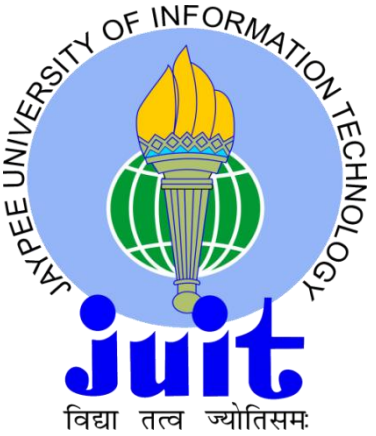
REGRESSION- OPTIMIZATION AND ESTIMATION

Enrollment no - 122213

Name of Student - Mrityunjay Sharma

Name of Supervisor - Mr. Suman Saha



May-2014

Submitted in Partial fulfillment of the Degree of

Master of Technology

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT, DIST. SOLAN, (H.P.), INDIA

Contents

CERTIFICATE

This is to certify that the work titled **"REGRESSION -OPTIMIZATION AND ESTIMATION"** submitted by **" MRITYUNJAY SHARMA"** in partial fulfillment for the award of degree of Master of Technology in Computer Science of Jaypee University of Information Technology, Waknaghat has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor:- .....

Name of Supervisor:- Mr Suman Saha

Designation::- Assistant Professor (Grade-II)
Jaypee University of Information Technology

Date::- .....

DECLARATION

I declare that this thesis entitled "REGRESSION -OPTIMIZATION AND ESTIMATION",submitted by me for the award of degree of Master of Technology in Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat is original and it has not submitted previously to this or any other University for any degree or diploma.

Signature of Student:- .....

Name of Student:- Mrityunjay Sharma

Date::- .....

ACKNOWLEDGEMENT

Nothing can be accomplished without a 'Guru'.I would like to express my earnest gratitude to my Project Guide **Mr.Suman Saha** sir,Assistant Professor, Department of Computer Science and Engineering, Jaypee University of Information technology,Waknaghat, for his constant help and guidance. Sir helped me in formulating the idea and collection of the related material for the project proposal. His continuous monitoring to support me and my research work encouraged me a lot for doing my thesis in a smooth manner.

Last but not the least,I would like to extent my appreciation towards my parents and my friends for always being there through all phases of work,for their encouragement and patience and giving me their valuable support without which I would never be where I am today.

Signature of Student:- .....

Name of Student:- Mrityunjay Sharma

Date::- .....

ABSTRACT

Linear regression is an important and widely used approach amongst the wealth of machine learning techniques.This approach is used for prediction and forecasting by using massive datasets that have high dimensionality. The results obtained are beneficial for both the organization and the consumer for better decision making. Numerous different models have been proposed for selecting the best regression model i.e. is the model having less dimensionality and high degree of accuracy.The criteria for selection of attributes in these models is based only on their significance without considering the linear relationship between the attributes involved in the model (multicollinearity) which violates the assumptions of linear regression . This thesis is devoted to developing an algorithm which builds a model for the prediction considering high accuracy and low multicollinearity for every subset formed. It is not specific to a particular domain and can be applied to any dataset. The attributes are added to the model based on the above mentioned features.The Variance Inflation factor(VIF) and the sum of the summation of Cooks statistic, square root of mean square error, absolute value of difference of fits and Press residual of each observation are taken into consideration to keep the multicollinearity low and accuracy high.

The proper choice and number of predictors used for prediction of response variable plays an important role.Multicollinearity affects the model accuracy hence emphasis has to make to keep it as low as possible. In addition to all this the approach has to be pictorial or graphical in nature for better comprehensibility which is unique. This paper keeps in mind all these factors and proposes a technique for building a best (accurate, minimum multicollinearity, graphical) regression model.Incorporating all these qualities the thesis gives a graph based approach for minimum multicollinearity highly accurate regression model explaining maximum variability

List of Figures

List of Tables

Chapter 1

Introduction

Science has taken the advantage of computers which can store large amount of data.Biology was the first,achieving the proficiency to measure gene expression in DNA microarrays which produces datasets having high dimensionality,along with protein transcription data.Even other sciences such as astronomy uses high resolution images to store the observatories of the night sky:around a tera byte per night.Medical sciences also stores the results of various medical tests such as the blood tests and Magnetic Resonance Imaging(MRI) scans.The sudden outburst of data is well known.The main challenge is to do something useful with it. The data that is collected and extracted has high complexity and large size.Thus humans are unable to extract useful conclusion from it.The proliferation of database has also given impetus to the overwhelming massive collection of data.

1.1 Learning

The key concept is to make the machines learn from the abundant data that we have.It is easy to understand this term in human behavior aspect as to learning from experience and that is why humans and animals are termed as intelligent.The whole

procedure of learning basically involves three steps namely remembering,adapting and generalization.Generalization is establishing the correlation between different events so that decisions applied at one situation can be applied at a similar one.We recognize last time when we were in a particular situation(saw the data)we took some action(gave output) and it was correct(it worked), so we will take the same action or that action was not desired so we will try something else or will try it again.Reasoning and logical deduction which uses symbols to reflect the environment are the basis of Artificial Intelligence (also called symbolic processing)Machine learning involves no symbolic manipulations or symbols and hence sometimes called subsymbolic. Learning of machines are significant when

1)Human expertise is not sufficient and enough(navigating on Saturn)

2)Humans lack the ability to demonstrate their expertise (speech recognition)[1]

3)Solution is stochastic or is temporal (stock market or gold price)

4)Solution needs to be adjusted to specific classes (user biometrics)

Data is abundant and cheap where as the knowledge is expensive and scarce.The main objective is to build a general model by studying specific examples.When we talk about learning we mean building a model that is useful and good approximation of the data and decipher important and unique features in the dataset.a very good example is in retail studying the behavior of the customer by keeping the track of his/her transactions.

1.2 Machine Learning

Machine Learning is making the computer learn and adapt(making predictions or forecasting events)so that the actions that are performed achieve higher accuracy in the course of time.We measure the accuracy as how the actions chosen by the

computer match with the real one.It is perceived that when we play a game(eg chess) against a computer.We might win every game in the beginning,but after couple of games we start loosing and reach a situation where we never win.The computer will not start from scratch when it plays against some other opponent but infact use the strategies it used to defeat you.This is termed a generalization.

It is a natural outgrowth of the intersection of Statistics and Computer Science.Statistics involves inference from data in addition to some modeling assumptions.Whereas computer science has emphasized on manually programming computers machine learning focuses primarily on how to program computer system to program themselves(from some initial structure,data or experience)[2].Machine learning deals with which algorithms and computational architecture should be used in order to retrieve,merge,store,index the data and how diverse learning subtasks can be orchestrated in a bigger system, and solves questions of computational tractability.

Human and animal learning in Neuroscience and Psychology are closely related to machine learning. The way how computers can learn and how humans or animals learn have highly intertwined answers.Due to the weak state of our perception and understanding of Human Learning the cognizance that machine learning has gained from studies of Human Learning are much weaker than those it has abstracted from Statistics and Computer Science.But as both machine and human learning close neighbors in the landscape of some core scientific questions it is reasonable to expect synergy between the two in the coming years.

Other fields, from economic,biology and control theory also have a core interest as to how they can automatically optimize and adapt their environment and machine learning has exchange of ideas with these fields. For example, economics is interested in how effect the stock market would have on the current GDP of the country .And control theory, mainly the adaptive control theory, is interested in how a servo-control system can make its control strategy better through experience. The mathematical models used in these fields are different from those commonly used

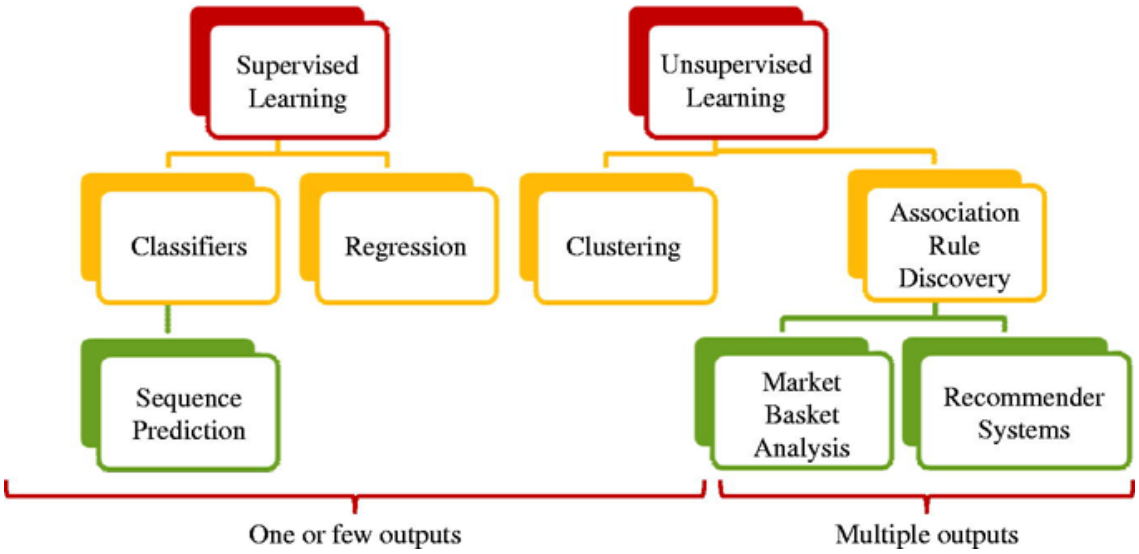


FIGURE 1.1: Classification of Machine Learning

in machine learning, suggesting expressive potential for cross-fertilization of models and theories. Thus we can now potray the basic essence of machine learning as a field that explains the basic and fundamental laws that govern all learning processes and implements the techniques by which we can build computer systems that automatically improves with experience.It is also used to optimize a performance parameter using past experience or example data.Machine learning algorithms are data analysis techniques which search the data sets for characteristic structures and patterns.

1.3 Fundamental types of Machine Learning

1.3.1 Supervised Learning

A training set of examples that contain the correct value of the class attribute(targets) are provided and on the basis of the training set,the algorithm generalizes to respond accurately to all the possible inputs.This is also termed as learning from exemplars. Regression and classification are examples of this type.

1.3.2 Unsupervised learning

In this technique the correct value of the response is not provided.The algorithm tries to analyze the similarity between the inputs so that the inputs having something in common are categorized together.Density estimation is the statistical approach to unsupervised machine learning.Clustering is example of this type

1.3.3 Reinforcement Learning

This lies somewhere in between supervised and unsupervised machine learning covering both labeled and unlabeled examples.The algorithm is told when it is wrong but not told how to correct it.Thus it has to use clever explorations by applying different combinations and possibilities so that it gets the correct answer.Due to this reason it is sometimes called learning with critic as thoughthe monitor scores the answer,it does not suggest improvements.

1.3.4 Evolutionary Learning

Evolution in biology can be seen as a effective learning process.Biological organisms accustom to improve theirchance of having offsprings and improve survival rates in the envirnment.This is done by using the idea of fitness which harmonizes with how good the current solution is.

1.4 Scope and Applications of Machine Learning

Machine learning develops an efficient algorithm that is made to note changes that takes place in the data and to incoporate in its design the recent finding and explorations. When it is applied to predictive analytics, this quality mentioned has

significant impact on the activities normally undertaken to develop, test, and refine an algorithm for a specific purpose.

Sophisticated pattern recognition :– Along with noting important and significant relationships it is able to determine a variety of attribute and quantify it as well. This happens on each and every relationship that takes part in forming the pattern. It selects the features that should be selected to determine the requisite output or the response variable with high accuracy which is termed as feature extraction.

Intelligent decisions :– Along with the proficiency to note data that is irrelevant , and form a hierarchy of the relative significance of variables in the model it also makes resolution that are either aided by the user or not. This becomes clear when we have to predict a event that is rare or infrequent. The solution can manage to discern between subclasses and make a judicial decision on what type of data be considered and what should not be with least instruction.

Self modifying :– Clearlyhaving the ability to add,drop or tweak various aspects of a particular algorithm to describe the variability in a specific model is a time saver. This is an significant feature as it helps to make the system robust and predict the correct output responses for a specific model.

Multiple iterations :– To make the model more robust a number of multiple iterations are done to make a final model that has high level of accuracy and maintains the finest fit to the data.

Some of the recent applications of machine learning are as follows:-

Automating Employee Access Control

Amazon, a main pioneers of machine-learning based price discrimination algorithms and recommendation engines , floated a machine learning competition on Kaggle to determine whether it is able to automate employee revocation and access granting . Amazon has a large dataset of employee access levels and employee roles . They are

trying to make a computer algorithm that will make a good prediction as to which employees should be granted access to what type of resource. According to Amazon, "These models which are auto-access seek to reduce the human involvement that is required to revoke or grant employee access."

Protecting Animals

Cornell University is working on an efficient algorithm to spot whales in the ocean which is based on audio recordings so that ships can abstain hitting them. Also, Oregon State University is also working on software that will predict which bird species is/are on a particular audio recording collected in field conditions.

Predicting Emergency Room Wait Times

Healthtech companies and healthcare organizations use a new technique called Discrete Event Simulation for the prediction of wait times for patients who are in waiting room in emergency department . The models use features such as patient data,staffing levels, emergency department charts, and also the layout of the room itself to predict wait times.

Stock Market Prediction

Stock Market has always intrigued the researchers and accurate techniques and algorithms for its prediction are at hype these days.The main issue is as to which approach to be applied so as to increase the accuracy.Here accuracy is of more concern than the speed such as in application involving prediction of natural disasters where the prediction has to be fast as well as accurate[3].

Identifying Heart Failure

IBM researchers have found a technique to extract heart failure diagnosis evaluation from free-text physician notes. They developed a machine learning algorithm that combs through physicians free-form text notes (in the electronic health records) and analyze the text using a technique called "Natural Language Processing" (NLP).

Similar to the way a cardiologist can see through some another physician's notes and analyze or figure out whether a patient has heart failure, computers systems can now do the same work.

Predicting Strokes and Seizures

Singapore-based startup Healint flated an application called JustShakeIt that gives user an ability to send an emergency alert to emergency contacts and caregivers by shaking the phone with one hand. The program uses an efficient machine learning algorithm to make difference between actual emergency shakes and regular jostling. In addition to the JustShakeIt application, Healint is also working on a model that analyzes and evaluates patients' cell phone data from accelerometer to help analyse warning signs for chronic neurological disorders.

Predicting Hospital Readmissions

Additive Analytics, is working on a machine learning model that identifies and predicts which patients are at a higher risk of readmission. Using our efficient and proprietary predictive model, hospitals can make a prediction of emergency room admissions before they happen—improving care results and reduce costs. In the 90s and 2000s, software as well as internet transformed the way companies did business. Cutting-edge and tech-savvy companies such as Amazon and Google came up rapidly. Old and stodgy companies like Blockbuster and Borders were not able to keep up and failed.

In the 2010s and 2020s, strong analytics and machine learning is transforming industries again, just as software transformed the whole world over the past 30 years.

1.5 Problem Statement

There are various issues that can hamper the prediction and size of the dataset involved for prediction of response variable from a set of explanatory variables.One

such issue identified was multicollinearity.Generally while building a model only the significance of variable in predicting a given response variable is regarded as the most important parameter on which the variables are entered in the model.A set of variables that have a relationship between them may all be significant and thus they are included in the model but this results in redundancy and increase in the number of explanatory variables.

In this thesis an approach is proposed that does not include the predictors that have linear relationships keeping high predictive power.This has been done by taking into account certain statistical parameters such as Cook's statistics, Square root of mean square error,Difference of fits and Press residual.Chossing these parameters helps to build a model with high accuracy and minimum multicollinearity. The greater the value of the VIF(Variance Inflation factor)the greater is the multicollinearity.The VIF is kept below 10 as its value above 10 means the introduction of severe multicollinearity.The proposed algorithm is compared with some of the other aproaches on the basis of accuracy and Condition Index. It has been seen in the previous approaches that the main focus was on selecting a model which had high significance or F statistics.It was later observed that multicollinearity cannot be ignored and has to be taken care .It may lead to various adverse effects in prediction of the model, parameter estimation and other problems.So an important line of research is to develop a regression model in which the predictors are entered in the model based on least multicollinearity and low mean square error. Building such a model will help to achieve both the objectives.This will build a model that is free from variables that convey the same information and reduce the duplication of information conveyed by a specific variable. We have proposed a technique of building best regressor variable having minimum multicollinearity and high accuracy.

1.6 Motivation

A data set may have large dimensionality.From this some predictors are chosen so as to build a model which involves variables that have high significance , greater accuracy and reduce the size of data set for prediction.It has been observed that the main aim of the best regressor models is high accuracy but while selecting a model there may be attributes that are closely related or have some linear relationship between each other.The primitive models take all these regressors in the model as thay are significant but these explanatory variables pose duplicacy in the prediction model with many variables containing the same information.Building such a model may also increase the size of the regression model.Here a technique is proposed whereby we build a model for regression that has minimum multicollinearity maintaining the accuracy.

Doing this makes all the variables in the data set independent of each other.The data value that are entered in the dataset may come from various typical and complex experiments so by applying such a technique,we can in future collect only the data values of independent variables without entering those values which are closely related to them.Keeping all these issues in mind we have proposed a technique for building a regression model that has minimum multicollinearity or in other terms it has minimum linear relationship between the variables/explanatory variables involved in the model for prediction of the values that are not far from the true ones.

1.7 Organization of Thesis

This thesis is organized into five chapters. The Chapter1 describes machine learning,its type and application and the problem statement.Chapter 2, explains simple linear regression,multiple linear regression and criteria for selection of best regressor model. Chapter 3, describes multicollinearity-its effect and the techniques by which

we can deal with it.Chapter 4, explains a robust regressor technique for selection of best regressor model having minimum multicollinearity along with the results obtained.Chapter 5,describes the graph Based Approach for minimum multicollinearity highly accurate regression model explaining maximum variability along with the results and its analysis.

Chapter 2

Premliminaries and Background

Regression is a supervised machine learning method which advances by building a equation by which we can estimate the response variable. Data preprocessing methods have been applied so as to improve the predictive performance.Forecasting is the objective of machine learning and various methods can be used for it one of which is regression.This technique can be applied in a number of research fields .The values of response is estimated through a equation which can be linear,non linear ,quantile or polynomial.The proper choice and number of predictors used for prediction of response variable plays an important role.Regression is a statistical technique which is used to study the relationship between and dependent attribute (such as the burnt area of forest fire) and one or more independent attribute (such as the humidity, rain, speed of wind etc).It has found its applications in various fields such as economics, social science and engineering in predicting stock market, forest fire prediction, heating and cooling load prediction of a room and variety of other ample applications[4, 5, 6]. Regression makes an equation with certain selected attributes having high significance that explain approximately total variability in the model to predict the class attribute. Calculating the significance i.e. which attributes to consider to get a high predictive power of the model have also attracted researchers Least square method is used to determine the coefficient used

in the regression equation. Thus the main objective of this technique is to build a robust model with high predictive power[7]. The robustness plays an important role especially when we have a time series data which may sometimes contain infrequent or unusual occurrences.

2.1 Simple Linear Regression

It is a model with a single independent variable or regressor(X) and a single dependent variable or response(Y) A simple linear regression model is:-

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

where Y is the response variable to be predicted

X is the explanatory or the independent variable

β_0 is the intercept

β_1 is the slope

ε_i is the random error component

2.1.1 Assumptions

There are some basic assumptions on the linear regression model which are as follows:-

1. ε_i is random variable with mean zero and varianceσ^2.

2.There exist no relation between ε_i and ε_j that is both of them are uncorrelated.

3. ε_i is a normally distributed variable with mean $\equiv 0$ and variance $\equiv \sigma^2$.

$$\varepsilon_i \sim N(0, \sigma^2)$$

The value of β_0 and β_1 are calculated by least square method such that the regression line that is fitted by this method is the one that makes the sum of squares as small as possible. that is

$$\sum_{i\equiv 1}^{n} e_i^2 \;\; is \;\; minimum$$

where $e_i = Y_i - \widehat{Y_i}$ and

Y_i is the orignal value of the response.

By taking the partial derivative of the least square estimators β_0 and β_1 and equating to zero we get their value such that the sum of square due to error is minimized.

$$S = SS_{res} = \sum_{i\equiv 1}^{n} e_i^2 = \sum (Y - \widehat{Y_i})^2$$

$= \sum (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$is minimum

By taking the derivative and equating it to zero it can be seen that $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are solution of the equation.These equations are also called the normal equations

$\sum (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i \equiv 0$

$\sum X_i (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i \equiv 0$

by which we get the value of $\widehat{\beta_0}$

$$\Rightarrow \sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i - n\widehat{\beta_0} - \widehat{\beta_1} \sum_{i=1}^{n} X_i)^2 = 0$$

$$\Rightarrow n\widehat{\beta_0} = \sum_{i=1}^{n} Y_i - \widehat{\beta_1} \sum_{i=1}^{n} X_i)^2$$

$$\Rightarrow \widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$$

where

$$\overline{X} = \frac{X_i}{N} \; and \; \overline{Y} = \frac{Y_i}{N}$$

$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1} X$ Eq 1

The value of $\widehat{\beta_1}$ can be calculated by using the second normal equations as follows

$$\sum X_i(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i \equiv 0$$

By using Eq 1 we get

$$\sum X_i(Y_i - \overline{Y} + \widehat{\beta_1}\overline{X} - \widehat{\beta_1} X_i = 0$$

$$\Rightarrow \sum X_i(Y_i - \overline{Y}) - \widehat{\beta_1} \sum (X_i - \overline{X})X_i$$

$$\Rightarrow \widehat{\beta_1} = \frac{\sum (Y_i - \overline{Y})X_i}{\sum (X_i - \overline{X})X_i}$$

$$\Rightarrow \frac{S_{XY}}{S_{XX}}$$

So we get the value of

$$\widehat{\beta_1} = \sum \frac{(Y_i - \overline{Y})X_i}{(X_i - \overline{X})X_i}$$

These are used to fit the linear regression model.the line that is fitted has the least residual sum of square.

2.1.2 Useful Properies of Least Square Fit

1)For any regression model having intercept β_0 the sum of residuals is always zero. $e_i = \sum Y_i - \widehat{Y}_i = 0$

2) $\sum Y_i = \sum \widehat{Y}_i$

3) $\sum X_i e_i = 0$

4) $\sum \widehat{Y}_i e_i = 0$

It can also be proved that $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the unbiased estimator of β_0 and β_1. The regression model will be explained in detail by taking the multiple linear regression in consideration.

2.2 Multiple Linear Regression

A regression model that has more than 1 regressor variable and a response variable is termed as multiple linear regression.This model consists of linears of unknown parameters β_0,$\beta_1 \ldots\ldots\ldots \beta_{k-1}$ and having the same assumptions as the simple linear regression.

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

$$\mathrm{Y} = \begin{pmatrix} \mathrm{Y}_1 \\ \mathrm{Y}_2 \\ \mathrm{Y}_3 \\ .. \\ .. \\ \mathrm{Y}_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ .. \\ .. \\ \beta_n \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ .. \\ .. \\ \varepsilon_n \end{pmatrix}$$

Vector of Observations $\quad Vector\ of\ Parameters \quad Vector\ of\ errors$

$$X = \begin{pmatrix} 1 & \mathrm{X}_{11} & \mathrm{X}_{12} & .... & ... & \mathrm{X}_{1(k-1)} \\ 1 & \mathrm{X}_{21} & \mathrm{X}_{22} & .... & ... & \mathrm{X}_{2(k-1)} \\ 1 & \mathrm{X}_{31} & \mathrm{X}_{32} & .... & ... & \mathrm{X}_{3(k-1)} \\ .. & .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. & .. \\ 1 & \mathrm{X}_{n1} & \mathrm{X}_{n2} & .... & ... & \mathrm{X}_{n(k-1)} \end{pmatrix}$$

If the value of i ranges from 1 to n we have

$$(Y_i, X_{i1}, X_{i2}, X_{i3}.....X_{i(k-1)})$$

and the final model can be expressed as

$$Y = X\beta + \varepsilon$$

where Y is the vector of observations β the vector of the parameters and ε is vector of the random error.

By constructing the normal equation by partially differentiating sum of square of residual SS_{res} and equating it to zero we get the value of

$$\widehat{\beta} = (X^{'}X)^{-1}X^{'}Y$$

We build an ANNOVA table which is based on the following parameters The ANNOVA Table (or Analysis Of Variance) table gives us the following information:

1. Degrees Of Freedom $\Rightarrow$ Degree of freedom are basicaaly the unknown parameters whose value is not known.

2. Residual Sum Of The Squares$\Rightarrow$ It does not take into account the degree of freedom and is given by the formula.It can also be understood as the variablility

not explained by the model.

$$S_{res} = \sum_{i=1}^{n} (Y_i - \widehat{Y})^2$$

3. The Mean Square ⇒It takes into account the degree of freedom and hence is considered as an important measure or error.

$$MS_{res} = \frac{Residual\ Sum\ of\ square}{Degree\ of\ Freedom}$$

4. F Statistic ⇒ The F ratio is calculated as the ratio of Mean square of regression and the mean square of residual error.

$$Fratio = \frac{MS_{reg}}{MS_{res}}$$

5.Sum of Square Regression⇒This denotes the variability that is demonstrated by the regression model.

$$SS_{reg} = \sum \widehat{Y} - \overline{Y}$$

6.Sum of Square Total⇒This explains the total variability in the model.A model is best when $SS_{reg} = SS_T$ that is the total variability is explained by the regression equation.

$$SS_T = \sum_{i=1}^{n} (Y_i - \overline{Y}$$

Data preprocessing methods have been applied so as to improve the predictive

Source of variation	Degree of Freedom	Sum of Square (SS)	Mean Square (MS)	F
Regression	k-1	SS_{reg}	$MS_{reg} \equiv \frac{SS_{reg}}{k-1}$	
Residual	n-k	SS_{res}	$MS_{res} \equiv \frac{SS_{res}}{n-k}$	$F \equiv \frac{MS_{reg}}{MS_{res}}$
Total	n-k	SS_T	$F \equiv \frac{MS_{reg}}{MS_{res}}$	

performance.Forecasting is the objective of machine learning and various methods can be used for it one of which is regression.This technique can be applied in a

number of research fields .The values of response is estimated through a equation which can be linear,non linear ,quantile or polynomial.

2.3 Criteria for evaluation

The criteria by which model is evaluated are as follows:-

1.Coefficient of multiple Determination $\Rightarrow$ R_p^2 is the coefficient of multiple determination for a subset of regression model with (p-1) regressors and intercept β_0

$$R_p^2 = \frac{SS_{res}(p)}{SS_{total}}$$

The main benefit of using this statistical parameter is that it takes into account only (p-1) regressors and not the whole model.

2.Residual Mean square $\Rightarrow$ $MS_{res}p$ decreases as the number of variables increases so this parameter is introduced which may increase or decrease with increase in the number of variables in the model.

$$MS_{res}(p) = \frac{SS_{res}(p)}{n-p}$$

Increase in $MS_{res}(p)$ takes place when the reduction in $SS_{res}p$ for adding a regressor to the model is not sufficient to compensate the loss of one degree of freedom in the denomnator.

3.Mallow's Statistics $\Rightarrow$ This measures the overal bias or mean square error in the fitted model.

$$\frac{\sum_{i=1}^{n} E(\widehat{yi} - E(\widehat{yi}))^2}{\sigma^2}$$

where $\widehat{yi}$ is the i^{th} fitted value $E(y_i$ is the expected response for the regression model

$$C_{(p)} = \frac{SS_{res}p}{MS_{res}^{full}} - n + 2p$$

when this value of Mallow's statistic is equal to the number of current regressors in the model it is considered as the best model.In other words lower value of this statistical parameter is desirable.

Techniques such as random forests have also been used in order to estimate Heating and Cooling Load in an Energy Efficient Building[8].the dataset used in this approach is also evaluated by our procedure.Complex experiments and their evaluation on 768 diverse residential buildings show that we can predict HL and CL with low mean absolute error deviations from the ground truth.Agus et.al [9] proposed that increasing the dataset inputs.The in between data in the time series is filled and it also explains that greater the training data greater is the predictive power.Detrend and Deseasonalization is done so that the data can be divided into trend and seasonality.In this approach the prediction is done by neural network and support vector machine[10].The neural network uses the multilayer perceptron with a main aim to reduce the error and no focus is laid on the multicollinearity. Non parametrized kernel regression is also used so as to smooth the data.The neural network adjusts the weigh so that the mean square error is least.Multiple Linear Regression has been used to forecast the value of an assets such as gold [11].This model predicted the gold price based on considering some of the economic parameters such as the movement of the currency rate and inflation in the market.The results obtained were compared with the "'forecast1" which is a naive method.This method uses the observation that is most recent.The main application which needs high accuracy is stock exchange because it can bring laurels to a person or it may leave him fully shattered and depressed with large amount of monetary loss.The past data is present in large repositories that are maintained by various companies and this training data is used to predict the future values outcomes.Knowledge discovery in databases is basically used in making of a pattern on which the future values can be predicted.Regression analysis has been used as a data mining technique by researcher.The moving average method has been applied by researchers to time series regression which reduces the fluctuations in the data that might occur

in the due duration of time [12].The main approach in moving average is calculating the average for a particular period and time instances and putting these values in centre of the time span.Here the number of years consisting of the date,months and year is taken as the trend value.The main point in moving average method is the number of instances for which the average is to be calculated in this paper the average value over 3 months is calculated[13].

Many researches has been done in order to build a model that forecasts value that is not far away from the true value.Stock exchange is considered as a complicated prediction because this process is time varying and it has variations daily with varying statistical behavior.Stock market is thus considered as a random walk process.Prediction is also possible using binary logistic regression which does not involve the normality condition of predictors as in the linear model.Model has been applied to the Indian stock market and the results show that it can predict the the shares out performing [14]. In order to calculate the error in the linear regression the sum of squared error is not considered as a good measure of the error as it does not involve the degree of freedom that is associated with the variable.So it is advisable that we use mean square error as it takes mean square error into consideration[15].Neural Network back propagation method has been applied in order to predict the stock to which the learning rate has to be supplied in advance [16].This paper also focuses that use of only a single type of error for the evaluation is not a good approach so this techniques uses 5 errors which are Mean square,Root mean square error,Percentage mean absolute deviation,mean absolute error and mean absolute percentage error.The results show that this technique helps in reducing the error. Research using Back propagation algorithm which has multilayer feed forward network Back propagation is the Genetic algorithm that has been used here for learning[17].Every neuron employs an activation function and this iterative procedure carries on till we have a convergence of he errors or we can also supply our own threshold value.When this threshold value comes the procedure stops and the current model is considered as the best model for prediction.The dataset from

Microsoft corporation has been used in this experiment. Researches have also been carried out to see and evaluate which is better is it neural network or regression.It was observed that the Neural Network Standard feed-forward back prop (FFB) algorithm was better in prediction accuracy for the tehran stock exchange data which is being used here.But we should also lay stress that the size of the dataset should also reduce.This has been observed mainly in the micro array data sets such as DNA micro array dataset where the biologist have to find a disease based on on selecting the main features of interest so that we have a small,non redundant and relevant dataset built[18].This uses dataset of domain specific knowledge,gene is considered as the output label a there is no output variable.Regression can be Linear as well as non linear.Generally the techniques and models that can be applied to the linear model can not be applied to non linear model so we need to build a model which will be applicable to all models .Regression has also been used in order to study the Climate based on some parameters and to establish the relationship between the the set of explanatory variables and the climate variation of interest which can be like humdity,temperature,rain etc.This paper builds a model that is flexible and can fit for any type of regression model.It uses the concept of Contour regression which basically regularizes the area between the two Cumulative distribution[19].Presorting the predictor matrix is the main preprocessing step that is applied here.The aim of the multiple linear contour regression is based on the reduction of the sum squared error and it does the regularization of the parameters which increase the accuracy to a great extent.Initial parameters of forecasting models can be predicted by linear regression efficiently.But if the model is non linear it is unable to build a model that is optimal[20].It has been shown that by using certain optimization on the smoothing constants which are common for time series analysis.It has been shown by the research done that when the smoothing constants are optimized in a non linear manner the linear demand results in better value of the constants and the initial parameters than the linear method.

2.4 Selecting Best Regression Model

The datasets are massive and have high dimensionality.The selection of attributes is done on the basis of their significance in the model and considering some important parameters such as those mentioned in previous section.The main of a regression model is to have minimum number of explanatory variables such that they explain the maximum variability in the model[21].The following are the methods generally used for the selction os the best regressor model.

2.4.1 All Possible Selection

This method is quite cumbersome as it requires fitting the regression equation for every set of attributes.thus if we have n number of predictors the total equations that can be formed is 2^n.this requires high computation.If n is even a small number 10 then also $2^{10} = 1024$ equations may be required,which is a large number.The three criteria by which we select the variables in the model are as follows:- **1)**The value of the R-Sqrd that is achieved from the least square method.

2)The SS_{res} (Residual mean square) value.

3)The value of the Mallows statistic C_p.

We need to consider all the regression equation when we have $(k-1)$ regressors variables:-

Regressor	No. of models	Equation
0	$k-1 choose 0$	$Y = \beta_0 + \varepsilon$
1	$k-1 choose 0$	$Y = \beta_0 + \beta_1 X_1 + \varepsilon$
2	$k-1 choose 0$	$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
....		
k-1	$k-1 choose k-1$	
total	2^{k-1}	

2.4.2 Sequential Selection

2.4.2.1 Forward Selection

This technique starts by considering that there are no regressors or the predictors in the model.At every step we add a new regressor in the model if follows certain criteriaseber2012linear.This approach is an incremental approach.The main drawback of this approach is that when the data has predictors that have high correlation among themselves then it may select the explanatory variables that show high significance in the global F test but they are not significant in the partial F test.this arises due to the linear relationship between the different attributes which bring the duplicacy in the information that is conveyed by the attributes. The algorithm proceeds in the following manner:-

1. Initially we consider that there are no regressor in the model.

2.Considering the model with one regressor we consider all possible models with one regressor and then commpute the F statistic value for each predictor or the regressor.

A regressor having the highest value of F statistic is added to the model if

$$F > F_{\alpha,1,resdf}$$

3.Partial F statistic value is calculated of the remaining regressor with the previous regressor already in the model.The model obtained by the combination of regressors such that it has the highest F statistic is added if and only if

$$F > F_{\alpha,1,resdf}$$

.

4.This algorithm terminates when the value of the highest F statistic does not exceed the F tabulated value or the last regressor is added to the model.

5.Backward elimination algorithm terminates when the smallest value of $F > F_{out}$

2.4.2.2 Backward Elimination

The backward Algorithm is selects the best regressors by taking the whole model into consideration first.This method does not give good results when the variables have high correlation among themselves because matrix multiplication becomes a tough task and the computer software are unable to calculate it or sometimes give unstable answers.the approach proposed in this thesis also explains that the backward elimination method could not be performed on the datasets having high correlation. The algorithm proceeds in the following manner:-

1. Start with the full model having all the explanatory variables or the independent variable.

2.Compute the partial F statistic value for each regressor in the presence of the other regressor already selected in the model.

3.The regressor that has the minimum value of the partial F statistic is removed from the model if $F < F_{out}$

4. Partial F statistic is calculated for the new model and the process iterates .

5.Backward elimination algorithm terminates when the smallest value of $F > F_{out}$

We calculate and calculate the partial and F statistic as follows:-

$$F_{1/234} = \frac{SS_{reg}(1,2,3,4) - SS_{reg}(2,3,4)}{MS_{reg}(1,2,3,4)}$$

This means we are calculating the F statistic of attribute 1 in the presence of 2,3,4 .this formula takes in account the SS_{reg} or the variability explained by(1,2,3,4) and variability shown when attribute 1 is not in the model.The calculation of partial F statistic is done in the same manner for the rest of the regressors.

2.4.2.3 Stepwise Selection

Stepwise regression method starts off by selecting an equation which contains the single best predictor and then builds up the addition of subsequent predictors one at a time as long as they are significant.Significance is calculated by the F statistic.The value of the highest partial F statistic value obtained is compared with F to enter value.After the addition of any predictor the equation is examined so that some other variable need to be deleted.This method is considered better than the forward and the backward regression methods as this also check the significance of the predictor entered in the model after it has been added unlike the previous methods.The procedure of the stepwise regression proceeds as follows:-

1. Initially we consider that there is no regressor in the model.

2.All the possible models containing only one variable is considered and the F statistic value is computed for it.We add the predictor that has the highest F statistic value.

3.Partial F statistic value are calculated for the regressors left in the presence of previously chosen regressors and the one that yields the highest value of the F statistic is added to the model if $F < F_{\alpha,1,resdf}$

4. Partial F statistic is used to evaluate all the variables in the model to check if one is still significant and at this step any regressor that is no longer significant is removed from the model.

5. This method terminates when no other regressor variable yields a partial F value that is greater than the threshold value that is choosen and all the regressors that are present in the model are all significant.

2.5 Conclusion

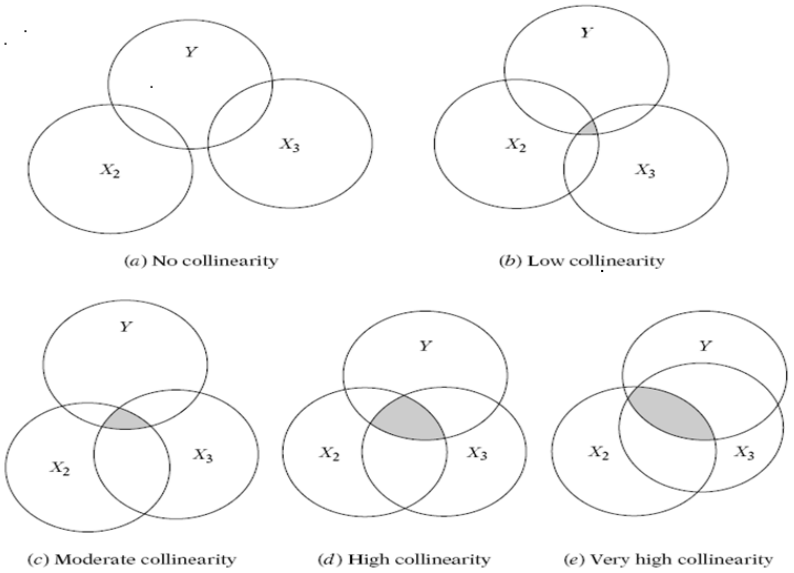
Regression is a approach to predict the response variable by using a number of predictors.Various methods are employed for selection of the best predictors that would yield the highest accuracy.the primitive methods such as forward selection,backward elimination and stepwise selection builds a model that has inherent multicollinearity.All these techniques calculate the F statistic to calculate the significance of the predictor.Therefore the line of research is to make a model that takes multicollinearity into account as this leads to a weak model with redundancy of information shown by the predicotrs.

Chapter 3

Multicollinearity

Multicollinearity is said to occur in a dataset when there exists a strong linear relationship between the variables in the model or in other words we can say that the variables have strong coorelation .Another implication of this problem is that $X^T X$ is singular and hence inverse of it cannot be performed.

FIGURE 3.1: Types of Multicollinearity



Some simple methods have also been proposed such as collection of additional data and remoaval of the variables that have high VIF.Collecting additional data might be difficult for various experiments such as the one involving complex experiments which may require performing the same experiment again in the same conducive environment. In pure statistical mathematicical basis, results are not biased by multicollinearity, but multicollinearity can multiply (by order of magnitude)if there are other problems which could introduce bias.The multicollinearity is studied by the correlation matix here[22].It is analyzed that one of the effects of multicollinearity is the shifting of signs (from positive to negative or from negative to positive) in the specific model built.The following consequences were analyzed after studying the road accident dataset [23].

⇒ Large covariances and variances of the estimatorsinvolved in the model.

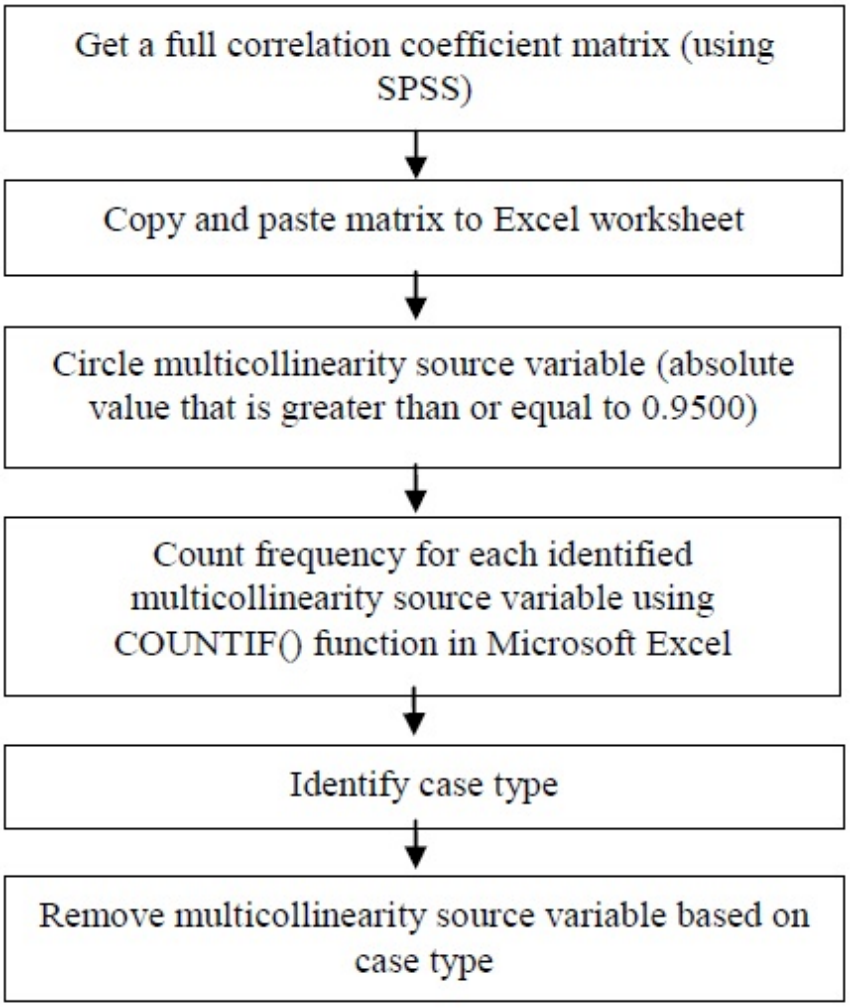
⇒ Low estimation precision Due to high standard error of estimates Low estimation precision

⇒ Acceptance of the null hypothesis due to wider confidence interval.

Multicollinearity has been given due focus and researches has been done for its elimination and estimation. A technique has been adopted by . Zainodin et.al for removal of multicollinearity which is not manual and can be explained as follows by a flowchart as explained in Figure 3.2

There are some instances where results obtained from multiple regression results may seem paradoxical. For a very low overall P , P values are high for all of the individual variables . This conveys that the model fits the data well, even though statistically none of the X variables have a significant impact on predicting the value of Y. This is due to the high correlation between the independent variables. In this case, neither of them may contribute significantly to the model after the other one is included. But together they may contribute a lot hence significant for model. The fit would be much worse if we remove both variables from the model.thus best

FIGURE 3.2: Technique that is proposed by Zainodin et.al [24]



solution is to remove multicolinearity.Consequences of multicollinearity as shown by research is [25]

1.When we consider a two-variable model where multicollinearity is present, we may observe that estimated standard error for the coefficients will be large. This occurs because there is a multiplying factor in the form of l/(l-r2) in the coefficient variance formula, where r is the correlation coefficient between two variables, and its value lies in the range -1 and +1. This factor is called variance inflation factor.When r = 0, there is no effects of multicollinearity is observed, and the inflation factor equals

to 1. As the value of r increases in absolute terms, the variances for the estimated coefficients have an increase in their magnitude. As r tends to +1, we can observe that the variance inflation factor approaches infinity.

2. Even if some little changes are made to the data the estimated coefficients may become sensitive because of correlation among the attributes.

3 It is difficult for user to access relative performance of variables when the estimated coefficients have large standard errors and are unstable.

4. Sometimes the researcher might not include a variable in the moddule to low t statistics.

It has been observed that high degree of multicollinearity exists in student evaluation of teaching (SET) which covers different dimensions of teaching[26].It is seen that the VIF for all the attributes is greater than 10 so they are highly correlated. The impact of the instructional attributes on TEVAL score can be easily quantified by regresssion analysis.A large body of literature recognises that linear regression is not appropiate when we observe that the dependent variable is categorical, or if it is qualitative.In this approach as adopted by the paper we give numbers to each category of the data and give them integer value respectively.

The main aim of this research is its use of individual responses of student and departure from the aggregative type of analysis which relies on class averages. Only for one thing, a disaggregated analysis involving individual data can capture the underlying heterogeneity that exists in a group of respondents while analysis that is based on class averages masks it. Ordered probit analysis to a subset of the data (second-level undergraduate courses) shows that an instructor may need to lay importance on improving the explanation,presentation and organization of the lecture materials.

3.1 Problems or Effects

The multicollinearity in the dataset implies that there is a strong linear relationship between the attributes.We need not include such variables in the model as this would introduce the redundancy of information given by the variable[27].The model proposed in the thesis makes effort to do away with this as this would sometimes also alter the accuracy of the model and may also allow introduction of more number of predictors in the model with more than one conveying the same information.If a predictor has a value $'x'$ and the other predictor has a value $2x + 1$ thay are said to be correlated and exhibit multicollinearity as if we know the value of one variable we would definately know the value of the second thus it poses an obstacle when we are building a robust time series prediction model[28, 29].

1.Strong multicollinearity betweeen the predictors accounts for the large value of the varience and covariance of the regression equation coefficient. This can be avoided by applying some techniques such as scaling and centering.Strong correlation constant is observed when there is strong multicollinearity.

2.Multicollinearity produces the estimators of the least square $\widehat{\beta}$ with values that are too far from the true value that should be there.

3. Model coefficients where a positive sign is expected is ornamented with a negative sign and vice versa.This ia a threat because it makes the regre4ssion equation prone to errors and disturbances.

4.High significance is shown by a predictor in the global test but it in the partial f test none of them are significant. this is so because more than one regressor divulge the same information.

5.Application of different model selection procedure give different results. This means that the resulst obtained by backward elimination,forward selection and some other method would be different.

3.2 Techniques for detection

(i) Examination of correlation matrix : Inspect the diagonal elements of the correlation matrix $X^{'}X$.If $|r_{ij}| > 0.9$ where r_{ij} is the correlation among the attributes i and j.This can be done my various tools such as minitab or can be programmed in matlab.It has been observed that the examining the correlation matrix is not helpfull in detecting the linear independence that exists between more than two regressor.

(ii) Eigen system analysis of $(X^{T}X)$**:** Multicollinearity can also be seen to exist from the eigen values of the correlation matrix $X^{'}X$.For a model having $(k-1)$ regressors there will exist $(k-1)$ eigen values $\lambda_1, \lambda_2, \lambda_3.....\lambda_{k-1}$.If there exists linear dependencies in the data then it will be observed than one or more of the eigen values will be significantly small. We define the Condition number of the correlation matrix $X^{'}X$ as k which is given as

$$k = \frac{Maximum\ eigen\ value}{Minimum\ eigen\ value}$$

It is considered as a thumb of rule that the following information is conveyed by the k value called condition index

k$\leq$ 10 multicollinearity is not a serious problem

$100 \leq k \leq 1000$ moderate to strong multicollinearity

$k \geq 1000$ multicollinearity is a serious problem We calculate the condition indices for the correlation matrix as

$$k = \frac{\lambda_{max}}{\lambda_{min}}$$

The largest value of the condition number is called the condition index.The eigen system analysis approach can be used to identify the nature of the linear dependencies that exist in the data.

(ii) Variance Inflation Factor: If R_i^2 is the coefficient of multiple determination where the explanatory variable X_i is regressed on remaining regressors.

$$VIF_i = \frac{1}{1 - R_i^2}$$

Large value of VIF_i indicates the presence of multicollenarity associated with a regressor X_i.In general it has been seen that

$VIF_i \geq 5$ indicates possibility of multicolliearity $VIF_i \geq 10$ indicates that multicolliearity is a severe problem

3.3 Dealing with Multicollineaity

It is important to do way with the problem of multicollinearity as it makes our model weak and prone to the disruption by the outliers that may come in due course especially when we are dealing with the time series data.

(i) Collect additional data : Apart from the regressors that are already in the model we may require more regressors which can predict the response effectively.This has been suggested as the best method in literature.Additional data can even add to the accuracy of the model.

(ii)Remove regressors from the model : If two regressors are having a linear relationship this means that both of them are conveying the same information.Thus it is better than we pick one variable out of the two.the main disadvantage of this model is that this may ruin the accuracy that is there in the model.

3.4 Conclusion

Multicolliearity has many adverse effects on the regression model.It has to be therefore eliminated.Transformation of the data has also considered as an effective approach to remove it.High correlation is observed among variables exhibiting multicollinearity.In the presence of this problem we are unable to build a robust model.

Chapter 4

Robust Best Regressor Model with Mininum Multicollinearity

Ragnar Frisch introduced the word "Multicollinearity" in the year 1934.It refers to the strong relationship between the predictors in the model i.e. the regressors are highly correlated.Numerous different machine learning techniques can be used with regression such as neural networks and support vector machine [16, 30] .Back-propogation and Feed forward network are used to predict the stocks efficiently.An approach employing Neural Network and Support vector regression increased the accuracy by adding values in dataset but it was observed that the dimensionality increased to five time [9].This posed an overhead of storing such huge amount of information.When the demand model is not linear ,non linear models are used, but the main aim of all the models constructed is high F value or significance and does not take into account the multi collinearity [12, 20, 31].A possible affect of it can be shifting of sign from positive to negative and vice versa in the model[11, 23].Combination of nearest neighbour and multiple linear regression is used where even the use of the regularization methods such as the ridge method , lasso method and their relatives, cannot improve the prediction accuracy much.The analysis has been done on the auction of cars and it has been seen that there is increase in the prediction

accuracy[32].Eliminating the variable with VIF greater than a threshold value in an iterative way may be used. [33].

The main problem associated with this approach was that it built a model with less multicollinear variables but the accuracy was low. A good model is which has high accuracy in addition to solving multicollinearity..It has been suggested that a)Dropping a variable. b)Transformation of variable. c)Additional or new data can help remove multicollinearity [25]. It is observed as a problem in variables measuring the student evaluation of teaching(SET) which considers the different teaching dimensions which are highly correlated [34]. While choosing the features from a high dimensional microarray dataset ,redundancy and similarity in attributes can lead to high dimensionality[18].The solution to multicollinearity has been proposed by Zainodin et al [24] which is based on observing the correlation matrix and deleting variables that have their correlation coefficient value greater than 0.95.This was done based on three cases based on the frequency of elements which had value greater than 0.95.All this was done by pasting the correlation matrix in Excel and then applying the cases.

4.1 Observations Leading to our Approach

It has been seen in the previous approaches that the main focus was on selecting a model which had high significance or F statistics.It was later observed that multicollinearity cannot be ignored and serious attention has to be taken as it may lead to various adverse effects in prediction of the model .So an important line of research is to develop a regression model in which the control variables or predictors are entered in the model based on low mean square error and least multicollinearity. In the proposed approach not only the final model obtained fulfills these two properties but taking any subset of the regressors obtained satisfy them.Building such a model will help to achieve both the objectives.This will build a model that

is free from variables that convey the same information and reduce the duplication of information conveyed by a specific variable.

4.2 Statistical Parameters Used

In order to calculate the value of 'stat' we considered the following statistical parameters which are as follows :-

4.2.1 's'-Square Root of Mean Square Error

SS_{res} which is the residual sum of square i.e. SS_{res}=$(y_i$- $\widehat{y_i})^2$ where y_i is the observed value and $\widehat{y_i}$ is the fitted value.It is not a good measure because it does not take into consideration the degree of freedom associated with the model.Due to which it has been observed that the value of SS_{res} (p) always decreases with the increase in value of 'p',where 'p' is the number of regressors considered in a given model. The proposed algorithm considers 's' i.e. square root of mean square error (MS_{res}(p))

$$MS_{res} = \frac{SS_{res}}{n-p}$$

which takes into consideration the degree of freedom and not necessarily decrease when 'p' increases. Increase in (MS_{res}(p)) occurs when the reduction in SS_{res} (p) for adding a regressor to the model is not sufficient to compensate for the loss of one degree of freedom.

4.2.2 PRESS(Predicted Residual Sums Of Squares) Residual(P)

Unlike the SS_{res} which uses the fitted value $\widehat{y_i}$, Press statistic uses sample predicted values $\widehat{y}$(i) and is calculated by the following formula

$$(PRESS)e_{(i)} = y_i - \widehat{y_{(i)}}$$

Where $\widehat{y_{(i)}}$ represents the fitted value of i^{th} response based on all observation except the i^{th} one. The value of $e_{(i)}$ is given as

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}$$

where h_{ii} is the diagonal element corresponding to row i of the Hat Matrix.

$$H = X(X^{'}X)^{-1}X^{'}$$

For a point if h_{ii} is large then it is a leverage point. PRESS statistic is calculated using the observation which are not present in the model hence it is considered as a good measure to prevent the overfitting of the model designed. A smaller PRESS statistic value interprets a good model and hence a high value of R-sqr predicted.

4.2.3 Absolute Value Of Difference Of Fits(|DFFIT|)

For the proposed model we have considered the absolute value of the summation of the DFFIT obtained for the given regressor set. It basically investigates the deletion influence of the ith observation on the fitted values. Values of |DFFIT| larger than $2 * (k/n)^{\frac{1}{2}}$ represents a highly influential observation .

$$DFFIT_i = \frac{(\widehat{y_i} - \widehat{y_{(i)}})}{\sqrt{MS_{res(i)} * h_{ii}}}$$

Where $MS_{res(i)}$ is the value of the mean square that is obtained when i^{th} observation is not present in the model.

4.2.4 Cook's Statistics (D_i)

It is the difference between values of the predicted response when all the regressors are present and the predicted response when ith observation is not present in the model.

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{y_j} - \widehat{y_{j(i)}})}{p * MS_{res}}$$

If the value of D_i is much larger than the others it is an indication of influential observation. For this model we have taken the summation of the D_i values obtained for each observation when a specific regressor set is chosen.

4.2.5 Variance Inflation Factor(VIF)

The Variance Inflation Factor (VIF) is the reciprocal of tolerance and is given by the formula $1/(1 - R_i^2)$.Commonly a VIF of 10 is used as an critical threshold indicator of multicollinearity.When a certain model reaches this value the researchers think of reducing the multicollinearity so that it does not effect the results of the regression analysis as though a model has acceptable F value does not mean that the model is statistical significant[35].Among other tests for detection of multicollinearity are examination of correlation matrix and condition number test.

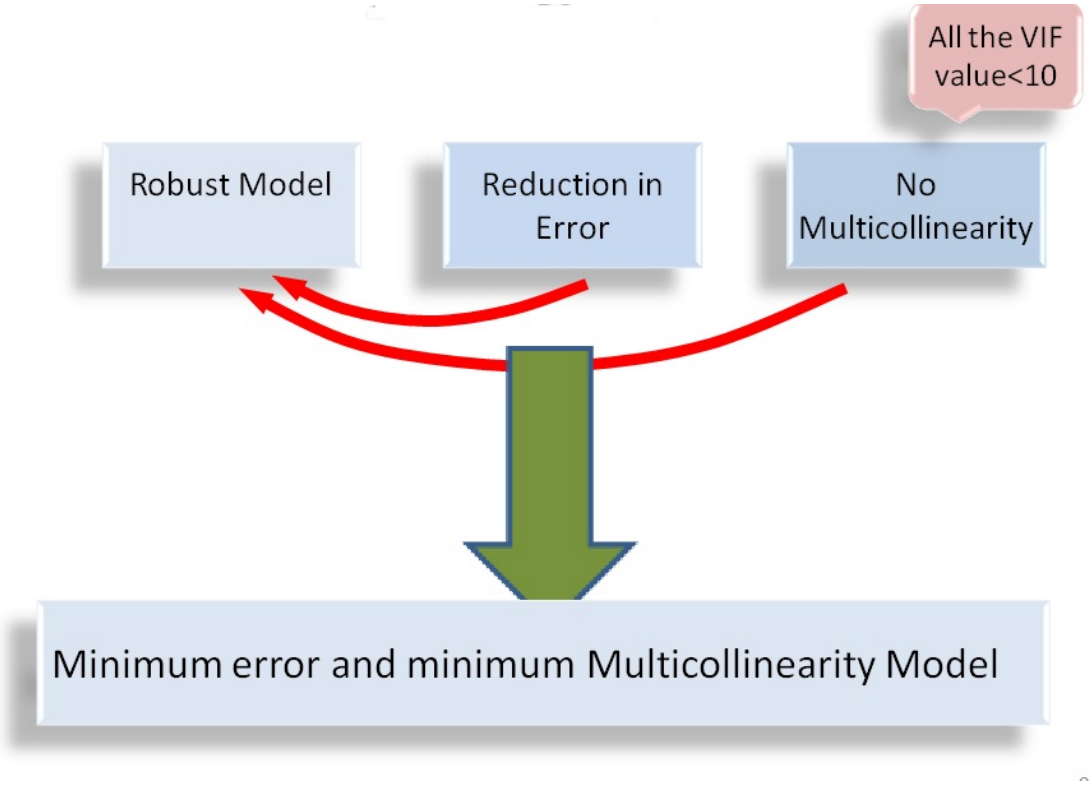


FIGURE 4.1: Reasons for selecting Square Root of Mean Square Error,PRESS residual,VIF and $|DFFIT|$

4.3 Reason for Selecting Above Statistical Parameters

These parameters are chosen to get a model that is robust and have least multicollinearity i.e. there is least correlation between the various predictors involved in the model maintaining the R-sqr prediction as high as possible and nearly equal to that predicted by primitive models such as stepwise selection method, backward elimination and forward selection. The former models just laid stress on the global f-test without taking into account the multicollinearity in which the model would be a victim of the above referred problems..The proposed model will have all predictor's variance inflation factor less than 10. VIF greater than 10 is not desirable as it adversely affects the regression result hence leading in overfitting in the regression

analysis [23].The aim of the model proposed is " less error less multicollinearity less overfitting ".

4.4 Increasing Accuracy and Decreasing Multicolliearity

A dataset having n instances
MIN_MULT_COL ($stat_i$,VIF_i)
1)No regressor in the model.
2) All possible models with one regressor is considered. The regressor with least 'stat' value is added to the model.
3) Stat' value is computed for all the remaining regressor in the presence of regressors selected previously and are arranged in ascending order if they satisfy Eq.1.

$$VIF_{new} < 10 \qquad Eq.1$$

4) Starting from the model having least VIF_i a regressor is added only if

$$stat_i < stat_{i-1} \qquad Eq.2$$

5) If two regressor have the same value of VIF ,the one with minimum 'stat' value and satisfying Eg.2 is added to the model.
6) Step 3 and Step 4 are performed iteratively. **7)** The algorithm terminates when after arranging all the VIF in ascending order.
(i) No regressor can be added.
(ii) All models do not satisfy Eq. 1 and Eq.2.
Output:A best regressor model that has least multicollinearity

FIGURE 4.2: Proposed algorithm

4.4.1 Data Collection

We crawled the UCI Machine learning repository,created as a ftp archive in 1987 by David Aha to collect the datasets that were required for the study [36].This repository contains a huge collection of data generators,databases and domain theories

TABLE 4.1: Meaning of symbols used in the algorithm

Symbol	Meaning
P	$\sum_{i=1}^{n} Press\ residual$
s	$\sum_{i=1}^{n} Mean\ square\ error$
D	$\sum_{i=1}^{n} Cook's\ statistic$
f	$\sum_{i=1}^{n}$ |DFFIT|
VIF_{new}	VIF of new regressor added
stat	P+s+D+f
$stat_i$	stat value for 'i' regressors

which are useful for empirical analysis of machine learning algorithms. We used the following publicly available datasets of different sizes:-

Energy efficiency Data Set : This multivariate dataset having integer and real type of attributes has 768 samples and 8 features, aims to predict two real valued responses [8].

Parkinsons Telemonitoring Data Set : This multivariate dataset having integer and real attributes has 5875 instances and total 16 biomedical voice measures with two output variable [37].

Hald Cement Data Set : This dataset helps to predict heat(cal/grm) as function of various chemical components.It has 13 instances, 4 predictors and one output variable [15].

Iris Data Set : This is a popular dataset that is commonly used in pattern recognition literature.For considering this multivariate dataset for regression we have considered the classes Iris Setosa,Iris Versicolour,Iris Virginica as 1,2,3.It consists of real attributes with 150 instances and 4 attributes [38].

Longley Data Set : It consists of 1 response variable,6 explanatory variables and has 16 Observations.A macroeconomic data set which provides a well-known example for a highly collinear regression.This dataset consists of 7 predictors and 1 response variable , observed yearly from 1947 to 1962 and has 16 instances [39].

STEP 1

		Variance inflation factor	average of vif	S value	R-Square(%)	R-Sq(adj)(%)	R-Sq(pred)	press residual	COOK'S STATISTICS	DIFFT	absolute ofDIFFT
WITH ONE REGRESSOR	STATS										
1	669285	1	1	10.6717	0.6	0.5	0.49	669270	0.944341	-2.97007	2.97007
2	669917	1	1	10.6772	0.4	0.4	0.39	669902	0.799866	-3.19176	3.19176
3	670224	1	1	10.6792	0.4	0.4	0.35	670210	0.917079	-2.31655	2.31655
4	670231	1	1	10.6797	0.4	0.4	0.35	670217	0.806333	-2.52278	2.52278
5	670223	1	1	10.6792	0.4	0.4	0.35	670209	0.917087	-2.31751	2.31751
6	667227	1	1	10.6557	0.8	0.8	0.79	667212	0.819289	-3.2183	3.2183
7	666373	1	1	10.6488	1	1	0.92	666358	0.820671	-3.1871	3.1871
8	668718	1	1	10.6674	0.6	0.6	0.57	668704	0.857449	-2.86829	2.86829
9	668243	1	1	10.6639	0.7	0.7	0.64	668229	0.801823	-2.96297	2.96297
10	663129	1	1	10.6228	1.5	1.4	1.4	663114	0.853038	-3.5887	3.5887
11	668718	1	1	10.6674	0.6	0.6	0.57	668704	0.857457	-2.86837	2.86837
12	670387	1	1	10.6813	0.4	0.4	0.32	670374	0.707336	-2.08169	2.08169
13	655299	1	1	10.5596	2.6	2.6	2.57	655282	0.915016	-5.5808	5.5808
14	656443	1	1	10.5687	2.5	2.4	2.4	656429	0.97383	-2.02923	2.02923
15	664342	1	1	10.6321	1.3	1.3	1.22	664329	0.972558	-1.01839	1.01839
16	656557	1	1	10.5699	2.4	2.4	2.38	656543	0.898265	-2.06941	2.06941

STEP 2

		Variance inflation factor	average of vif	S value	R-Square(%)	R-Sq(adj)(%)	R-Sq(pred)	press residual	COOK'S STATISTICS	DIFFT	absolute ofDIFFT
with two regressors											0
13,1	653972	1.838	1.838	10.5482	2.9	2.8	2.76	653954	0.923358	-6.2114	6.2114
13,2	652384	1.996	1.996	10.536	3.1	3	3	652367	0.803076	-5.90161	5.90161
13,3	653700	1.699	1.699	10.5458	2.9	2.9	2.81	653683	0.959444	-5.45503	5.45503
13,4	653185	1.782	1.782	10.5418	3	2.9	2.88	653168	0.933064	-5.58523	5.58523
13,5	653701	1.699	1.699	10.5458	2.9	2.9	2.81	653684	0.959391	-5.4559	5.4559
13,6	652776	2.795	2.795	10.5389	3	3	2.94	652759	0.854429	-5.84805	5.84805
13,7	653617	2.809	2.809	10.5456	2.9	2.9	2.82	653600	0.866542	-6.08543	6.08543
13,8	651633	2.561	2.561	10.5295	3.2	3.2	3.11	651617	0.873264	-4.97062	4.97062
13,9	651849	2.667	2.667	10.5316	3.2	3.1	3.08	651832	0.818237	-5.3147	5.3147
13,10	655424	2.533	2.533	10.5601	2.6	2.6	2.55	655406	0.867611	-6.60455	6.60455
13,11	651633	2.561	2.561	10.5296	3.2	3.2	3.11	651617	0.873261	-4.97052	4.97052
13,12	652235	1.881	1.881	10.535	3.1	3.1	3.02	652219	0.763548	-4.28476	4.28476
13,14	652566	1.768	1.768	10.5365	3.1	3	2.97	652550	0.98381	-4.53647	4.53647
13,15	636567	1.092	1.092	10.4067	5.4	5.4	5.35	636548	0.943587	-8.09955	8.09955
13,16	653733	2.357	2.357	10.5464	2.9	2.9	2.8	653715	0.881005	-6.67983	6.67983

STEP 3

		Variance inflation factor	average of vif	S value	R-Square(%)	R-Sq(adj)(%)	R-Sq(pred)	press residual	COOK'S STATISTICS	DIFFT	absolute ofDIFFT
with three regressors											
13,15,1	635703		3.342	10.3985	5.6	5.6	5.48	635683	0.993251	-8.89622	8.89622
13,15,2	636044		3.67	10.4021	5.5	5.5	5.43	636024	0.877579	-8.40153	8.40153
13,15,3	635433		3.1355	10.3961	5.7	5.6	5.52	635414	1.02328	-8.06332	8.06332
13,15,4	634189		3.273	10.3859	5.8	5.8	5.71	634169	1.03612	-8.2772	8.2772
13,15,5	635433		3.1355	10.3961	5.7	5.6	5.52	635414	1.02324	-8.06441	8.06441
13,15,6	630767		4.9945	10.3588	6.3	6.3	6.22	630748	0.885674	-7.65766	7.65766
13,15,7	631815		5.0415	10.3673	6.2	6.1	6.06	631796	0.89846	-7.74265	7.74265
13,15,8	629476		4.6005	10.3481	6.5	6.5	6.41	629458	0.90651	-6.896	6.896
13,15,9	629452		4.8105	10.3482	6.5	6.5	6.41	629434	0.856743	-7.17169	7.17169
13,15,10	636451		4.444	10.4051	5.5	5.4	5.37	636431	0.930501	-8.42464	8.42464
13,15,11	629476		4.6005	10.3481	6.5	6.5	6.41	629458	0.906507	-6.89589	6.89589
13,15,12	625569		3.839	10.3166	7.1	7	6.9	625552	0.813875	-5.54626	5.54626
13,15,14	633816		3.244	10.3831	5.9	5.8	5.76	633798	0.990473	-7.07585	7.07585
16	629757		4.327	10.3503	6.5	6.4	6.37	629736	0.920115	-9.31814	9.31814

TABLE 4.2: Illustrating the first three steps of the proposed approach on Parkinsons Telemonitoring Data Set [37]

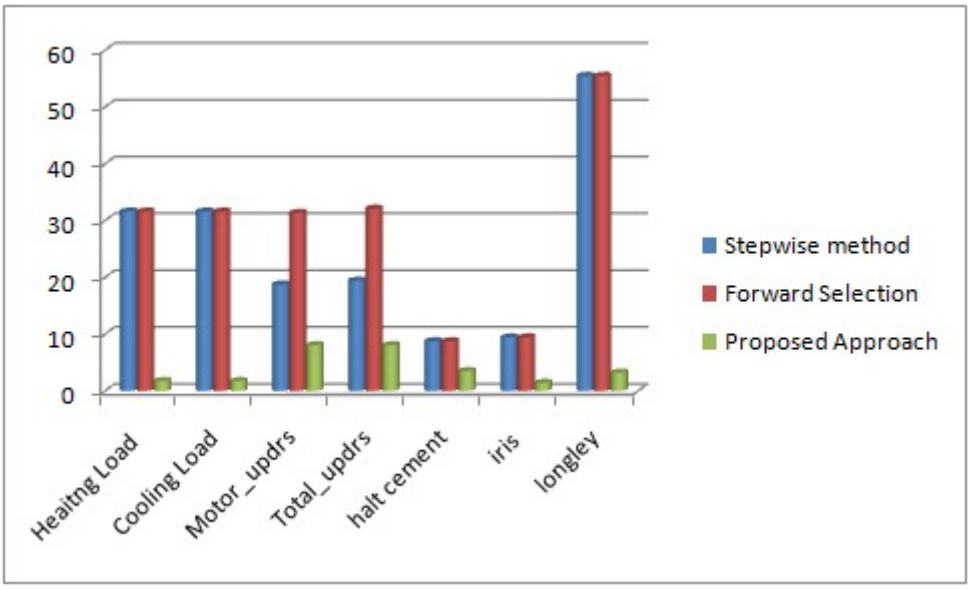


FIGURE 4.3: Comparison of Proposed method with other methods

4.5 Experimental Results

TABLE 4.3: Comparison with other methods

No.	Dataset	Prediction Attribute	Stepwise selection		Forward Selection		Proposed Algo	
			R-Sqrd	C.I	R-Sqrd	C.I	R-Sqrd	C.I
1	Energy Efficiency	Heating Load(Y1)	91.6	31.62	91.6	31.62	91.3	6.14
		Cooling Load(Y2)	88.8	31.62	88.8	31.62	88.4	6.14
2	Parkinsons Telemonitoring	Motor_UPDRS	10.6	18.79	10.7	31.37	10.3	8.10
		Total_UPDRS	10.0	19.45	10.0	32.12	9.9	8.10
3	Hald Cement	Heat in cement (cal/grm)	98.2	8.78	98.2	8.78	98.1	3.55
4	Iris	Class of iris plant	93	9.46	93	9.46	92.2	1.45
5	Longley	Total derived Employment	98.1	55.495	98.1	55.495	98.5	3.25

TABLE 4.4: Comparison with some basic solution proposed considering Energy Efficiency dataset [8]

Research	Attri-butes	R-sqrd	Max VIF	Condition Index
Zainodin et al.(2013)	Y1	91.2	18.447	8.6026
	Y2	88.2	21.675	10.35
Gujarati and Porter (2009)	Y1	91.5	211.938	10.35
	Y2	88.8	211.938	31.62
Proposed Approach	Y1	91.3	9.626	6.14
	Y2	88.4	9.626	6.14

Fig 4.3 presents the performance results on the above mentioned datasets.The green line show the forward regression,blue line the stepwise regression and red the result

obtained by our proposed solution. It can be visualized from the figure that the Proposed algorithm always builds a model with VIF$\leq$ 10 for all attributes in the model.From Fig 4.4also shows the comparison of results on Energy Efficiency dataset and it shows that the proposed approach keeps the multicollinearity minimum.

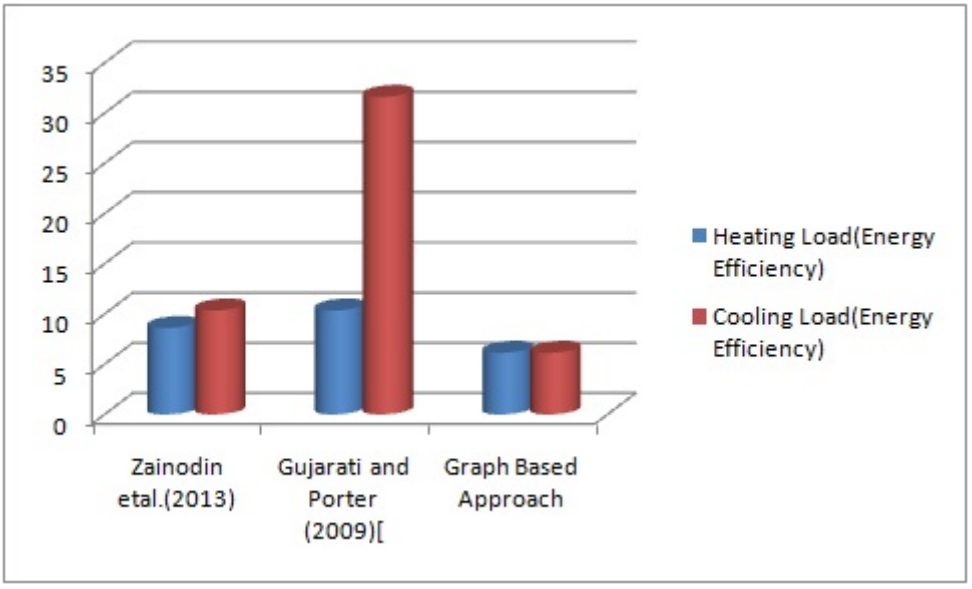


FIGURE 4.4: Comparison of Proposed method on Energy Efficiency Dataset with State of Art methods

When backward elimination is applied on datasets having high multicollinearity such as the Energy Efficiency and Parkinsons Telemonitoring through software such as minitab or matlab it generates a error and is not successful in building a model because of multicollinearity.Coefficient calculation is done by using matrix inversion. If singularity exists in the matrix, the inversion process is impossible and if multicollinearity exists the inversion obtained is unstable.It displays error as the X matrix is nearly singular.

4.5.0.1 Attribute Selection and Proposed Approach

We compared our method with some well known methods and the comparison results are shown in Table 4.3.The values of the $R - sqrd$ and the Condition Index(CI) is

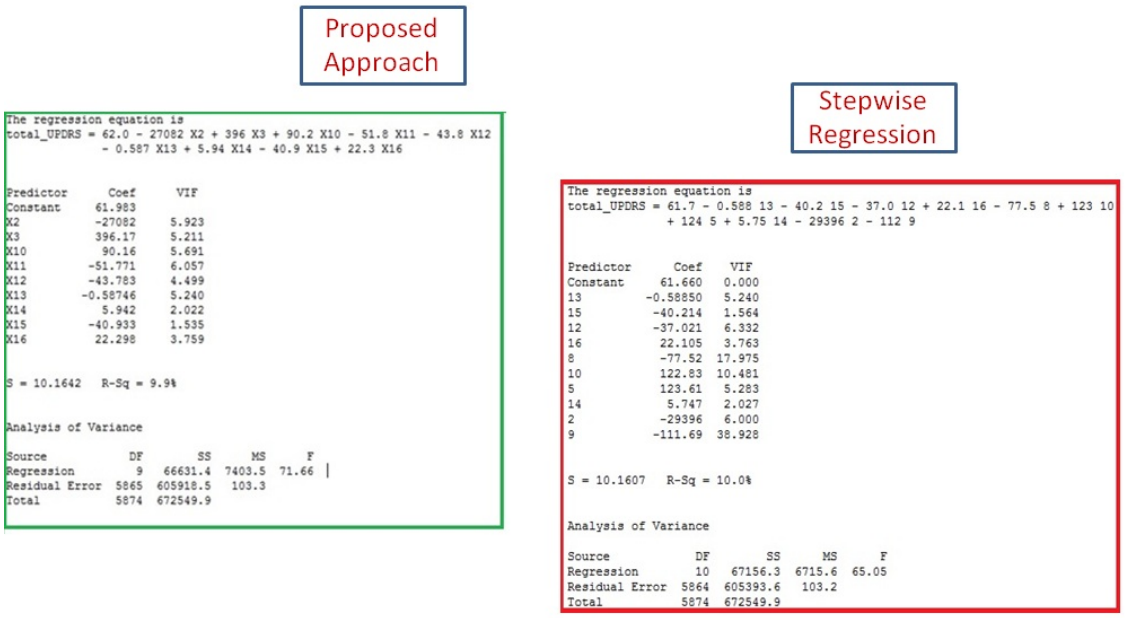


FIGURE 4.5: The results obtained in minitab for the Parkinsons Telemonitoring Data Set [37]

tabulated.CI is obtained from the eigen value of the attribute matrix.

$$k = \frac{Maximum\ Eigen\ Value}{Minimum\ Eigen\ Value}$$

and there is moderate to strong multicollinearity if it is between 100 to 1000.CI is given as the $\sqrt{k}$.CI value between 10 and 30 indicate moderate to strong multicollinearity whereas if its value is greater than 30 then severe multicollinearity is indicated and should be taken care of.Fig 4.5 shows the output by minitab of Parkinsons Telemonitoring Data Set [37].It can be seen that the VIF of 3 predictors are greater than 10 indicating multicollinearity but the proposed approach does away with it,keeping all VIF less than 10. From the Table 4.3 it is clear that the proposed model always achieves a model with $CI \leq 10$.It has also been observed that for models where the CI is already less than 10 the proposed algorithm tries to reduce the CI more as seen in Table 4.3 as seen in Hald cement dataset which reduces the value 8.78 to 3.55 of CI.

4.5.0.2 Comparison of Proposed Technique with Some Basic Methods for Reducing Multicollinearity

Table 4.4shows the comparison of the proposed approach with other models of the same kind.Gujarati and Porter [40] suggested that removing one of the variable with high multicollinearity may build a good model.Zainodin et al. (2013) [24] proposed that good model could be built considering the correlation matrix.Three cases based on the output of the matrix is used to remove the variables .It is clear from the comparison that the proposed approach builds a model that has the best accuracy and lowest VIF value which is considered as an indicator for multicollinearity.

4.6 Conclusion

We proposed a methodology for building a best regressor model whereby our main aim was to add predictors in the model based on low values of both multicollinearity and mean square error and build a model that is better than the earlier models that focused only on the significance.The accuracy is obtained does not differ much from the earlier models.Our method uses VIF value greater than 10 as the technical indicator of multicollinearity.Experimental studies show that it builds a model with less multicollinearity over some benchmark corpora.This work can be extended to include all regressors involved in the model rather than some of them so as to increase accuracy.

Chapter 5

Graph Based Approach for Minimum Multicollinearity Highly Accurate Regression Model explaining Maximum Variability in Model

Machine Learning deals with making the computer system learn and improve its performance based on a set of data that is given. The main research that surrounds this subject is the ability of the computer system to recognize certain complex patterns or data values and predict or forecast events. A simple example would be a computer system that recognizes the handwritten letters of english. The machine learns from a set of examples .In literature supervised machine learning has been termed as synonym of classification. When we point to various fruits time and again and tell a child that this is apple, mango etc. The next time after a certain period when he sees the same he would recognize it. This is the similar case with supervised machine learning where the machine is trained with a set of labelled objects .Our

main aim would be the prediction based on the previous events that were used to train or make the machine learn. SVM, decision trees, neural networks, naive bayes and regression are the common ones. In this approach we will use regression as a technique for prediction.

Regression is a statistical technique which is used to study the relationship between and dependent attribute (such as the burnt area of forest fire) and one or more independent attribute (such as the humidity, rain, speed of wind etc).It has found its applications economics, social science and engineering in predicting stock market, forest fire prediction, heating and cooling load prediction of a room and variety of other ample applications. Regression makes an equation with certain selected attributes having high significance that explain approximately total variability in the model to predict the class attribute. Calculating the significance i.e. which attributes to consider having high predictive power of the model have also attracted researchers Least square method is used to determine the coefficient used in the regression equation. Thus the main objective of this technique is to build a robust model with high predictive power. The robustness plays an important role especially when we have a time series data which may sometimes contain infrequent or usual occurrences.

Regression has been used in a variety of applications ranging from detecting natural calamities, DNA binding techniques, stock market prediction,predicting the highly stochastic stock exchange and climate modelling [41, 42]. This technique has also being applied with other machine learning techniques such as SVM or PCA [43]. Reducing the linear relationship between variables has also attracted researchers .Thus developing a model that works with high multicollinearity has also been proposed which works in forward stepwise manner and takes in account the Mallow statistic Cp. Researchers have proposed an approach whereby by studying the correlation among the attributes the predictors are removed from the model according to cases which are based on the variables among which there exists strong correlation.

5.1 Observations Leading to Our Approach

Unlike the previous approaches of model selection where main emphasis was given on the significance of the regressor or the independent variable when they were included in the model the proposed approach has the main aim of building a robust model. Efforts have been made to keep high accuracy and keep minimum multicollinearity to achieve this objective [44]. Former models calculated F statistics at every step when a new regressor was included considering only significance of that specific variable with respect to previously selected one .This lacked the quick understanding and did not take multicollinearity into consideration.

The beauty of the proposed algorithm lies in the way that it graphically shows the attributes which should be included in the model by modeling it with the help of a graph which can be easily understood and if some change or the value of the parameter is required it can be judged .This approach is also good because it shows clearly why we are including the approach and what we are doing in each subsequent step.

The parameters are scaled between 0-10 so that one parameter does not outweigh the other. If we did not scale the parameter then if one of the statistical parameter would have a greater value that would play the deciding factor.If for e.g the PRESS residual had a large value though its R-sqrd(variability explained in the model) would be high,the model would not include that regressor in the model.But the graph based approach that is proposed takes this into account and thus scales the parameters.This achieves three fold benefits of model having less regressor, high accuracy and minimum multicollinearty .

5.2 Statistical Parameters Used

In order to calculate the value of 'stat' we considered the following statistical parameters which are as follows :-

5.2.1 R-Sqrd

It determines the total variability that is reflected by the regression model that is built. So a higher value of R-sqrd is desirable to get a good predictive model. It is the ratio of the variability explained by the model SS_{reg} to the total variability in the model SS_total.

$$R - sqrd = \frac{SS_{reg}}{SS_total}$$

or

$$R - sqrd = 1 - \frac{SS_{res}}{SS_total}$$

It can be visualized from the equation that to achieve R-sqrd 100 percent either the residual sum of square $SS_{res} = 0$ or the regression equation should explain the full variability i.e. $SS_{reg} = SS_{total}$ .In the proposed approach we minimize $(1 - Rsqrd)$ where we have scaled R-sqrd in range 0 to 10.

5.2.2 Variance Inflation Factor

This is used to explain the extent to which multicollinearity is present in the model. A VIF greater than 10 denotes a model in which multicollinearity is a serious concern and has to be taken care of. It is high when there exists high correlation among the attributes or R_i^2 (coefficient of multiple determination when xi is regressed on the remaining regressors) is high.

$$VIF = \frac{1}{1 - R_i^2}$$

combination of regressors	Calc	Rsquared	S value	Variance inflation factor	PRESS	1- r sqrd	vif scaled to 10	scaled press residual
12	73.71132	49.1	7.20919	61.991	40021.3	0.509	6.1991	4.00213
13	8.47788	74.1	5.13979	1.043	20360.9	0.259	0.1043	2.03609
14	10.20896	80.8	4.42675	4.079	15112.1	0.192	0.4079	1.51121
15	8.80115	83.2	4.13708	3.176	13200.7	0.168	0.3176	1.32007
16	14.33947	38.7	7.9089	1	48175.7	0.613	0.1	4.81757
17	13.20846	46	7.42419	1	42442.7	0.54	0.1	4.24427
18	14.22289	39.5	7.85953	1	47583.6	0.605	0.1	4.75836
21	73.71132	49.1	7.20919	61.991	40021.3	0.509	6.1991	4.00213
23	7.56876	78.8	4.65043	1.04	16663.3	0.212	0.104	1.66633
24	10.98676	78.8	4.65043	4.458	16663.3	0.212	0.4458	1.66633
25	9.40342	83.3	4.12847	3.794	13139.5	0.167	0.3794	1.31395
26	13.62981	43.3	7.60693	1	44558.8	0.567	0.1	4.45588
27	12.47844	50.6	7.10164	1	38828	0.494	0.1	3.8828
28	13.5113	44.1	7.55558	1	43967.2	0.559	0.1	4.39672
31	8.47788	74.1	5.13979	1.043	20360.9	0.259	0.1043	2.03609
32	7.56876	78.8	4.65043	1.04	16663.3	0.212	0.104	1.66633
34	7.62176	78.8	4.65043	1.093	16663.3	0.212	0.1093	1.66633
35	6.6091	83.7	4.07839	1.086	12817.1	0.163	0.1086	1.28171
36	17.01525	20.8	8.99347	1	62297.8	0.792	0.1	6.22978
37	15.94663	28	8.57032	1	56563.1	0.72	0.1	5.65631
38	16.9051	21.5	8.95008	1	61700.2	0.785	0.1	6.17002
41	10.20896	80.8	4.42675	4.079	15112.1	0.192	0.4079	1.51121
42	10.98676	78.8	4.65043	4.458	16663.3	0.212	0.4458	1.66633
43	7.62176	78.8	4.65043	1.093	16663.3	0.212	0.1093	1.66633
45	24.91368	79.1	4.61593	18.443	16457.5	0.209	1.8443	1.64575
46	8.40628	74.3	5.12438	1	20249	0.257	0.1	2.0249
47	6.97511	81.6	4.33902	1	14520.9	0.184	0.1	1.45209
48	8.26398	75	5.04784	1	19661.4	0.25	0.1	1.96614
51	8.80115	83.2	4.13708	3.176	13200.7	0.168	0.3176	1.32007
52	9.40342	83.3	4.12847	3.794	13139.5	0.167	0.3794	1.31395

FIGURE 5.1: Preparation of table from which the $Calc_{(i,j)}$ values is calculated

5.2.3 PRESS(Predicted residual sum of square) Residual

It is used to study the deletion influence of an observation in the model. For an observation that has substantial difference between the regular residual and the PRESS residual denotes a influential observation.

$$(PRESS)e_{(i)} = y_i - \widehat{y_{(i)}}$$

where Y_i is the response value when all the observations are considered and $\widehat{y_{(i)}}$ is the response value when we do not take ith observation into account. Intuitively we can define

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}$$

where h_{ii} is the i^{th} diagonal element of the hat matrix

$$H = X(X^{'}X)^{-1}X^{'}$$

. This parameter also helps us to explain how well the model can predict the new data.

5.2.4 's'-Square root of mean square error

This statistical quantity takes into account the degree of freedom so in spite of taking the SS_{res} as a measure of error we have taken this quantity.

$$MS_{res} = \frac{SS_{res}}{n-p}$$

,where, n is the total instances and p is the number of attributes. SS_{res} always reduces when we increase the number of attributes but MS_{res} may not always reduce.

5.3 Analysis of Algorithm

The algorithm has scaled parameter so that one does not outweigh the other.All the parameters used have to be kept minimum due to the following reasons a)VIF greater than 10 denotes high multicollinearity so its value much be kept low. b)Root mean square error measures how far we are from the true value of the response so it should also be kept low. c)PRESS residual measures a influential observation which denotes an instance in dataset that has unusual X and Y coordinate.We dont have to take instances that have large value of it so as these observations are not included. d) R-squared has to be keep high so that maximum variability is shown by the model.So we need to minimize (1- R-sqrd)[45, 46].All these are the collection of parameters that would guarantee minimum multicollinearity,maximum accuracy,less attributes explaining the maximum variability in the model.The execution of the algorithm is explained in step by step manner in Fig 5.1 and the way how we calculate $Calc_{(i,j)}$ is shown in Fig 5.3

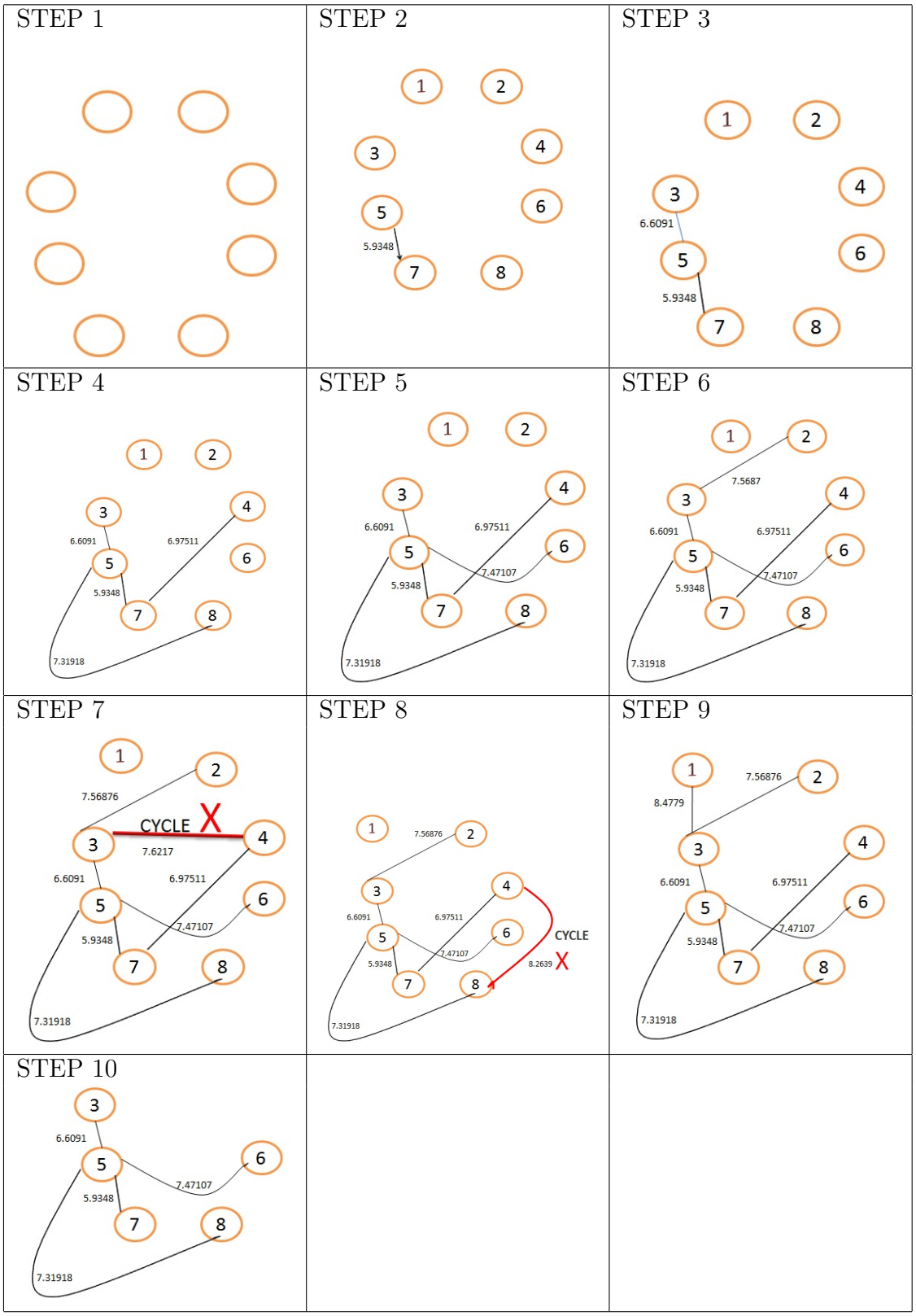


TABLE 5.1: Illustrating the steps of the graph based approach on Energy Efficiency dataset [8]

TABLE 5.2: Algorithm: Graph Based Approach for Minimum Multicollinearity Highly Accurate Regression Model explaining Maximum Variability in Model

Input: A dataset having n instances for which the best model has to be selected by using graph. $MIN\_MULT\_ACC(Calc_{(i,j)}G(V,E))$
Output: Best regressor model that has least multicollinearity
begin
1) Let $G(V,E)$ be the graph with edges 'E' equal to number of parameters
and no edges.
2) Give mnemonic name or number to each node marked.
3) Calculate $Calc_{(i,j)}$ when regressor i and j are in the model.
4) Select minimum $Calc_{(i,j)}$ and mark it as weight of edge (i,j)if no cycle is
formed.
5) Repeat step 4 until all nodes are connected
6) N_{high} ←node having maximum degree
7) Delete the node j from the edge $(N_{high}, j)$where VIF is greater than 10.
8) N_{high} and the nodes left after performing step 7 is the best regressor model.
9) If we have more than one N_{high} select the one that has max R- sqrd, step 7 and 8 already done.
7 and 8 already done.
end

The nodes with minimum of CALC value is joined as the edge between them denotes that the nodes it connect optimize all these functions mentioned above. When we build a model we take the node that has maximum degree as this is the node that has maximum accuracy and explains variability with the nodes connecting it.The nodes that are not directly connected ensure that they have high correlation between them.These variables are thus not included in the model .

5.4 Results and Discussions

It can be observed from Table 5.3 and Table 5.4 that the proposed approach in addition to having a comprehensive graph based approach has better results than the other approaches .the graph approach has better comprehensibility.All the datasets

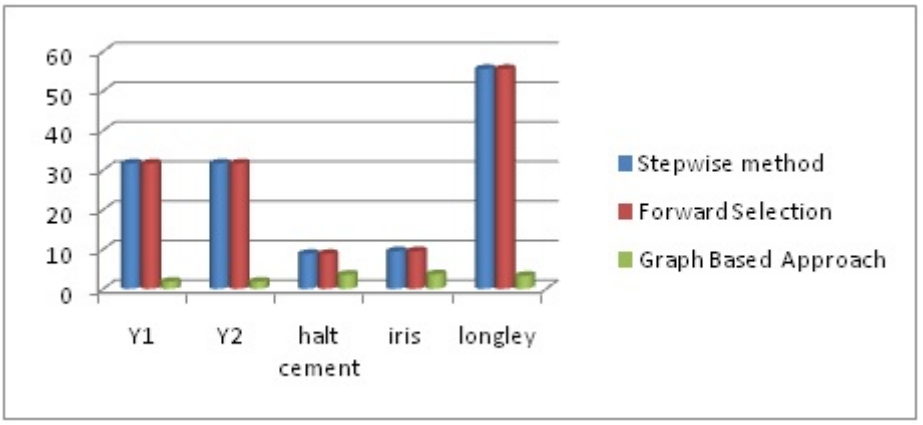


FIGURE 5.2: Comparison of Graph based approach with other methods

TABLE 5.3: Comparison of graph based method with other methods

No.	Dataset	Prediction Attribute	Stepwise selection		Forward Selection		Proposed Algo	
			R-Sqrd	C.I	R-Sqrd	C.I	R-Sqrd	C.I
1	Energy Efficiency	Heating Load(Y1)	91.6	31.62	91.6	31.62	91.3	1.78
		Cooling Load(Y2)	88.8	31.62	88.8	31.62	88.4	1.78
3	Hald Cement	Heat in cement (cal/grm)	98.2	8.78	98.2	8.78	98.1	3.55
4	Iris	Class of iris plant	93	9.46	93	9.46	92.2	3.67
5	Longley	Total derived Employment	98.1	55.495	98.1	55.495	98.5	3.25

TABLE 5.4: Comparison of graph based approach with some state of art method considering Energy Efficiency dataset [8]

Research	Attri-butes	R-sqrd	Max VIF	Condition Index
Zainodin et	Y1	91.2	18.447	8.6026
al.(2013)	Y2	88.2	21.675	10.35
Gujarati and	Y1	91.5	211.938	10.35
Porter (2009)	Y2	88.8	211.938	31.62
Proposed	Y1	91.3	9.626	1.78
Approach	Y2	88.4	9.626	1.78

on which the proposed approach has been implemented consists of varied number of attributes so that the results can be verified for datasets.These datasets are avaliable for research in the UCI repositry. Backward selection could not be performed on dataset such as Longley dataset and energy efficiency having high multicollinearity .This is due to the singularity that exists between the matrix which poses difficulty in matrix inversion. High multicollinearity leads to highly unstable inversion.

The results have been compared with some state of art methods in Fig 5.3 and some

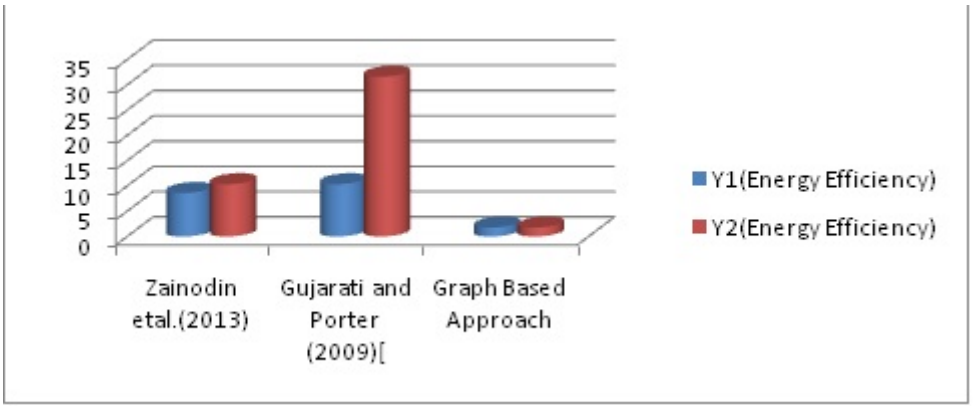


FIGURE 5.3: Comparison of Graph based approach on Energy Efficiency Dataset with State of Art methods

basic methods in Fig 5.2 where it can be seen that the proposed approach works appreciably better. It has a lower Condition index (CI) value.CI index is the ratio of the maximum Eigen value to the minimum Eigen value of the correlation matrix and is used to check multicollinearity.CI value between 100 and 1000 is considered to have moderate to strong multicollinearity. It can be seen that the proposed in addition to be a good graphical approach decreases the CI even if it is less than 10.

5.5 Conclusion

Regression is a supervised machine learning method which advances by building a equation by which we can estimate the response variable. The proper choice and number of predictors used for prediction of response variable plays an important role.Multicollinearity affects the model accuracy hence emphasis has to make to keep it as low as possible. The coefficients of the regression equation are calculated by $X^T X$.In case of strong correlation amongst attributes it is difficult for software to calculate this term. In addition to all this the approach has to be pictorial or graphical in nature for better comprehensibility which is unique. This approach keeps in mind all these factors and proposes a technique for building a

best (accurate, minimum multicollinearity, graphical) regression model.This chapter proposes a simplified and easily understandable approach which builds a best regression model. The results show that it performs better than basic methods and some state of art methods. The approach has scaled all the statistical parameters used so as to give equal weightage to all of them and to reach a collective goal i.e minimum multicollinearity,high accuracy,and maximum variability. This approach can be extended to applying some transformation or combining some other machine learning technique so as to show more variability in the model.

References

[1] P. Nguyen, "Techware: Speech recognition software and resources on the web [best of the web]," *Signal Processing Magazine, IEEE*, vol. 26, no. 3, pp. 102–105, May 2009.

[2] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control*. John Wiley & Sons, 2013.

[3] I. Bose and R. K. Mahapatra, "Business data mining—a machine learning perspective," *Information & management*, vol. 39, no. 3, pp. 211–225, 2001.

[4] H. Zhang, X. Han, and S. Dai, "Fire occurrence probability mapping of northeast china with binary logistic regression model," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 6, no. 1, pp. 121–127, Feb 2013.

[5] C. Li, "Social post-evaluation of world bank projects in yanhe basin based on ridge regression and support vector machines," in *Intelligent Systems and Applications (ISA), 2011 3rd International Workshop on*, May 2011, pp. 1–3.

[6] A. Haque, M. Nehrir, and P. Mandal, "A hybrid intelligent model for deterministic and quantile regression approach for probabilistic wind power forecasting," vol. PP, no. 99, 2014, pp. 1–10.

[7] S. Chatterjee and A. S. Hadi, *Regression analysis by example*. John Wiley & Sons, 2013.

[8] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings*, vol. 49, pp. 560–567, 2012.

[9] A. Widodo, M. Fanani, and I. Budi, "Enriching time series datasets using nonparametric kernel regression to improve forecasting accuracy," in *Advanced Computer Science and Information System (ICACSIS), 2011 International Conference on*, 2011, pp. 227–232.

[10] G. P. Zhang and M. Qi, "Neural network forecasting for seasonal and trend time series," *European journal of operational research*, vol. 160, no. 2, pp. 501–514, 2005.

[11] Z. Ismail, A. Yahya, and A. Shabri, "Forecasting gold prices using multiple linear regression method," *American Journal of Applied Sciences*, vol. 6, no. 8, p. 1509, 2009.

[12] S. A. S. Olaniyi, K. S. Adewole, and R. Jimoh, "Stock trend prediction using regression analysis–a data mining approach," *vol*, vol. 1, pp. 154–157, 2010.

[13] S. Shekhar and H. Xiong, "Moving average regression," in *Encyclopedia of GIS*. Springer, 2008, pp. 732–732.

[14] A. Upadhyay, G. Bandyopadhyay, and A. Dutta, "Forecasting stock performance in indian market using multinomial logistic regression," *Journal of Business Studies Quarterly*, vol. 3, no. 3, pp. 16–39, 2012.

[15] N. R. Draper and H. Smith, *Applied Regression Analysis (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 1998.

[16] D. A. Kumar and S. Murugan, "Performance analysis of indian stock market index using neural network time series model," in *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2013 International Conference on*. IEEE, 2013, pp. 72–78.

[17] B. Pradhan and S. Lee, "Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling," *Environmental Modelling & Software*, vol. 25, no. 6, pp. 747–759, 2010.

[18] M. Hasan, M. M. Tanvee, M. Abdul Mottalib *et al.*, "Selecting features from high dimensional datasets using regression analysis," in *Knowledge and Smart Technology (KST), 2013 5th International Conference on.* IEEE, 2013, pp. 1–4.

[19] B. Li, H. Zha, and F. Chiaromonte, "Contour regression: a general approach to dimension reduction," *Annals of statistics*, pp. 1580–1616, 2005.

[20] S. Sarker and M. Hossain, "Increasing forecasting accuracy of trend demand by non-linear optimization of the smoothing constant," *Journal of Mechanical Engineering*, vol. 41, no. 1, pp. 58–64, 2010.

[21] R. Hocking and R. Leslie, "Selection of the best subset in regression analysis," *Technometrics*, vol. 9, no. 4, pp. 531–540, 1967.

[22] D. E. Farrar and R. R. Glauber, "Multicollinearity in regression analysis: the problem revisited," *The Review of Economic and Statistics*, pp. 92–107, 1967.

[23] O. O. Awe, "How bad is multicollinearity? evidence from multiple linear regression analysis," *International Journal of Current Scientific Research*, vol. 2, no. 2, pp. 344–349, 2012.

[24] H. Zainodin and S. Yap, "Overcoming multicollinearity in multiple regression using correlation coefficient," in *AIP Conference Proceedings*, vol. 1557, 2013, p. 416.

[25] S. S. Kumari, "Multicollinearity: Estimation and elimination," *Journal of Contemporary Research in Management*, vol. 3, no. 1, 2012.

[26] P. A. Cohen, "Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies," *Review of Educational Research*, vol. 51, no. 3, pp. 281–309, 1981.

[27] A. E. McKellar, P. P. Marra, S. J. Hannon, C. E. Studds, and L. M. Ratcliffe, "Winter rainfall predicts phenology in widely separated populations of a migrant songbird," *Oecologia*, vol. 172, no. 2, pp. 595–605, 2013.

[28] A. Garg and K. Tai, "Comparison of regression analysis, artificial neural network and genetic programming in handling the multicollinearity problem," in *Modelling, Identification Control (ICMIC), 2012 Proceedings of International Conference on*, June 2012, pp. 353–358.

[29] W. Qian, "Notice of retraction remedy to severe multicollinearity through ridge regression: A study on relationship bwtween tqm practice and performance," in *E -Business and E -Government (ICEE), 2011 International Conference on*, May 2011, pp. 1–4.

[30] K. Abhishek, A. Khairwa, T. Pratap, and S. Prakash, "A stock market prediction model using artificial neural network," in *Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on*. IEEE, 2012, pp. 1–5.

[31] M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," *arXiv preprint arXiv:0912.3924*, 2009.

[32] H. Hirose, Y. Soejima, and K. Hirose, "Nnrmlr: A combined method of nearest neighbor regression and multiple linear regression," in *Advanced Applied Informatics (IIAIAAI), 2012 IIAI International Conference on*. IEEE, 2012, pp. 351–356.

[33] S. Das and S. Chatterjee, "Multicollinearity problem–root cause, diagnostics and way outs," *Diagnostics and Way Outs (April 29, 2011)*, 2011.

[34] M. Alauddin and H. S. Nghiemb, "Do instructional attributes pose multicollinearity problems? an empirical exploration," *Economic Analysis and Policy*, vol. 40, no. 3, p. 351, 2010.

[35] B. Render, *Quantitative Analysis For Management*. 1st, 2011.

[36] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[37] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, I. M. Moroz *et al.*, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 23, 2007.

[38] P. Zhong and M. Fukushima, "Regularized nonsmooth newton method for multi-class support vector machines," *Optimisation Methods and Software*, vol. 22, no. 1, pp. 225–236, 2007.

[39] J. W. Longley, "An appraisal of least squares programs for the electronic computer from the point of view of the user," *Journal of the American Statistical association*, vol. 62, no. 319, pp. 819–841, 1967.

[40] D. N. Gujarati and B. Handelshøyskolen, *Econometrics by example*. Palgrave Macmillan Hampshire, UK, 2011.

[41] F. Mordelet, J. Horton, A. J. Hartemink, B. E. Engelhardt, and R. Gordân, "Stability selection for regression-based models of transcription factor–dna binding specificity," *Bioinformatics*, vol. 29, no. 13, pp. i117–i125, 2013.

[42] A. Kavousi-Fard, H. Samet, and F. Marzbani, "A new hybrid modified firefly algorithm and support vector regression model for accurate short term load forecasting," *Expert Systems with Applications*, 2014.

[43] A. Upadhyay, G. Bandyopadhyay, and A. Dutta, "Forecasting stock performance in indian market using multinomial logistic regression." *Journal of Business Studies Quarterly*, vol. 3, no. 3, 2012.

[44] J.-B. Chatelain and K. Ralf, "Spurious regressions and near-multicollinearity, with an application to aid, policies and growth," *Journal of Macroeconomics*, 2013.

[45] M. Ueki and Y. Kawasaki, "Multiple choice from competing regression models under multicollinearity based on standardized update," *Computational Statistics & Data Analysis*, vol. 63, pp. 31–41, 2013.

[46] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (Pt.1)*. Wiley-Interscience, 2000.

LIST OF PUBLICATION

1) Sharma Mrityunjay and Saha Suman,"Robust Best Regressor Model with Minimum Multicollinearity" in IEEE International Conference of Convergence of Technology,April 2014.

2)Sharma Mrityunjay and Saha Suman, "Graph Based Approach for Minimum Multicollinearity Highly Accurate Regression Model Explaining Maximum Variability" (Accepted in International Conference on Emerging Research in Computing, Information, Communication and Applications,ERCICA),2014).