β : Least Square Method[12]

$$\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

With the estimated regression coefficient, dependent variables can be predicted for independent variables. Generally, requires that $p = rank(\mathbf{X}\mathbf{X}^T) < N$, or $\mathbf{X}\mathbf{X}^T$ is invertible. However, this condition may not always be satisfied.

Among the linear regression techniques, multivariate linear regression (MLR) is one of the most important methods. MLR simultaneously models the relationship between multiple dependent variables and a set of independent variables [10]. It can be easily derived as a maximum likelihood estimator under the assumption that the errors are normally distributed. As a result, the model possesses a unique global minimum, which can be given explicitly. Because of the simplicity, MLR has been regarded as a basic tool in the social and natural sciences.

B. Principal Component Regression Model

The second technique is Principal Component Regression (PCR), which makes use of the principal component analysis. Figure 3 shows the steps of principal components regression model, PCR consists of three steps [13].The computation of the principal components, 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. By reducing the dimensions, one selects features for the regression model.

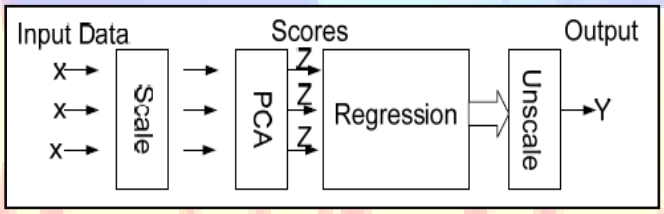

Figure 3. Steps of principal components regression

$$X = U \,.\, S \,.\, V^T$$

The PCA is computed using singular value decomposition (SVD) [14], 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.

$$\boldsymbol{z} = U \,.\, S$$

Or

$$\boldsymbol{z} = X \,.\, V$$

U = Orthonormal (*MxM*) matrix of

S = Diagonal (*MxN*) matrix

where *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 eigen values.

V = Orthonormal matrix (*NxN*) of the eigenvectors, called the loadings vectors or the Principal Components:

z is an *MxN* matrix called the score matrix, ***X*** is an *MxN* matrix of original data, and ***V*** is an *NxN* 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. His whole process is one of projecting the data matrix ***X*** onto the new coordinate system ***V***, resulting in scores ***z***. ***X*** can be represented as a linear combination of *M* orthonormal vectors $\boldsymbol{V}_i$:

Transforming to principal components [15]:

$$x = \sum_{i=1}^{n} z_i v_i + \sum_{i=n+1}^{u} r_i v_i$$

or

$$X = z_1 v_1^T + z_1 v_1^T + \ldots + z_M V_M^T + E$$

The correlation coefficient between the scores of the PCs (*U*S* or *X*V*) and the response variable *y* is computed, and the variables with the strongest correlations are used to build the model. The correlation coefficients are values sorted out by their absolute values (irrespective of sign) and the PCs are entered in this order. It may interest the modeller to transform back into the original *X* transformation with the elimination of features (PCs) that are irrelevant for the best prediction before performing the regression analysis [16].

C. Partial least Square Regression Model

Another predictive data-mining technique is the Partial Least Squares (PLS). PLS is a method of modelling input variables (data) to predict a response variable. Figure 4 shows the

steps of partial least square regression, it 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[17]. A linear mapping (b) is performed between the score vectors t and u. 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).

In contrast to principal component analysis (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.

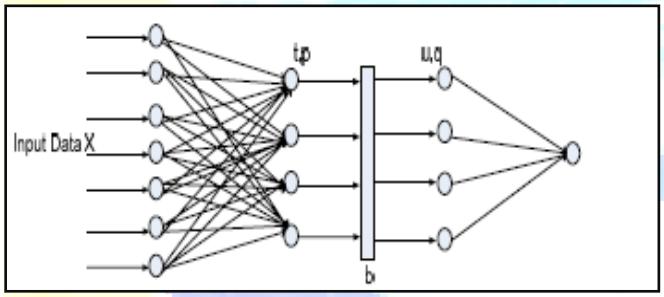

Figure 4. Steps of partial least square regression

As can be seen in Figure 4, b would represent the linear mapping section between the t and u scores.

The general underlying model of PLS is[18]

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{\mathbf{T}} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\,\mathbf{Q}^{\mathbf{T}} + \mathbf{F}$$

- Where X is an $n \times m$ matrix of predictors,
- Y is an $n \times p$ matrix of responses
- T and U are $n \times l$ matrices that are, respectively,
  projections of X (the *X score)* and projections of Y (the *Y scores*);
- P and Q are, respectively, $m \times l$ and $p \times l$ orthogonal *loading* matrices

- E and F matrices are the error terms
- The decompositions of X and Y are made so as to maximize the covariance of T and U.

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 eigen factors. The eigen factors are equivalent to the singular values or the explained variation in the PC selection and are normally called the Malinowski's reduced eigen value [19]. When the reduced eigen values are basically equal, they only account for noise.

III. REVIEW OF PREDICTIVE DATA MINING REGRESSION MODELS

The conventional regression model captures individual differences in intercept and slope; it is not always realistic to assume that a single-population model can account for all kinds of individual differences. Regression mixture models described here are a part of a general framework of finite mixture models [20] and can be viewed as a combination of the conventional regression model and the classic latent class model. For this kind of regression data modeling following papers are reviewed which are based on mixture model.

1. Predictive data mining-a relative study of linear techniques by Abhishek Taneja and Chauhan R.K. [11] 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. 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.

MLR and factor analysis give stable result since R2 will be consistent with respect to scaling whereas PLS or ridge can be affected to estimators due to scaling. Factor analysis and ridge reduces the output prediction error considerably, since R2 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. 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.

2. Comparison of principal component regression (PCR) and Partial least square (PLS) methods in prediction of raw milk Composition by vis-nir spectrometry. Application to Development of on-line sensors for fat, protein and lactose Contents by RocíoMuñiz, Miguel A. Pérez, Jesús A. Baro,[21] studies and compares the potential use of PCR and PLS statistical methods to obtain the values of milk nutrients composition in milk, and present the application to the development of on-line sensors for those nutrients.

Experimental results have demonstrated the capability of method to predict fat and lactose content of milk with high explanation of variance. PLS-1 and PCR method produce similar results for three studied output variables (fat, lactose and total protein), although PLS-1 uses fewer input components to predict lactose content.

3. A comparative study of principal component regression and partial least squares regression with application to FTIR diabetes data by P. Venkatesan, C. Dharuman and S. Gunasekaran, [22] a FTIR diabetes dataset is used in order to examine the performance of the PCR and PLS regression models on prediction. PCs that follow a numeric sequence, depending on "how strong this relationship is". One of the major advantages of PLS is that PCs are modeled not only on the predictors, but also on the responses, so that it is possible to minimize the variance of both X and Y co- ordinates of the model. PLS differs from other multivariate calibration models such as PCR, because the utilization of the responses data set is accomplished in an active way during the statistical calculations.

There is also modification and extensions of partial least square the SIMPLS algorithm. On any case PLS has become an established tool in Chemo-metrics modeling. PLS is still evolving as a statistical modeling tool. Further studies are needed to standardize PCR and PLS for the FTIR spectral data analysis. It is concluded that for prediction, PCR and PLS provides similar results which require substantial verification that may claim superior to any of the two biased regression methods.

4. A comparison of partial least squares regression with other Prediction methods by ÄOzgÄurYeniay, AtillaGÄokta [24], found that partial least squares regression yields somewhat better results in terms of the predictive ability of models obtained when compared to the other prediction methods.

Two of the most used methods, namely PCR and RR, require a large amount of computation when the number of variables is large. PCR handles the collinearity problem with few factors. However, PLS may even overcome the collinearity problem with fewer factors than PCR. Meanwhile simulations tend to show that PLS reaches its minimal mean square error (MSE) with a smaller number of factors than PCR. Hence, PLS gives a unique way of choosing factors, contrary to PCR, and it requires fewer computations than both PCR and RR. In the following sections theoretical aspects of those methods have been presented.

IV. PROPOSED SYSTEM ARCHITECTURE

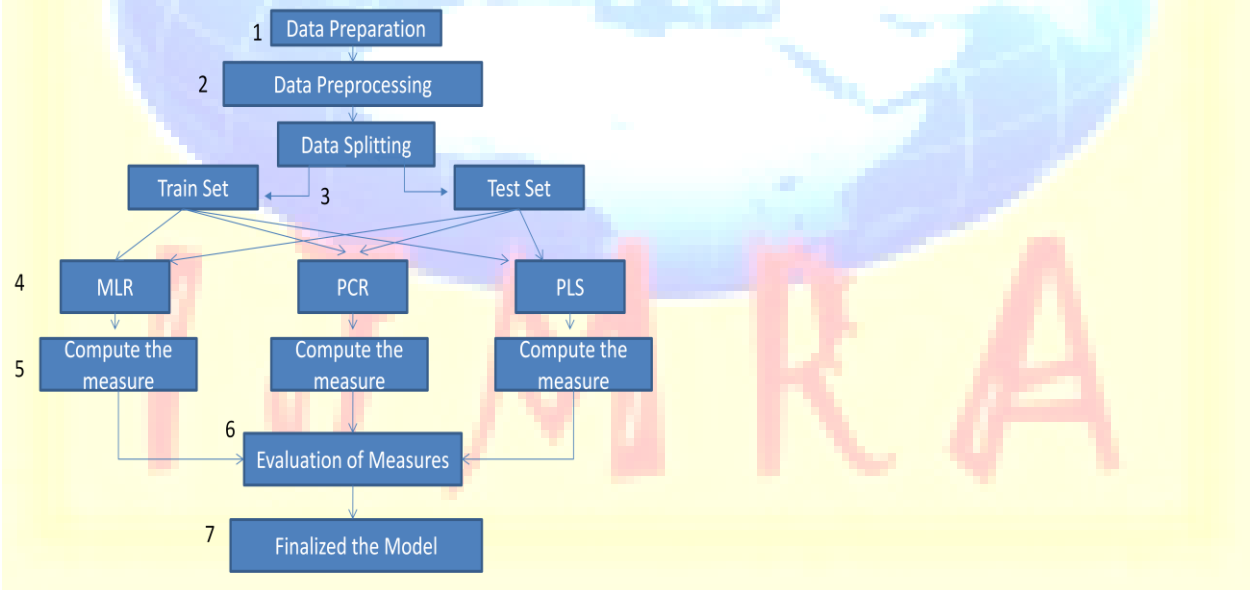

Figure 3.3 System architecture

1. Data Preparation: Database as their nature contains large amount of data. To extract information from the database Structured Query Languages are used [24]. SQL commonly used for the aggregation of large volumes of data. With the help of aggregation details in one table can

be aggregated with details in another table. Aggregation functions play a major in the summarization of tables.. Normal SQL aggregation functions are sum ( ), avg ( ), min ( ), max ( ) and count ( ). Data aggregation can be user-based [25]. For example personal data aggregation services offer the user a single point for collection of their personal information from other Web sites.

Vertical aggregations: Vertical aggregation is similar to standard SQL aggregations. This produces results in a vertical format and contains more rows. There are some approaches which produce results in vertically aggregated form.

Horizontal aggregations: Horizontal aggregations are also similar to standard SQL aggregations but this can produce results in horizontal tabular format.

There is a data preparation framework [26] for efficiently preparing dataset for analysis. There are four steps for the dataset preparation. Dataset preparation starts with data selection. In data selection, the analyst wants to perform analysis on the available data and select appropriate data for analysis. Second step is data integration. In data integration, data collected from different source are combined and stored inside a table. Third one is the data transformation. In data transformation the analyst wants to transform data into the format required for each operation. The last step is the data reduction. Here the data is compressed for the easiness of the analysis.

2. Data Preprocessing: 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 proposed neighbor-based feature scaling (NBFS) method [27] works as follows:

‘‘A’’ dataset consists of *n* features and *m* data. Y is the output

variable of A dataset. Feature set is $x_1(1),\ldots, x_n(1),\ldots, x_1(m), \ldots, x_n(m)$ and Y output variable is $Y_1(1), Y_2(2), \ldots, Y_n(m)$.

After NBFS applied to dataset, the new being features $X_1(1),$

$\ldots, X_n(1), \ldots, X_1(m), \ldots, X_n(m)$ are computed as follows:

$$x_{i,j} = \frac{x_{i,j}}{\text{The Total Euclidean Distance (belong to each feature in dataset)}},$$

Where, $X_{i,j}$ is the new feature value of dataset, $X_{i,j}$ is the old feature value of dataset for i = 1, 2, …, n and j = 1, 2,.... ,

m(i: raw value of feature, j: number of features).

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 [28].

Given a sample $\{(X_i, Y_i), 1 \leq i \leq n\}$, then the formula can be written as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Where $\bar{X}$ *and* $\bar{Y}$ is the average of each sample.

This is followed by finding the singular value decomposition (SVD) of the data sets transforming them into principal components [29]. This also will be helpful in checking the relationship between the variables in each data set.

3. Data Splitting: 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.

Systematic data splitting method [30] is a semi-deterministic method, in which every k^{th} sample from a random starting point is selected to form the training, test and validation datasets. In implementing systematic sampling in this study, the data are first ordered in increasing values along the output variable dimension. Then the sampling interval is determined based on the training and test data proportions specified by the user. Thereafter, a starting point is randomly selected and training samples are drawn first, followed by the test samples. Finally, unsampled data are allocated into the validation set.

4. 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. For example, Multiple Linear Regression has two methods associated with it in this study: the full model regression model and the model built selecting the best correlated variables to the output variables. This model is validated by using the test validation data set. Model adequacy criteria are used at this stage to measure the goodness of fit and adequacy of the prediction.

5. The results are presented in tables. The results of the train sets are not presented in this study because they are not relevant. This is because only the performance of the model on the test data set or entirely different (confirmatory) data set is relevant.

6. The model is expected to perform well when different data sets are applied to it. In this thesis work, 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.

7. Finally, all the methods of all the techniques are compared (based on their results on each data set) using four very strong model adequacy criteria. The best result gives the best prediction technique or algorithm for that particular type of data set.

V. CRITERIA FOR MODEL EVALUATION

Many criteria can be used to evaluate the predictive abilities of the different DM techniques. For the purpose of this work, about five criteria will be used in comparing different methods within each technique

1. Root Mean Square Error: The MSE of the predictions is the mean of the squares of the difference between the observed values of the dependent variables and the values of the independent variables that would be predicted by the model [31][33]. It is the mean of the squared difference between the observed and the predicted values or the mean square of the residuals.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}}$$

Where the sum is over all samples for a given Y- variable. The particular MSE obtained depend upon which data Y_i and fitted (predicted) values $\hat{Y}_i$ are used.

2. Mean Absolute Error (MAE): This error measure is similar to MSE, except that it uses absolute error values instead of the squared errors [32] [33], i.e.

$$MAE = \frac{\sum_{i=1}^{N}|Y_i - \hat{Y}_i|}{N}$$

3. Coefficient of Efficiency: This has been used in many fields of science for evaluating model performance. According to Nash et al. [33], the coefficient of efficiency can be defined as

$$COE = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}$$

$$COE = 1 - \frac{MSE}{Variance\ of\ Observed}$$

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.

4. Coefficient of determination: The First criteria which have used in our study are 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 [8].

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

$$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST}$$

Where SST is corrected sum of square response

$$SST = \sum_{i=1}^{N}(Y_i - \bar{Y}_i)^2$$

SSR is the residual sum of square

$$SSR = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

And SSE is the is the sum of squares due to error

$$SST = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y}_i)^2$$

Where $\bar{Y}_i$ denotes the mean of the dependent variable and $\hat{Y}_i$ is the i^{th} fitted value. Method to compare the two models would be to compare the R^2 corresponding to the models: the model with the highest R^2 provides the closest fit. According to the R^2 criteria [34], one should choose the model which has the largest R^2. R^2 lies by definition between 0 and 1 and reports the fraction of the sample variation in *y* that is explained by the *x*.

5. Number of features or variable used: The number of variables included in a model determines how good the model will be. A good predictive DM 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 or variables that will account for the most information needed to explain or build the model.

VI. CONCLUSION AND DISCUSSION

Before actual procedures for selecting models, there are a few general considerations which must be taken into account. First of all, we need to define the maximum model, that is, the model containing all explanatory variables which could possibly be present in the final model. Note that this includes interaction terms that might affect the response variable. Thus, any possible model for the data is a restriction of the maximum model, in the sense that it can be achieved by omitting a number of the explanatory variables from the maximum model. Let k denote the maximum number of feasible explanatory variables (including appropriate interaction terms). Then, the maximum model is given by

$$Y_i = \beta_0 + \beta_1 x_i, 1 + \beta_2 x_i, 2 + \cdots + \beta_p x_i, p + \varepsilon_i$$

where $x_i, 1, \ldots, x_i, k$ are the explanatory variables, and ε_i are independent, normally distributed random error terms with zero mean and common variance.

The most careful selection procedure is the all possible models procedure in which all possible models are fitted to the data and the selection criterion is used on all the models in order to find the model which is preferable to all others. Note that one has to choose the selection criterion carefully, as different selection criteria can result in different 'best' models!

For selecting the best model in the group, five measuring criteria were considered over the nine used in study. Models with condition numbers below 100 were chosen first. Then, those with the lowest RMSE among those that passed the condition number were chosen. If there were ties, the MAE was used to break the tie via models with lower MAE were chosen over others. If there was still a tie, then coefficient of efficiency was used, models with higher modified coefficient of efficiency were favoured over others at this stage. Still there is any contradiction in

decision of finalising the model coefficient of determination are used. The closer the model fits the data, the larger R^2 will be. Thus, an intuitive, but crude, method to compare the two models would be to compare the R^2s corresponding to the models with number of variables, factors or PCs that made the model was used to select the best model. Models with fewer variables, factors or PCs were favoured over others. MAE is especially useful in resolving the problem RMSE has with outliers.

REFERENCES

[1]. Usama, M. Fayyad, "Data-Mining and Knowledge Discovery: Making Sense Out of Data", Microsoft Research IEEE Expert, Vol. 11, No. 5, pp. 20-25, 1996.

[2]. Giudici P., "Applied Data-Mining: Statistical Methods for Business and Industry" West Sussex, England: John Wiley and Sons, 2003.

[3]. Joyce Jackson, "Data Mining: A Conceptual Overview", Communications of the Association for Information Systems, Volume 8, pp. 267-296, 2002.

[4]. Debahuti Mishra, Asit Kumar Das, "Predictive Data Mining: Promising Future and Applications", Int. J. Of Computer and Communication Technology, Vol. 2, No. 1, 2010.

[5]. Cabena P., Hadjinian P., Stadler R., Verhees J., Zanasi A., "Discovering Data Mining: From Concept to Implementation", Upper Saddle River, NJ: Prentice Hall, 1998.

[6]. Fayyad U., Piatetsky-Shapiro G. and Smyth R, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, 27-34, 1996.

[7]. Ruppert D., Wand M. P., Caroll R.J., "Generalized Regression Models", Journal of the American Statistical Association, STAT902, pp. 1-25, 2003

[8]. Joseph G. Eisenhauer, "Regression through the Origin", Teaching Statistics. Volume 25, No. 3, Autumn 2003.

[9]. Ming Yuan, Yi Lin, "Model selection and estimation in regression with grouped variables", J. R. Statist. Soc., Part 1, pp. 49–67, 2006.

[10]. John O. Rawlings, Sastry G. Pantula, David A. Dickey, "Applied Regression Analysis A Research Tool", Springer Texts in Statistics, 1998.

[11]. Abhishekh Taneja, "Predictive Data Mining-A Relative Study of Linear Techniques: Multiple Linear Regression vs. Factor Analysis", International Journal of Innovative Technology & Creative Engineering, Vol.1 No.4, April 2011.

[12]. Ya Su, Xinbo Gao, Xuelong Li and Dacheng Tao, "Multivariate Multilinear Regression", IEEE Transactions On Systems, Man, and Cybernetics—Part B: Cybernetics, 2012.

[13]. Yazid M. Al-Hassan, "A Monte Carlo Comparison between Ridge and Principal Components Regression Methods", Applied Mathematical Sciences, Vol. 3, No. 42, 2085 – 2098, 2009.

[14]. Wall, Michael E., Andreas Rechtsteiner, Luis M. Rocha."Singular value decomposition and principal component analysis". A Practical Approach to Microarray Data Analysis, Kluwer: Norwell, pp. 91-109, 2009.

[15]. Shih-Ming Huang and Jar-Ferr Yang, "Improved Principal Component Regression for Face Recognition under Illumination Variations", IEEE Signal Processing Letters, Vol. 19, No. 4, April 2012.

[16]. Abu Jafar Mohammad Sufian, "Analyzing Collinear Data by Principal Component Regression Approach — an Example from Developing Countries", Journal of Data Science 3, pp.221-232, 2005.

[17]. Svante Wold, Michael Sjo¨stro¨ma, Lennart Eriksson, "PLS-regression: a basic tool of chemometrics", Chemometrics and Intelligent Laboratory Systems 58, pp. 109–130, 2001.

[18]. Aniruddha Kembhavi, David Harwood, "Vehicle Detection Using Partial Least Squares", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 6, June 2011.

[19]. Hong Su and Gangtie Zheng, "A Partial Least Squares Regression-Based Fusion Model for Predicting the Trend in Drowsiness", IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems And Humans, Vol. 38, No. 5, September 2008.

[20]. Cody S. Ding, "Using Regression Mixture Analysis in Educational Research", Practical Assessment Research & Evaluation, Vol. 11, No 11, December 2006.

[21]. Rocío Muñiz, Miguel A. Pérez, Jesús A. Baro, "Comparison of principal component regression (PCR) and Partial least square (PLS) methods in prediction of raw milk Composition by vis-nir spectrometry", IMEKO World Congress Fundamental and Applied Metrology, pp.6-11, September 2009.

[22]. P. Venkatesan, C. Dharuman and S. Gunasekaran, "A comparative study of principal component regression and partial least squares regression with application to FTIR diabetes data", Indian Journal of Science and Technology, Vol. 4, No. 7, July 2011.

[23]. ÄOzgÄur Yeniay, Atilla GÄokta, "A comparison of partial least squares regression with other Prediction methods", Journal of Mathematics and Statistics, Volume 3, pp. 99-111, 2008.

[24]. Kai-Uwe Sattler, Eike Schallehn, "A Data Preparation Framework based on a Multidatabase Language," IEEE Trans. Knowledge and Data Eng., 2001.

[25]. Jincy Annie V.V, J. A. M. Rexie, "Efficient Tabular Dataset Preparations by the Aggregations in SQL: A Survey", International Journal of Computer Applications, Vol. 58, No.15, November 2012.

[26]. C. Ordonez, "Horizontal Aggregations for Building Tabular Data Sets", IEEE Trans. Knowledge and Data Eng, Vol. 24, No. 4, April 2012.

[27]. Kemal Polat, "A novel data preprocessing method to estimate the air pollution (SO2): neighbor-based feature scaling (NBFS)", Neural Computation and Applications, pp.1987–1994, 2012.

[28]. Xu Jian, Lu Xiaolin, "Correlations among direct input coefficients and its applications to update IO tables: an empirical investigation", Journal ofInternational Input-Output Association, June 2011.

[29]. Alter O., Brown P.O., Botstein D.," Singular value decomposition for genome-wide expression data processing and modelling", Proc. Natl. Acad. Sci. USA, 2000.

[30]. Wenyan Wu, Robert May, Graeme C. Dandy and Holger R. Maier, "A method for comparing data splitting approaches for developing hydrological ANN models", International Environmental Modeling and Software Society, 2012.

[31]. ElMoustapha Ould-Ahmed-Vall, James Woodlee, Charles Yount, Kshitij A. Doshi, "On the Comparison of Regression Algorithms for Computer Architecture Performance Analysis of Software Applications", Performance Analysis of Systems & Software, 2007.

[32]. Sergio Cleger-Tamayo, Juan M. Fernández-Luna, Juan F. Huete, "On the Use of Weighted Mean Absolute Error in Recommender Systems", Proceedings of the Workshop on Recommendation Utility Evaluation: Beyond RMSE, 2012.

[33]. Cort J. Willmott*, Kenji Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance", Journal of Climate Research, Vol. 30, pp. 79–82, 2005.

[34]. Li-Shan Huang and Jianwei Chen, "Analysis of Variance, Coefficient of Determination and F-Test for Local Polynomial Regression", The Annals Of Statistics, Vol. 36, No. 5, 2085–2109, 2008