Using Data Mining Technique to Analyze Export Goods and Services of Bangladesh

MD. kamruzzaman
Roll No.: 140120 Reg. No.: 101598
Session: 2013 - 2014

Supervised By

Md. Mahmudul Hasan
Assistant professor, Department of Computer Science and Engineering
Pabna University of Science and technology

A thesis has been submitted to the Department of Computer Science and Engineering for the fulfillment of the requirement of Bachelor Degree in Computer Science and Engineering

Course Title: Project / Thesis
Course Code: CSE-4200

Department of Computer Science and Engineering
Pabna University of Science and Technology
Pabna-6600

June 2019

CERTIFICATE

I am pleased to certify that Md.kamruzzaman, Roll No: 140120, Reg. No: 101598, Session: 2013-14 has performed a thesis work entitled **"Using Data Mining Technique to Analyze Export Goods and Services of Bangladesh"** under my supervision for the requirement of the completion of course entitled 'Project/Thesis'. So far as I concern this is an original thesis that has been carried out for one year in the Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna-6600, Bangladesh.

To the best of my knowledge, this paper has not been duplicated from any other paper or submitted to elsewhere prior submission to the department.

[Md. Mahmudul Hasan]
Assistant professor,
Department of Computer Science and Engineering
Pabna University of Science and Technology, Pabna-6600.
Bangladesh.

DECLARATION

In accordance with rules and regulations of Pabna University of Science and Technology following declarations are made:

I hereby declare that this thesis has been done by me under the supervision of Md. Mahmudul Hasan, Assistant professor, Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna-6600.

I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for awarding of any degree and any material reproduced in this thesis has been properly acknowledged.

Signature of student

ACKNOWLEDGEMENT

All praise for Allah who has created us and given a greatest status among his all creations. First of all I express my gratefulness to the Almighty Allah for enabling me to perform this task successfully. I would like to express my deepest sense of gratitude to my honorable supervisor Md. Mahmudul Hasan, Assistant professor, Department of Computer Science and Engineering (CSE), Pabna University of Science &Technology (PUST), for his scholastic supervision, valuable guidance, adequate encouragement and helpful discussion throughout the progress of this work. I am highly grateful to him for allowing me to pursuing this study under his supervision.

I am deeply thankful to all the respectable teachers of the Department of Computer Science and Engineering, Pabna University of Science & Technology, Pabna-6600, Bangladesh, for their encouragement and help in the last few months that enabled me to acquire a lot of knowledge relevant to my research work.

Finally, I am much grateful to my family members especially to my parents, all of my friends and well-wishers for their encouragement and supports.

July , 2019 Author

Abstract

With the weak economic process, the exportation has become a hot issue of Bangladesh. In this paper we investigated the impact of exports goods and services on the economic growth of Bangladesh using the data of the period 1989-2012. This paper also find some important reasoning factors which is related to exportation and also find relationship between those reasoning factors and exportation. Despite structural limitations in the Bangladesh economy, the export sector performed well throughout the 1990s. In this paper, linear regression model is used to forecast future export goods and services trends of Bangladesh. Then the linear regression model is trained on this dataset. The results show that there is a bidirectional long run relationship between exports goods and services and its reasoning factor (Age dependency ratio, Arable land, Export value index, internal debt stocks, Merchandise imports, industry value added, Official exchange rate). Cross validation is used to measure the effectiveness of the model. These results provide evidence that growth of exports goods and services in Bangladesh is increased by developing those independent reasoning factor.

Table of Contents

CHAPTER THREE

3. System Architecture

CHAPTER FOUR

CHAPTER FIVE

CHAPTER SIX

CHAPTER SEVEN

List of Tables

List of figures

CHAPTER ONE

Introduction

In this chapter we will introduce our thesis overview, background and objective. In section 1.1 we will talk about our thesis overview; in section 1.2 we will describe background of our thesis; in section 1.3 we will discuss about our thesis objectives**.**

1.1 Overview

It has been theoretically argued export play a crucial role in economic development. A South Asian nation physically located near economic powerhouses India and China, the People's Republic of Bangladesh shipped $42.2 billion worth of goods around the globe in 2018. Bangladesh is world's second-biggest apparel exporter after China. Garments including knit wear and hosiery account for 80% of exports revenue; others include: jute goods, home textile, footwear and frozen shrimps and fish, merchandise, Knit or crochet clothing, freight, insurance, transport, travel, royalties, license fees, and other such as business, personal, and government services. Despite many difficulties faced by the sector over the past years, it continued to show robust performance, Competitive strength. Exclusive Export Processing Zones (EPZ) are established to attract foreign direct investment and export promotion.

1.2 Background

Haydory Akbar Ahmed, Md. Gazi Salah Uddin (2015) describe Export, Imports, Remittance and Growth in Bangladesh. Annual data on Real GDP, exports, imports, implicit GDP deflator and remittance from 1976 to 2005 are used for this paper. Real GDP, export, import and implicit GDP deflator (base year 1990) data are collected from UN Statistical Division website3 [1].

Sayef Bakari, Mohamed Mabrouki examined Impact of exports and imports on economic growth new evidence from panama .This paper investigates the relationship between exports, imports, and economic growth in Panama. In order to achieve this purpose, annual data for the periods between 1980 and 2015 The data set entails of observation for GDP (current US$), exports of goods and services (current US$), and imports of goods and services (current US$). All data set have brought from World Development Indicators 2016[2].

Afaf Abdull J. Saaed and Majeed Ali Hussain (2015) examined Impact of Exports and Imports on Economic Growth: Evidence from Tunisia. The analysis used in this study cover annual time series of 1977 to 2012 or 36 observations which should be sufficient to capture the short run and long run correlation between Export, Import and economic growth .The data set consists of observation for GDP, exports of goods and services (current US$), and imports of goods and services (current US$). All data set are taken from World Development Indicators 2014[3].

Md. Tareq Ferdous Khan and Nobinkhor Kundu(2012) attemted to find out Future Contribution of Export and Import to GDP in Bangladesh: A Box-Jenkins Approach. This paper describes the growth rates of GDP, imports and exports product over the last three decades . The paper used the data of the three indicators such as GDP, EXPORT and IMPORT for the past three decades starting from year 1981 to 2010 from the "World Development Indicators 2001" published on April, 2011 by The World Bank[4].

Ullah et al (2009) investigated Export-led-growth by time series econometric techniques (Unit root test, Co-integration and Granger causality through Vector Error Correction Model) over the period of 1970 to 2008 for Pakistan. In this paper, the results reveal that export expansion leads to economic growth.

Mohammad Mafizur Rahman describes The Foreign Trade of Bangladesh: Its Composition, Performance, Trend, and Policy. This paper describes some issues which are very much related to exporting such as Exports Performance Compared to Imports, Composition and Performance of Imports of Bangladesh, Comparative Performance of Bangladesh's Export and Import Sectors, Region-wise Exports of Bangladesh, Directions of Bangladesh's Exports and Imports, Export Policy and Reform Programme, Tariff Rationalization etc[5].

1.3 Thesis objectives

The aim of the study is to find out the factor that is responsible for exporting goods and services from Bangladesh. Also try to find out the how to develop those factor so that exportation from Bangladesh will be increased. To find out a smooth, hassle free and effective exportation process. Here, I tried my best to utilize every part of my knowledge gained through learning and practical works done under different subjects, relevant books, research-studies, articles, journals, and websites regarding this matter. I have tried my best to find all possible factor regarding the exportation process and their viability, so that reader can get a clear picture of exportation relating factors of goods and services and can take decision which is best for their context.

CHAPTER TWO

Literature Review

In this chapter, we conduct several studies related to Data mining, the problem background, the platform and several techniques. In section 2.1 we discuss about data mining; in section 2.2 we discuss about application of data mining; in section 2.3 we talk about advantages of data mining then in section 2.4 we discuss about disadvantages of data computing; in section 2.5 data mining tools and techniques is discussed; in section 2.6 we explain some data mining tools .

2.1 Data Mining

Data mining is the process that extracts information from the enormous data sets and attempts to produce meaningful patterns. In other words, data mining is the techniques that make the collected data useful for us [6]. Data mining sometimes called knowledge discovery from data (KDD) is simply the discovery of patterns among data. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [7]. Data mining is the techniques for discovering interesting patterns hidden in large-scale data sets, focusing on issues relating to effectiveness and efficiency. The ultimate goal of data mining is to derive information and knowledge from the data in order to help users make intelligent decisions about complex problems.

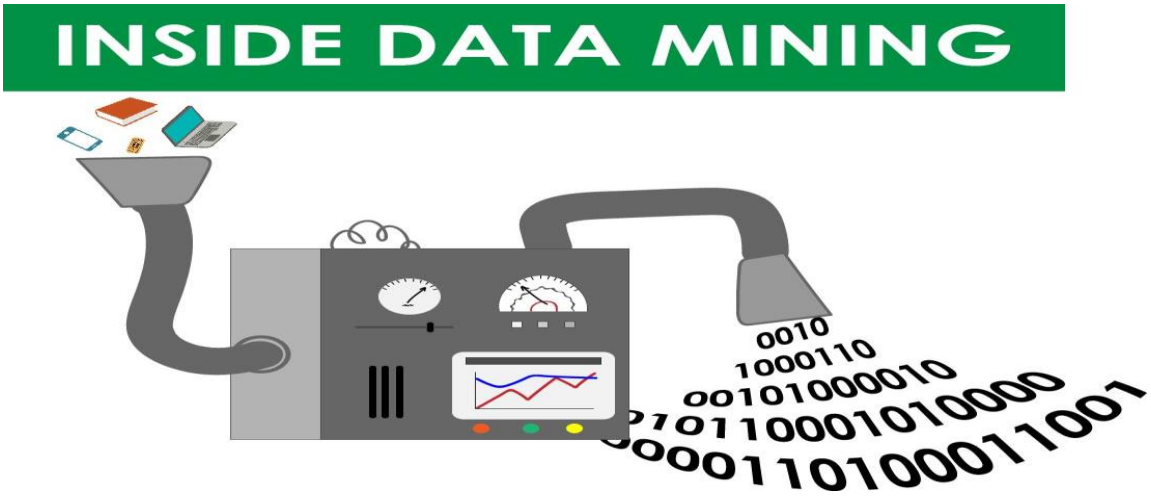

Figure 2.1: Data mining concept

2.2 Applications of Data Mining

Data mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field. Here we explore some fields of data mining applications.

2.2.1 Education

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

2.2.2 Engineering

Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

2.2.3 CRM

Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyze the information. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

2.2.4 Fraud Detection

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful

patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent.

2.2.5 Intrusion Detection

Any action that will compromise the integrity and confidentiality of a resource is an intrusion. The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection. Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common everyday network activity. Data mining also helps extract data which is more relevant to the problem.

2.2.6 Customer Segmentation

Traditional market research may help us to segment customers but data mining goes in deep and increases market effectiveness. Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the customers. Market is always about retaining the customers.

2.2.7 Financial Banking

With computerized banking everywhere huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers. The managers may find these information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

2.2.8 Surveillance

Corporate surveillance is the monitoring of a person or group's behaviour by a corporation. The data collected is most often used for marketing purposes or sold to other corporations, but is also regularly shared with government agencies. It can be used by the business to tailor their products desirable by their customers. The data can be used for direct marketing

purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails.

2.2.9 Analysis

History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research.

2.2.10 Multimedia Data Mining

Multimedia data mining is just what it sounds like. The practice looks to extract relevant data from text, hypertext, audio, video, still images and other content, and then convert that data into a numerical representation of knowledge. Multimedia data mining can be used to identify associations, clustering and classification, as well as perform similarity search.

2.2.11 Investigation

Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring and detecting crimes and their relationships with criminals. Text based crime reports can be converted into word processing files. These information can be used to perform crime matching process.

2.2.12 Informatics

Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction [8].

2.2.13 Data Mining In Computer Security

It concentrates heavily on the use of data mining in the area of intrusion detection. The volume of data dealing with both network and host activity is so large that it makes it an ideal candidate for using data mining techniques [9].

2.2.14 Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business [10].

2.3 Advantages of data mining

Data mining has a lot of advantages when using in a specific industry. Besides those advantages, data mining also has its own disadvantages e.g., privacy, security, and misuse of information. We will examine those advantages and disadvantages of data mining in different industries in a greater details.

2.3.1 Marking/Retailing

Data mining can aid direct marketers by providing them with useful and accurate trends about their customers purchasing behavior. Based on these trends, marketers can direct their marketing attentions to their customers with more precision. For example, marketers of a software company may advertise about their new software to consumers who have a lot of software purchasing history. In addition, data mining may also help marketers in predicting which products their customers may be interested in buying. Through this prediction, marketers can surprise their customers and make the customer's shopping experience becomes a pleasant one.

2.3.2 Banking/Crediting

Data mining can assist financial institutions in areas such as credit reporting and loan information. For example, by examining previous customers with similar attributes, a bank can estimated the level of risk associated with each given loan.

2.3.3 Law enforcement

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

2.3.4 Researchers

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing those more time to work on other projects [11].

2.3.5 Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipment and determine optimal control parameters. For example, semiconductor manufacturers have a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are a lot the same and some for unknown reasons even has defects.

2.3.6 Governments

Data mining helps government agency by digging and analyzing records of the financial transaction to build patterns that can detect money laundering or criminal activities [12].

2.3.7 Minimizes clients involvement

Most of the time while gathering information about certain elements, products and services, one used to depend on their clients for some additional information. But these data mining processes change everything and that is because of the help of such inclusion of technology in the data mining process.

2.3.8 Customer satisfaction:

One of the main nature of working which is involved in the mining techniques are from their informational matters. Most of the people seek for others help while making some decision. But it is not always easy to follow any one suggestion. And that is why with the help of data mining one can be confident enough to make their own decision [13].

2.4 Disadvantages of data mining

And while involvement of these mining systems, one can come across several disadvantages of data mining and they are as follows.

2.4.1 Violates user privacy

It is a known fact that data mining collects information about people using some market-based techniques and information technology. And these data mining process involves several numbers of factors. But while involving those factors, data mining system violates the privacy of its user and that is why it lacks in the matters of safety and security of its users. Eventually, it creates Mis-communication between people.

2.4.2 Additional irrelevant information:

The main functions of the data mining systems creates a relevant space for beneficial information. But the main problem with these information collection is that there is a possibility that the collection of information process can be little overwhelming for all.

2.4.3 Misuse of information

As it has been explained earlier that in the data mining system the possibility of safety and security measure are really minimal. And that is why some can misuse this information to harm others in their own way.

2.4.4 An accuracy of data

Most of the time while collecting information about certain elements one used to seek help from their clients, but nowadays everything has changed. And now the process of information collection made things easy with the mining technology and their methods. One of the most possible limitations of this data mining system is that it can provide accuracy of data with its own limits.

Finally the bottom line is that all the techniques, methods and data mining systems help in discovery of new creative things. And at the end of this discussion about the data mining methodology, one can clearly understand the feature, elements, purpose, characteristics and benefits with its own limitations [14].

2.5 Data mining methods and techniques

The data mining tasks can be classified generally into two types based on what a specific task tries to achieve. Those two categories are descriptive tasks and predictive tasks. Predictive Task uses some variables to predict unknown or future values of other variables. It might determine what might happen in future. Descriptive task find human interpretable patterns that describe the data. It describe what happened past.

Data mining methods:

1. Association (Descriptive)

2. Classification. (Predictive)

3. Clustering. (Descriptive)

4. Regression. (Predictive)

5. Sequential Patterns. (Descriptive)

6. Decision trees. (Predictive)

2.5.1 Association

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk".

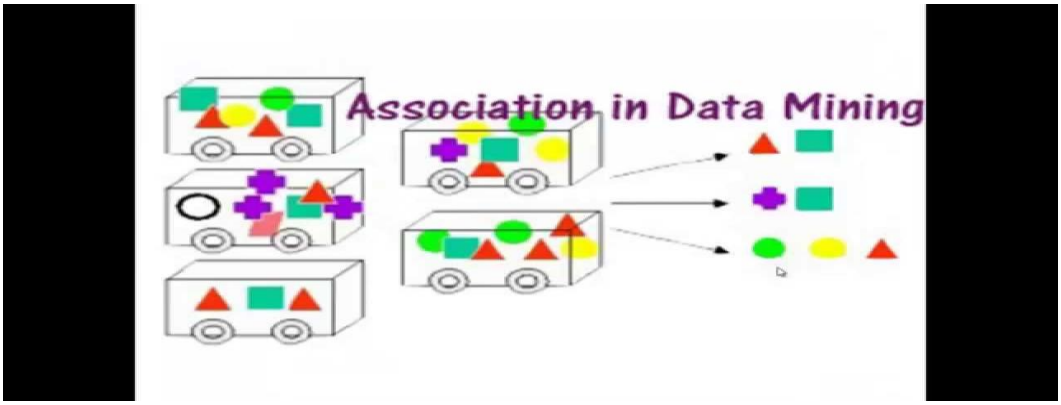

Figure 2.2(a): Association concept

2.5.2 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Examples: A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.

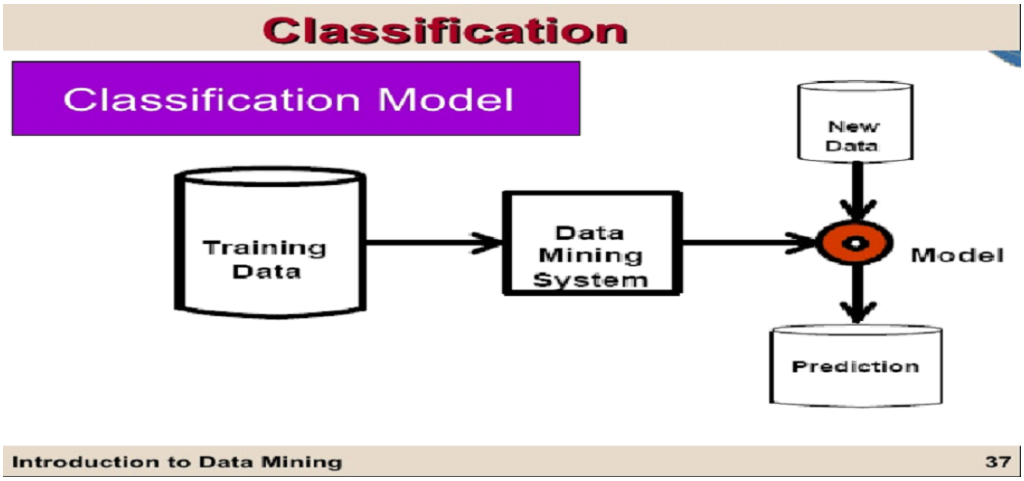

Figure 2.52(b): classification process

2.5.3 Clustering

Clustering is a process which partitions a given big data set into homogeneous groups based on given features such that similar objects are kept in a group whereas dissimilar objects in different groups.

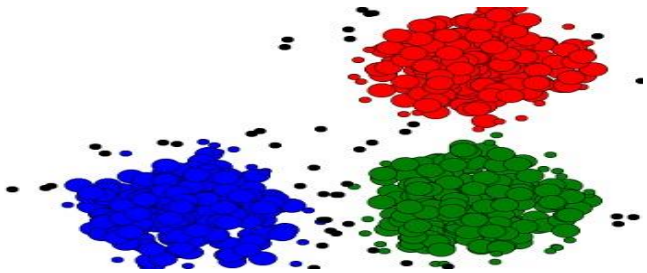

Figure 2.2(c): clustering concept

2.5.4 Regression

Regression is used to predict a numeric or continuous value while classification assigns data into discrete categories. Regression and classification are data mining techniques used to solve similar problems, but they are frequently confused. Both are used in prediction analysis.

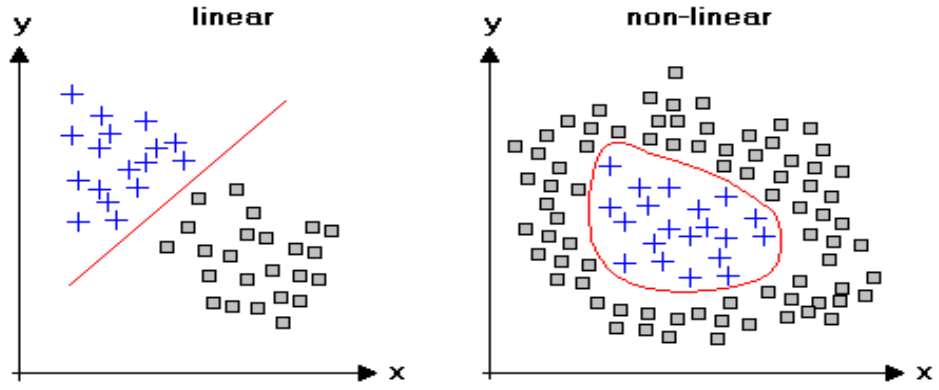

Figure 2.2(d): Regression analysis

2.5.5 Sequential Pattern

Pattern mining consists of discovering interesting, useful, and unexpected patterns in databases.

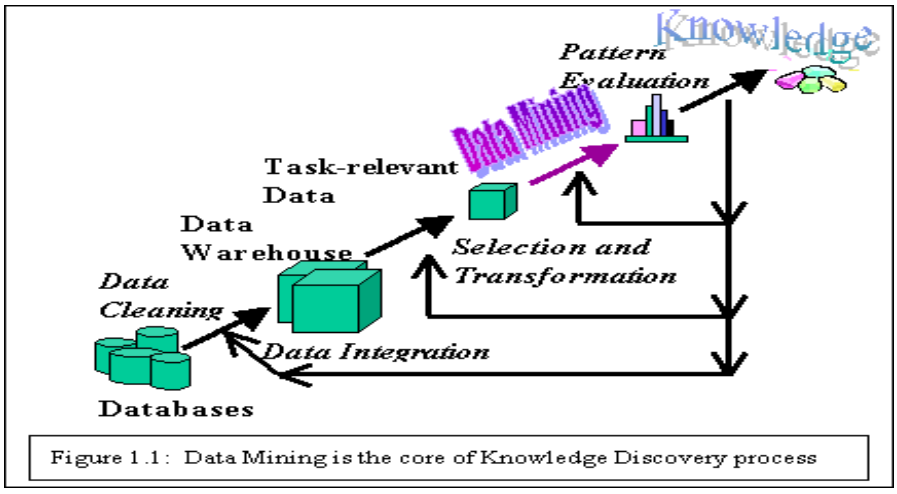

Figure 2.5(e): sequential pattern analysis

2.5.6 Decision Tree

Decision tree is used to take a problem with multiple possible solution and represent it easy to understand hierarchical structure formats. Sometimes it is also called classification tree.

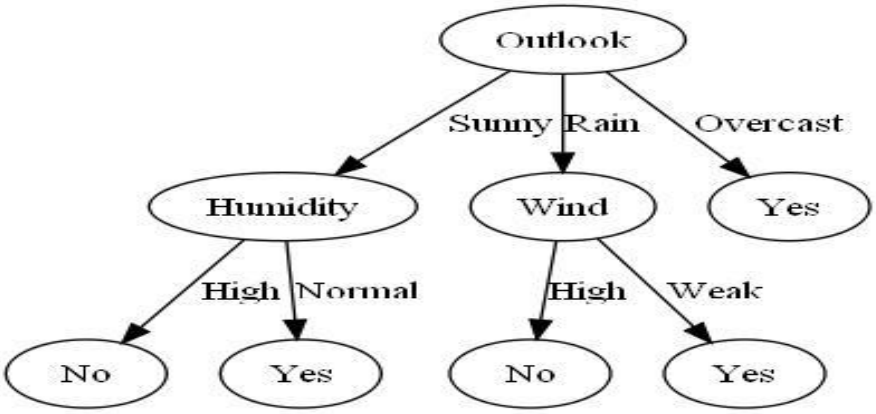

Figure 2.2(f): Decision tree concept

2.5 Data mining tools

2.6.1 Weka

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

2.6.2 Orange

Python users playing around with data sciences might be familiar with Orange. It is a Python library that powers Python scripts with its rich compilation of mining and machine learning algorithms for data pre-processing, classification, modelling, regression, clustering and other miscellaneous functions.

2.6.3 R

R is a free software environment for statistical computing and graphics written in C++. R Studio is IDE specially designed for R language.

2.6.4 Knime

Primarily used for data preprocessing—i.e. data extraction, transformation and loading, Knime is a powerful tool with GUI that shows the network of data nodes.

2.6.5 Rattle

Rattle, expanded to 'R Analytical Tool to Learn Easily', has been developed using the R statistical programming language. The software can run on Linux, Mac OS and Windows, and features statistics, clustering, modelling and visualization with the computing power of R.

2.6.6 RapidMiner

RapidMiner tends to be the preferred choice for startups and next gen "smart plant" manufacturers. Mobile apps and Chabot tend to depend on this software platform for machine learning, rapid prototyping, app development, and text mining and predictive analytics for customer experience.

2.6.7 H2O

If you want to get out on the cutting edge, start learning H2O. In its less than five years, it's been installed thousands of times, with applications for fraud detection at PayPal and customer metrics for the popular WordPress plugin Share This.

2.6.8 Kaggle

Kaggle is the world's largest data science community. Companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models. Kaggle is a platform for data science competitions. It help you solve difficult problems, recruit strong teams, and amplify the power of your data science talent.

2.7 Data resources

Every great data visualization starts with good, clean data. If we don't already have data you want to work with, where should you start? The reality is that there are thousands of free datasets available, ready to be analyzed and visualized. We just need to know where to look.

2.7.1 WDI

World Development Indicators (WDI) is the World Bank's premier compilation of international statistics on global development. Drawing from officially recognized sources and including national, regional, and global estimates, the WDI provides access to approximately 1,600 indicators for 217 economies, with some time series extending back more than 50 years. The database helps users find information related to development, both current and historical.

2.7.2 The World Factbook

The World Factbook provides information on the history, people, government, economy, geography, communications, transportation, military, and transnational issues for 267 world entities.

2.7.3 Open Data for Africa

Here you can visualize Socio-Economic indicators over a period of time, gain access to presentation-ready graphics and perform comprehensive analysis on a Country and Regional level.

2.7.4 Google Scholar

Google Scholar provides a simple way to broadly search for scholarly literature and academic studies.

2.7.5 The New York Times Developer Network

Search Times articles from 1851 to today, retrieving headlines, abstracts and links to associated multimedia. You can also search book reviews, NYC event listings, movie reviews and more.

2.7.6 Amazon Web Services

Browse Amazon Web Services' Public Data Sets by category for a huge wealth of information.

2.7.7 Wikipedia: Database

Wikipedia offers free copies of all available content to interested users. These databases Can be used for mirroring, personal use, informal backups, and offline use or database queries.

2.7.8 W3Schools

Fantastic set of interactive tutorials for learning different languages. Their SQL tutorial is second to none. You'll learn how to manipulate data in MySQL, SQL Server, Access, Oracle, Sybase, DB2 and other database systems.

2.7.9 Hadoop

The Definitive Guide – As a data scientist, you will undoubtedly be asked about Hadoop. So you'd better know how it works. This comprehensive guide will teach you how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. Make sure you get the most recent addition to keep up with this fast-changing service.

2.7.10 DataCamp

Learn data mining from the comfort of your home with DataCamp online courses. They have free courses on R, Statistics, Data Manipulation, Dynamic Reporting, Large Data Sets and much more.

2.7.11 Coursera

Coursera brings you all the best University courses straight to your computer. Their online classes will teach you the fundamentals of interpreting data, performing analyzes and communicating insights. They have topics for beginners and advanced learners in Data Analysis, Machine Learning, Probability and Statistics and more.

2.7.12 Udemy

With a range of free and pay for data mining courses, you're sure to find something you like on Udemy no matter your level. There are 395 in the area of data mining! All their courses are uploaded by other Udemy users meaning quality can fluctuate so make sure you read the reviews.

2.7.13 Udacity

Master a new skill or programming language with Udacity's unique series of online courses and projects. Each class is developed by a Silicon Valley tech giant, so you know what your learning will be directly applicable to the real world.

2.7.14 Treehouse

Learn from experts in web design, coding, business and more. The video tutorials from Treehouse will teach you the basics and their quizzes and coding challenges will ensure the information sticks. And their UI is pretty easy on the eyes.

CHAPTER THREE

System Architecture

In this chapter we describe the basic model of our proposed system. In section 3.1 we discuss about proposed system architecture, in section 3.2 we explain the model building process to follow.

3.1 Proposed system structure

This paper proposed system structure is a summary that details an outline of work. It identifies a problem and clearly states all the questions that will be researched as well as describes the resources and materials that one need. There are considered several steps to build a Regression model that analyse export goods and services and its responsible factors. There are considered several steps to analyse and manipulate required dataset. First, we have collected data that pre-process and extract some features for analysing further manipulation of it. Then we used linear regression algorithms to classify them visualize predicted model with the appropriate figure. Cross validation model is also used to test the effectiveness of the model. After the Anova test and the linear regression we find 7 responsible factors which are directly related to the outcome(export goods and services) and those factors are Age dependency ratio, Arable land, Export value index, External debt stock, Merchandise imports, Industry value added, Official exchange rate .For easy understand renaming the factor age dependency ratio as ADR, Arable land as AL, Export value index as EVI, External debt stocks as EDS, Merchandise imports as MI, Industry value added as IVA, Official exchange rate as OER

The usual flow of this proposed structure is as follows: Outline, Prepare visuals, data pre-processing, Describe methodology, Conclusions drawn from the data and References.

3.2 Model building

The model building based on applying anova test and linear regression model to the modified dataset. This process find some significant factors that is related to the outcome.

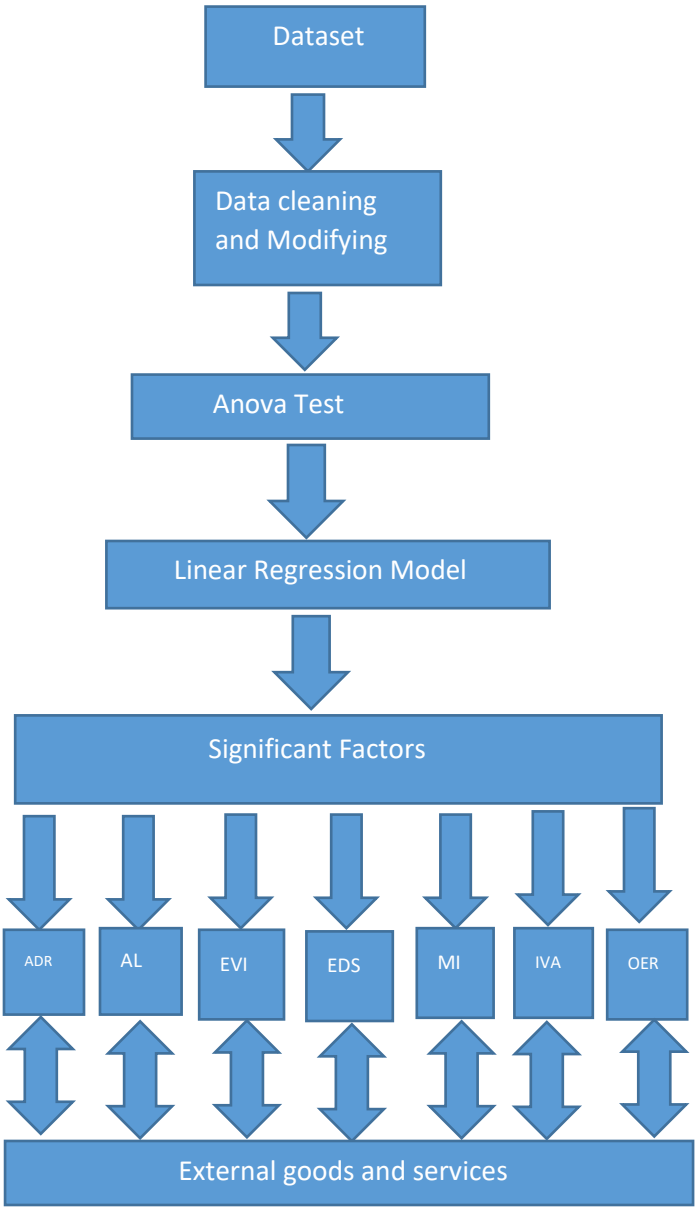

Figure 3.1: proposed system structure

Chapter Four

Implementation

To implement our task we first we need to use some tools. These are described here. In section 4.1 we discuss about tools used in implementation, in section 4.2 we briefly talk about the implementing steps of the model.

4.1 Tools

To implement our thesis we need to use some tools. These tools are very important for us, because we can't implement our thesis without these tools. Tools those are use here are-

4.1.1 Excel sheet

All the data used here are numeric type so excel sheet plays a significant role here. Excel sheet also used for calculating different type of mathematical terms.

4.2.2 Rstudio

Rstudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics[15]. The one-way analysis of variance (ANOVA), also known as one-factor ANOVA, is an extension of independent two-sample t-test for comparing means in a situation where there are more than two groups. If we use RStudio this is a nice way to see the data in spreadsheet format.

Rstudio is used here for calculating the one way anova test. In Rstudio it is easy to compare the value of the mean and t-test among different groups. Intercept value are easily calculated and easily understandable and more accurate.

4.3.3Weka

Weka is data mining software that uses a collection of machine learning algorithms. These algorithms can be applied directly to the data or called from the Java code. Weka is a collection of tools for Regression, Clustering, Association, Data pre-processing, Classification, and Visualization [16]. Weka is used for cross validating the model .Cross Validation is used to assess the predictive performance of the models and to judge how they perform outside the sample to a new data set also known as test data.

4.2 Implementation steps

The four implementation steps are data collection and preprocessing, Anova test, implementation of linear regression model and cross validation those four steps are discussed below.

Step1: Data collect and preprocessing

The first step of implementation was to collect data and preprocess that data. Data collection is a very important step of building prediction model because it is the base of all your predictions, a minor error in the data cause a blunder in the prediction. So collected data must be accurate. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. The dataset use for this research is collected from the world development indicators 2015(WDI).Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Data preprocessing is a proven method of resolving such issues. Data quality is explained in terms of accuracy, consistency, completeness, believability, interpretability and timeliness. These qualities are assessed by the usage of the data. In this study, we removed several tuples which have multiple unclear, duplicate and missing values from existing data. The world development indicators published data in the year between1960-2015. After cleaning datasets get the data of the year between 1989- 2012.

Step2: Anova Test

The second step of implementation is ANOVA test. The Analysis of Variance table is also known as the ANOVA table . Anova test is sometimes called as significance test. In order to find out the significant properties of the variables under consideration (Exports Goods And Services) we carry out Anova test. The purpose of anova test is to find out how the input factors affect the outcome factors. There is variability in the response variable. It is the uncertainty that would be present if one had to predict individual responses without any other information. The best one could do is predict each observation to be equal to the sample mean.

Step3: Model selection and implementation

The third step of implementation is the model selection and implement that model on the dataset.In order to quantitatively forecast the status of export goods and services, different data mining techniques can be used. The associated task for the dataset used in this paper is regression. Therefore, liner regression model is used to forecast the status of export goods and services. The linear regression model is simple and provides enough description of how the input affects the output. It predicts a variable Y (target variable) as a liner function of another variable X (input variable/features), given m training examples of the form (x1,y1), (x2,y2), …, (xn,yn), where xi ε X and yi ε Y.

Since the ANOVA test signifies that the variables are significant individually, now it is time to test whether they are co-integrated or not using Linear Regression equation. Linear regression model is used for determining the co-integration between the input factors and the outcome.

Significant factors are those which are directly related to the outcome factors (Exports goods and services). After the Anova test and the linear regression find 7 input factors and those factors are Age dependency ratio, Arable land, Export value index, External debt stock, Merchandise imports, Industry value added, Official exchange rate .For easy understand renaming the factor age dependency ratio as ADR, Arable land as AL, Export value index as EVI, External debt stocks as EDS, Merchandise imports as MI, Industry value added as IVA, Official exchange rate as OER .The paper is based on the following hypotheses for testing the co-integration relationship between Exports goods and services and its related independent factor.

1. Whether there is bi-directional relationship between export goods and services and its independent factors (ADR, AL, EVI, EDS, ML, IVA, and OER).
2. Whether there is unidirectional causality between ADR, AL, EVI, EDS, MI, IVA and OER.
3. Whether there exists a long run relationship between Exports goods and services and ADR, AL, EVI, EDS, MI, IVA, OER in Bangladesh.
4. Whether there is no causality between Exports goods and services and ADR, AL, EVI, EDS, MI, IVA, OER in Bangladesh.

Linear regression is a basic and commonly used type of predictive analysis. These regression estimates are used **to** explain the relationship between one dependent variable and one or more independent variables.
The linear regression model function here is used:

Fcalculated=k+a1*fADR+a2*fAL+a3*fEVI+a4*fEDS+a5*fML+a6*fIVA+a7*fOER.

Here, k=intercept, a1=intercept of ADR, a2=intercept of AL, a3=intercept of EVI, a4=intercept of EDS, a5=intercept of ML, a6=intercept of IVA, a7=intercept of OER.

fADR is the actual value of ADR, fAL is the actual value of AL ,fEVI is the actual value of EVI, fEDS is the actual value of EDS ,fML is the actual value of ML, fIVA is the actual value of IVA, fOER is the actual value of OER.

Now it is time to put those value on the linear regression equation to get the calculated value for the model. The proposed models calculated value is described below

Fcalculated1(1989) =(-113.1)+ 84.68334*(0.8882) + 73.38865*(0.3903) + 20.39124*(-0.04571)+ 619.1613*(-0.03147)+ 3.65E+09*(0.0000000006187)+ 20.28788*(0.9052)+ 32.27*(0.4464)=5.3704

Fcalculated2(1990)=(-113.1)+ 83.33352*(0.8882) + 72.64347*(0.3903) + 26.39155*(-0.04571)+ 577.2708*(-0.03147)+ 3.62E+09*(0.0000000006187)+ 20.69696*(0.9052)+ 34.56881*(0.4464)= 6.3138

Fcalculated3(1991)=(-113.1)+ 82.26793*(0.8882) + 72.09803*(0.3903) + 26.39155*(-0.04571)+ 593.2396*(-0.03147)+ 3.41E+09*(0.0000000006187)+ 21.7382*(0.9052)+ 36.59618*(0.4464)= 6.359591212

Fcalculated4(1992)=(-113.1)+ 80.96754*(0.8882) + 66.13659*(0.3903) + 32.783*(-0.04571)+ 500.4864*(-0.03147)+ 3.73E+09*(0.0000000006187)+ 22.47782*(0.9052)+ 38.95076*(0.4464)= 7.42318875

Fcalculated5(1993)=(-113.1)+ 79.498*(0.8882) + 65.34532*(0.3903) + 35.59743*(-0.04571)+ 444.7181*(-0.03147)+ 3.99E+09*(0.0000000006187)+ 23.81543*(.9052)+ 39.56726*(0.4464)= 9.083600621

Fcalculated6 (1994)=(-113.1)+ 77.94962*(0.8882) + 64.59246*(0.3903) + 41.57945*(-0.04571)+ 421.3013*(-0.3147)+4.6E+09*(0.0000000006187)+ 24.32645*(.9052)+ 40.21174*(0.4464)= 9.004415301

Fcalculated7 (1995) =(-113.1)+ 76.3753*(0.8882) + 64.56173*(0.3903) + 54.79731*(-0.04571)+ 334.4808*(-0.3147)+6.69E+09*(0.0000000006187)+ 24.55986*(.9052)+ 40.27832*(0.4464)= 11.25748542

Fcalculated8 (1996) =(-113.1)+ 74.99962*(0.8882) + 64.20066*(0.3903) + 66.50493*(-0.04571)+ 319.6096*(-0.3147)+7.03E+09*(0.0000000006187)+ 22.80733*(.9052)+ 41.79417*(0.4464)= 9.126937318

Fcalculated9 (1997)=(-113.1)+ 73.58258*(0.8882) + 64.45418*(0.3903) +75.62999*(-0.04571)+ 253.8403*(-0.3147)+7.26E+09*(0.0000000006187)+ 23.00293*(.9052)+ 43.89212*(0.4464)= 10.8764254

Fcalculated10 (1998) =(-113.1)+ 72.13925*(0.8882) + 64.94584*(0.3903) +80.15339*(-0.04571)+ 260.2409*(-0.3147)+7.5E+09*(0.0000000006187)+ 23.8329*(.9052)+ 46.90565*(0.4464)= 11.61822315

Fcalculated11 (1999) =(-113.1)+ 70.67338*(0.8882) + 64.80756*(0.3903) +86.0385*(-0.04571)+ 259.7778*(-0.3147)+8.33E+09*(0.0000000006187)+23.49542*(.9052)+49.0854*(0.4464)= 11.19262458

Fcalculated12 (2000)=(-113.1)+ 69.19632*(0.8882) + 64.14688*(0.3903) +100*(-0.04571)+ 213.8613*(-0.3147)+8.88E+09*(0.0000000006187)+23.31361*(.9052)+52.14167*(0.4464)= 11.97091177

Fcalculated13 (2001)=(-113.1)+ 68.06691*(0.8882) + 63.78582*(0.3903) +95.16356*(-0.04571)+ 215.8685*(-0.3147)+9.02E+09*(0.0000000006187)+23.76936*(.9052)+55.80667*(0.4464)= 13.11688937

Fcalculated14 (2002) =(-113.1)+ 66.87892*(0.8882) + 63.40171*(0.3903) +96.24354*(-0.04571)+ 237.3487*(-0.3147)+8.59E+09*(0.0000000006187)+24.04238*(.9052)+57.888*(0.4464)= 12.09912188

Fcalculated15 (2003)=(-113.1)+ 65.66006*(0.8882) + 63.20965*(0.3903) +109.4068*(-0.04571)+ 226.4685*(-0.3147)+1.04E+10*(0.0000000006187)+23.73511*(.9052)+58.15004*(0.4464)= 11.66075834

Fcalculated16 (2004)=(-113.1)+ 64.43361*(0.8882) + 62.55666*(0.3903) +129.989*(-0.04571)+ 210.5687*(-0.3147)+1.2E+10*(0.0000000006187)+24.01593*(.9052)+59.51266*(0.4464)= 11.72974627

Fcalculated17 (2005)=(-113.1)+ 63.20223*(0.8882) + 60.77437*(0.3903) +145.5157*(-0.04571)+ 162.9434*(-0.3147)+1.39E+10*(0.0000000006187)+24.59065*(.9052)+64.32748*(0.4464)= 14.54546335

Fcalculated18 (2006)=(-113.1)+ 62.29122*(0.8882) + 60.53622*(0.3903) +184.7275*(-0.04571)+ 151.4345*(-0.3147)+1.6E+10*(0.0000000006187)+25.3972*(.9052)+68.93323*(0.4464)= 16.32638983

Fcalculated19 (2007)=(-113.1)+ 61.32581*(0.8882) + 60.04456*(0.3903) +194.9163*(-0.04571)+ 140.6353*(-0.3147)+1.86E+10*(0.0000000006187)+25.73827*(.9052)+68.87488*(0.4464)= 17.01892969

Fcalculated20 (2008)=(-113.1)+ 60.31973*(0.8882) + 59.94469*(0.3903) +240.5635*(-0.04571)+ 129.603*(-0.3147)+2.39E+10*(0.0000000006187)+25.95879*(.9052)+68.59828*(0.4464)= 17.67998558

Fcalculated21 (2009)=(-113.1)+ 59.29565*(0.8882) + 59.89091*(0.3903) +236.0748*(-0.04571)+ 139.5787*(-0.3147)+2.18E+10*(0.0000000006187)+26.42566*(.9052)+69.03907*(0.4464)= 16.00591654

Fcalculated22 (2010)=(-113.1)+ 58.26268*(0.8882) + 59.8525*(0.3903) +300.4289*(-0.04571)+ 108.4024*(-0.3147)+2.78E+10*(0.0000000006187)+26.14449*(.9052)+69.64929*(0.4464)= 16.83573653

Fcalculated23 (2011) = (-113.1)+ 57.0606*(0.8882) + 58.98441*(0.3903) +382.52*(-0.04571)+ 100.1191*(-0.3147)+3.62E+10*(0.0000000006187)+26.39095*(.9052)+74.1524*(0.4464)= 19.36336958

Fcalculated24 (2012)=(-113.1)+ 55.93046*(0.8882) + 58.92295*(0.3903) +393.2864*(-0.04571)+ 101.9163*(-0.3147)+3.42E+10*(0.0000000006187)+26.74366*(.9052)+81.86266*(0.4464)= 20.28537805

So those are the calculated value of the linear regression equation. It is also necessary to calculate the absolute error, mean absolute error, relative error to know the amount of error the model holds. Error calculation is also needed to know the inaccuracy of a model.

Absolute error is the difference between the measured value and the actual value. If the measured value is defined by x_0 and the actual value is define by x then the absolute error is given by,

$$\text{Absolute error, } \Delta x = x_0 - x$$

The absolute error of the sum or difference of a number of quantities is less than or equal to the sum of their absolute errors. Absolute error is a measure of how far 'off' a measurement is from a true value or an indication of the uncertainty in a measurement. It is one way to consider error when measuring the accuracy of values.

the mean absolute error is an average of the absolute errors. MAE has a clear interpretation as the average absolute difference between x_0 and x. we want to know this average difference because its interpretation is clear.

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n} |x_i - x|$$

MAE stands for mean absolute error. Here n is the measurement numbers. We divede the entire data ser into 4 fold so here n value is 6.

Relative error is a measure of the uncertainty of measurement compared to the size of the measurement. Let the true value of a quantity be x and the measured or inferred value x_0. Then the relative error is defined by,

$$\delta x = \frac{\Delta x}{x} = \frac{x_0 - x}{x} = \frac{x_0}{x} - 1,$$

Where Δx is the absolute error. The relative error of the quotient or product of a number of quantities is less than or equal to the sum of their relative errors. The percentage error is 100% times the relative error.

Table 4.1 shows the calculated model result and the error result below

Year	Actual value	Calculated value	Absolute error	Mean absolute error	Relative error
1989	5.540144	5.370413384	0.16973	0.185166743	0.030636
1990	5.908316	6.313874452	0.405558		0.068642
1991	6.662612	6.359591212	0.303021		0.045481
1992	7.586677	7.42318875	0.163488		0.021549
1993	9.017269	9.083600621	0.066332		0.007356
1994	9.001544	9.004415301	0.002872		0.000319
1995	10.86463	11.25748542	0.392851	0.401148622	0.036159
1996	9.706508	9.126937318	0.57957		0.059709
1997	10.52037	10.8764254	0.356055		0.033844
1998	11.75733	11.61822315	0.139108		0.011832
1999	11.75864	11.19262458	0.566018		0.048136
2000	12.3442	11.97091177	0.373289		0.03024
2001	13.38656	13.11688937	0.26967	0.262176506	0.020145
2002	12.40997	12.09912188	0.310847		0.025048
2003	11.43115	11.66075834	0.22961		0.020086
2004	11.14651	11.72974627	0.583237		0.052325
2005	14.39284	14.54546335	0.152621		0.010604
2006	16.35346	16.32638983	0.027074		0.001656
2007	16.99533	17.01892969	0.023595	0.412176121	0.001388
2008	17.65886	17.67998558	0.021127		0.001196
2009	16.94013	16.00591654	0.934216		0.055148
2010	16.02411	16.83573653	0.811624		0.05065
2011	19.92207	19.36336958	0.558705		0.028045
2012	20.16159	20.28537805	0.123789		0.00614

Table 4.1: proposed model Calculated value and error result

After the implementation of the linear equation on the actual model it signifies that the calculated model is very close to the actual model. That means the calculated model fits the actual model.

Step 4: Cross validation

Cross-validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately this proposed model will perform in practice.

K fold cross validation model is used in this proposed model. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. Here a specific value of 4 k is chosen that is why the model is called as 4-fold cross-validation. The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds.

The general procedure is as follows:

1. Shuffle the dataset.
2. Split the dataset into 4 groups
3. For each unique group:
    1. Take the group as a hold out or test data set
    2. Take the remaining groups as a training data set
    3. Fit a model on the training set and evaluate it on the test set
    4. Retain the evaluation score and discard the model

4 .Summarize the skill of the model using the sample of model evaluation scores

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

By doing all these we identify the responsible factor with the outcome (export goods and services). Thus we calculate the error rate, relative error, mean absolute error and implementation is done.

CHAPTER FIVE

Result and Discussion

All the tests, model implementation results are described in this chapter. Section 5.1 describes anova test result, in section 5.2 we discuss about the linear regression model result.

5.1 anova test result

We first test the significance properties of the variables under consideration i.e. their order of integration, then test for co-integration among the variables. An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help us to figure out if we need to reject the null hypothesis or accept the alternate hypothesis. Basically, we're testing groups to see if there's a significance between the output (Export goods and services) and the responsible input factors.

In order to find out the significant properties of the variables under consideration (Exports Goods and Services) we carry out a one way Anova Test .Anova test is performed at 5% level of significance. Table-5.1 show the result of the ANOVA test below.

Factor	Probability(P value)	Level of significance
Age dependency ratio	7.497e-13	***
Arable land	1.054e-09	***
Export value index	1.288e-10	***
External debt stocks	3.338e-10	***
Merchandise imports	4.041e-10	***
Industry, value added	8.106e-11	***
Official exchange rate	1.939e-13	***

Table 5.1: ANOVA test result

To determine whether any of the differences between the means are statistically significant, compare the p-value to our significance level to assess the null hypothesis. The null hypothesis states that the population means are all equal. Usually, a significance level (denoted as α or alpha) of 0.05 works well. A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

If P-value ≤ α: The differences between some of the means are statistically significant. If the p-value is less than or equal to the significance level, we reject the null hypothesis and conclude that not all of population means are equal. So from the table it is clear that all the p-value is less than the significance level of 0.05 (denoted as α or alpha) we can reject the null hypothesis and conclude that some of the factors have different means. That is why we accept the alternative hypothesis. So we conclude that every factors are highly significant with the outcome factor (Exports Goods and Services). This is why these factor are directly responsible for the exporting goods and services from our country. Anova test result shows the great relationship between the input factors and the output factors.

5.2 Linear regression model result

Since the variables are significant individually, we can test whether they are co-integrated or not (Linear Regression). We test for the number of co-integrating relationships using the approach of Linear Regression model. Linear regression, also known as simple linear regression or bivariate linear regression, is used when we want to predict the value of a dependent variable based on the value of an independent variable. Here we have two or more independent variables, rather than just one, we need to use multiple regression.

Table 5.2 show linear regression model result below

Factor	Intercept	Standard error	Pr (>|t|) P value	Significant
Age dependency ratio	8.88e-01	1.47e-01	1.74e-05	***
Arable land	3.90e-01	1.23e-01	5.53e-06	**
Export value index	-4.5e-02	1.57e-02	0.01051	*
External debt stocks	-3.1e-02	4.29e-03	1.68e-06	***
Merchandise imports	6.1e-10	1.65e-10	0.00179	**
Industry value added	9.05e-01	2.14e-01	0.00066	***
Official exchange rate	4.46e-01	6.70e-02	5.53e-06	***

Table 5.2: Linear regression model result

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Here all the factors p-value less than 0.05 so it can easily said that the predictor's value is greatly significant with the outcome (export goods and exports).Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

In the output below, we can see that the predictor variables of all ADR, AL, EVI, EDS, MI, IVA and OER are significant because both of their p-values are well less than the 0.05. However, it is also clear that when the p values increases when towards significance value is greater than the significant level also decreases. From the table 5.2 result it is evident that four factors (ADR, EDS, IVA and OER) are highly significant, two factors (AL, MI) are less significant and one factor (EVI) is only significant with the outcome (Exports goods and services).

The sign of a regression coefficient tells you whether there is a positive or negative correlation between each independent variable the dependent variable. A positive coefficient (ADR, AL, MI, IVA, and OER) indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. A negative coefficient (EVI, EDS) suggests that as the independent variable increases, the dependent variable tends to decrease.

There needs to be a linear relationship between the dependent and independent variables. Whilst there are a number of ways to check whether a linear relationship exists between two variables, we suggest creating a scatterplot, where we can plot the dependent variable against your independent variable. we can then visually inspect the scatterplot to check for linearity.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the rest data set. Here the datasets are normally distributed, so the points in the QQ-normal plot lie on a straight diagonal line .The data are from a distribution of the same type (up to scaling and location), a reasonably straight line is being observed.

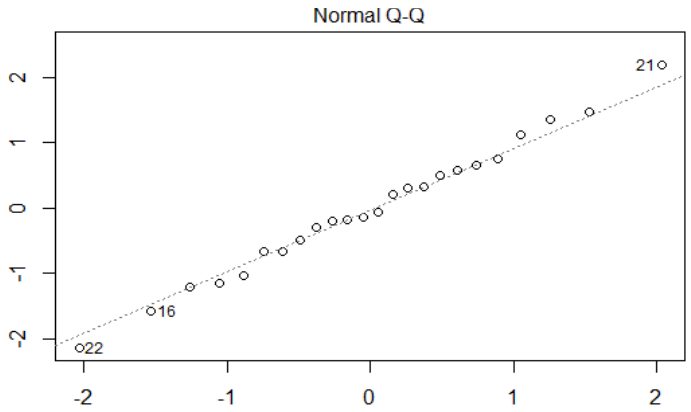

Figure 5.2(a): Normal Q-Q plot

If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data otherwise, a non-linear model is more appropriate.

When conducting a residual analysis, a "residuals versus fits plot" is the most frequently created plot. It is a scatter plot of residuals on the *y* axis and fitted values (estimated responses) on the *x* axis. The plot is used to detect non-linearity, unequal error variances, and outliers. The residual vs fitted graph below describes how the model is more linearly appropriate.

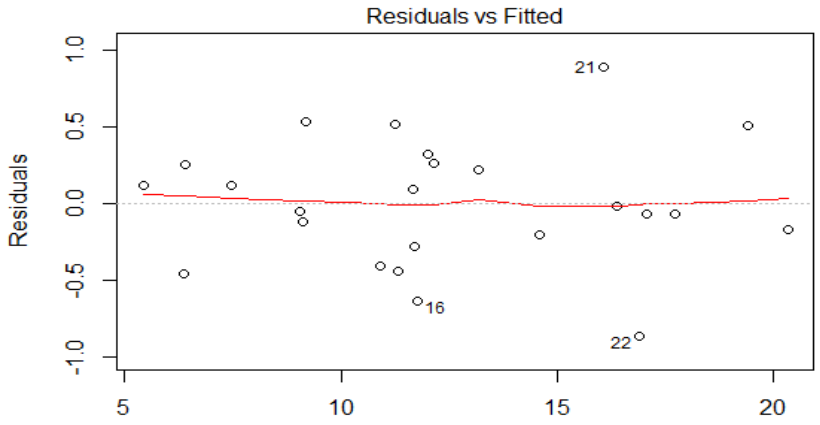

Figure 5.2(b): Residuals vs fitted plot

A "good" residuals vs. fitted plot should be relatively shapeless without clear patterns in the data, no obvious outliers, and be generally symmetrically distributed around the 0 line without particularly large residuals.

If we find equally spread residuals around a horizontal line without distinct patterns, that is a good indication we don't have non-linear relationships. So the Residuals vs fitted plot signifies that there is a linear relationship between the input factors and the outcome (Export goods and services).

There are edits built into the data capture application to check the entered data for unusual values, as well as to check for logical inconsistencies. Whenever an edit fails, the interviewer is prompted to correct the information (with the help of the respondent when necessary). For most edit failures the interviewer has the ability to override the edit failure if necessary.

K-fold cross validation is used to detect the error of the model and also used for model evaluation. The total dataset is splits into 4 set so 4 fold cross validation is performed. The result is performed at significance level of 0.05 with paired T-Tester (correlated). After the cross validation the first model shows an error of 8.42%, the second model shows an error of 13.42%, the third model shows an error of 11.04% the error of last one is 10.96%. So from the result it was assumed that the model performed well at practice.

Results of Linear Regression tests suggest the existence of at least seven co-integrating relationships among the variables in the series at 5% level of significance. This implies that the series under consideration are driven by at least seven common trends. We save the Residual standard error (0.4871) which is the difference between a set of observed and predicted values. The residual standard error calculates how much the data points spread around the regression line. Which are used as the error-correction term in the subsequent tests for Linear Regression model.

We also calculate multiple R-squared (.9906) and Adjusted R-squared (0.9865) which is really fantastic. The higher the values the greater the model. R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. The value of R^2 of this model is 0.9906, then approximately .9906 of the observed variation can be explained by the model's inputs.

CHAPTER SIX
Limitation and Future Work

We try to build our model that has less error as possible. But we have some limitation in this proposed models. In future we can solve this problem with the help of other system, these are described in this chapter. In section 6.1 we discuss the limitation of our system, in section 6.2 we discuss future works to solve our limitations.

6.1 Limitations

The limitations of the study are those characteristics of design or methodology that impacted or influenced the interpretation of the findings from our research Note again that discovering a limitation can serve as an important opportunity to identify new gaps in the literature and to describe the need for further research.

Firstly the sample size used here was too small, it will be difficult to find significant relationships from the data. As the size of sample is small, it is difficult to find a trend and a meaningful relationship.

Secondly because Lack of prior research or studies on this topics, it may be required to develop an entirely new research typology which seems to be very difficult.

Thirdly all the fractional point are not taken as consideration and also we take the fractional as approximate value, thus the error rate may slightly increased.

Fourthly other methods (association, classification, regression, clustering, and decision tree) maybe also fitted for this research but that is not checked in this implementation.

Fifthly the model was fit on linearly. No non-linear model is built on this dataset.

Sixthly if the dataset is large then more fold can be formed as a result the cross validation model becomes more accurate.

6.2 Future work

To overcome our limitation we have some future works. Firstly require a larger sample size to ensure a representative distribution of the population and to be considered representative of groups of people to whom results will be generalized or transferred.

Secondly the model will be built in a way that should fit both linearly and non-linearly.

Thirdly, an alternative evolutionary algorithm can be used in our system for calibrating our model to improve further.

Fourthly cross validation model should be more precise and more accurate in future research.

Chapter seven

Conclusion

7. 1 Conclusion

In this paper, data mining technique is used to forecast future exportation from Bangladesh .The aim of this study is to find the factor that are related to Exporting goods and service of Bangladesh using the data during the period 1989-2012.The Anova test and Linear Regression model are applied to investigate the relationship between the export goods and services and those factors that are responsible for exporting. Anova test finds the individual's relationship between the outcome factors and the responsible input factor .Linear Regression model finds the co-integration relationship between the input responsible factors and outcome factor. For this purpose, linear regression model is trained by the data of previous years. After training linear regression, different types of exportation reason found. All the results are shown through the table 5.1 to table 5.2.table-4.1 shows how accurate our calculated model and also show the calculated value of absolute error, mean absolute error, relative error. Table-5.1 shows the result of anova test. Table 5.2 shows that, how accurate the liner regression model is to forecast future exportation trends from Bangladesh, This finding clarified that export goods and services are found co-integrated with seven factor and those are Age dependency ratio, Arable land, Export value index, External debt stock, Merchandise imports, Industry value added, Official exchange rate .different types of exportation reason are found from this research .From the experimental result it is also seen that, those reason are greatly signified with the outcome. The results of this data mining could potentially be used to increase exportation for the forthcoming years.

Referances

[1] Haydory Akbar Ahmed1 Md. Gazi Salah Uddin(2015).Export, Imports, Remittance and Growth in Bangladesh.

[2] Sayef Bakari, Mohamed Mabrouki. IMPACT OF EXPORTS AND IMPORTS ON ECONOMIC GROWTH: NEW EVIDENCE FROM PANAMA .

[3] Afaf Abdull J. Saaed and Majeed Ali Hussain(2015). Impact of Exports and Imports on Economic Growth: Evidence from Tunisia.

[4] Md. Tareq Ferdous Khan and Nobinkhor Kundu(2012).Future Contribution of Export and Import to GDP in Bangladesh: A Box-Jenkins Approach.

[5] Mohammad Mafizur Rahman(2008), The foreign trade of Bangladesh: its composition, performance, trend, and policy.

[6].Computer Science for Data Mining, https://www.kth.se/en/om/internationellt/projekt/swb/2.49254/computer-science/computer-science-for-data-mining-1.354606, 20 June, 2019.

[7].Data mining, https://en.wikipedia.org/wiki/Data_mining, 20 June 2019.

[8].top 14 useful applications for data mining, http://bigdata-madesimple.com/14-useful-applications-of-data-mining/, 20 June, 2019.

[9]. Applications of Data Mining in Computer Security, https://www.springer.com/gp/book/9781402070549, 20 June, 2019.

[10].Data Mining - Applications & Trends, https://www.tutorialspoint.com/data_mining/dm_applications_trends.htm, 25 June 2019.

[11].advantages of data mining, https://xiangyun86.wordpress.com/2006/12/05/advantages-disadvantages-of-data-mining/, 25 June, 2019.

[12] Advantages and disadvantages of data mining, https://www.zentut.com/data-mining/advantages-and-disadvantages-of-data-mining/ , 26 June, 2019.

[13].purpose of data mining, https://content.wisestep.com/data-mining-purpose-characteristics-benefits-limitations/ , 26 June, 2019.

[14]. Limitations or Disadvantages of Data Mining Techniques, https://content.wisestep.com/data-mining-purpose-characteristics-benefits-limitations/, 26 June, 2019.

[15]. Rstudio**,** https://en.wikipedia.org/wiki/RStudio, 26 June, 2019.

[16]. an introduction to weka**,** https://opensourceforu.com/2017/01/an-introduction-to-weka/, 26 June, 2019.