ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF ELECTRICAL ENGINEERING AND COMPTING

DEPARTMENT OF COMPUTING

Discovering Knowledge From complex Data: The Case of Ethiopian Revenue and Customs Authority

By: - BABESHA KENAW

JANUARY, 2018

ADAMA, ETHIOPIA

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF ELECTRICAL ENGINEERING AND COMPTING

DEPARTMENT OF COMPUTING

Discovering Knowledge From complex Data: The Case of Ethiopian Revenue and Customs Authority

A Thesis Submitted to the department of computing at Adama Science and Technology University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Systems.

By: - Babesha Kenaw

ADVISOR: - Mr. Bahiru Shifaw (PhD Candidate)

JANUARY, 2018

ADAMA, ETHIOPIA

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF ELECTRICAL ENGINEERING AND COMPTING

DEPARTMENT OF COMPUTING

Discovering Knowledge From complex Data:The Case of Ethiopian Revenue and Customs Authority

By: - Babesha Kenaw

Name and Signature of Members of the Examination Board

Mr. Bahiru Shifaw _____

Advisor Signature

_____ _____

External examiner's name Signature

_____ _____

Internal examiner's name Signature

_____ _____

Chairperson's name Signature

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Babesha Kenaw

January, 2018

The thesis has been submitted for examination with my approval as university Advisor

---

Mr. Bahiru Shifaw

January, 2018

Acknowledgement

First of all, I would like to forward my gratefully and deepest thanks to the almighty God for his guidance, help and support by adding days to my age in completion of this research. Next, my heartfelt thanks go to my research advisors Mr. Bahiru shifaw for his valuable advice, patience and constructive criticisms on this research work. Thirdly, I would like to thanks all ERCA IT staffs for their honesty, suggestion and cooperation by giving necessary data and information for this research. I am thankful to my lovely wife w/ro Nestanet Sileshi and my son Yared, and my daughter Arsema Babesha, who have been always with me through this study by encouraging and asking me about the progress of my thesis work.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ACRONOMY

ASCII	American Standard code for Information Interchange
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASYCUDA	Automation System Customs Data
CIF	Cost, Insurance and Freight
DM	Data Mining
EBCDIC	Extended Binary Coded Decimal Interchange Code
ERCA	Ethiopian Revenue and Custom Authority
GDP	Gross Domestic product
IMF	International Monetary Fund
KDD	Knowledge Discovery in Database
KNN	K-nearest Neighbors Algorithm
ML	Machine Learning
MLP	Multi-Layer Perceptron
MMH	Maximum Margin Hyperplane
OLR	Ordinary Linear Regression
PCA	Potential Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Square Regression
RDMS	Relational Database Management Systems
RF	Random Forest
RMSE	Root Mean Square Error
SMO	Sequential Minimal Optimization
SQL	Structured Query Language

SVC	Support Vector Classifier
SVM	Support Vector Machine
SVR	Support Vector Regression
UNDP	United Nations Development Programme
USD	United States Dollars

Abstract

The research area of the thesis is data mining in Revenue and customs sector. We applied data mining in imported Customs items data set by using machine learning techniques. The object of the research is to evaluate models trained by using machine learning algorithms and compare the results which would increase the efficiency of data analysis..

In this thesis, we collected a total number of 100,310 imported Items described with 15 attributes from ERCA. And 10% of this data set used in the experiment by random sampling method. The dataset for the study collected from ERCA and the data pre-processing and resampling techniques are explained in order to improve the performance of the training model. During the implementation of machine learning algorithms, three typical models (Ordinary Linear Regression, SVM and Random Forest) have been implemented by using the different packages in R on the given large datasets.

The experiment result shows almost 1 for multiple R2 and adjusted R2 for Ordinary Linear Regression that shows as there is existence of 100% of the Variation in total import costs. The 10-fold cross validation result on the test set for OLR model shows 2952689 and 0.8113806 for the smaller RMSE and maximum R2 respectively. When we compare the result of RF with the OLR model, the minimum RMSE and maximum R2 we can get from the results are 464212.8 and 0.9928060 which shows better performance than OLR.

The quantitative and visual results of our practical machine learning implementation show the feasibility for the large datasets under the random forest algorithms. The research results of the work revealed new opportunities in the application of data mining methods by using machine learning in the Revenue and customs domain.

Key words: Machine Learning, Regression Model, Ordinary Linear Regression, support Vector Machine, Random Forest

CHAPTER ONE

INTRODUCTION

1.1 Background and motivation

It is generally accepted that developing economies require increasing quantities of Imported Goods that they cannot produce themselves efficiently, especially capital goods, intermediate goods and many raw materials. Frankly speaking, import substitution only tempers the nation's economy, but does not eliminate the growing demand for custom imported goods. Imports and export are a main part of international trade and the custom imports of capital goods are vital to economic growth. Imported capital goods directly affect inward investment which, in turn, constitutes the engine of economic growth [46].

Ethiopia is one of the developing countries that acquire most of substantial goods from abroad. Ethiopia is recently rated to be one of the countries in the world that have achieved the fastest annual economic growth since 2004. The economic growth cannot be achieved without the importation of capital goods, raw materials and intermediary goods, which in turn increases the openness of the national economy to the international economy. According to UNDP, the total import for the country was estimated 9.4 billion USD in the first seven months of 2014/15 only. This shows 19.7 percent increase compared to its preceding year. The trade deficit that was 8.4 billion USD in 2013/14 increased to 10.4 billion USD in 2014/15. The major imports were capital goods that are account for 39.7%, consumer goods 28.1%, semi-finished goods 15.5% and fuel 13.7% [46].

Data mining is the process of discovering of knowledge insightful, interesting, and novel patterns. In addition, DM is descriptive, understandable, and predictive models from large-scale of data [27]. DM is the extension of the Relational Database Management System (RDMS) that helps to organize highly complex data beyond the capacity of human mind. The difference between RDMS and data mining is that while RDMS requires users to provide queries and extract information from the database, data mining automatically produces results from the patterns that it detects in the information. Moreover, Data mining is an interdisciplinary field that combines artificial

intelligence (AI), database management, data visualization, machine learning, mathematics algorithms, and statistics. Data mining, also known as knowledge discovery in databases (KDD)

The key task for data mining is the process of modeling the dataset. Data mining is the process of extracting or detecting hidden patterns or information from large databases. With an enormous amount of custom imported Goods data, data mining technology can render business intelligence to process new results for decision making [2]. In this research we planned to use machine learning techniques to make the result better than old data mining techniques. Machine learning systems automatically learn programs from data. This is often a very attractive alternative to manually constructing them, and in the last decade the use of machine learning has spread rapidly throughout computer science and beyond. Machine learning is used in Web search, spam filters, recommender systems, and placement, credit scoring, fraud detection, stock trading prediction, drug design, and many other applications. A recent report from the McKinsey Global Institute asserts that machine learning will be the driver of the next big wave of innovation [29].

The customs and revenue sector can be considered as important and relevant development sector that drive huge investment and currency for the country. The sector is also important and relevant in providing information system and knowledge creation because the importers, tax payers, public, government and researchers need the information in the area for different purposes. Especially, as far as the researchers knowledge there are huge data sets created every year in both import and export and yet not appropriately interpreted to gain hidden knowledge for decision. This is the reason why the researcher motivated to conduct this research in the area.

The Ethiopian Revenues and Customs Authority (ERCA) is one of government organization which was established by the proclamation No .587/2008 for the purpose of enhancing the mobilization of government custom and revenues. The main purpose of the establishment of ERCA was to increase the public revenue generation function by bringing the relevant concerned agencies under the umbrella of the central revenue collector body.

This establishment of ERCA was aimed at improving public service delivering, facilitating international trade, enforcing the tax and customs regulations and thereby enhancing mobilization of Government revenue in sustainable manner. Also it improved key business processes and has

come across improving inefficient organizational structure and unnecessary complicated procedures that hinder to give proper service delivery through business re-engineering. The import and export of goods procedures were subjected to change due to their inefficiency delay of goods at exit or entry points which is disfavor for importers or exporters. The authority came to existence to bring clear customs procedure that aligned with international trade principles and efficient and effective system to control tax process.

One of the duties and responsibility of the Authority as indicated on the establishment proclamation is to Collect, analyze and interpret information necessary for the control of custom import and export goods and the assessment and determination of taxes; compile statistical data on criminal offences and frauds relating to the sector, and disseminate the information to other concerned bodies as may be necessary [22]. In line with this, a predictive approach of a complex data mining by using machine learning algorithms research designed on the imported goods data set to come up with decision support output.

1.2 Statement of the research problem

Both import substitution and export promotion are the key element of Ethiopian development strategy in the custom and revenue sector. The export earnings provide the foreign exchange needed for the investments that requires Capital goods import [22]. According to IMF report, the Ethiopian Imports increase relatively compared to the previous year whereas, the export shows decline. In reality, the gap between import and export should be balance or either export should exceed imports. This problem indicates as there is a need for further interpretation of imported customs data in the area to minimize the gap [24].

A knowledge base machine learning algorithms of customs Import goods data set would support decision regarding trade agreement, logistic of imported goods, Customs clearance process, import substitution, predict import elasticity, and increasing revenue from import [11].

At present, the data collection, analysis, interpretation and dissemination techniques is made by using simple Statistical tools like MS-Excel, database and SPSS exist in the agency. The results also described in percentage, mean, median, mode and some other statistic methods, and decision made based on these traditional statistical techniques for descriptive data analysis purpose. The

other problem is that we found large amount of data collected through a long period of time (i.e. many years of data) not well organized and analyzed using analytical machine learning tools and applications to improve the decision making process in the sector.

The primary goal of this research work is to analyze the customs imported goods dataset by using machine learning algorithms for appropriate decision making activities. The datasets the researcher planned to use in this research was very huge and complex as mentioned before, and the simple statistical tools do not support to provide significant information and knowledge from these large collections of big data set for the purpose of planning, preparedness and decision making activities. Furthermore, the researcher plan to perform an effective analysis of the dataset with machine learning techniques to model descriptive and predictive mining for effective and efficient customs imported goods system. In general, the following listed points are major reasons to conduct this research: -

- Data or information creates per year from import and export goods are very big data to manipulate and use for decision support system by the existing DBMS.
- No research that can contribute towards the proper utilization of machine learning on the available data sets has been done so far.
- Conducting the research by using machine learning algorithms plays an important role in adding value in ICT support for decision in the area.

This study investigated the application of machine learning methods and techniques to extract useful and interesting knowledge from the large and complex data set of custom imported goods. This study aims to answer the following three basic research questions:

- Can we propose a model for evaluating performance of customs imported goods data set?
- How machine learning techniques can be used to analyze the customs imported data set?
- What will be the best predictive model of import goods from different algorithms employed?

1.3. Objectives

1.3.1 General objective

The general objectives of this research is evaluating predictive model for the Ethiopian imported custom goods data by using machine learning algorithms in order to discover useful information for effective decision making in the sector.

1.3.2 Specific objectives

To achieve the above mentioned general objective, the following specific objectives are formulated:-

- To collect customs imported goods Data from Ethiopian custom and Revenue authority.
- To Study the data set.
- To assess the potential machine learning algorithms and tools to process the data.
- To prepare the data that needed for machine learning technique from Custom import goods data set.
- To develop different prediction models using machine learning algorithms from customs imported goods data set.
- To compare the predictive models and suggest the best model.
- To report the results, conclusion and make necessary recommendations.

1.4. Scope and Limitation

The scope of this research limited to examine predictive model by using machine learning techniques and algorithms to support of Customs imported goods data set especially for Ethiopian Revenue and Custom Authority. Moreover, the research identified an interesting and critical informative pattern that is used for preparation of predictive model from imported Items data set. On the other hand, finding and identification of appropriate machine learning software, getting enough resources such as relevant literature related to the specified study can be considered as some of the limitation of this research.

1.5. Significance of the study

Even though the research was done for educational purpose, its findings and results would be applicable for better decision support activities in the Customs and revenue sector. It also provided

an input for a better data collection, analysis, dissemination, interpretation and presentation purpose to take measurable action for import elasticity in the country. It would also be a direction to a research in the customs and revenue sector and provide a comprehensive documentation for other researchers in the field.

Finally, the researcher provided recommendation with different alternatives to customs and revenue sector for data collection, dissemination, interpretation, presentation, analysis techniques and tools to the sector. As a result, quality data which can increase the effectiveness, efficiency and total quality management of the customs imported goods system for planning, managing, preparedness and decision-making processes will be obtained. Therefore, this quality data would be used for a better model development with analytical machine learning tools.

1.6. Organization of the Thesis

This thesis is divided into five chapters. The first chapter is an introduction part, which Contained background information to the research work, the statement of the problem that this research addressed, the purpose and objective of the research, and methodologies adopted for the study, scopes and limitation of the study, and organization of the thesis work.

The second chapter discusses about data mining, machine learning techniques methods/algorithms used in the study, and its application in the customs and revenue sector, also it considered the specific differences exist between data mining applications and machine learning algorithms by reviewing related literature. In addition, it explained what has been done so far in the sector.

The third chapter is devoted to the general methodology followed in the research. It gives further understanding about business understanding and data preprocessing. In this chapter the main issues are data understanding and preprocessing, data preparation, data quality measures, the data collection and analysis tools and techniques, data transformation, and feature selection processes.

The forth chapter provides experimentation about the different machine learning algorithms and steps that are undertaken in the research work. The issues considered here are model Development and testing results for machine learning, customs imported data clustering, Regression and prediction by using R studio programing software tool. Finally, the fifth chapter deals about the

final parts of this thesis work the findings, the concluding remarks and recommendations forwarded based on the research findings.

CHAPTER TWO

LITERATURE REVIEW

This chapter discussed data mining as a data analytics tool in which machine learning technique is the primary focus of this thesis. The focus is on the creation of a data mining process framework by using machine learning algorithms and not on the actual application of the data mining techniques. The principles, concepts, weakness and strengthens of selected machine learning algorithms are also discussed. In addition, the difference and similarities exist between Data mining and machine learning is indicated. The last paragraphs of this chapter provide us with a summary of the available related works in the area internationally and locally.

2.1. Data mining

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. Data produced at extreme rates which push to limit our data analytical capacity. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data.

Data mining is to discover knowledge that is of interest from large amounts of data stored in data repositories or data sets [10]. Data Mining refers to extraction of hidden information which has some sort of value; hence, data mining is to pull out valuable information from huge and complex data set. Numerous data mining tools and techniques are available in the market to predict future trends and assist decision making today. This practice further helps organizations to make proactive decisions by depending on past and present data [37]. Some of the application areas of data mining are marketing, sales, customer relationship management, banking, insurance, fraud detection, bioinformatics and many other more areas. In custom and revenue area also data mining applied in risk assessment, fraud detection, and import elasticity and so on [37].

Data mining (DM) tasks has not established well yet. However, information sources classified the tasks of data mining in classification, clustering, prediction, association analysis, visualization and task analysis [11]. And the methods of learning techniques based on three main categories mainly known as supervised, unsupervised learning techniques and others. Supervised learning technique

includes tasks such as classification and prediction whereas unsupervised learning technique includes tasks of clustering and association rule mining [11].

2.1.1 Data Mining Tasks

Data mining problems are always broken down to a number of tasks that can be used to group techniques and application areas. Even if the number of tasks and the names of the tasks vary slightly, they all cover the same concept; According to [17] defines the following data mining tasks:

- **Classification**. The task of training some sort of model capable of assigning a set of predefined classes to a set of unlabeled instances. The classification task is characterized by well-defined classes, and a training set consisting of pre classified examples.
- **Estimation**. Similar to classification but here the model needs to be able to estimate a continuous value instead of just choosing a class.
- **Prediction**. This task is the same as classification and estimation, except that the instances are classified according to some predicted future behavior or estimated future value.
- **Affinity Grouping or Association Rules.** The task of affinity grouping is to determine which things go together. The output is rules describing the most frequent combinations of the objects in the data.
- **Clustering.** The task of segmenting or dividing a heterogeneous population into a number of more similar subgroups or clusters.
- **Profiling or Description**. The task of explaining the relationships exists and represented in a complex database.

These tasks can divided into the two major categories *predictive* and *descriptive* tasks [17]. Classification, estimation and prediction are all predictive in their nature, while affinity grouping, clustering and profiling can be seen as descriptive.

- **Predictive tasks.** The objective of predictive tasks is to predict the value of a particular attribute based on values of other attributes. The variable to be predicted is commonly known as the *target* or the *dependent variable*, while the variables used for prediction are known as *explanatory* or *independent variables*.
- **Descriptive tasks.** Here the objective is to derive patterns such as correlations, trends, clusters, and anomalies that summarize the existing relationships in the data. These tasks

are said to be always exploratory in their nature and also frequently requires post processing techniques to validate and explain the results.

Prediction is the most important task seen from a business perspective as it is the only task dealing with future values. However, prediction is nothing else but a special case of classification and estimation and most of the techniques defined for these tasks can also be applied to prediction.

2.2 Machine Learning

Today intelligent machines are doing different interesting things through Learning. Well understanding of the input would help in taking and giving better decisions in a more optimized way and also help them to work in most effective and efficient method. Instead of building different and enormous machines with explicit programming and different algorithms it is better to train the machine to understand the virtual environment and based on their understanding the machine will take particular decision. This will gradually decrease the number of programming concepts and the machine will become independent and take necessary decisions on their own.

Machine learning the discipline that originated at the intersection of many other subjects such as statistics, database science, and computer science. It is a powerful tool, capable of finding actionable insight in large quantities of data. Still, caution must be used in order to avoid common abuses of machine learning in the real world [27]. Machine learning is a learning technique or discipline in the area of artificial intelligence which deals with making intelligent decisions and drawing conclusions based on features of the input data rather than preset rules. The idea is to be able to solve problems where the algorithm is unknown, or the problem becomes unsolvable due to constantly changing requirements. By analyzing existing data we will be able to detect patterns and relations in the data. From the conclusions made with existing data, the algorithm can then “learn" to make accurate predictions when encountering new data for the same problem by producing general rules that can be used [27].

Machine learning can be described in terms of its data-power as a computer algorithm that search for patterns in data and the perfect tool for describing these patterns is naturally embedded in statistics; this makes the relationship between ML and statistics to be fairly significant [27].

Most algorithms use the concept of pattern recognition to make the most optimized decisions. As result of this new interest in machine learning we are experiencing a new era in data analysis and prediction techniques and their applications. This research paper emphasizes on different types of machine learning algorithms and their most efficient use to make informative decisions on Ethiopian custom and revenue authority data set. Different kinds of algorithm gives machine several learning experience and enable them to adapt other things from the environment. Based on these algorithms the machine takes the decision and performs the specialized tasks. So it is very important for the algorithms to be the most optimized and complexity should be reduced to gain more efficient decisions that the machine makes [27]. Fundamentally and scientifically these algorithms depends on the data structures used as well as theories of learning cognitive and genetic structures. But still natural procedure for learning gives great exposures for understanding and good scope for variety of different types of circumstances.

Predictive analytics also happens to be among the underlying technologies behind ML. It is the scientific way the past is used to predict the future with the aim of deriving a desired outcome, this outcome can then be used to analyze questions about future events and also, provide answers to business related questions based on probability.

ML algorithms operate by discovering patterns in data and then use the discovered patterns to develop a model that provides the summary of the important properties in the data. It can also figure out how to perform important tasks by generalizing from examples. ML algorithms and ML models are not the same thing [27][36]. The algorithm consumes the dataset and detects patterns that map the variables to the target.

Many machine learning algorithms are being borrowed from current thinking in cognitive science and neural networks [30]. Overall we can say that learning is defined in terms of improving performance based on some sort of measurement. In order to know whether an agent has learned, we must define a measure of success. The measure is not always indicates how well the agent performs on the training experiences, but how well the agent performs for new experiences.

2.2.1 Application and use of Machine learning

Machine learning is most useful and successful when it augments rather than replaces the specialized knowledge of a subject-matter expert. It helps the medical doctors to fight and eradicate cancer, assists engineers and programmers with their efforts to create smarter homes and automobiles, and helps social scientists to build knowledge of how societies function. Over all, Machine learning applied in many areas of business, scientific laboratories, hospitals, and governmental organizations. Any organization that generates or aggregates data likely employs and uses at least one type machine learning algorithm to help make sense of it [34].

According to Lantz it is impossible to list every use case and application of machine learning, the following includes a survey of recent success stories of several main applications areas:

- Identification of unwanted spam messages in e-mail
- Segmentation of customer behavior for targeted advertising
- Forecasts of weather behavior and long-term climate changes
- Reduction of fraudulent credit card transactions
- Actuarial estimates of financial damage of storms and natural disasters
- Prediction of popular election outcomes
- Development of algorithms for auto-piloting drones and self-driving cars
- Optimization of energy use in homes and office buildings
- Projection of areas where criminal activity is most likely
- Discovery of genetic sequences linked to diseases

On other hand, [12] summarized the benefits of machine learning as:

- To understand and improve efficiency and effectiveness of human learning. For example, it uses to improve methods for teaching and tutoring people, as done in CAI – Computer-aided instruction.
- To discover new things or structure that is unknown to humans. Example: Data mining
- To fill in skeletal or incomplete specifications about a domain. Large, complex Artificial Intelligence system cannot be completely derived by hand and require dynamic updating to incorporate new information.

2.2.2 Types of Machine Learning Algorithms

As the development of machine learning techniques, there are a number of algorithms available we can try. By the learning style, the machine learning algorithms can be mainly divided into the following two types. This taxonomy of machine learning algorithms considers the training data during the model preparation process for the purpose of getting the best result.

In addition machine learning algorithms are divided into different categories according to their purpose and use. Understanding the categories of learning algorithms is an essential task towards using data to drive the desired action [34].

A predictive model is used for tasks and techniques that involve the prediction of one value using other values in the dataset. The learning algorithm attempts to discover and model the relationship exist between the target feature and the other features. Despite of the common use of the word "prediction" to imply forecasting, predictive models need not necessarily mean foresee events in the future. To state more formally, given a set of data, a supervised learning algorithm attempts to optimize the model to get the combination of feature values that result in the target output. The regularly used supervised machine learning task of predicting which category an example belongs to is known as classification. It is easy to think of potential benefits for a classifier.

The target in classification feature which is to be predicted is a categorical feature and known as the class, and also it is further divided into categories called levels. A class can have two or more levels with in it, and the levels may or may not be ordinal. Classification algorithm is so widely used in machine learning data processing and based on its strengths and weaknesses suited for different types of input data.

Supervised learners can also be employed to predict numeric data such as income, laboratory values, test scores, or counts of items. Although regression models are not the only type of algorithms or numeric models, they are still the most widely used. Regression methods are widely used for forecasting, as they quantify in exact terms the association between inputs and the target variables, including both, the magnitude and uncertainty of the relationship.

A descriptive model is very important for tasks that would benefit from the insight gained from summarizing data in new and interesting ways. In contrary, predictive models that predict a target of interest, in a descriptive model, no single feature is more important than any other. In fact, because there is no target variable to learn, the process of training a descriptive model is called unsupervised learning. For example, the descriptive modeling task which called pattern discovery is used to identify relevant associations within data.

The descriptive modeling task of dividing a dataset into homogeneous or similar groups is called clustering. For example, if we have five different clusters of shoppers at X grocery store, the marketing team will need to understand the differences exist among the groups in order to create a promotion that best suits each group. The following table lists few types of machine learning algorithms. These are the few of the entire sets of machine learning algorithms exist and only described in this research thesis.

Model	**Learning tasks**
Supervised learning algorithms	
Nearest Neighbor	Classification
Naive Bayes	Classification
Decision Trees	Classification
Classification Rule Learners	Classification
Linear Regression	Numeric prediction
Regression Trees	Numeric prediction
Model Trees	Numeric prediction
Neural Networks	Dual use
Support Vector Machines	Dual use
Unsupervised Learning Algorithms	
Association Rules	Pattern detection
k-means clustering	Clustering

Table 2.1: types of machine learning algorithms [34]

2.2.2.1 Supervised Learning

The supervised learning abstraction in ML tries to look out for the mapping between inputs and outputs using already known results. A good instance is when trying to predict some unknown answers using data that contains answers that are readily available. After the data has been derived, the data is split into two random sets, training dataset and testing dataset often in ratio 70:30 respectively. The 70 percent is used for training a robust model what it needs to learn from; while the 30 percent is used to test the model to be sure the model behaves well with new datasets [4].

After selecting and cleaning the dataset to conform with ML requirements, a point is selected in the dataset that shows this is the right answer which is also known as label or target, then an algorithm one or more is chosen to compute the outcome, and finally, run the program on the dataset to see if the correct answer is gotten from the testing data. The final output becomes a model that can be used against more data to get needed answers based on probability [27][36]. A good example of supervised ML algorithm is the classification algorithms. The algorithm is used to arrange data into different classes that can then be used to predict discrete variables based on some other attributes in the datasets.

A predictive model learns to predict one value by using other values in the data set to model the relationship among the target feature (the feature being predicted) and the other features. The model is given clear instruction on what to learn and how to learn it and therefore the training of a predictive model is called supervised [34]. One of the most common uses of supervised machine-learning tasks is predicting which category an example belongs to. This is known as classification and the model trained for this task is called a classifier. Supervised learning can be summarized in the following steps:

- Training - train the model with the labeled training set.
- Validation (optional) - tune the parameters of the model.
- Testing - test the performance of the model against the test set.
- Application - apply the model to real-world data.

2.2.2.1.1. Nearest neighbor

Nearest neighbor classifiers defined by their characteristic of classifying unlabeled examples. They are assigning the classifiers of the class of similar labeled examples. Despite the simplicity of this idea, nearest neighbor methods are extremely powerful. They have been used successfully for:

- Computer vision applications, including optical character recognition and facial recognition in both still images and video
- Predicting whether a person will enjoy a movie or music recommendation
- Identifying patterns in genetic data, perhaps to use them in detecting specific proteins or diseases

In general, nearest neighbor classifiers are well-suited for classification tasks, where relationships among the features and the target classes are numerous, complicated, or extremely difficult to understand, yet the items of similar class type tend to be homogeneous. Another way of putting it would be to say that if a concept is difficult to define, but you know it when you see it, then nearest neighbors might be appropriate. On the other hand, if the data is noisy and thus no clear distinction exists among the groups, the nearest neighbor algorithms may struggle to identify the class boundaries [34].

The nearest neighbors approach to classification is exemplified by the k-nearest neighbor's algorithm (k-NN). Although this is perhaps one of the simplest machine learning algorithms, it is still used widely. The strengths and weaknesses of this algorithm are as follows:

Strengths	**Weaknesses**
• Simple and effective • Makes no assumptions about the underlying data distribution • Fast training phase	• Does not produce a model, limiting the ability to understand how the features are related to the class • Requires selection of an appropriate *k* • Slow classification phase • Nominal features and missing data require additional processing

Table 2.2 strengths and weaknesses of KNN algorithm [34]

The k-NN algorithm gets its name from the fact that it uses information about an example's k-nearest neighbors to classify unlabeled examples. The letter *k* is a variable term implying that any number of nearest neighbors could be used. After choosing *k,* the algorithm requires a training dataset made up of examples that have been classified into several categories, as labeled by a nominal variable. Then, for each unlabeled record in the test dataset, k-NN identifies *k* records in the training data that are the "nearest" in similarity. The unlabeled test instance is assigned the class of the majority of the k nearest neighbors.

2.2.2.1.2 The Naive Bayes algorithm

The Naive Bayes algorithm describes a simple method to apply Bayes' theorem to classification problems. Although it is not the only machine learning method that utilizes Bayesian methods, it is the most common one. This is particularly true for text classification, where it has become the de facto standard. The strengths and weaknesses of this algorithm are as follows:

Strengths	Weaknesses
• Simple, fast, and very effective • Does well with noisy and missing data • Requires relatively few examples for training, but also works well with very large numbers of examples • Easy to obtain the estimated probability for a prediction	• Relies on an often-faulty assumption of equally important and independent features • Not ideal for datasets with many numeric features • Estimated probabilities are less reliable than the predicted classes

Table 2.3 strengths and weaknesses of Naive Bayes algorithm [34]

The Naive Bayes algorithm is named as such because it makes some "naive" assumptions about the data. In particular, Naive Bayes assumes that all of the features in the dataset are equally important and independent. These assumptions are rarely true in most real-world applications. For example, if we were attempting to identify spam by monitoring e-mail messages, it is almost certainly true that some features will be more important than others. The e-mail sender may be a more important indicator of spam than the message text. Additionally, the words in the message body are not independent from one another, since the appearance of some words is a very good indication that other words are also likely to appear. A message with the word Viagra will probably also contain the words prescription or drugs.

2.2.2.1.3 Decision Tree

Decision tree learners are one of the powerful classifiers, which utilize a tree structure to model the relationships exist among the features and the potential outcomes. The decision tree classifier uses a structure of branching for decisions, which channel leads into a final predicted class value. To understand how this classification algorithm works in practical way, let's discuss here the tree

that predicts whether a job offer accepted or rejected. A job offer to be considered starts at the root node, when it is then passed through decision nodes that require choices to be made based on the attributes of the job. These choices divide the data among branches that indicate potential outcomes of a decision which is yes or no outcome values. Even though, there may be more than two possibilities in some cases. In this case a final decision or result can be made after the tree is terminated by terminal nodes or leaf node that indicates the action to be taken as the result of the series of decisions. In a predictive model, the leaf nodes provide the expected result given the series of events in the tree.

A great benefit of decision tree algorithms is that the flowchart-like tree structure is not necessarily exclusively for the learner's internal use. After the model is developed, many decision tree algorithms output as the resulting structure in a human-readable format. This provides an insight into how and why the model works or doesn't work well for a particular task. This also makes decision trees algorithm particularly appropriate for applications in which the classification mechanism needs to be transparent for legal reasons, or the results need to be shared with others in order to inform future business practices. With this in mind, some potential uses include:

- Credit scoring models in which the criteria that causes an applicant to be rejected need to be clearly documented and free from bias
- Marketing studies of customer behavior such as satisfaction or churn, this will be shared with management or advertising agencies
- Diagnosis of medical conditions based on laboratory measurements, symptoms, or the rate of disease progression

Although the previous applications demonstrated the value of trees in informing decision processes, this is not to say that their utility ends here. In fact, decision trees are the single most widely used machine learning algorithm, and can be applicable to model almost any type of data with excellent out of the box applications. Sometimes, it is worth noting in few scenarios where trees may not exactly fit. One such case might be a task where the data has a large number of complex features with many levels or it also has a large number of numeric features. These cases may result in a very large number of decisions and an overly complex tree. They may also contribute to the tendency of decision trees to over fit data, though as we will soon see, even this weakness can be overcome by adjusting some simple parameters.

Decision trees built using a heuristic called recursive partitioning. This approach is also known as divide and conquers algorithm because it divides the data into subsets, which are then split repeatedly into even smaller and smaller subsets. Finally the algorithm determines the data within the subsets are sufficiently homogenous, or another stopping criterion has been met. There are numerous implementations of decision trees, but one of the most well-known implementations is the C5.0 algorithm. This algorithm was developed by computer scientist J. Ross Quinlan as an improved version of his prior algorithm, C4.5. The source code for a single-threaded version of the algorithm was made publically available, and it has therefore been incorporated into programs such as R and R studio.

The C5.0 algorithm has become the industry standard to produce decision trees, Compared to other advanced machine learning models and algorithms, such as *Neural Networks and Support Vector Machines*, the decision trees built by C5.0 generally perform nearly as well, but are much easier to understand and deploy. Additionally, as shown in the following, the algorithm's weaknesses are relatively minor and can be largely avoided:

Strengths	Weaknesses
• It is all-purpose classifier that does well on most problems • Highly automatic learning process, which can handle numeric or nominal features, as well as missing data • Excludes unimportant features • Can be used on both small and large datasets • Results in a model can be interpreted without a mathematical background (for relatively small trees) • More efficient than other complex models	• Decision tree models are often biased toward splits on features having a large number of levels • It is easy to over fit or under fit the model • It can have trouble modeling some relationships due to reliance on axis-parallel splits • Small changes in the training data can result in large changes to decision logic • Large trees can be difficult to interpret and the decisions they make may seem counterintuitive

Table 2.4 strengths and weaknesses of Decision tree algorithms [34]

2.2.2.1.4 Classification rule learners

Classification rule learners represent knowledge in the form of logical (if else) statements that assign a class to unlabeled examples. They are specified in terms of an antecedent and a consequent; these form a hypothesis stating that "if this happens, then that happens." For example a simple rule might state, "if the hard drive is making a clicking sound, then it is about to fail." Rule learners are often used in a manner similar to decision tree learners. Similar to decision trees, classification rule learners can be used for applications that generate knowledge for future action, such as:

- Identifying conditions that lead to a hardware failure in mechanical devices
- Describing the key characteristics of groups of people for customer segmentation
- Finding conditions that precede large drops or increases in the prices of shares on the stock market

On the other hand, rule learners offer some distinct advantages over trees for some tasks. Unlike a tree, which must be applied from top-to-bottom through a series of decisions, rules are propositions that can be read much like a statement of fact. Additionally, for reasons that will be discussed later, the results of a rule learner can be more simple, direct, and easier to understand than a decision tree built on the same data.

2.2.2.1.5 Linear Regression

In mathematics, linear regression is a statistical model to evaluate the linear relationship exists between a dependent variable *y* and one or more independent variables *X*. Given that a dataset $\{x_{i1}, x_{i2}, \ldots, x_{ip}\}_{i=1}^{n}$ of *n* observations, the linear regression model takes the form:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i = X_i^T \beta + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{2.1}$$

Where *yi* represents the continuous numeric response for the ith observation, β_j is the regression coefficient for the jth variable, x_{ij} shows the jth variable for the ith observation, and ε_i is called the random error or the noise that is not able to be explained by the linear model. The above equation can also be written in vector form as follow:

$$Y = X\beta + \varepsilon \tag{2.2}$$

The common objective of the linear regression models is to find estimates of the regression coefficient vector β so that the mean squared error (MSE) can be minimized, according to the Variance-Bias trade-off. In general, the first advantage of this model is that it possesses high interpretability of the regression coefficients, relationship between each regression coefficient and the last response, even between different regression coefficients, can be clearly interpreted in this kind of model. The second is that as long as certain assumptions about the model residuals' distribution are met, we can directly make use of the existing statistical nature inside to get the standard errors of the regression parameters, and evaluate the performance of the predictive model. However, because of the high interpretability [31], it is required that relationship between each estimate of the parameter and the last response should fall along a flat hyperplane. For instance, if there is only one variable in the model, the relationship between the variable and the response should be linear in a straight line. Thus, the nonlinear relationship between the regression coefficients and the predicted response cannot be explained in this model.

2.2.2.1.5.1 Ordinary Linear Regression

The ordinary linear regression seeks to find appropriate estimates of the regression coefficients (i.e. the hyperplane) $\hat{\beta}$ so that the SSE (Sum of Squared Errors) or the bias between the predicted value $\hat{y}_i$ and the observed outcome y_i can be minimized, in which the definition of SSE can be shown as follow:

$$SSE = \sum_{i=1}^{n} \left(\hat{y}_i - y_i \right)^2 \tag{2.3}$$

The optimal regression coefficients $\hat{\beta}$ can also be described by the vector form:

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T y \tag{2.4}$$

The above equation is easy to implement, and is straightforward to tell that the estimates of the regression coefficients with minimized SSE are the ones with minimized bias. But it is worth to note that the matrix $\left(X^T X \right)^{-1}$ in the equation, which is proportional to the covariance matrix of the regression coefficients, is uniquely existed only under the circumstance that the number of the

observations is greater than that of the regression coefficients and the regression coefficients are with no relationship, i.e. independent from each other. But the unique set of the regression coefficients can still be gained by a conditional inverse of $(X^T X)$ or removing the linear relationship among the variables [14]. And if the number of the observations is not greater than that of the regression coefficients, the PCA (Potential Component Analysis) pre-processing can be conducted to reduce the dimension of the variables. As linear regression is not able to interpret the nonlinear relationship among the variables in the model, before implementing this model, we need to check if nonlinear or curvature relationship exist between the variables and the predicted response by some basic scatter plots of the observed outcome versus the predicted response and/or the residuals versus the predicted response.

The third problem with ordinary linear regression is that it is sensitive to the outliers, which are far away from the overall tendency of the majority dataset. Because the objective of the ordinary regression model is to find the estimates of the parameters with minimized SSE/bias, the model has to adjust the estimates of the regression coefficients to better fit the outliers, whose residuals between the observed outcome and the predicted response are extremely large. So that it is possible that a small number of outliers in the dataset have great influence on the performance of the linear regression model.

2.2.2.1.5.2 Partial Least Squares

As we have mentioned in last section, if the variables in the dataset are highly correlated or the number of the variables is greater than the number of the total observations, the ordinary linear regression model will not get a unique set of parameters with minimized bias, but still get high variance. In order to solve this problem, two methods were proposed [15]: (1) remove the highly correlated variables; (2) conduct PCA dimensional reduction. But the former may be unstable, and the latter just simply focuses on the variability of the variables without considering the predicted response, and it may reduce the interpretability of the new regression coefficients after PCA pre-processing. The Principal Component Regression (PCR) model [28], which is developed on the PCA, can only be used when the variability of the regression coefficients' space and that of the predicted response are correlated. Therefore, the Partial Least Squares (PLS) regression algorithm is recommended when the variables in the dataset are correlated but the linear regression model is

required. The main idea of the PLS regression model is to find a new set of potential components, which is able to explain the covariance between the matrix *X* and *Y* as much as possible, by decomposing both *X* and *Y* [1]. At first, the independent variables' matrix *X* is decomposed as follows:

$$X = TP^T + E \tag{2.5}$$

Where *T* is the projection of *X* (i.e. the X score matrix), *P* represents the orthogonal loading matrix (not orthogonal in PCR), and E is the error or noise term. Given that *B* is the diagonal matrix of the "regression weights", thus, the predicted response can be shown like the following:

$$\hat{Y} = TBC^T \tag{2.6}$$

In contrast with PCA, it just finds out the linear relationship that maximally gives out the variability of the variables, but PLS needs one more step to find out the linear components that maximally correlates with the response.

It is worth emphasizing that the attributes should be centered and scaled before implementing the PLS model, and the number of the components to retain, as the only one tuning parameter, can be determined by the resampling techniques used.

2.2.2.1.6 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a family of powerful nonlinear regression models inspired by the working principal of biological neural networks, which are capable of solving a wide variety of problems where the relationships may be quite dynamic or nonlinear. Similar to the Partial Least Squares in the linear regression models, the typical Artificial Neural Networks are organized by different layers, and each layer is made of a number of interconnected "units" that contain an "activation function". The input data are sent to the input layer, and processed in a forward direction through one or more hidden layers, and the last output of the ANN model is generated at the output layer [16].

Each unit in the hidden layers is a linear combination of some or all the variables in the previous layer. Each of the hidden units is not estimated, directly, but transformed by a nonlinear function (i.e. the activation function), e.g. logistic function:

$$h_k(X) = g\left(\beta_{0k} + \sum_{i=1}^{P} x_j \beta_{jk}\right), g(u) = \frac{1}{1+e^{-u}} \tag{2.7}$$

The coefficients jk represents the contribution of the jth variable on the kth hidden unit. After defining the number of the hidden units, the predicted response in the output layer can be shown as follows:

$$f(X) = \chi_0 + \sum_{k=1} \chi_k h_k \tag{2.8}$$

Giving that the number of the initial input variables is P , the number of the hidden units is H , therefore, the total number of the regression coefficients being estimated is

$$H*(P+1)+H+1. \tag{2.9}$$

As there are large numbers of regression coefficients in the model, the model is prone to over-fit; one approach to solve this over-fit problem is to regularize the model by adding a penalty for the large parameters. Thus, the objective of the optimization problem can be presented by the following mathematical equation [31]:

$$\min \quad \sum_{i=1}^{n}\left(y_i - f_i(x)\right)^2 + \lambda\left(\sum_{k=1}^{H}\sum_{j=0}^{P}\beta_{jk}^2 + \sum_{k=0}^{H}\chi_k^2\right) \tag{2.10}$$

The greater the regularization value is, the less likely the model to over-fit. Generally speaking, the given value of can be set between 0 and 0.1. During the data pre-processing, at first, there are two tuning parameters for the Artificial Neural Networks regression model: the value of the regularization parameter and the number of the hidden units. Secondly, all the variables in the dataset should be centered and scaled because the estimates of all the parameters are being summed. At last, the reasonable feature selection technique, such as Principal Component Analysis (PCA), should be conducted to remove the effect of the variables, which are highly correlated with other variables in the dataset.

2.2.2.1.7 Support Vector Machines

Support Vector Machines are considered as one of the best binary classifiers. They create a decision boundary such that most points in one category fall on one side of the boundary while most points in the other category fall on the other side of the boundary. For example consider an

n-dimensional feature vector x = (X1,…,Xn) [26]. We can define a linear boundary (hyperplane) as

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n = \beta_0 + \sum_{i=1}^{n} \beta_i X_i = 0 \tag{2.11}$$

Then elements in one category will be such that the sum is greater than 0, while elements in the other category will have the sum be less than 0. With labeled examples, $\beta_0 + \sum_{i=1}^{n} \beta_i X_i = y$, where y is the label. In our classification, $y \in \{-1, 1\}$. We can rewrite the hyperplane equation using inner products.

$$y = \beta_0 + \sum \alpha_i y_i x(i) * x \tag{2.12}$$

The optimal hyperplane is such that we maximize the distance from the plane to any point. This is known as the margin. The maximum margin hyperplane (MMH) best splits the data. However, since it may not be a perfect differentiation, we can add error variables e1…en and keep their sum below some budget B. The important element is that only the points closest to the boundary matter for hyperplane selection; the rest all others are irrelevant. These points are known as the support vectors, and the hyper plane is known as a Support Vector Classifier (SVC) since it places each support vector in one class or the other.

The concept of the SVC is similar to another popular linear regression model, Ordinary Least Squares (OLS), but the two optimize different quantities. The OLS finds the residuals, the distance from each data point to the fit line, and minimizes the sum of squares of these residuals [26]. On the other hand, the SVC looks at only the support vectors, and uses the inner product to maximize the distance from hyper plane to the support vector. In addition, the inner products in SVC are weighted by their labels, but in OLS the square of residuals serves as the weighting. SVMs fix this by applying non-linear kernel functions to map the inputs into a higher-dimensional space and linearly classify in that space. A linear classification in the higher-dimensional space will be non-linear in the original space. The SVM replaces the inner product with a more general kernel function K which allows the input to be mapped to higher-dimensions. Thus in an SVM, $y = \beta_0 + \sum \alpha_i y_i K(x(i), x)$

2.2.2.2 Unsupervised Learning

In the unsupervised learning abstraction, developing ML solutions becomes much more difficult when compared with supervised learning because the ML algorithm is not provided with known input-data and known outcomes. It has to analyze the data by itself and get common patterns and structures to develop a predictive model. This technique could be likened to density estimates in statistics where there are attempts to find patterns in the input data. A good example of unsupervised learning algorithm is clustering algorithm and anomaly detection. Clustering algorithm discovers natural hidden patterns and classifications in datasets by grouping the data-points based on similarity. The discovered patterns and classifications can then be used to predict the class for a given variable. To better make clustering algorithm to be more functional, a vast amount of data that can form clusters is needed to objectively describe the data [27][36].

In contrary to predictive models that predict a target feature, descriptive models give no special importance to any single feature. Because there is no target to learn, the process of training a descriptive model is called unsupervised [34]. A descriptive model tries to summarize data by dividing it into homogeneous groups. This is known as clustering or cluster analysis. This can be used for segmentation discovery where groups that are generally similar in some way can be identified in the data set.

2.3 Machine Learning vs. Data Mining

Most data mining techniques come from the field of Machine Learning (ML), the field of machine learning (ML) is concerned with the question of how to construct computer programs that automatically improve with experience. The algorithms and techniques used within ML exploit ideas from many fields including statistics, artificial intelligence, philosophy, information theory, biology, cognitive science, computational complexity and control theory. ML also has many other applications than data mining, e.g. information filtering and autonomous agents etc [49].

There is a significant overlap between ML and DM. These two terms are commonly confused because they often employ the same methods and therefore they overlap significantly. One of ML specialist [49] defined ML as a “field of study that gives computers the ability to learn without being explicitly programmed.” ML focuses on classification and prediction, which based on

known properties previously learned from the training data. ML algorithms need a goal (problem formulation) from the domain (e.g., dependent variable to predict). DM focuses on the discovery of previously unknown properties in the data. DM does not need a specific goal or objectives from the domain, but instead focuses on finding new and interesting knowledge.

ML approach in data mining usually consists of two phases: training and testing. Often, the following steps are performed:

- Identify class attributes (features) and classes from training data.
- Identify a subset of the attributes necessary for classification (i.e., dimensionality reduction).
- Learn the model using training data.
- Use the trained model to classify the unknown data.

Whereas, DM followed CRISP-DM six processes like business understanding, Data understanding, Data preparation, modeling, evaluation and finally deployment steps.
However, as this research is focused on how to combine different data sources and how to find meaningful Patterns, only ML techniques applicable to data mining tasks will be covered.

2.4 Problem faced in learning

Learning is a complex process since a lot of decisions are made and also it depends from machine to machine and from algorithm to algorithm. In order to understands a particular problem and on understanding the problem how it responds to that particular problem. Some of the issues make a complex situation for the machine to respond and react. These problems may make problem complex and it also affects the learning process of the machine. As the machine is dependent on what it perceives, the perception module of the machine learning should also focus on different types of challenges and environment which it will face, as different input can produce different outputs and the most appropriate and optimize output should be considered by the machine. Some of the common problems faced during the learning process are indicated as follows:-

Bias- the first problem we observed is bias. It is the tendency to prefer one hypothesis over another. For example consider the agents *X* and *Y*. Saying that a hypothesis is better than *X*'s or *Y*'s hypothesis is not something that is obtained from the data - both *X* and *Y* accurately predicts all of the data given in the experiment. Without a bias, an agent will not be able to make any sort

predictions on unseen examples. The hypotheses adopted in this example by *Y* and *X* will disagree on all further examples, and, if a learning agent cannot choose some other hypotheses as better, the agent will not be able to resolve this disagreement easily. To have any inductive process make predictions on unseen data, an agent requires a bias. What constitutes a good bias is an empirical question about which biases work best in practice; we do not imagine that either *Y*'s or *X*'s biases work well in practice.

Noise-In most real world situations, the data are not perfect. In the data some of the features have been assigned the wrong value and this considered as noise, the features given do not predict the classification, and often there are examples with missing features. One of the important properties of a learning algorithm is its ability to handle noisy data in all of its forms.

Pattern Recognition-This is another type of problem faced during machine learning process. In general, it aims to provide reasonable results for all possible inputs and to perform "closest to" or matching of the inputs by taking into account their statistical variations. This is different from pattern matching algorithms which match the exact values and dimensions. As algorithms have well-defined values like for mathematical models and shapes like different values for rectangle, square, circle etc. It becomes difficult for machine to process those inputs which have different values. This is the major problem faced by most of the machine learning process and algorithms.

2.5 Related works

2.5.1 International related works

The first paper reviewed in using machine learning for exploratory data analysis and prediction model on large set of datasets [47] is discussed about supervised learning one of machine learning methods. In his research he gave more emphasis to classical type of regression model like linear regression, nonlinear regression and regression tree. He used to improve the performance of the training model through data pre-processing and resampling techniques, including data transformation, dimensionality reduction and k-fold cross-validation. During the implementation of machine learning algorithms, he used three typical models such as Ordinary Linear Regression, Artificial Neural Networks and Random Forest. These have been trained by the different packages in R on the given large datasets. He also built several small models on small number of samples in the dataset to get the reasonable tuning parameters, and the optimal models were chosen by the value of RMSE and R2 among several training models. The quantitative and visual results of

practical implementation showed the feasibility for the large datasets under the artificial neural network and random forest algorithms. Finally when he compared with the ordinary linear regression model, the performance of the artificial neural network and random forest models are greatly improved. He suggested that the right choice between different models relies on the characteristics of the dataset and the goal, and also depends upon the cross-validation technique and the quantitative evaluation of the models.

The second international paper reviewed in customs and revenue area was a master thesis by [7] a data mining framework to detect tariff code circumvention in Turkish customs database in order to detect "Tariff Code Circumvention". For this study purpose, four types of products, which are the top circumvented goods in the Turkish customs, have been chosen by the researcher. The researcher First, tried to identify with the help of Risk Analysis Office, the significant features. Then, ranking these identified features by expert by using algorithm. Finally, KNN machine learning algorithm is applied on the Turkish customs database in order to identify the circumvented goods automatically. The results show that the developed framework is able to find such circumvented goods successfully. And the researcher recommend as SVM (Support Vector Machines) could be performed on customs data set.

The third research work reviewed here is a research that focuses on international trade. It is deals on GDP Forecasting through Data Mining of Seaport Export-Import Records [49]. In this paper the problem of GDP forecasting was analyzed with Seaport Export-Import data records as the learning resource. The usability of Learning Algorithms and problem formulation for these methods has also been discussed. The importance of GDP forecasting has been emphasized in the paper. It can also be viewed as a quantitative indicator of the effect of Export Import injection on the economy of a nation. The learning algorithms discussed are the Support Vector Machines, Fuzzy Set Based Genetic Learning Algorithms and also the Artificial Neural Network methods. The dataset for the exercise consists of daily export and import at a given port. Several ports in the country of interest are then considered [49].

The fourth paper focuses on the particular problem of speedy and accurate processing of customs declarations [50]. It present a novel use of graph based spreading activation algorithm for the automated processing of customer declarations for commercial goods, based on supervised

learning. The researcher used the method to allow him build recommender systems for use by customs officers, traders, carriers and insurers. He examine the particular use case of the recommendation to assign or not assign an armed escort to a shipping vehicle in cases of elevated risk of theft. In contrast to the usual risk based approach, this algorithm is trained solely on shipment data rather than on traditional risk indicators. For the research learning is done in two distinctive stages. First of all, He builds the network from data and secondly, he use graph algorithms to find patterns in the network. The feasibility of the approach was tested by application to 2500 customs records collected during a continuous period of one month at eight border checkpoints between Russian Federation and two EU countries. The algorithm achieved 100 % accuracy under experimental conditions [50].

2.5.2 Local related Research

The first local research reviewed was done on Student Performance Prediction Model using Machine Learning Approach [8]. The researcher used dataset for their study that taken from the Wolkite university registries office for college of computing and informatics from 2004 up to 2007 E.C. Data collected were related student's transcript data that included their final GPA and their grades in all courses. After pre-processing the data, the researchers applied the machine learning methods such as neural networks, Naive Bayesian and Support Vector Machine (SMO). Finally, they built the model for each method, evaluated the performance and compared the results of each mode. The result showed Naive Bayesian had higher performance when they compared with the other two selected methods (SMO and MLP).

In addition to the above researches that used machine learning the researcher tried to review research works that has been done in Ethiopian customs and revenue area so far. Most of the related works in this area are done by using data mining techniques. The first work in this regard is a research that was conducted on the application of data mining on Fraud protection [35]. The study exploring the potential applicability of the data mining technology in developing models which can detect and predict fraud happening in tax claims with a particular emphasis to Ethiopian Revenue and Custom Authority. The research was tried to apply the clustering algorithm followed by classification techniques for developing the predictive model, K-Means clustering algorithm is employed to get the natural grouping in to fraud and non-fraud results. The result the researcher got from cluster was then used in developing the classification model. The classification task of

this study was carried out using the J48 decision tree and Naïve Bayes algorithms to create model that best predict fraud suspicious tax claims [35]. The model developed using the J48 decision tree algorithm was showed highest classification accuracy of 99.98%. This model was then tested by using 2200 dataset and the result scored accuracy of 97.19% prediction. The results of the study had showed that the data mining techniques are important for tax fraud detection [35].

The other research conducted in this area was knowledge discovery in ERCA customer segmentation [9]. The research aimed at the test of clustering and classification of data mining techniques to support Customers Relation Management. The research collected different customer information from Existing ASQUDA database. By using k-mean clustering algorithm the data was clustered and the output used for classification model by J48 decision tree. The researcher approved the classification model performance as it showed 99.95% accuracy [9].

2.6 Summary

In this chapter, the concepts of data mining introduced, followed by what machine learning is and its application. In addition, types of machine learning algorithms including K-Nearest Neighbors (KNN), Decision tree, Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), linear regression such as Ordinary Linear Regression (OLR), and Partial Least Squares (PLS) are discussed. The basic principal, strengths and weaknesses of each particular model are also illustrated in this section.

Also the problem faced during machine learning and the difference between data mining and machine learning is elaborated. Finally, related researches from international and local are reviewed. This research aimed in mining the hidden knowledge in very large volume and complex customs imported goods dataset by using machine learning algorithms. It plans to use numerical regression and prediction method to show the best model and measures its performance of the imported goods data set patterns in the customs imports goods area.

CHAPTER THREE

METHODOLOGY AND DATA INTRODUCTION

The main purpose of this chapter is to design the methodology to carry out the study. This is the procedure used in evaluating the various predictive machine learning techniques and algorithms using the imported goods data sets from ERCA. The chapter also deals with the explanation of the software tool employed to apply different algorithms for data preprocessing, feature selection, dimensionality reduction, classification, regression and prediction. In addition, the instruments used to know the business environment are data collection using interview and review of relevant literature including similar studies.

3.1 Research design

The research design and methodology used in this study is based on the design science paradigm in Information System research [21]. The design science research is problem-solving paradigm that seeks to create viable artifacts in the form of a construct, a model, or a method, which provide solutions for problems. This could be problem identification, defining the objective for the solution, Design and development, and finally evaluation of the result. Based on this methodology the chapter examines the dataset used, steps of preprocess employed, algorithms used for data machine learning tools, evaluation metrics and performance measure of the algorithms with regard to the feature selection and processes to be carried out. The following framework depicts the research methodology the researcher employed throughout this research.

3.2. Literature Review

The researcher conducted literature review to assess Data mining, machine learning algorithm, patterns and research work in this field. Different books, journals, articles and papers from the Internet consulted to understand the practice of customs imported goods data, the potential applicability of machine learning algorithms and tools in the Revenue and custom area.

3.2 The data set

3.2.1 Data set collection

This study used data obtained from Ethiopian Revenue and custom authority. The total dataset for the study contains three year's Imported goods datasets from 2014-2016. We decided to collect for three years because of two reasons. The first reason is the items imported may be vary from time to time and secondly, the measuring for only one year dataset may not show the intended result. Based on the availability of the data, for this specific study a total number of 130,000 imported Items described with 15 attributes were collected. According to the variable definitions for dataset, this dataset has information related to date of import, commodity harmonization code, description of imported goods, the origin of the goods, their quantity, net and total weight, CIF in Ethiopian birr and dollar, total tax in Ethiopian birr and USD.

3.2.2 Sampling Techniques

The researcher has found easy to use MS Excel features for preprocessing; i.e. data cleaning and preparation is conducted while the data is in MS Excel format. After cleaning the data set, a total number of 100 340 customs imported items description data becomes ready for the process. Since this amount of data is very large for the machine learning process, we decide to take 10% of the data set by random sampling method. The researcher decides to use this small sample population to avoid computer memory overload for vector data during machine training.

3.2.3 Characteristics of datasets

In some data processing the dataset used could be large and complex. When we are talking about large and complex data here, we mean that the data is too large to fit in memory on a common computer or the machine learning tool. Either the data in its original form is too large and complex to fit in memory or the data is too big after necessary pre-processing and feature expansion. When the data becomes too large, training the model is simply not feasible. In this case, the most common approach will be to use aggregated features during preprocessing. In this research also we reduce the dimension of this complex and large customs imported goods data to the size that is suitable for R and Rstudio processing capacity.

3.2.4 Cleaning the data

In ML, data makes majority of the process during the experiment. So it is very important to understand how to format and prepare data to meet the best requirements for the ML tools and services. As is common in many real world data sets, the data set we use in this study contains several instances that have *missing values* for some attribute. According to domain experts in the agency data can be missing for several reasons; the value was not recorded, or the value could be missing due to some other error. Whatever the cause, it is very important to realize that missing values have impact on the final model depending on how they are handled. Most machine learning algorithms are based on the assumption that no data is missing and require that missing values are removed prior to modeling.

There are several strategies that could be taken to handle missing values. Instances with missing values could be removed, missing values can be replaced with a certain value not present it the data or they can be replaced with a value that is representative for the data set. However all strategies have their own drawbacks and which one to choose has to be decided from case to case. In this research we decide to remove all missing values because since it is large data replacing the missing value is time taking process.

In addition, before loading dataset into any ML service, the dataset could be saved in either a database, excel spreadsheet, or in the comma-separated value (.csv) format. The .csv file format is like a table in a database with rows and columns. Each observation contains attributes that are separated by a comma. Most times the first row is used for headers to describe the contents in the columns. The .csv file must satisfy the following conditions [32]:

- It must consist of one observation per line.
- Each observation must be terminated with an end-of-line character.
- The attribute values in each observation must be separated by a comma.
- It must be in plain text using character set like EBCDIC, Unicode, and ASCII.
- Each observation must have the same number of attributes and in the same sequence.
- End-of-line characters must not be included in attribute values, even if the attribute value is enclosed in double quotes.

With this dataset, ML model developed that will try to identify potential class of data set that is most likely to respond favorably to custom imported goods. This will help to know how to focus on international trade of what to import towards the predicted class of Goods. To do this the following activities were done:

- Even though it is possible to preprocess the data like data cleaning, normalization the data, and data size reduction etc. in R and Studio, the researcher has found easy to use MS Excel for preprocessing; i.e. data cleaning and preparation is conducted while the data is in MS Excel format. To view some outlier value and /or very little values, the Data Filter feature of MS Excel is used.
- Then, the data is saved as CSV file format in which the values are saved in comma delimited format file.
- The data that was converted into CSV file format has been imported in to R and Studio and used for the experimentation.

3.2.5 Features of Data

In examining the data on these custom imported goods, the issues, such as the availability of necessary records, standard data entry and so on, are taken into account. While deciding the features, the Knowledge of domain experts from ERCA was also considered. As a result of evaluation, it was decided to use the features available in Table 3.1.

Attributes	**Description**	**Data type**
IM_YEAR	Year in which the goods imported	date
IM_MONTH	Month in which the goods imported	date
CPC	Customs Procedure Code of the items	Character
HS_CODE	Harmonized System Code of the items	Character
HS_DESCRIPTION	Harmonized System Description of the items	Character
ORIGIN_COUNTRY	Country of origin of the imported items	Character
CONSIGN_COUNTRY	Country of Consignment	Character
QUANTITY	Quantity of items imported	numeric
UNIT	Unit of items imported	character
GROSS_WT_KG	Gross weight of items imported in KG	numeric

Attributes	Description	Data type
NET_WT_KG	Net weight of items imported in KG	numeric
CIF_VALUE_ETB	CIF Value of the items in USD	numeric
CIF_VALUE_USD	CIF Value of the items in USD	numeric
T_TAX_ETB	Total tax of the imported items in ETB	numeric
T_TAX_USD	Total tax of the imported items in USD	numeric

Table 3.1: Features in the Dataset

3.3 Modeling tools

The issue of choosing or selecting of the right tool for data processing tasks depends on dataset structure and the problem to be solved. According to Karina et al (2010) opinion that two parameters are essential for choosing the right tool with the right techniques which include main goal of the problem to be solved and the structure of the available data sets. Ralf and Markus [40] are also proposed criteria for tool selection based on different user groups, data structures, data processing tasks and methods, visualization and interaction styles, import and export options for data and models, platforms, and existing license policies. On the other hand, Abdullah et al (2011) has conducted a comparative study between a number of some of the free available data processing and knowledge discovery tools and software packages. The performance of the tools for the classification algorithm is affected by the kind and type of dataset used and by the way the classification Algorithms were implemented within the tools.

Tools selection has been done based on the analysis of problems to be solved in the custom and revenue sector. The most significant factor in tool selection is the fact that most of the public or government organizations have very limited financial resources and cannot afford to pay for commercial licenses of statistical readily available solutions. Now days, the prices for licensed software are very high and because of that the decisive factor in tool selection is that the solution is built on open-source products. The other reason in tool selection is that existing ASQUDA-database system stores data and the major analytical problems relate to processing data from this database and visualization of the analytical results. After the investigation of available statistical tools, it was found out that the R-environment and Rstudio fits the requirements well.

3.3.1 R environment

For the purposes of this study R software package was used, that is free open software [41]. R environment incorporates all types of the standard statistical tests, models, and analyses, as well as it provides a comprehensive language for managing and manipulating data. It allows anyone to use and, importantly, to modify it. R plays well with many other tools, importing data, for example, from CSV, SAS, and SPSS, or directly from Microsoft Excel, Microsoft Access, Oracle, and MySQL. It can also produce graphics output in PDF, JPG, PNG, and SVG formats. R is licensed under the GNU General Public License, with copyright held by The R Foundation for Statistical Computing. R language provides a great range of data processing packages.

3.3.2 R Studio

R studio is an IDE for user to run R in more efficient and easy manner. It is open source software. R compiles and runs on various platforms of UNIX, Mac OS and Windows.

3.4 Machine learning Algorithms used

There are many learning algorithms useful for data set classification or regression purposes. Such algorithms may be more appropriate for certain application domain and type or size of data than the other. In order to select or evaluate the best algorithm the following lists are helpful [18]:

- Predictive accuracy of classifier: This refers to the ability of the model/algorithm to correctly predict the class label of new or previously unseen data. This is the most common criterion.
- Speed of learner and classifier, which refers to the computation costs involved in building the classifier and classifying a new document respectively
- Robustness: the ability of the model developed to make correct predictions on given noisy data or data with missing values
- Its scalability, which refers to the ability to construct model efficiently given large amounts of data.
- Interpretability: the level of understanding and insight that is provided by the model.

The study dataset is an N dimensional data vector, where N is the number of classes of commodities considered. Each dimension of the data vector class is a volume imported. Hence, we have several data vectors for several months and years under consideration.

This research work has carried out experiments in order to evaluate the performance and usefulness of different classification or regression algorithms for predicting the customs imported data of ERCA. Tests were conducted using three algorithms tests such as Ordinary Linear Regression (OLR), Support Vector Machine and Random forest tree [43][44]. These algorithms were selected to use in the experiment based on the goal of this research. In this research we plan to evaluate regression predictive model by using continuous variables and the indicated algorithms can handle better as we assess it in the chapter two literature part.

3.6 Validation Techniques

Once the model is created, different methods can be applied to evaluate its performance. Simple split, holdout, bootstrapping and cross-validation are among the most used evaluation methods. In this research we used 10 fold cross-validation methods.

Cross-validation [33] is a statistical method that is used to estimate the performance of the algorithm. The method divides data into two sets: the training data set that is used to train the model and the validation data set that is used to test the model. One of the forms of cross-validation is k-fold cross-validation. In k-fold cross-validation the data is randomly partitioned into the k equally sized parts or folds. One part is used for testing the model and the remaining k-1 parts are used for the training. The process is repeated k times such that each of k-parts of the data is used exactly once for the testing. A recall, precision and F-score are calculated for each run. The advantage of this method is that it is not influenced by how the data is divided because every data point is involved in both the testing and the training process. The variance of the resulting estimation decreases as the value of k increases. The disadvantage of this method is that the algorithm needs to be rerun k times, but this allows us to obtain a more precise estimation. Various numbers of tests were done on different datasets in order to define the best choice of the fold's number. As a result, it was known that ten folds is the right number even though there are no strong theoretical explanations for this [48].

CHAPTER FOUR

EXPERIMENTATION AND RESULTS

This chapter presents steps and procedures followed during experimentations. In the initial experiment the researcher took 8 of 15 attributes such as IM_YEAR, HS_CODE, HS_DESCRIPTIN, ORIGION_COUNTRY, QUANTITY, NET_WT_KG, CIF_VALUE_ETB, and T_TAX_ETB for model building purpose. The selection of these attributes is made using subjective judgment with the support of domain experts' opinions from ERCA. To use the data for numeric regression and build model the .csv format of the data set was imported in to Rstudio and after training the machine the following result was obtained.

4.1 Exploring and understanding data

4.1.1 Exploring Linear data

Measuring the mean and median of the data set provides to summarize the values, but these measures of central tendency tell us little about whether or not there is diversity in the data measurements. In addition to measuring the diversity, we need to employ different type of the summary statistics to know how tightly or loosely the values are spaced in the different quartiles. We use the summary () function to get result from the dataset. Knowing about the spread provides a sense of the data's highs and lows and whether most values are like or unlike the mean and median. And on the other hand, variance and standard deviation indicates how data set is spread out around the median and mean. The standard deviation indicates, on average, how much each value differs from the mean.

	QUANTITY	NET_WT_KG	CIF_VALUE_ETB	T_TAX_ETB
Minimum (Min.)	1	0	101	6
First quartile (1st Qu.)	3	16	7543	3673
Median	26	142	42654	20919
Mean	51879	10679	811377	381889
Third quartile (3rd Qu.)	500	1500	257801	130603
Maximum (Max.)	75240000	9212093	245634010	65021765
variance	1.540462e+12	12946112622	2.289744e+13	3.70091e+12
Standard deviation	1241154	113781	4785127	1923775

Table 4.1 Measurement of the central tendency of the dataset

To know how the custom imported data set is organized in terms of data frames, vectors, or lists the str() function were run and we learn the data set has 10031 observations and Eight (8) a ttributes. Moreover, it indicates the data structure of the dataset.

```
> str(impdata)
 data.frame':  10031 obs. of  8 variables:
 $ IM_YEAR         : Date, format: "2014-01-01" ...
 $ HS_CODE         : chr  "37011000" "40116200" "94036000" "85287290" ...
 $ HS_DESCRIPTION : chr  "PHOTOGRAPHIC" "NEW PNEUMATIC TYRES OF RUBBER" "WOODE
 $ OIGIN_COUNTRY  : chr  "United States" "China" "China" "China" ...
 $ CONSIGN_COUNTRY: chr  "Germany" "United Arab Emirates" "China" "China" ...
 $ QUANTITY        : num  16270 1752 9257 5449 6790 ...
 $ NET_WT_KG       : num  42576 118175 340864 90201 9179 ...
 $ CIF_VALUE_ETB  : num  17447822 12460063 7889753 7101316 6532670 ...
 $ T_TAX_ETB       : num  872391 4878115 5321638 5957649 3579903 ...
```

Fig 4.1 the data structure of the dataset.

Quantile

	0%	25%	50%	75%	100%
QUANTITY	1	3	26	500	75240000
NET_WT_KG	0.05	16.00	142.50	1500.00	9212093.00
CIF_VALUE_ETB	101.28	7543.26	42654.33	257801.07	245634009.51
T_TAX_ETB	6.310	3672.645	20919.28	130603.06	65021764.93

Table 4.2 the quantile of dataset

From the above table we observed the data distribution and spread with in the first quantile (minimum), second quantile, median, mean, third quantile and the maximum value. We see for all the four variables the existence of big spread between the third quantile and maximum value. The generic function in R quantile produces sample quantiles corresponding to the given probabilities. The smallest observation corresponds to a probability of 0 and the largest to a probability of 1. Before fitting a regression model to this data set, it can be helpful in this research to determine how the independent variables are related to the dependent variable and each other. Here the researcher uses a correlation matrix to provide a quick overview of these relationships. To create

a correlation matrix for the four numeric variables in the import goods data frame, we use the cor() command to gain the following result:

VARIABLES	QUANTITY	NET_WT_KG	CIF_VALUE_ETB	T_TAX_ETB
QUANTITY	1.00000000	0.1371658	0.09480638	0.1027245
NET_WT_KG	0.13716577	1.0000000	0.44490090	0.4515751
CIF_VALUE_ETB	0.09480638	0.4449009	1.00000000	0.7448453
T_TAX_ETB	0.10272453	0.4515751	0.74484532	1.0000000

Table 4.3 correlation matrix for the four numeric variables of the data set

As we observed from the table 4.3, intersection of each row and column pair and the Correlation is listed for the variables indicated by that row and column. The diagonal is always 1.0000000 since there is always a perfect correlation between a variable and itself. In addition, the values exist above and below the diagonal as we observed from the picture are identical, because correlations are symmetrical in its behavior. In other words, cor (x, y) is equal to cor (y, x). The correlations in the matrix above are seen as strong with some notable associations for Total tax and CIF value. Quantity and CIF value appear to have a weak positive correlation. There is also a moderate positive correlation exist between quantity and Net weight and, quantity and Total tax, and Net weight and CIF value, and net weight and total tax. These associations imply that as quantity, Net weight, and CIF value increase, the expected cost of total tax also goes up. We will try to test out these relationships more clearly when we build our final regression model.

The researcher also tried to visualize the data set or the four-number summary by using boxplot also known as a box-and-whiskers plot. The boxplot displays the center and spread of a numeric variable in a format that allows us to quickly obtain and know a sense of the range and skew of a variable or compare it to other variables. We use R's graphical capabilities to create a scatterplot matrix for the four numeric features: QUANTITY, NET_WT_KG, CIF_VAUE_ETB, and T_TAX_ETB. The pairs() function is provided in a default R installation and provides basic functionality for producing scatterplot matrices.

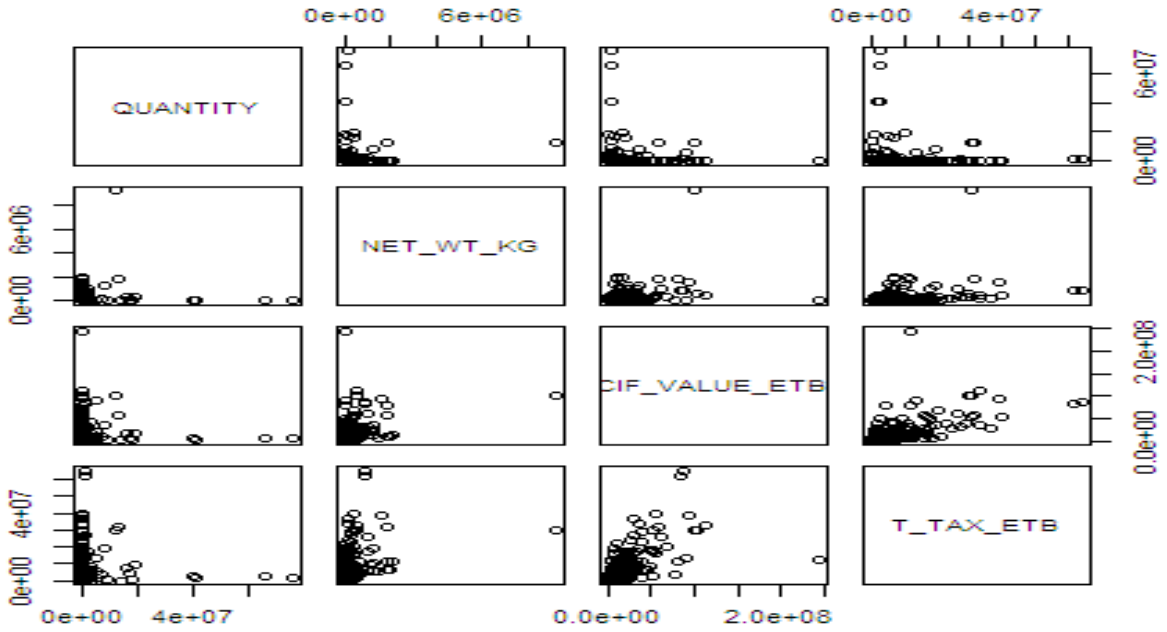


Fig 4.2 Scatterplot matrices.

We can also enhanced the scatterplot matrix indicated in fig. 4.2 by using pairs.panels() function in the psych package by loading psych package for R. Then, we created a scatterplot matrix as follow:

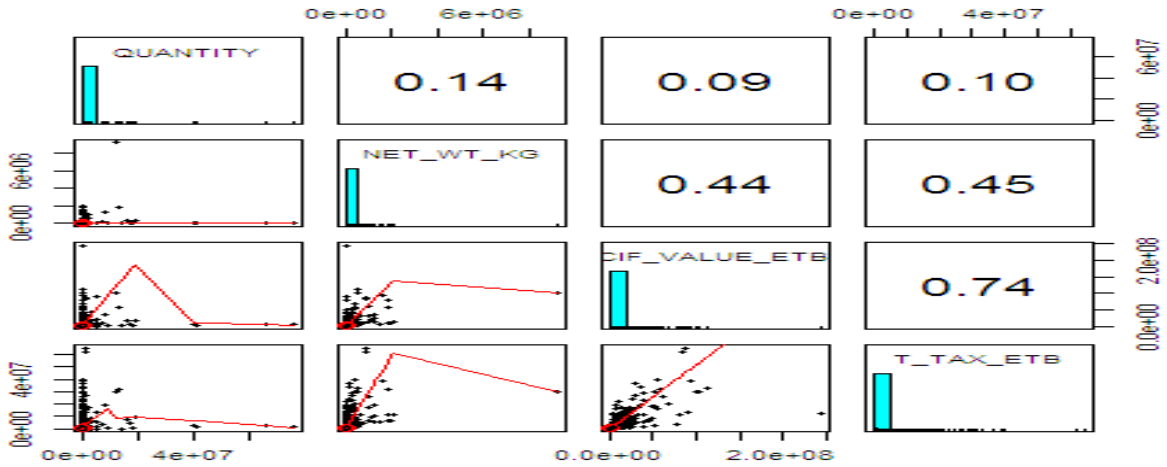


Fig 4.3 enhanced scatterplot matrix

In figure 4.3 the diagonal; the scatterplots have been replaced with a correlation matrix. In addition, a histogram is showing the distribution of values for each feature on the diagonal. Finally, the scatterplots below the diagonal are presented with additional visual information.

4.1.2 Exploring categorical variables

From the eight dataset variables we choose for experiment IM_YEAR, HS_CODE, HS_DESCRIPTION and ORIGIN_COUNTRY are categorical variables. Categorical data is typically examined using tables rather than summary statistics. We visualize the model by using table() and plot()functions each of the variables as indicated in figure 4.4

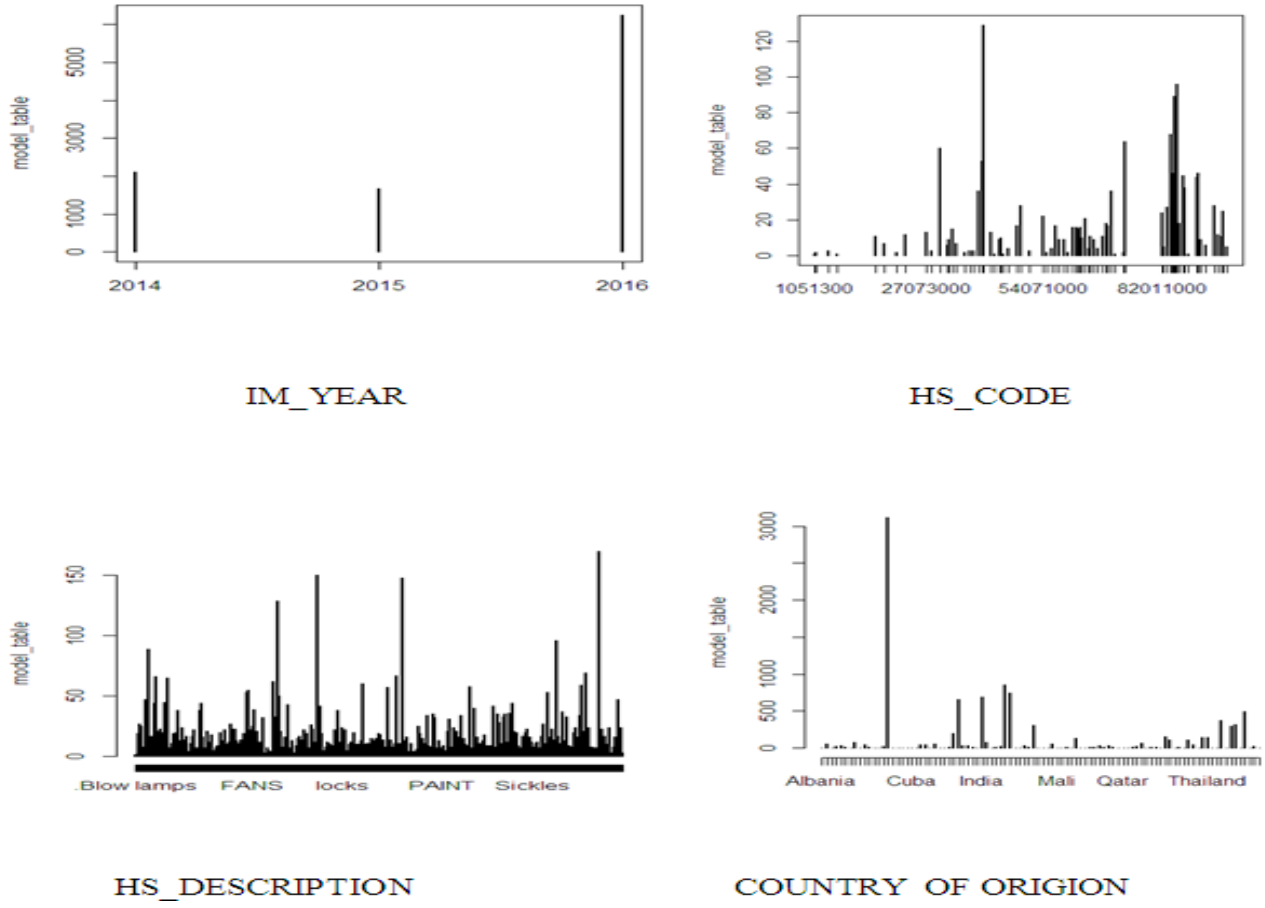


Fig. 4.4 Visualization of Categorical Variables

In categorical data a mode is another measure of central tendency. It is often used for categorical data, since the mean and median are not defined for nominal variables. And from figure we can observe the highest value for year is 2016 with frequency 6257, for HS_CODE 40169300 with frequency 129, For HS_DESCIPTION petroleum and oil with frequency 149 and for

COUNTRY OF ORIGION china with the frequency 3121. From the data also we can learn the most imported Good is petroleum and oil and most goods come from China.

4.2 Multiple ordinary linear regressions

4.2.1 Ordinary Linear Regression Model

In this research work the ordinary linear regression is built by the lm () function with all the 7,018 samples data in the training set. The model parameters and statistics value is obtained by the summary () function, which can be seen in Figure 4.5. The estimated regression coefficients in the multiple ordinary linear regression models represent that the value of the dependable variable can be increased when the value of the independent variable is changed by one unit. And the estimated coefficient is significantly different from zero with the p-value < 2e-16. In summary, about 100% of the variance in the dataset can be explained by the model, and the value of the residual standard error is 7.353e-10 on 7018 degrees of freedom.

```
> imp_model <- lm(T_TAX_ETB ~ QUANTITY + NET_WT_KG + CIF_VALUE_ETB ,
data = impo1_train)
> summary(imp_model)

Call:
lm(formula = T_IMORT_COST ~ QUANTITY + NET_WT_KG + CIF_VALUE_ETB +
    T_TAX_ETB, data = impdata.train)
Residuals:
       Min         1Q     Median         3Q        Max
-2.908e-08 -9.320e-11 -7.100e-12  7.820e-11  1.852e-08

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  -3.705e-11  8.967e-12 -4.132e+00 3.63e-05 ***
QUANTITY      8.012e-18  6.046e-18  1.325e+00    0.185
NET_WT_KG     1.000e+00  1.577e-16  6.340e+15  < 2e-16 ***
CIF_VALUE_ETB 1.000e+00  2.758e-18  3.626e+17  < 2e-16 ***
T_TAX_ETB     1.000e+00  7.901e-18  1.266e+17  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.353e-10 on 7018 degrees of freedom
Multiple R-squared:      1,    Adjusted R-squared:      1
F-statistic: 1.241e+35 on 4 and 7018 DF,  p-value: < 2.2e-16
```

Fig. 4.5 multiple ordinary regression result for training set

The key difference between the regression modeling and other machine learning approaches is that regression typically leaves feature selection and model specification to the user [34]. Hence, we can use this information to inform the model specification and potentially improve the model's performance by adding nonlinear terms.

```
> set.seed (1)
> train=sample(8000,2031)
> lm.fit=lm(T_IMORT_COST~QUANTITY+NET_WT_KG+CIF_VALUE_ETB+T_TAX_ETB,data=i
mpdata,subset =train)

> summary(lm.fit)

Call:
lm(formula = T_IMORT_COST ~ QUANTITY + NET_WT_KG + CIF_VALUE_ETB +
    T_TAX_ETB, data = impdata, subset = train)

Residuals:
       Min         1Q     Median         3Q        Max
-1.482e-08 -7.650e-11 -2.330e-11  3.280e-11  1.004e-08

Coefficients:
               Estimate Std. Error    t value Pr(>|t|)
(Intercept)    1.672e-13  1.471e-11  1.100e-02    0.991
QUANTITY      -1.117e-16  2.603e-17 -4.292e+00 1.86e-05 ***
NET_WT_KG      1.000e+00  4.183e-16  2.391e+15  < 2e-16 ***
CIF_VALUE_ETB  1.000e+00  1.051e-17  9.513e+16  < 2e-16 ***
T_TAX_ETB      1.000e+00  1.719e-17  5.817e+16  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.451e-10 on 2026 degrees of freedom
Multiple R-squared:      1,    Adjusted R-squared:      1
F-statistic: 2.372e+34 on 4 and 2026 DF,  p-value: < 2.2e-16
```

Figure 4.6 the result of improved multiple ordinary regression on training set

From figure 4.6 we learn that the new model fit statistics help us to determine whether our changes improved the performance of the regression model. Relative to our first model, the R-squared value has stayed the same as 1. Similarly, the adjusted R-squared value, which takes into account the fact that the model grew in complexity, also remained as the original 1. Our model is now explaining 100 percent of the variation in import total tax costs. 100% of the variability observed in the total cost of import items is explained by the assessed values. Hence, the assessed values of imported items variables contribute a lot of information about the total imported items cost.

4.2.2 Measuring Performance in OLR Model

In this research Cross-validation method is used to evaluate the model. In the k-fold cross-validation, the training set is divided into k subsets. Each of k subsets is selected as the hold-out set; the other k-1 subsets are selected as the training set. Since the regression model is only developed on the training set, the hand-out grouped is used to train the model further. The performance of the hand-out groups will better reveal the model generalizability. In our implementation, the crossval() function is used to perform the 10-fold cross-validation to check the R2 statistics result mentioned above.

```
> library(caret)
Loading required package: lattice
Loading required package: ggplot2
> library(rpart)
> train_control<- trainControl(method="cv", number=10)
> model<- train(T_IMORT_COST ~ QUANTITY + NET_WT_KG + CIF_VALUE
_ETB +                    T_TAX_ETB,  data=impo1, trControl=trai
n_control, method="rpart")
> predictions<- predict(model,impo1)
> impo1<- cbind(impo1,predictions)
> confusionMatrix<- confusionMatrix(impo1$predictions,impo1$T_T
AX_ETB)
> model

CART
10031 samples
    4 predictor

Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 9029, 9030, 9028, 9027, 9028, 9027, ...
Resampling results across tuning parameters:

  cp          RMSE     Rsquared   MAE
  0.1213470   2952689  0.8113806   899325.6
  0.1502264   3692543  0.6364825  1212612.6
  0.6001346   5001748  0.5414442  1568234.2

RMSE was used to select the optimal model using  the
 smallest value.
The final value used for the model was cp = 0.121347.
```

Fig 4.7 cross validation result for OLR

We can see from Figure 4.6 the original R2 based on the training set without cross-validation is 1, and the new R2 with 10-fold cross-validation in fig 4.7 is 0.8114. Thus, the change between the original R2 and the 10-fold cross-validated R2 is only 0.1886. The smaller R2 change is able to tell us that the model owns better generalizability.

For regression Model predicting continuous numeric outcomes. The Root Mean Squared Error (RMSE) and coefficient of determination (R2) are usually used to evaluate the performance of the model. RMSE is the square root of the difference between the observed values and predicted outcomes. Whereas, the simplest explanation of R2 is the square of correlation coefficients between the observed and predicted outcomes. Its value tells the percentage of the information or variation in the outcome explained for OLR model trained for customs imported dataset, but not the accuracy. Even though, the linear regression model with a 100% R2 is optimistic, the average distance between the observed and the predicted values is quite large, which means that the linear regression model shows poor predictive accuracy.

4.2.3 Regression Diagnostics

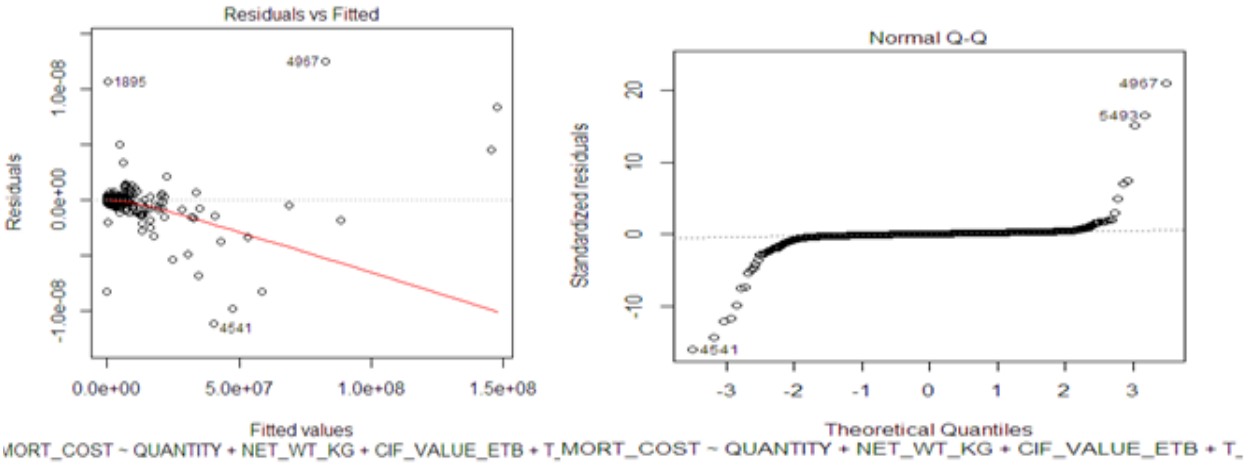
Before having fully confidence in the performance of the linear regression model our model should be evaluated and diagnosed after the regression model is built. Regression diagnostics are a set of useful tools to evaluate and judge the performance of the regression model. As we can observe in Figure 4.9, the diagnostic plots for the multiple ordinary linear regression model can be got by applying the plot() function to the object fit, which is returned by the lm() function. Through analyzing the four graphs indicated in figure 4.8, we can get the information about the satisfaction degree to the statistical assumptions underlying the OLR model.

- **Linearity:-** we observe from the picture If the predicted variable is linearly related with the independent variables in the dataset, there should be no relationship between the predicted values and the residuals. In another words, all the variance, except for the random noise, in the dataset should be captured by the model. However, according to the Residuals vs Fitted graph (upper left), all the points are apparently not randomly distributed.
- **Normality:-** in the statistical assumptions of the OLR model, the predicted variable should be normally distributed in the case of fixed values of the independent variables. But in the Normal Q-Q graph (upper right), a probability graph of the standardized residuals against the theoretical quantiles, a majority of the points on this graph do not fall on the straight

45-degree dash line, thus we can get the result that the normality assumption is not satisfied in this model.

- **Homoscedasticity:-** Homoscedasticity represents that the variance of the predicted variable does not change while the levels of the independent variables (quantity, Net weight in kilograms, CIF values in ETB and total tax in ETB) are changed. If the constant variance assumption is met, we should have a randomly distribution of the points around a horizontal line on the Scale Location plot (bottom left). But for this model, it is not. Thus, the homoscedasticity assumption here is also not met in this model.

  Hence as the linearity, normality and homoscedasticity are not met in this model; it means that we cannot have fully confidence for the results of this multiple linear regression model. This leads us to try other types of algorithm in the next sections.



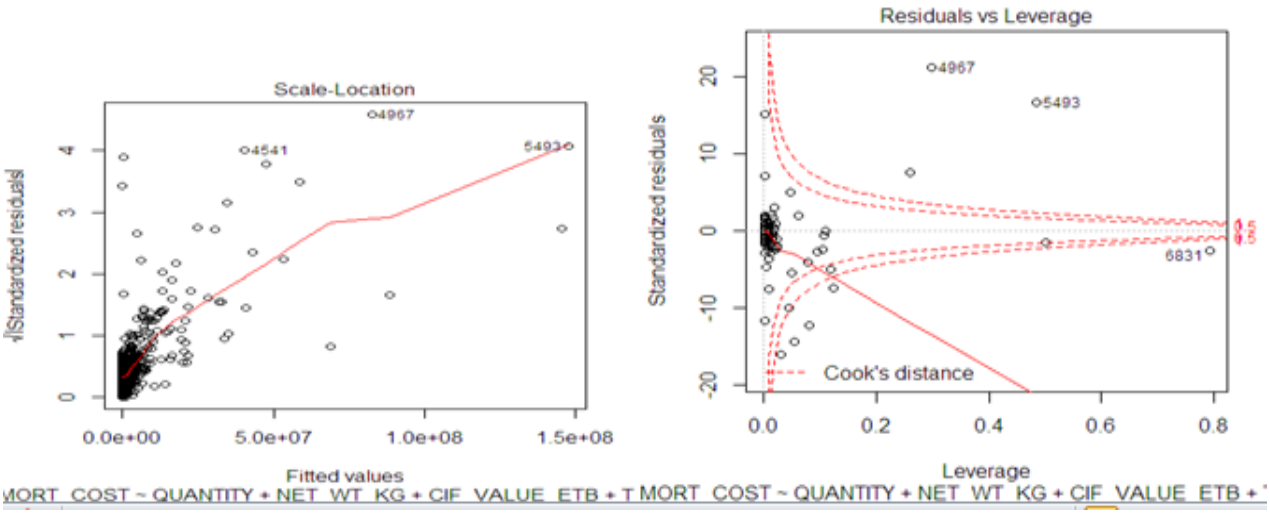


Fig 4.8 Diagnostic Plots for Ordinary Linear Regression Model

4.3. Random Forest

Random forest is one of an ensemble machine learning method in the case of regression that is mainly implemented by building a number of decision trees during the training time and outputting the averaging forest's prediction of the individual trees. The randomForest package in R can be used to build the model for classification and regression [83]. The core function randomForest() in the package requires several tuning parameters: number of trees to grow tree n , number of variables at each random split selection mtry and so on.

4.3.1 Choosing Tuning Parameters

The random forest algorithm is computationally intensive, the train() function in the caret package is still suggested to be used on the small number of samples in the training set, starting from 1 to 15 (total number of the predictors). The number of trees for the random forest is also specified. The larger the random forest, the more time we will spend on training and building the model. Therefore, the default value 500 trees is used in our experiment as a starting point. Then we can train over this parameter in Figure 4.9 as follows:

```
> library(randomForest)
> library(MASS)
> mysample<-impdata.train[sample(1:nrow(impdata.train),3000,replace = FALSE),]
> impdata.forest <- train(T_IMORT_COST ~ QUANTITY+ NET_WT_KG+ CIF_VALUE_ETB+T_T
AX_ETB,  data = mysample,
+ method="rf",
+ tuneGrid=data.frame(.mtry=1:15),
+ trcontrol=trainControl(method = "cv"),
+ maxit=100)
```

Fig 4.9 train() Function for Choosing RF Tuning Parameters

The random forest regression model developed here is also non-deterministic algorithm, the model randomly selected variables at each split probably which give rise to totally different predictions. Thus, the train() function run 15 times independently in order to select the table tuning parameter with minimum RMSE and maximum R2. RMSE is also used in the single train

() function to select the optimal model with the smallest value, for instance in Figure 4.10, the f inal value used for the current training model is mtry =3.

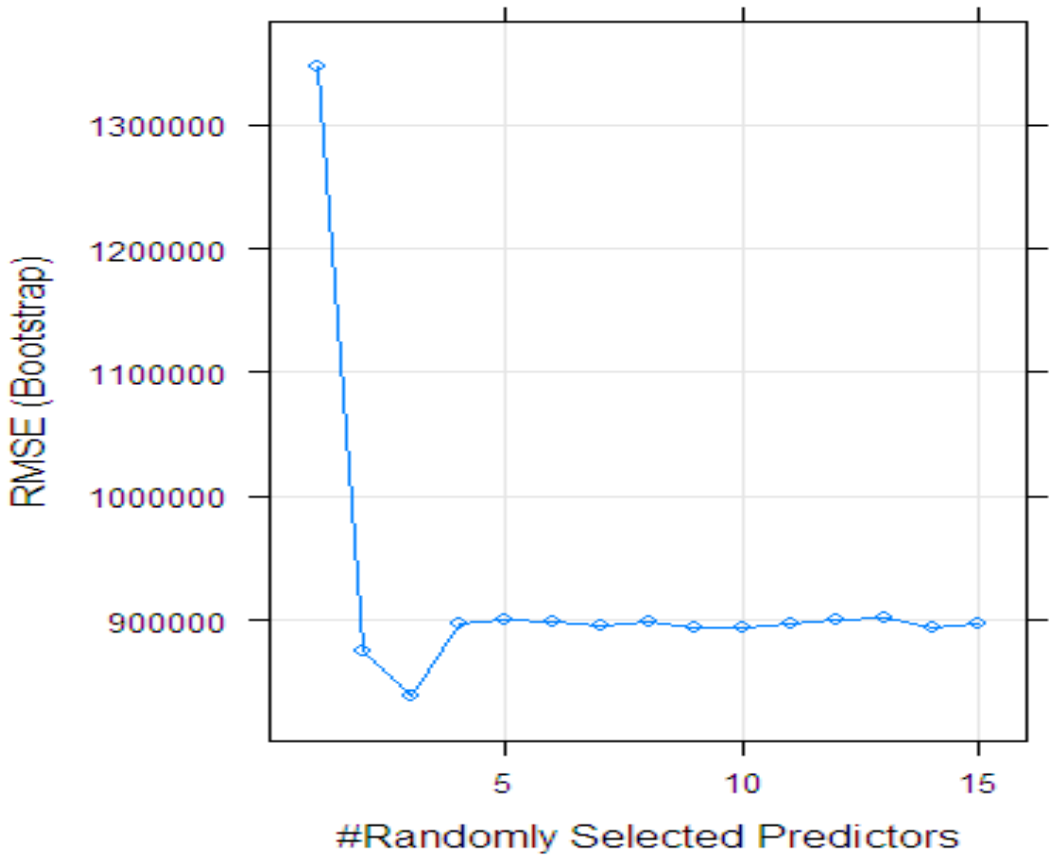


Fig 4.10 RMSE Profiles for RF model by train() function

All the fifteen optimal results for the instances of the train() function can be seen in the following Table 4.11. The optimal values of the number of randomly selected variables to choose from at each split ranges from 1 to 15, and the minimum RMSE and maximum R2 which appears in the third model with the values 838464.5 and 0.9801928, respectively. According to the optimal results of the train() function in Table 4.5, the tuning parameters of the Random Forest model by the randomForest() function in the randomForest package can be chosen as mtry = 3 and ntree = 500.

Sequence Number	ntree	mtry	RMSE	Rsquared	MAE
1	500	1	1347332.4	0.9554767	174443.92
2	500	2	874500.4	0.9800716	81675.92
3	500	3	838464.5	0.9801928	71000.54
4	500	4	897862.0	0.9767498	81614.08
5	500	5	899771.1	0.9764491	81895.25
6	500	6	899357.3	0.9764883	81740.83
7	500	7	896186.4	0.9764650	81248.34
8	500	8	899527.6	0.9762756	81703.19
9	500	9	894067.7	0.9765566	81223.19
10	500	10	893424.8	0.9766703	81228.27
11	500	11	896859.9	0.9767097	81226.13
12	500	12	900585.2	0.9765785	81462.25
13	500	13	901236.0	0.9764684	81851.09
14	500	14	894108.2	0.9765951	81398.62
15	500	15	897047.2	0.9765461	81053.81

Table 4.4 Optimal Results of the train() Function for RF

4.3.2 Building RF Model

After we choose the best tuning parameters, we implemented the random forest regression model as it can be seen in Figure 4.11. Apart from the two tuning parameters mtry = 3 and ntree = 500, the option of importance = TRUE represents that the variable importance scores can be accessed, importance = FALSE means that they are not calculated as the calculation is time consuming.

```
> library(randomForest)
> library(MASS)

> mysample<-impdata.train[sample(1:nrow(impdata.train),3000,rep
lace = FALSE),]
> impdata.rf <- randomForest(T_IMORT_COST ~ QUANTITY+ NET_WT_KG
+ CIF_VALUE_ETB+T_TAX_ETB,  data = mysample,
+ importance=TRUE,
+ ntree =500,
+ mtry=3)
```

Fig 4.11 Random Forest Model

After building the Random Forest model, we use the print() function to get the information. Call shows the original call to randomForest; Type illustrates the type of the training model, and it is a regression model; Number of trees means 500 trees grown in this model, and 3 predictors or variables sampled for splitting at each node; the mean of squared residuals is about 97.22% of variance in the training set has been explained by the RF model.

```
Print(impdata.rf)
Call:
 randomForest(formula = T_IMORT_COST ~ QUANTITY + NET_WT_KG +
CIF_VALUE_ETB + T_TAX_ETB, data = mysample, importance = TRUE,
ntree = 500, mtry = 3)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 839061751977
                    % Var explained: 97.22
```

Fig 4.12 Summary of the Random Forest Model

4.3.3 Measuring Performance in RF Model

In the following section we tried to see the performance of the random forest regression model t hrough visualization of the importance of predictor variables and also by plotting the predicted v s. observed graphs.

4.3.3.1 Visualization of variable importance scores

The importance () function gives us a textual representation of how important each of the v ariables were in classifying the import data set, while the varImpPlot () method on the other h and gives us a graphical representation of it. We see from the Variable importance plot figure, th

at the most important variables are CIF_VALUE_ETB, and T_TAX_ETB. Whereas, the rest tw o other variables QUANTITY and NET_WT_KG have very less importance.

```
> importance(impdata.rf)
                %IncMSE IncNodePurity
QUANTITY       3.304017   3.848478e+14
NET_WT_KG      4.913281   3.896391e+14
CIF_VALUE_ETB 44.929994   6.403259e+16
T_TAX_ETB     15.694351   2.578321e+16
```

Figure 4.13 Variable Importance Scores for the 4 Predictors in the RF Model

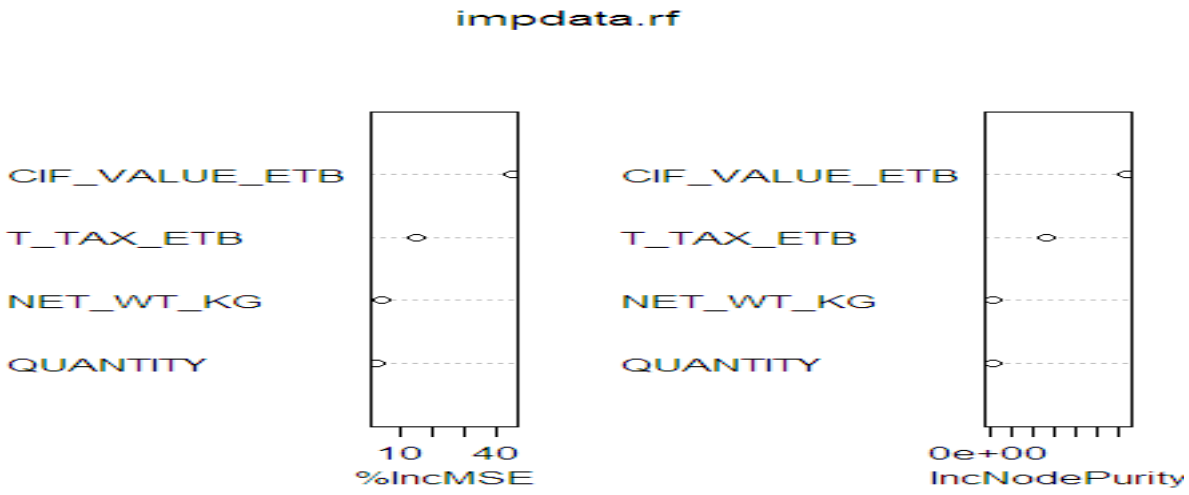


Fig. 4.14 Dot-chart of Variable Importance Scores

As we can learn from figure 4.14 the first measure is computed from permuting data for each tree, the prediction error on the out-of-bag portion of the data is recorded (error rate for MSE for regression). Then the same is done after permuting each predictor variable. The difference between the two are then averaged over all trees, and normalized by the standard deviation of the differences. If the standard deviation of the differences is equal to 0 for a variable, the division is not done (but the average is almost always equal to 0 in that case).

4.3.3.2 Quantitative Measures of Performance

The randomForest() function has been run 15 times to get the best model with minimum RMSE on the testing set. To get the quantitative results of the ten RF cross validation model developed as can be seen in figure 4.15.

```
> library(caret)
> library(randomForest)
> control <- trainControl(method="repeatedcv", number=10, repeats=10, search=
"grid")
> set.seed(100)
> tunegrid <- expand.grid(.mtry=c(1:15))
> imp.rf <- train(T_IMPORT_COST ~ QUANTITY+ NET_WT_KG+ CIF_VALUE_ETB+T_TAX_ET
B, data=mysample1, method="rf",  tuneGrid=tunegrid, trControl=control)
```

Fig. 4.15 Cross validation model on the test set using RF

From the RF cross validation model indicated in figure 4.15 we can generate the result that the values of RMSE and R2 are very stable, RMSE varying from about 464212.8 to 468828.3 and R2 varying from 0.9927242 to 0.9928062, respectively. The minimum RMSE and maximum R2 that we can get from the results are 464212.8 and 0.9928060 in the fifth model. The worst model is the fourteenth model, in which the RMSE and R2 are 468828.3 and 0.9927330, respectively. Comparing with the corresponding results in the ordinary linear regression model (RMSE = 2952689, R2 = 0.8113806), the quantitative performance of the RF model is better than OLR models

```
> print(imp.rf)
Random Forest

3000 samples
   4 predictor

Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 2700, 2700, 2700, 2700, 2700, 2700, .....
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared   MAE

   4     467553.2  0.9927350   60984.52
   5     464212.8  0.9928060   60427.24
   6     464317.8  0.9927242   60915.60
   7     465408.3  0.9928062   61020.28
   8     464726.6  0.9928378   60568.79
   9     468345.4  0.9927353   61046.73
  12     466893.7  0.9927253   60781.02
  13     468417.8  0.9927398   60869.34
  14     468828.3  0.9927330   61088.16
  15     466229.1  0.9927989   60847.76

RMSE was used to select the optimal model using
 the smallest value.
The final value used for the model was mtry = 5.
```

Fig. 4.16 Cross validation model results on the test set using RF

4.3.3.3 Graphical Measures of RF model Performance

In order to visualize the performance of RF model we use the following function.

```
> library(randomForest)
> library(miscTools)
> library(ggplot2)
> impdata.rf <- randomForest(T_IMORT_COST ~ QUANTITY + NET_WT_KG + CIF_VALUE_ETB+T_
TAX_ETB, data=mysample, ntree=20)
> (r2 <- rSquared(mysample$T_IMORT_COST, mysample$T_IMORT_COST - predict(impdata.rf
, mysample)))            [,1]
[1,] 0.9731079
> (mse <- mean((mysample$T_IMORT_COST - predict(rf, mysample))^2))
[1] 811787509053
> p <- ggplot(aes(x=actual, y=pred),
+             data=data.frame(actual=mysample$T_IMORT_COST, pred=predict
(impdata.rf, mysample)))
> p + geom_point() +
+     geom_abline(color="red") +
+     ggtitle(paste("RandomForest Regression in R r^2=", r2, sep=""))
```

Fig 4.17 Code of the RF Model Fit

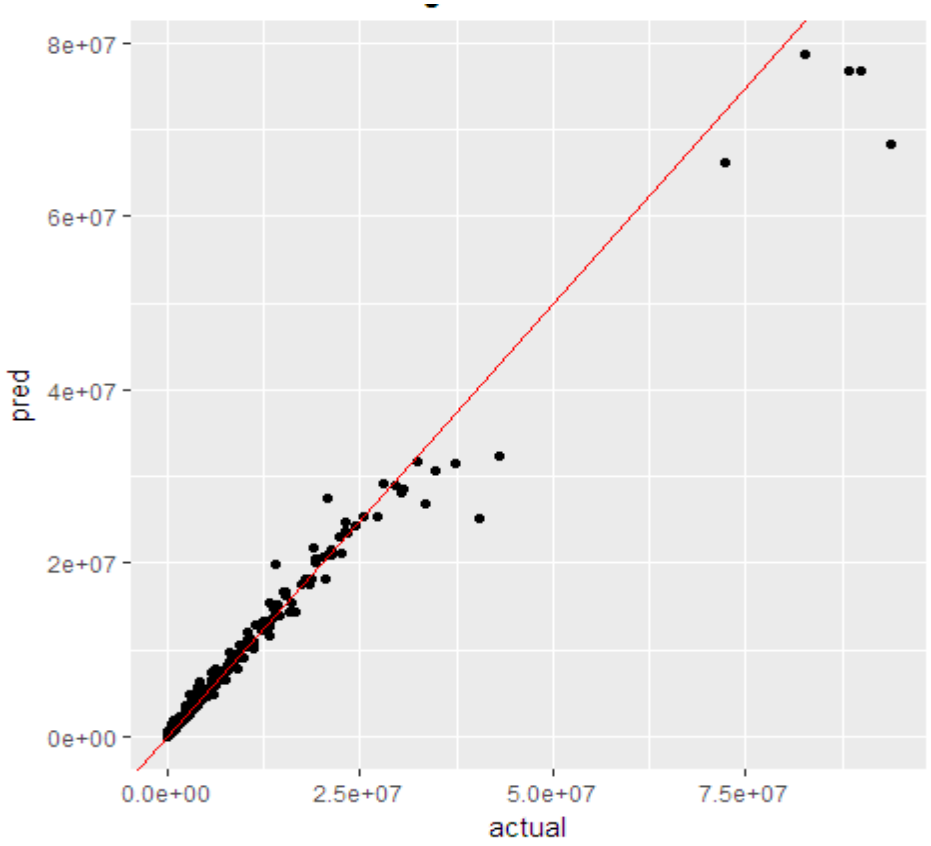


Fig. 4.18 Visualizations function of the RF Model Fit

As indicated in Figure 4.18, the plot of observed values vs. predicted outcomes in RF Model 5 where the RMSE and R2 for the testing dataset are 464212.8 and 0.9928, respectively. As we can see there is a slight difference in the model fit and test results in cross validation. But, it has a tendency to over-predict the small observed values, and all the other points are also mainly located around the diagonal line.

4.4 Support Vector Machine (SVM)

In the next section we tried to train SVM algorithm. Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier detection. The advantages of support vector machines are effective in high dimensional data to fit model and predicted outcomes. Support Vector Regression (SVR) works on similar ways and principles as Support Vector Machine (SVM) classification. SVR is the adapted form of SVM when the dependent variable is numerical rather than categorical. One of the major advantages of using SVR is that it is a non-parametric technique. Instead the SVR technique more depends on kernel functions. Another advantage of SVR is that it permits for construction of a non-linear model without changing the explanatory variables, helping in better interpretation of the resultant model. Here the regression can also be penalized using a cost parameter, which becomes handy to avoid over-fit.

4.4.1 Model development for SVM

R package "e1071" is required to call SVM function and after we install this package we developed the model by using R code is as follows:

```
> library(e1071)
> library(caret)
> library(Metrics)
> attach(impdata.train)
> x <- subset(impdata, select=-T_IMORT_COST)
> y <- T_IMORT_COST
> svm_model <- svm(T_IMORT_COST ~QUANTITY+NET_WT_KG+CIF_VALUE_ETB+T
_TAX_ETB+HS_CODE , data=mysample)
> svm_model

Call:
svm(formula = T_IMORT_COST ~ QUANTITY + NET_WT_KG + CIF_VALUE_ETB +
    T_TAX_ETB + HS_CODE, data = mysample)

Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.0009433962
    epsilon:  0.1
Number of Support Vectors:  190
```

Fig 4.19 Code that shows SVM model

The gamma parameter we mentioned here defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The gamma parameters also can be seen as the inverse of the radius of influence of samples selected by the model as support vectors. Whereas, the C parameter trades off misclassification of training examples exist against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly by giving the model freedom to select more samples as support vectors. In the figure 4.20 we find results for cost 1, gamma 0.0009433962 and epsilon as 0.1. The following figure 4.21 shows the pictorial representation of the SVM Model.

```
> plot(T_IMORT_COST~QUANTITY+NET_WT_KG+CIF_VALUE_ETB+T_TAX_ETB,
+               data=impdata.train,
+            pch=20,
+            col="blue",
+            xlab="actual",
+            ylab="predicted",
+           main="model_svm")
> points(impdata.train$T_IMORT_COST, impdata.train$predicted.svm, col = "r
ed", pch=20)
```

Fig 4.20 plot function for svm model

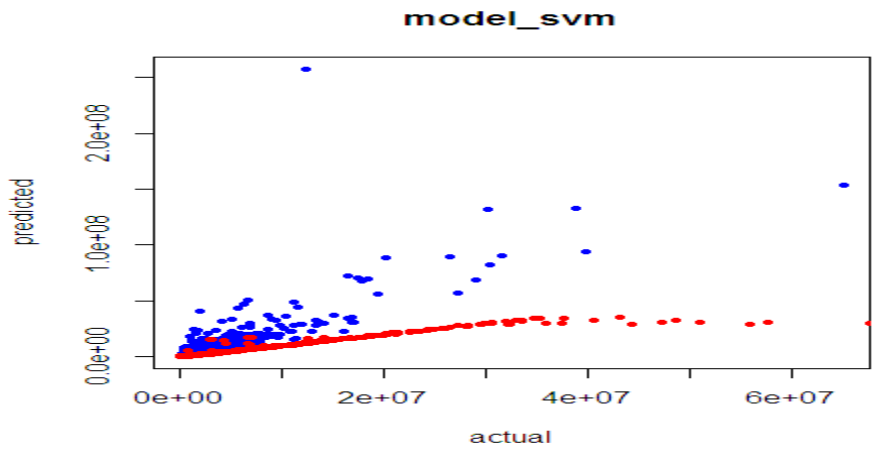


Fig 4.21 visualization of svm model

From figure 4.21 we learn the predictions are still far to the real values. In order to improve the performance of the support vector regression model we will need to select the best parameters f or the model.

In our above example, we performed an epsilon-regression, we did not set any value for **epsilon** (ϵ), but as we see from figure 4.19 it took a default value of 0.1. There is also a **cost** parameter which we can change in the model to avoid over fitting. In this model optimization we will train a lot of models for the different couples of ϵ and cost, and choose the best one. The RMSE of originally developed SVM model was "svr RMSE = 4099734.43159052" which is not better than OLR and Random forest model.

4.4.2 Choosing Tuning Parameters

In order to optimize the SVM model we run the following code by using tune() function.

```
> imptuneResult <- tune(svm, T_IMORT_COST~QUANTITY+NET_WT_KG+CIF_V
ALUE_ETB+T_TAX_ETB,  data = impdata.train,
+                    ranges = list(epsilon = seq(0,1,0.1), cost =
2^(2:9))
+ )
> print(imptuneResult)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 epsilon cost
     0.1   64

- best performance: 1.45232e+13
```

Fig. 4.22 SVM model optimization code

By using the code indicated in figure 4.22 above first we use the tune method to train models w ith ϵ=0, 0.1, 0.2, ...,1 and cost = 2^2,2^3,2^4,...,2^9 that means it will train about 88 models. From t he model we learnt that the best parameters epsilon is 0.1 and cost is 64 while the best performa nce is 1.45232e+13. The tune model RMSE shows much improvement from the original 409973 4.4 to 500752.1.

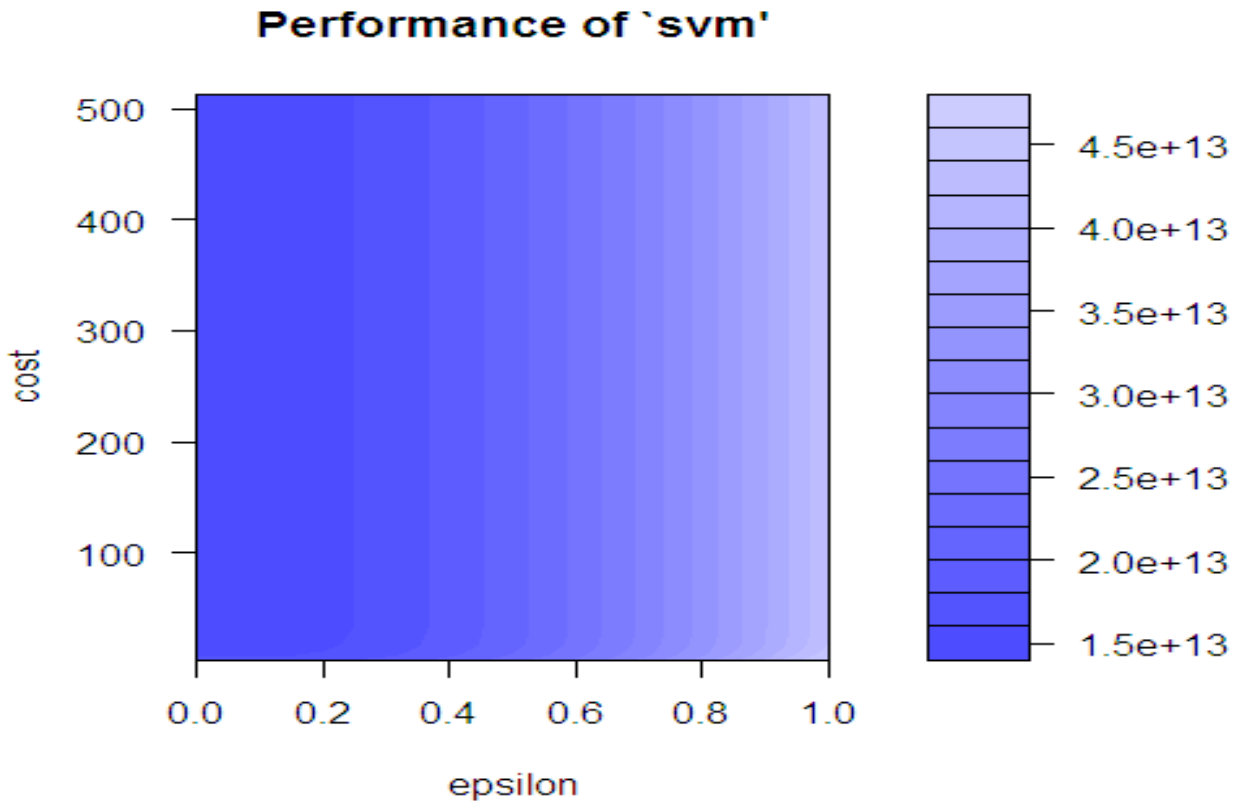


Fig 4.23 SVM performance using color coding

The above plot on figure 4.22 shows the performance of various models in SVM using color coding. Darker regions imply better accuracy in the model. The use of this plot is just to determine the possible range where we can narrow down our search to and try further tuning if required. For instance, this plot shows that we can run tuning for epsilon in the new range of 0 to 0.2 but going further lower or higher may lead to over fitting.

4.4.3 Measuring Model performance

In the following section we tried to see the performance of the support vector machines model through cross validation test.

```
library(caret)
> dim(imp_train); dim(imp_test);
[1] 7022    7
[1] 3009    7
> anyNA(impo1)
[1] FALSE
> trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
> svm_Linear <- train(T_IMORT_COST ~ QUANTITY + NET_WT_KG + CIF_VALUE_ETB + T
TAX_ETB , data = imp_train, method = "svmLinear",
+                     trControl=trctrl,
+                     preProcess = c("center", "scale"),
+                     tuneLength = 10)
> svm_Linear
Support Vector Machines with Linear Kernel

7022 samples
   4 predictor

Pre-processing: centered (4), scaled (4)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 6321, 6318, 6319, 6321, 6319, 6319, ...
Resampling results:

  RMSE      Rsquared   MAE
  627866.5  0.9999961  626128.4

Tuning parameter 'C' was held constant at a value of 1
```

Fig 4.24 cross validation result for SVM

We can see from Figure 4.24 the new R2 with 10-fold cross-validation is 0.9999968. Which is better than R2 of OLR=0.81 and R2 of Random forest 0.992 respectively. Thus, in terms of RMSE 604748.2 for SVM, 2952689 for OLR and 464212.8 for random forest and RMSE for random forest shows better performance compared to SVM and OLR. The smaller R2 change is able to tell us that the model owns better generalizability.

4.5. Validation of OLR, RF and SVM predictive models

Data set that we use in this experiment is necessary to undertake the intended machine learning process. However, the data is not readily available in the database in the format that is suitable for machine learning algorithms requirements. The fields of the data contain outliers, missing values, inconsistent data type such as wild characters, links etc. even with in a single field and many other anomalies. Hence, the researcher take considerable time to clean, integrate and transform the data on the MS Excel during data preprocessing task that is suitable for R and Rstudio machine learning process.

As discussed in the experiment section before the regression and prediction models developed by using OLR, SVM and RF algorithms were selected because the minimal RMSE and the maximum R2 indicated relatively good model during comparison of different model. In addition to these the domain experts from ERCA consulted on the result and they suggest as the model developed for SVM and RF algorithms are good to predict the Customs imported items data. Therefore, the model developed by SVM and RF are chosen as the final model for the study.

4.6 Summary

In this chapter, first we explored and understood the character of data by using summary () function to know how data set is spread out around the median and mean. str() function is also used to know the structure of the data is organized in terms of frames, vectors or lists. Before we fit our models, we tried to run cor() function to determine how the independent variables are related to the dependent variable and each other. From the cor() function in the matrix also we observed as there is a strong correlation, with some notable associations for T_TAX_ETB and CIF_VALUE_ETB. QUANTITY and CIF_VALUE_ETB appear to have a weak positive correlation. There is also a moderate positive correlation between QUANTITY and NET_WT_KG, QUANTITY and T_TAX_ETB and NET_WT_KG and CIF_VALUE_ETB, and NET_WT_KG and T_TAX_ETB. These associations imply that as QUANTITY, NET_WT_KG, CIF_VALUE_ETB and T _TAX_ETB increase, the expected total import cost increases proportionally that shows positive correlation. And also we visualized these correlation distribution of the data set by using pair() functions.

Secondly, three typical models such as Ordinary Linear Regression, Support Vector Machine and Random Forest algorithms have been implemented on customs imported goods dataset in to get the quantitative and visual performance of the built models which is analyzed in details. The Ordinary linear regression model is built by using the lm() function. The result shows almost 1 for multiple r2 and adjusted r2 with Residual standard error 7.759e-10 on 5561 degrees of freedom. This shows 100% of the Variation in total imported items costs. But, the regression diagnostics made through graphs and numeric performances explain the poorly predictive ability of the OLR model for this specific data set. The 10-fold cross validation result on the test set for OLR model shows 2952689 and 0.8113806 for the smaller RMSE and maximum R2 respectively.

Next, the random forest regression model is constructed by the randomForest() function. When we compare with the OLR model, the minimum RMSE and maximum R2 we can get from the results are 464212.8 and 0.9928060. Also the variable importance scores and number of nodes for each tree is easily got by the random forest model. Thirdly, we are using e1071 package in order to train SVM model. The new R2 with 10-fold cross-validation is 0.9999968. Which is better than R2 of OLR=0.81 and R2 of Random forest 0.9928 respectively. Thus, in terms of RMSE 604748.2 for SVM, 2952689 for OLR and 464212.8 for random forest and RMSE for random forest shows better performance compared to SVM and Ordinary Linear Regression, but the Random Forest and Support Vector Machine models training process is more complex and more time consuming than OLR model.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

This days big, large data exist everywhere that requires appropriate analytical mechanisms. Machine learning has been widely used in many industries and organization, which is enabling to get computers to learn without being explicitly programmed.

As we stated in chapter 1, the goal of this research work is to evaluate the predictive models trained by using OLR, RF, and SVM to see the characteristics of the imported customs datasets. The researcher follow training the machine on the dataset through selection of proper tuning parameters, evaluation of the performance of the models based on the test data and finally comparing the performance achieved by the three mentioned algorithms.

The dataset pre-processing and resampling techniques, including data transformation, dimensionality reduction and k-fold cross-validation, are discussed in theory which can be used to effectively improve the performance of the training model. During the implementation of machine learning algorithms, three typical models (Ordinary Linear Regression, Random Forest and support Vector machine) have been implemented by using the different packages exist in R on the given large size dataset. In addition to the model training, the regression diagnostics are conducted to explain the poorly predictive ability of the ordinary linear regression model. Because of the non-deterministic characteristic of the random forest and support vector machine models, several models with small scale samples in the dataset are built to get the reasonable tuning parameters, and the optimal models with minimum RMSE and maximum R2 are chosen among several training models.

The result shows almost 1 for multiple R2 and adjusted R2 for Ordinary Linear Regression that shows as there is existence of 100% of the Variation in total import costs. The 10-fold cross validation result on the test set for OLR model shows 2952689 and 0.8113806 for the smaller RMSE and maximum R2 respectively. When we compare the result of RF with the OLR model, the minimum RMSE and maximum R2 we can get from the results are 464212.8 and 0.9928060 which shows better performance than OLR.

The new R2 with 10-fold cross-validation for SVM model is 0.9999968. Which is better than R2 of OLR=0.81 and R2 of random forest 0.9928 respectively. Thus, in terms of RMSE 604748.2 for SVM, 2952689 for OLR and 464212.8 for random forest and RMSE for random forest shows better performance compared to SVM and Ordinary Linear Regression, but the Random Forest and Support Vector Machine models training process is more complex and more time consuming than OLR model.

5.2. Recommendation

This research is mainly done for academic purpose. Now days, the use & implementation of appropriate data analytic tools become crucial in every sector. Based on the findings of this study, the following recommendations are forwarded:

- The Ethiopian Revenue and Customs Authority is one of the main economy sectors that established for generating the country's revenue. In this course, the agency produces a large data to collect taxes from import and export of goods as well as from in land different revenues. Hence, the Agency should handle the data resources in convenient way for research and use.
- The study also recommends further research work by using different methodology or alternative machine learning techniques that can be a basis for comparison. Such will enhance the reliability and validity of the model developed result.
- This study would also invite any interested researchers to explore more on the application of machine learning techniques in customs and revenue areas or else in related and similar settings for the future.
- The data set and attributes are selected from the agency's ASQUDA database. However, it can also be possible to utilize other feature selection algorithms for the purpose of comparison.

References

[1]. Abdi, H., 2003. Partial least squares regression (PLS-regression). Thousand Oaks, CA: Sage. p. 792-795.

[2] Ahmed, S. R., 2004. Effectiveness of neural network types for prediction of business failure. Information Technology: Coding and Computing, 2, 455–459.

[3] Alpaydin, Ethem, 2014. Introduction to machine learning. MIT press.

[4]. Alpaydin, E., 2015. Introduction to machine learning.: MIT press.

[5]. Ayodele, T.O., Types of machine learning algorithms. 2010: INTECH Open Access Publisher.

[6] Barry de Ville., 2006. Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner. SAS Institute Inc., Cary, NC, USA.

[7] Bastabak, Burcu., 2012. Data mining framework to detect tariff code circumvention in Turkish customs database, the Middle east Technical university.

[8] Birihanu, Ermiyas Belachew and Akmel, Feidu Gobena, 2017. Student Performance Prediction Model using Machine Learning Approach: The Case of Wolkite University, International Journal of Advanced Research in Computer Science and Software Engineering, vol.7 no.2.

[9] Beiazen, Belete, 2011. Knowledge Discovery for Effective customer segmentation: the case of Ethiopian Revenue and custom Authority, Department of Information Science, Faculty of Informatics, Addis Ababa University, Addis Ababa.

[10] Bhise, R. B., S. S. Thorat, and A. K. Supekar. "Importance of data mining in higher education system." IOSR Journal Of Humanities And Social Science (IOSR-JHSS) ISSN (2013): 2279-0837.

[11] Chen, H. Fuller, S.S., Friedman, C., and Hersh, W., 2006. Medical informatics: Knowledge management and data mining in biomedicine, springer.

[12] Dyer, C. R. Machine Learning. Lecture Notes in Computer Science. Madison: University of Wisconsin. http://www.cs.Wisc.edu/%7Edyer/cs540/notes/learning/html, n.d.

[13] Ehsan H., Hamed Davari Ardakani and Jamal Shahrabi, 2010. Applications of data mining techniques in stock markets: A survey. Journal of Economics and International Finance vol. 2(7), pp. 109-118.

[14]. Graybill, F.A., 1976. Theory and applications of the linear model.

[15]. Haenlein, M. and A.M. Kaplan, 2004. A beginner's guide to partial least squares analysis. Understanding statistics, 3(4): p. 283-297.

[16]. Hakan Arslan, M., 2009. Application of ANN to evaluate effective parameters affecting failure load and displacement of RC buildings, Natural Hazards and Earth System Science, 9(3): p. 967-977.

[17]. Han J. and Kamber M, 2006. Data Mining: Concepts and Techniques, 2nd Edition, The University of Illinois at Urbana-Champaign, Morgan Kaufmann.

[18] Han, Jiawei, and Mitcheline Kamber. 2001. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers,

[19] H. Witten and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2 edition.

[20]. Hendrickx I. and Vanden B. , 2005. Hybrid Algorithms with Instance-based Classification. ILK, Tilburg University, PP. 158-169.

[21] Hevner, A., March, S., Park, J., and Ram, S. , 2004. "Design Science in Information Systems Research," MIS Quarterly (28:1), pp. 75-105.

[22] http://www.erca.gov.et

[23] https://www.r-project.org

[24] IMF, 2014. The Federal Democratic republic of Ethiopia, selected issue paper, Washington, DC.

[25] J. Nayak, B. Naik, and H. S. Behera, "A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges," International Journal of Database Theory and Application, vol. 8, pp. 169-186.

[26]. James, G., et al., 2013. An introduction to statistical learning.: Springer.

[27] Jeff Barnes, 2015. Azure machine learning: Microsoft azure essentials. ISBN: 978-0-7356-9817-8.

[28]. Jolliffe, I., 2005. Principal component analysis.: Wiley Online Library.

[29] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers, 2011. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute.

[30]. Kodratoff, Y. and R.S. Michalski, 2014. Machine learning: an artificial intelligence approach. Vol. 3.:Morgan Kaufmann.

[31]. Kuhn, M. and K. Johnson, 2013. Applied predictive modeling. Springer.

[32] Kumar, S. Arun, Rajan, S. sundau. 2015. Discovering Informative knowledge in complex data using pre and post processing mining, villupuram District, IJRS.

[33] L. Liu and M. T. Zsu , 2009. Encyclopedia of Database Systems. Springer Publishing Company, Incorporated,

[34] Lantz, Brett, 2015. Machine learning with R. second edition, Packt Publishing

[35] Mamo, Danil, 2013. Application of Data mining technology to support fraud protection: the case Ethiopian Revenue and Custom Authority, Department of Information Science, Faculty of Informatics, Addis Ababa University, Addis Ababa.

[36] Nick Wilson. 2016. Machine Learning Throwdown: ttps://blog.bigml.com/2012/08/02/machine-learning-throwdown. Accessed.

[37] Padhy, Neelamadhab, Dr Mishra, and Rasmita Panigrahi., 2012. "The survey of data mining applications and feature scope." arXiv preprint arXiv:1211.5723.

[38] R. C. Blattberg, B.-D. Kim, P. do Kim, and S. A. Neslin. , 2008. Database marketing: analyzing and managing customers.

[39] Robert Gentleman Rafael A. Irizarry Vincent J. Carey Sandrine Dudoit Wolfgang Huber Editors Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer.

[40] Ralf, Mikut and Markus, Reichl. 2011. Data mining tools. John Wiley & Sons, Inc.

[41]. R and Data Mining: Examples and Case Studies 1 Yanchang Zhao.

[42]. Satish Kumar et al , 2015. Analysis Clustering Techniques in Biological Data with R International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6 (2) .

[43] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, 2000. Improvements to the SMO Algorithm for SVM Regression, IEEE, vol. 11, pp. 1188-1193, SEPTEMBER.

[44] Taiwo Oladipupo Ayodele, 2010. Types of Machine Learning Algorithms, New Advances in Machine Learning, Yagang Zhang (Ed.), ISBN: 978-953-307-034-6, InTech.

[45] Talwar, Anish and Kuma, Yogesh, 2013. Machine Learning: An artificial intelligence methodology, International Journal of Engineering and Computer Science, Volume 2 Issue 12, Dec.

[46] UNDP, 2014 Annual Report of United Nations Development Programme Ethiopia.

[47] Xiao, Chengwei , 2015. Using Machine Learning for Exploratory Data Analysis and Predictive Models on Large Datasets. University of Stavanger.

[48]. Witten I. and Frank E. 2005, Data Mining: Practical Machine Learning Tools and Techniques. Second editions, Morgan Kaufmann, Massachusetts.

Appendixes

Appendix1: – Source Code

import the original files into R.

check the summary of the observed outcome

```
> library(Hmisc)
> describe(outcome)
>Summary()
>library(caret)
>nzv <- nearZeroVar(dataset,saveMetrics = FALSE)
>dataset.filtered.1<- dataset[,-nzv]
## calculate the correlation matrix
>cor.matrix <- cor(dataset.filtered.1)
#to build OLR model
>library (caret)
>imp_model <- lm(T_TAX_ETB ~ QUANTITY + NET_WT_KG + CIF_VALUE_ETB ,  data
= impdata.train)
> summary(imp_model)



##to improve the performance
> set.seed (1)
> train=sample(8000,2031)
> lm.fit=lm(T_IMORT_COST~QUANTITY+NET_WT_KG+CIF_VALUE_ETB+T_TAX_ET
B,data=impdata,subset =train)
> summary(lm.fit)

## Cross validation for OLR
> library(caret)
> library(rpart)
> train_control<- trainControl(method="cv", number=10)
> model<- train(T_IMORT_COST ~ QUANTITY + NET_WT_KG + CIF_VALUE_ETB +
T_TAX_ETB,  data=impo1, trControl=train_control, method="rpart")
> predictions<- predict(model,impdata)
> impdata<- cbind(impdata,predictions)
> confusionMatrix<- confusionMatrix(impdata$predictions,impdata$T_IMORT_COST)
> print(model)

##to build RF model
> library(randomForest)
> library(MASS)
> mysample<-impdata.train[sample(1:nrow(impdata.train),3000,replace = FALSE),]
```

```
> impdata.forest <- train(T_IMORT_COST ~ QUANTITY+ NET_WT_KG+ CIF_VALUE_ET
B+T_TAX_ETB,  data = mysample,
+ method="rf",
+ tuneGrid=data.frame(.mtry=1:15),
+ trcontrol=trainControl(method = "cv"),
+ maxit=100)

##to tune SVM model
> library(e1071)
> library(caret)
> library(Metrics)
> attach(impdata.train)
> x <- subset(impdata, select=-T_IMORT_COST)
> y <- T_IMORT_COST
> svm_model <- svm(T_IMORT_COST ~QUANTITY+NET_WT_KG+CIF_VALUE_ETB+T_
TAX_ETB+HS_CODE , data=mysample)
> svm_model

##to build the best model
> library(randomForest)
> library(MASS)
> mysample<-impdata.train[sample(1:nrow(impdata.train),3000,replace = FALSE),]
> impdata.rf <- randomForest(T_IMORT_COST ~ QUANTITY+ NET_WT_KG+ CIF_VALUE
_ETB+T_TAX_ETB,  data = mysample,
+ importance=TRUE,
+ ntree =500,
+ mtry=3)
Print(impdata.rf)

##to measure graphically Rf model
> library(randomForest)
> library(miscTools)
> library(ggplot2)
> impdata.rf <- randomForest(T_IMORT_COST ~ QUANTITY + NET_WT_KG + CIF_VALU
E_ETB+T_TAX_ETB, data=mysample, ntree=20)
> (r2 <- rSquared(mysample$T_IMORT_COST, mysample$T_IMORT_COST - predict(impdata
.rf, mysample)
> (mse <- mean((mysample$T_IMORT_COST - predict(rf, mysample))^2))
> p <- ggplot(aes(x=actual, y=pred),
+           data=data.frame(actual=mysample$T_IMORT_COST, pred=predict
(impdata.rf, mysample)))
> p + geom_point() +
+     geom_abline(color="red") +
+     ggtitle(paste("RandomForest Regression in R r^2=", r2, sep=""))
```

```
##cross validation for RF
> set.seed(100)
> library(caret)
> library(randomForest)
> control <- trainControl(method="repeatedcv", number=10, repeats=10, search="grid")
> set.seed(100)
> tunegrid <- expand.grid(.mtry=c(1:15))
> imp.rf <- train(T_IMORT_COST ~ QUANTITY+ NET_WT_KG+ CIF_VALUE_ETB+T_TA
X_ETB, data=mysample1, method="rf",  tuneGrid=tunegrid, trControl=control)

## to plot the model
> plot(T_IMORT_COST~QUANTITY+NET_WT_KG+CIF_VALUE_ETB+T_TAX_ETB,
+             data=impdata.train,
+           pch=20,
+           col="blue",
+           xlab="actual",
+           ylab="predicted",
+          main="model_svm")
> points(impdata.train$T_IMORT_COST, impdata.train$predicted.svm, col = "red",
pch=20

##to choose best tuning parameter for  SVM model
> imptuneResult <- tune(svm, T_IMORT_COST~QUANTITY+NET_WT_KG+CIF_VALUE_E
TB+T_TAX_ETB,  data = impdata.train  ranges = list(epsilon = seq(0,1,0.1), cost = 2^(2:9)) )
> print(imptuneResult)

##cross validation of SVM
> library(e1071)
> library(caret)
> dim(impdata.train); dim(impdata.test);
> trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
 > svm_Linear <- train(T_IMORT_COST ~ QUANTITY + NET_WT_KG + CIF_VALUE_ETB
+ T_TAX_ETB , data = impdata.train, method = "svmLinear",
+                                 trControl=trctrl,
+                                 preProcess = c("center", "scale"),
+                                 tuneLength = 10)
> svm_Linear
```

Appendix 2: Partial List of the imported custom goods dataset

IM_YEAR	HS_CODE	QUANTITY	NET_WT_KG	CIF_VALUE_ETB	T_TAX_ETB
2015	1051300	30,462	1,440.00	741,158.61	265,520.06
2014	1069000	4,000	2.00	8,177.94	2,684.38
2015	1069000	2,000	1.00	3,930.08	1,290.03
2014	4029900	2	377.54	3,299.79	1,154.91
2015	4029900	1,700	22,835.00	2,092,740.93	732,459.31
2015	4029900	19,500	19,500.00	1,811,992.75	634,197.46
2016	6021000	600	12.00	1,878.55	672.97
2014	15079010	1	361.17	11,922.58	8,041.76
2015	15079010	17,904	196,585.98	5,287,327.63	3,365,648.33
2015	15079010	6,600	72,468.02	1,905,164.63	1,212,732.51
2014	15079090	114	113,465.00	3,118,296.32	647,046.47
2015	15091010	1,965	739.45	274,220.55	184,961.74
2014	15099010	573	629.00	72,589.36	48,961.50
2015	15099010	540	840.00	51,318.76	34,614.48
2016	15099010	30	300.00	24,318.20	16,402.61
2014	15119090	15	14,800.00	286,951.75	109,041.66
2015	15119090	185,000	185,000.00	3,687,315.26	977,138.53
2015	15119090	127,680	127,680.00	2,126,871.06	563,620.82
2014	15121910	1,793	29,220.00	1,108,425.03	747,632.66
2014	15121910	228,960	208,353.60	6,537,319.16	4,409,421.75
2014	15121910	3,589	59,990.84	2,048,710.62	1,381,855.29
2014	15121910	5,700	27,288.00	937,168.17	632,119.92
2014	15121910	1,500	1,380.00	56,191.33	37,901.02
2015	15121910	6,350	69,952.00	2,471,102.85	1,572,980.50
2015	15121910	26,400	24,209.00	878,582.32	592,603.75
2015	15121910	30,552	71,519.55	2,814,067.82	1,898,088.66
2015	15121910	26,001	27,470.00	1,029,936.07	694,691.83
2015	15121910	145,260	134,414.40	4,956,620.14	3,343,240.21
2015	15121910	72,000	65,160.00	2,254,405.24	1,435,041.63
2015	15132910	1	1.00	327.97	172.17
2015	15149910	38	224.00	9,799.18	6,609.52
2015	15151100	60	60.00	6,001.55	3,288.83
2016	15155010	810	8,778.90	717,280.66	483,805.76
2014	15155010	10	58.00	12,546.21	8,462.40
2014	17041000	500	2,638.00	883,015.39	651,444.58
2014	17041000	2,300	28,750.00	2,268,510.12	1,673,593.32
2015	17041000	9,000	67,500.00	2,181,497.66	1,609,399.88
2015	17041000	1,300	20,800.00	1,034,766.26	763,398.79

2015	17041000	1,026,000	9,811.12	240,402.81	177,357.15
2015	17041000	9,000	67,500.00	2,205,447.35	1,627,068.77
2015	17041000	9,000	67,500.00	2,237,494.95	1,650,711.88
2014	20091200	1	5.00	219.73	148.18
2014	20094100	15	18.00	2,832.96	1,910.80
2014	20094100	150	975.00	21,114.43	14,241.66
2014	20095000	180	2,101.43	29,440.85	19,857.83
2016	20096100	222	2,275.20	61,888.02	41,743.45
2014	20097100	3,240	3,394.00	49,685.18	33,512.63
2014	22021000	140	1,443.00	111,686.25	139,616.15
2014	22030000	18,088	143,256.96	2,150,965.46	3,423,530.38
2014	22030000	1,429	17,148.00	420,093.97	668,632.03
2014	22030000	27,132	214,855.44	3,083,030.40	4,907,028.25
2015	22030000	1	7.00	124.92	198.81
2014	22042100	76	135.13	5,105.00	8,125.23
2015	22042100	339	3,263.00	259,260.28	412,645.08
2015	22042100	3,180	2,729.00	276,643.77	440,313.11
2014	22042900	60	542.63	14,213.61	22,622.71
2015	22042900	2,517	33,109.33	1,064,964.74	1,695,024.47
2015	22042900	89	64.22	87,462.06	139,206.79
2015	22043000	1	1.00	144.54	230.02
2014	22051000	132	439.16	23,154.58	36,158.76
2015	22059000	144	263.00	22,137.08	34,569.79
2014	22072000	7	20.00	9,297.94	929.79
2014	22082000	1	50.00	3,432.33	8,393.73
2015	22082000	4,990	1,564.58	179,882.60	439,902.84
2015	22082000	443	935.08	87,637.46	214,317.40
2014	22083000	372	3,418.42	935,525.08	1,489,005.08
2014	22083000	1,200	13,665.60	2,120,432.55	3,374,933.43
2014	22083000	277	7,062.00	1,028,860.93	1,637,560.74
2016	22083000	1,518	1,440.55	329,478.69	524,406.46
2015	22083000	10,632	14,850.00	401,772.43	639,471.03
2015	22083000	850	9,679.80	3,463,042.64	5,511,865.22
2015	22083000	4,758	4,514.73	1,371,125.30	2,182,317.27
2014	22084000	720	685.60	45,465.30	111,185.36
2015	22084000	55	337.00	70,843.01	173,246.56
2015	22084000	120	113.76	14,064.77	34,395.35
2014	22085000	26	241.25	48,753.33	119,226.23
2014	22085000	1,200	1,132.80	185,426.03	453,459.35
2014	22085000	6,660	6,301.80	843,863.08	2,063,667.10
2015	22085000	2,562	2,424.34	364,257.80	890,792.39

2015	22085000	100	999.72	196,897.62	481,513.09
2014	22086000	160	1,827.00	402,801.09	985,050.05
2014	22086000	204	572.62	30,191.07	72,926.51
2014	22086000	332	7,138.00	567,063.72	1,386,754.31
2014	22086000	840	793.26	160,607.58	392,765.81
2014	22086000	16,728	26,430.24	1,059,029.48	2,589,856.56
2015	22086000	16,512	26,088.92	1,053,907.95	2,577,331.86
2015	22086000	5,400	5,142.00	227,047.43	555,244.47
2015	22086000	360	99.63	76,410.46	186,861.76
2014	22087000	70	667.20	87,617.50	214,268.57
2014	22087000	88	1,710.00	41,353.02	101,128.79
2014	22087000	120	127.93	26,037.03	63,673.54
2014	22087000	534	2,208.39	116,436.65	281,252.70
2014	22087000	225	2,543.00	347,506.88	849,828.04
2014	22087000	17	12.94	1,851.86	4,528.71
2015	22087000	130	2,956.15	95,084.97	232,530.26
2015	22087000	150	2,000.00	111,236.88	272,029.75
2015	22087000	180	151.20	44,184.71	108,053.68
2015	22087000	630	3,176.00	125,270.83	302,591.68
2015	22087000	85	764.15	236,376.36	578,058.31
2015	22087000	280	1,971.55	403,815.44	987,530.60
2014	22089000	120	114.12	12,994.66	31,778.42
2015	22089000	10	79.00	27,905.02	68,241.70
2014	22090000	1,375	3,576.65	269,721.86	198,987.19
2014	22090000	50	860.00	15,631.91	11,532.42
2014	22090000	258	81.00	4,009.89	2,837.99
2014	22090000	300	102.00	10,304.89	7,602.42
2014	22090000	68	25.27	4,138.04	3,052.80
2014	27073000	720	1,579.00	65,273.70	23,384.28
2015	27073000	720	1,466.00	55,322.29	19,819.19
2014	27079900	1	947.00	202,036.90	72,379.70
2015	27079900	6,700	6,444.00	605,891.35	217,060.51
2015	27079900	5	124.00	7,588.44	2,490.89
2014	27101910	4,250	442.00	100,582.59	30,174.77
2015	27101910	50	291.30	7,565.19	3,744.76
2014	27101940	53,040	53,040.00	1,506,531.21	444,426.68
2015	27101940	336	336,960.00	6,774,543.51	1,795,254.01
2015	27101940	26	1,756.00	1,053.15	310.66
2014	27101990	10	50.00	968.61	174.34
2014	27101990	3,513	157,954.00	6,427,910.01	964,186.49
2014	27101990	14,920	542,016.00	22,425,067.58	4,036,512.12

2014	27101990	6,800	276,608.00	10,444,742.32	1,880,053.56
2014	27101990	23,992	567,066.01	27,338,918.82	4,100,837.78
2014	27101990	78,080	77,200.00	2,925,890.29	526,660.22
2015	27101990	1	1.00	3,990.75	718.33
2015	27101990	18,582	19,007.99	1,431,337.23	214,700.57
2015	27101990	141,525	141,525.00	7,635,207.11	1,374,337.27
2015	27101990	68	60,490.00	1,364,032.72	245,525.87
2015	27101990	1,004	17.80	74,914.72	13,484.64
2015	27101990	170	254.00	17,388.99	3,130.00
2014	28042100	6	135.00	6,049.26	1,436.66
2015	28042100	660	1,805.69	474,696.41	112,740.39
2015	28042100	1,980	1,980.00	764,485.95	181,565.39
2014	28043000	600	238.95	76,810.55	18,242.49
2014	30021000	201	152.10	1,406,199.30	210,929.89
2014	30022000	10,000	1,340.00	1,307,025.65	39,210.76
2014	30022000	2,000	28.00	173,274.26	5,198.22
2015	30022000	10,000	113.00	183,560.25	5,506.80
2014	30023000	1	34.00	10,555.45	316.66
2014	30023000	143	135.00	7,410.20	222.30
2015	30029010	102	33.00	267,761.73	101,749.45
2015	30029090	2	2.42	15,349.61	2,762.92
2014	30031000	100	216.80	490,773.71	39,261.87
2016	30031000	22,713	6,681.00	1,307,025.65	12,281,700.47
2015	30031000	123,182	10,457.00	5,646,341.63	282,317.06
2014	30032000	16	46.00	169,183.14	8,459.15
2015	30032000	97	4.60	22,458.02	1,796.64
2015	30033100	6	167.00	33,054.47	7,850.35
2014	30033900	41,257	2,904.18	2,264,368.48	181,149.47
2014	30041000	1	1.00	10,802.01	2,565.47
2014	30041000	960,000	42,600.00	11,606,533.05	928,522.60
2014	30041000	16,475	244.68	85,693.83	6,855.50
2014	30041000	140,040	18,672.00	1,466,394.40	117,311.55
2015	30041000	9,833	18,846.00	2,151,056.78	172,084.53
2015	30041000	50,000	5,000.00	2,902,773.36	232,221.86
2014	30042000	1	3.00	9,488.47	2,253.50
2014	30042000	1	1.00	5,164.06	1,226.45
2014	30042000	241,920	19,300.00	4,616,426.04	369,314.08
2014	30042000	1	14.00	5,464.68	437.17
2015	30042000	24,060	4,419.24	819,337.77	65,547.01

APPENDIX III

DISCUSSION POINTS FOR VALIDATING THE PROPOSED MACHINE LEARNING TRAINED MODEL

The following are questions/ discussion points intended for validating ‘‘**Discovering Knowledge from Complex data: the case of Ethiopian Revenue and Customs Authority’’**

1. Do you think the machine learning models trained in this research paper are relevant for your organizational data set?
2. What benefits would the trained model have or what problems would it solve?
3. Do you think it is possible to integrate with other models recommended for the agency so far? What challenges do you see in this regard?
4. Which of the machine learning model do you think should be improved or corrected?
5. What is your overall opinion on the three machine learning algorithms trained in this thesis?