


Sudan University of Science and Technology

College of Graduate Studies

Supply Chain Demand Forecasting based on Odoo ERP System, Using Linear Regression and Decision Trees

Case of a medium-sized company

A dissertation Submitted in Partial Fulfillment of the Requirements for MSc Degree in Information Technology

Submitted by

Omnia Yasir Izzeldin Makki

Supervised by

Dr. Wafaa Faisal Mukhtar

November 2020

Introductive

بسم الله الرحمن الرحيم

قال تعالى: {وَقُل رَّبِّ أَدْخِلْنِي مُدْخَلَ صِدْقٍ وَأَخْرِجْنِي مُخْرَجَ صِدْقٍ وَاجْعَل لِّي مِن لَّدُنكَ سُلْطَانًا نَّصِيرًا}

سورة الإسراء – آية 80

Dedication

To my loved ones, the ones who I know will always support me and believe in me.

To my parents, siblings, and family members,

> Your constant support and prayers made this work possible. I owe you the fruit of many days of persistent work and major mental breakdowns.

To Ahmed and Israa,

> Thank you for continuously pushing me forward when I was too tired to move. Thank you for always reminding me of how strong I am.

To Fatima, Elham and Rasha, my companions in this journey,

> I couldn’t have asked for better colleagues and study buddies.

I dedicate this piece of work to you.

Acknowledgement

I would like to thank Uz. Abdelhameed Kadafor for supporting me into applying for the Information Management Graduate Program in Sudan University of Science and Technology.

I would like to also thank the Data Mining community around the globe and its numerous contributions in sectors such as health, education, business…etc.

The published papers of this community were of tremendous help into shaping this thesis.

My thanks to College of Computer Science and Information Technology for its outstanding staff and facilities.

And finally, my most sincere thanks to Dr. Wafaa Faisal for constant guidance.

Abstract

The transactional data coming from a software system's transactional database provides more value than the day-to-day operations. Data mining techniques may be used to draw more value from these transactional data in order to enhance decision making process.

Demand Forecasting is an undoubtedly essential strategic tool to any profit-seeking organization who are seeking to decrease their operational costs. This study conducted a conjunction between CRISP-DM process and Ralph Kimball's data warehouse dimensional modelling methodology; in order to forecast sales demand in a given time frame. The application of CRSIP-DM went through two consecutive investigations. The first model used the Multiple Linear Regression which showed limitations in the model due to the mixed nature of ERP's data. This has led to the second model where Decision Tree C4.5 was used. The forecasting model showed remarkable accuracy in the forecasting of the sales demand.

مستخلص الدراسة

القيمة التي توفرها البيانات المستخصلة من قواعد البيانات الخاصة بالتطبيقات الموجهة للمعاملة تتعدى تلك المخصة بالعمليات اليومية. يمكن استخدام عمليات تنقيب البيانات في هذه البيانات لتحسين عملية اتخاذ القرار.

توقعات الطلب المستقبلي بلا شك تمثل أداة استراتيجية للشركات ذات الطابع الربحي التي تركز على تخفيض التكاليف. قامت هذه الدراسة بدمج طريقة تنقيب البيانات مع النموذج المتعدد الأبعاد المستخدم في مستودعات البيانات لتوقع الطلب المستقبلي في فترة زمنية محددة. تطبيق عملية تنقيب البيانات تم في مرحلتين متتاليتين. المرحلة الأولى أظهرت بعض القيود في النموذج المستخدم وهو نموذج الإنحدار الخطي، نتيجة لوجود بيانات ذات طابع مختلف في نظام تخطيط موارد المؤسسة المستخدم. قاد هذا إلى استخدام نموذج اخر تم فيها استخدام شجرة القرار. نموذج التوقعات المسقبلي أظهر نتائج جيدة في توقع طلبات المبيعات.

Table of Content

List of Tables

List of Figures

List of Abbreviations

BP	Back-propagation
CRISP-DM	Cross-Industry Standard Process of Data Mining
DBMS	Database Management System
DM	Data Mining
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load
IT	Information Technology
KDD	Knowledge Discovery in Databases
LAN	Local Area Network
MAE	Mean Absolute Error
OLTP	Online Transactional Processing
OODBMS	Object Oriented Database Management System
ORD	Object Relational Database
PoS	Point of Sale
RAM	Random Access Memory
RDBMS	Relational Database Management System
SCM	Supply Chain Management
SVR	Support Vector Regression
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER 1

INTRODUCTION

1.1 Research Background

The software systems in today's corporate worlds provide a rich ground for data. The transactional data such as sales orders, purchase orders, invoices, manufacturing data, inventory movements are usually stored in massive databases and used in day-to-day operations. These data may be used for further analysis and finding additional value to them. "This stems from the fact that finding useful information is challenging due to the large volume of the data or the nature of the data might make basic statistical analysis impossible". (Tan, Steinbach, and Kumar, 2006)

1.1.1 Enterprise Resource Planning Systems

Enterprise Resource Planning (ERP) is a set of applications (modules) for core business operations and back-end management that was originally developed for manufacturing and commercial companies. Figure 1.2 shows the different ERP system modules.

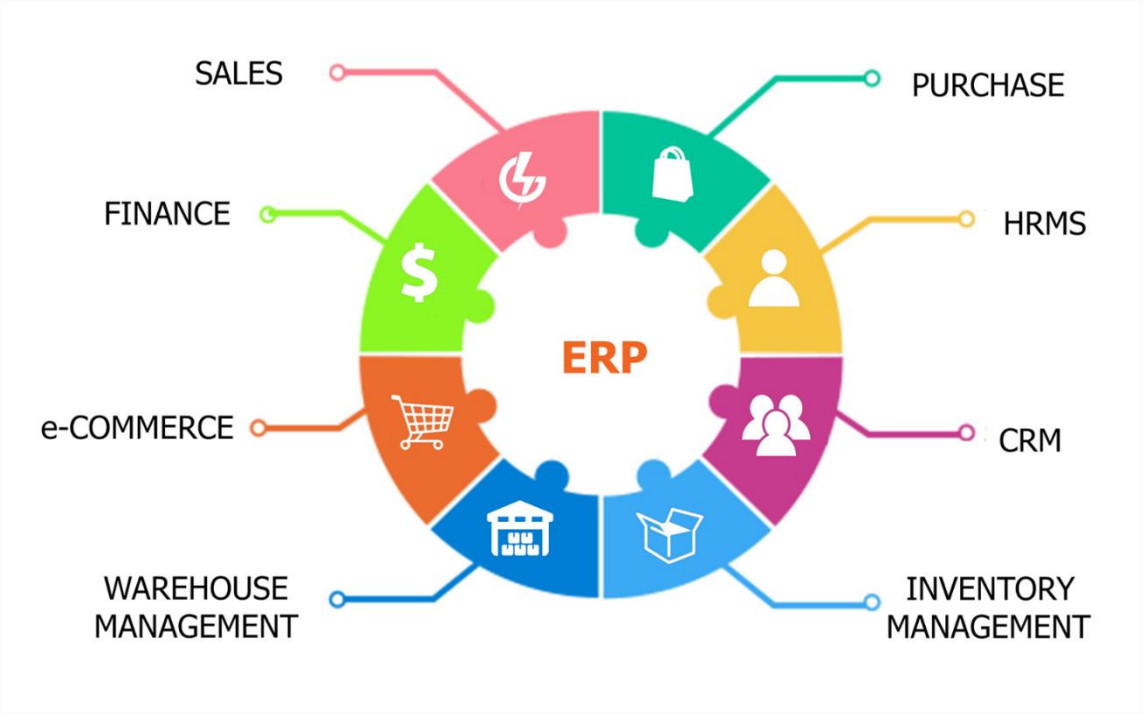

Figure (1.1) Different ERP Modules.

1.1.2 Supply Chain and Demand Forecasting

(Computerworld, 2001) defined Supply Chain management (SCM) as the management that allows an organization to get the right products and services to the location they required on time, in the suitable quantity and at a satisfactory cost. The management of this process involves effectively supervising connections with customers, suppliers. It also involves controlling the warehouses (inventory), forecasting demand.

A Typical SCM module list in ERP includes Sales, Purchases, Production, Inventory and warehouse. According to (Pontius, 2019), Inventory Management is a component of SCM that involves supervising non-capitalized assets, or inventory, and stock items.

According to (Kot, Grondys, and Szopa, 2011), typical cause of constantly increasing costs (of Inventory) is excessive inventory levels throughout the chain. The instability of the level of supply to the level of demand in the market results in stock surplus. Forecasting the demand in the market is the starting point for reduction in inventory levels.

Demand Forecasting is a technique for estimation of probable demand for a product or services in the future. It is based on the analysis of past demand for that product or service in the present market condition.

1.2 Problem Statement

Demand Forecasting imposes a headache for enterprises working in retail. As it can impose overheads resulting from surplus inventory, and that can increase the operational cost. Additionally, the software available now in the market does not provide accurate predictions for demand forecasting which makes these retail companies unable to predict their sales demand.

1.3 Scope and Objectives of the Study

The research will be studying the case of a single medium-sized company based in Khartoum, Sudan. Through its existing ERP system, and its integrated online transaction processing (OLTP) database.

The objectives of the thesis are:

- Accurately predicting the sales demand in a certain timeframe
- Make use of available sales information in improving the prediction model of demand

1.4 Methodology

This research is presented with a clear business case, which requires a side of business analysis to be implemented. A number of stakeholders are involved in this study, and therefore a means of effective communications and overall planning is required. According to (Wirth, 2000) “The generic CRISP-DM process model is useful for planning, communication within and outside the project team, and documentation”. Figure (1.3) describes the Cross-Industry Standard Process of Data Mining (CRISP-DM), which defines six standard phases for DM (Data Mining) process.

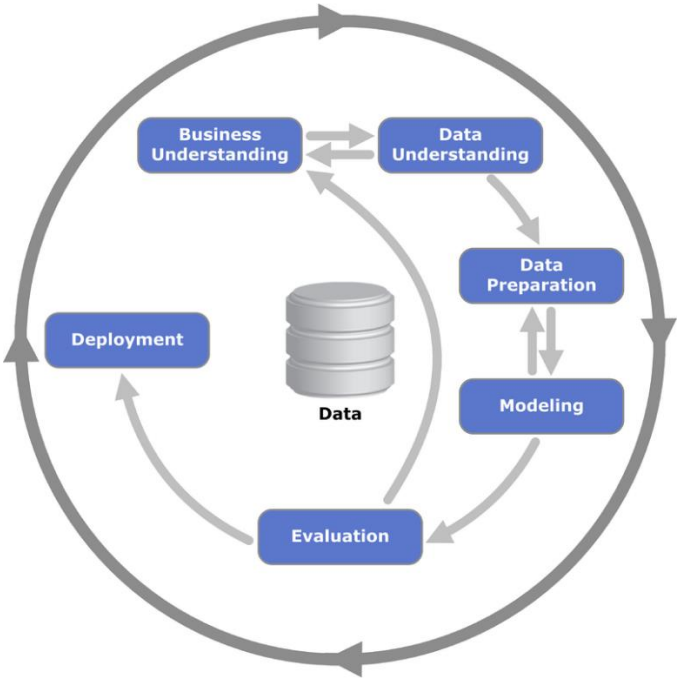

Figure (1.2) The Cross-Industry Standard Process of Data Mining (CRISP-DM).

1.5 Organization of Thesis

The thesis consists of five chapters. Chapter One performs as an introduction to the research, providing a background and defining main research problem and objectives. Chapter Two discusses the related work done by previous studies and was of affection to this thesis. Chapter Three shows in details how the proposed methodology was used to meet the research objectives. Chapter Four discusses results found in Chapter Three. Chapter Five concludes the research and addresses future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter discusses the literature concerned with Data Mining and Knowledge Discovery Forecasting techniques used in Supply Chain Management modules of Enterprise Resource Planning Systems. The chapter firstly talks about ERP systems, and Odoo ERP. Then deliberates how it is possible to integrate Data Mining Techniques in ERP Systems. Lstly, it discusses using Data Mining techniques in Supply Chain Management Forecasting and Analysis, both in general and in ERP Systems.

2.2 Enterprise Resource Planning Systems

According to (Elragal and Al-serafi, 2014), the use of enterprise resource planning (ERP) software has become increasingly more common in a lot of today's businesses as it is adopted in many firms in attempts of improving business performance. ERPs present a holistic view of the enterprise's operations all linked with one another. The processes are easier to monitor and reports are easier to generate.

(Saleem, 2017) states that an Enterprise Resource planning (ERP) integrates the functionality of all the business departments in an organization in a single system to carry out the particular needs of these different departments and allow efficient information sharing.

ERP provides two major benefits that do not exist in non-integrated departmental systems according to (Umble, Haft, and Umble, 2003). The first benefit is a unified enterprise view of the business that encompasses all functions and departments. The second benefit is an enterprise database where all business transactions are entered, recorded, processed, monitored, and reported.

(Nazemi and Tarokh, 2012) argues that an ERP system contribute to organizational efficiency by contributing to operational and delivery by providing a critical role in improving the way a firm takes customer orders and processes them into invoices; a process is also known as 'Order Fulfillment'.

2.3.1 Odoo ERP

Odoo is a suite of open source business apps that covers CRM, eCommerce, accounting, inventory, point of sale, project management, etc.

Figure (2.1) Odoo ERP Applications

Figure (2.1) shows the main applications in Odoo ERP. These are: Customer relationship management, Sales & distribution, Purchases, Manufacturing, Quality, Inventory, Maintenance, Human Resources, Accounting and Finance.

2.3.1.1 Programming Languages used in Odoo

Programming languages used in Odoo are Python, JavaScript, XML.

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding.

JavaScript is a scripting or programming language that allows you to implement complex features on web pages

XML stands for eXtensible Markup Language. XML is a markup language much like HTML. XML was designed to store and transport data.

2.3.1.2 Database used in Odoo

The database used in Odoo ERP is PostgreSQL.

PostgreSQL is an open source, object-relational, database system that uses and extends the SQL language combined with many features that safely store and scale the most complicated data workloads.

An Object-relational database (ORD) is a database management system (DBMS) that's composed of both a relational database (RDBMS) and an object-oriented database (OODBMS). ORD supports the basic components of any object-oriented database model in its schemas and the query language used, such as objects, classes and inheritance.

A relational database is a type of database. It uses a structure that allows us to identify and access data in relation to another piece of data in the database. The data is represented and stored in the form of tables.

An object-oriented database (OODBMS) or object database management system (ODBMS) is a database that is based on object-oriented programming (OOP). The data is represented and stored in the form of objects.

2.3 Data Mining and the Usage of Data Warehousing

(Jain and Srivastava, 2013) Define Data mining as extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as ‘Knowledge mining from data’ or “Knowledge mining”. Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.

Knowledge Discovery in Databases (KDD) is sometimes used as a synonym for DM, but, as a wider concept, it denotes a process which tries to convert raw data into useful information using DM, and also includes pre- and post-processing phases in addition to the actual data mining. (Han and Kamber, 2006)

According to (Amirthalingam et. al., 2014), Data warehousing helps set the stage for KDD in two important ways. The first way it helps set the stages of KDD is by data cleaning. As organizations are forced to think about a unified logical view of the wide variety of data and databases they possess, they have to address the issues of mapping data to a single naming convention, uniformly representing and handling missing data, and handling noise and errors when possible.

The second way datawarehousing helps set the stages of KDD is in data access. Uniform and well-defined methods must be created for accessing the data and providing access paths to data that were historically difficult to get to (for example, stored offline).

A Dimensional Model is a database structure that is optimized for online queries and Data Warehousing tools. Dimensional models implemented in relational database management systems are referred to as star schemas because of their resemblance to a star-like structure. It is comprised of "fact" and "dimension" tables.

(Kimball and Ross, 2013) argue that when a hierarchical relationship in a dimension table is normalized, low-cardinality attributes appear as secondary tables connected to the base dimension table by an attribute key. When this process is repeated with all the dimension table's hierarchies, a characteristic multilevel structure is created that is called a snowflake. Figure 1.1 demonstrates the structure of a dimensional model with a multilevel structure or a Snowflake Schema.

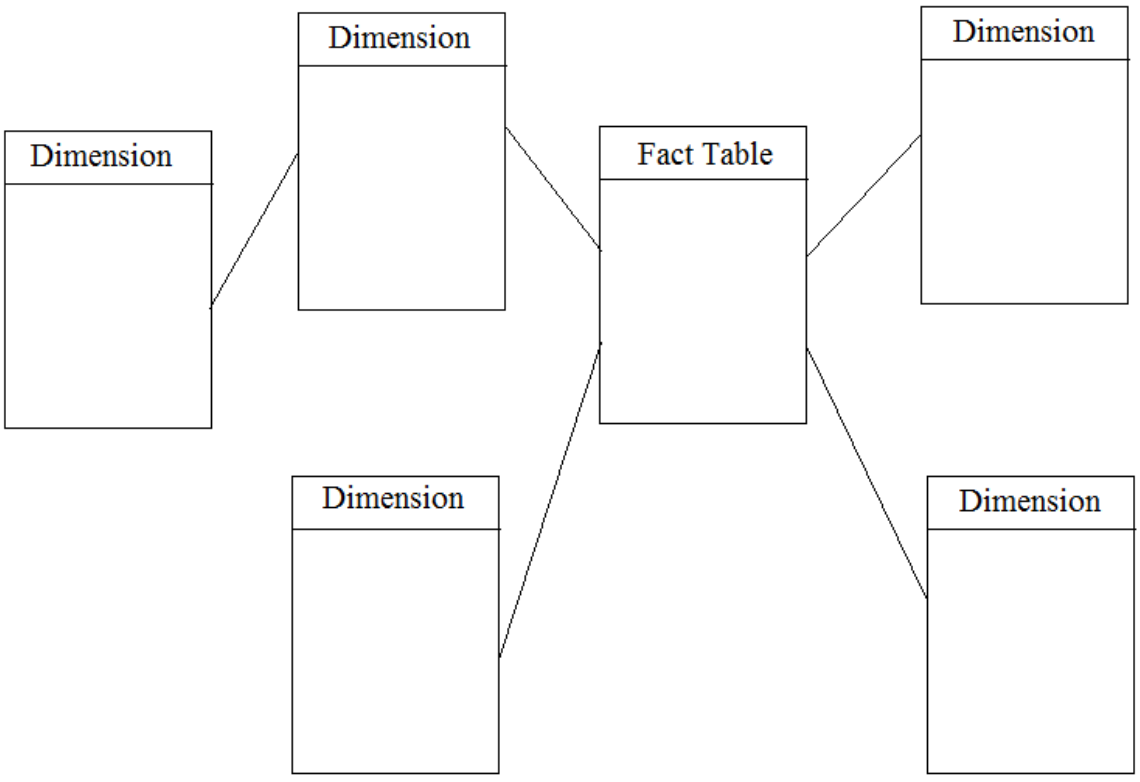

Figure (2.2) Snowflake Schema

Dimensional modeling is widely accepted as the preferred technique for presenting analytic data because it addresses two simultaneous requirements: Deliver data that's understandable to the business users and Deliver fast query performance. (Kimball and Ross, 2013)

Datawarehouse Creation (ETL Process)

According to (Simitsis, 2003), Extraction-Transformation-Loading (ETL) tools are pieces of software responsible for the extraction of data from several sources, their cleansing, customization and insertion into a data warehouse.

The most prominent tasks for ETL tools include: (a) the identification of relevant information at the source side; (b) the extraction of this information; (c) the customization and integration of the information coming from multiple sources into a common format; (d) the cleaning of the resulting data set, on the basis of database and business rules, and (e) the propagation of the data to the data warehouse and/or data marts. (Simitsis, 2003)

Talend Open Studio for Data Integration

Talend is an open source ETL tool for data integration. It's considered as a software tool with an open, scalable architecture. It allows faster response to business requests. The tool offers to develop and deploy data integration jobs faster than hand coding. It allows easily integrating data with other data warehouses or synchronize data between systems.

Data integration involves combining data stored in different sources and providing users with a unified view of these data. It helps you to manage various ETL jobs, and empower users with simple, self-service data preparation.

WEKA (Waikato Environment for Knowledge Analysis)

(Garner, 1995) defines WEKA as a workbench designed to aid in the application of machine learning technology to real world data sets. Each machine learning algorithm implementation requires the data to be present in its own format, and has its own way of specifying parameters and output.

2.4 Integrating Data Mining Techniques in ERP Systems

According to (Kolkas, El-Bakry, and Saleh, 2014) , The central transactional database of the ERP offers a rich source of data to apply analytical processing activities to gain benefits of ERP data. Table 1.1 provided by (Kolkas et. al., 2014) shows the various Data Mining Tasks that can be done to ERP data.

Table (2.1) Data Mining Tasks that can be performed on ERP systems

ERP Module	Data Mining Sample Tasks
Accounting and Finance Management	• Forecast total company profiles based on historical data • Predicting Cash Flow • Predicting overall profit/loss
Human Resources Management	• Select candidate employee based on historical data
Vendors and Purchase Management	• Determine best arrangement and quantities of purchase orders. (Purchase what of who and what amount)
Production Management	• Applying Classification/ Clustering technique to designs given designs parameters to find out if design may result in unacceptable defect percentage in final products
Customer Relationship Management	• Identify customers' behavior patterns. • Find people in similar life stages and may behave in the same way
Sales and Distribution Management	• Determine what items sold together more probably for PoS • Determine customer behavior over selling websites • Effectively segment customers into manageable groups • Focus marketing efforts on prospects more likely to purchase • Forecast sales for a given period of time • Discover which customers will respond to a given offer

According to (Amirthalingam, Shaheen, Kousar, and Bilfaqih, 2014), There are four major steps in integrating ERP with the multi-dimensional Data warehouse concept in order to have a solid ground for the data mining processes.

Firstly, the data mart of ERP must be defined. The overall methods and scopes during the evaluation stage of the design and determine using Schema or Snowflake in accordance with the requirements must all be set. It is an emphasis on the single business activity of the enterprise, such as importing, purchasing or ordering of goods.

Secondly a choice of a suitable fact for the particular data mining task must be made. The multidimension is built and completed using Fact. Therefore, Fact must be able to answer all the possible questions that may occur during the process of decision-making.

Thirdly, the use of simple and useful message in words. Codes, abbreviations and Null are all unfitted for dimensions. The explicated time, names or addresses allow more flexibility in inquiries. Each and every item of the Dimension Table carries multiple feedback capabilities in processing the Fact Table. When there are changes over the data, the new data will be added to the "newly added data row". Use the time to tell the track record at certain point. This method allows unlimited times of tracking the changes over data. The deficiency is that it must use time to identify the updated data row and increase the data rows of the dimension table. Use additionally built record column plus the time column to record the changes in time. The good point of it is that there is no need to build additional data row or to change the values architecture of dimension table.

Lastly, The design of aggregation. Aggregation is the advanced calculated total amount to increase the analysis speed when facing complicated inquiries.

(Moriya and Gosawi, 2015) provided a three views framework for Data mining intelligent systems based on ERP. An Outer View – consisting of the interaction with the ERP, Inner View – consisting of the single shared database by all requests coming from the outer view, Knowledge Discovery View – concerning with the central database having all kind of data saved from outer and inner views.

2.5 Using Data Mining Techniques in Supply Chain Management Forecasting

Data Mining uses many several predictive and statistical methods in order to explore and analyze data, according to (Kolkas, El-bakry, and Saleh, 2015). Such methods include association rule, linear regression, neural networks, regression trees, cluster analysis and classification trees. Below are research papers that used Forecasting in the field of Supply Chain Management.

2.5.1 Supply Chain Management Forecasting and Analysis

(Yu, Qi, and Zhao, 2013) presented an approach for newspaper/magazine sales forecast using Support Vector Regression (SVR). (Yu et al., 2013) states that recent theoretical studies in statistics proposed a novel method, namely support vector regression (SVR), to overcome over-fitting problem. In contrast to traditional regression model, the objective of SVR is to achieve the minimum structural risk rather than the minimum empirical risk. The experiment showed that SVR is a superior method in this kind of task in.

(Ozsaglam, 2015) has shown a method for providing sales forecast of an electronics store. Regression analysis and Naïve Bayes classifier were used. Sales forecasts were measured against real outcomes.

In their paper, (Pan, 2016) resolved a method of Inventory Prediction Based on the Improved BP Neural Network. The research collected 150 sets of data. The first 130 sets of data are the training data and the last 20 set of data sample data are the sample data. The comparison of the predicted results and the actual results were used in explaining that the results of the method are better than other algorithms.

A comparison of model forecasts of demand for multiple products made by (Valencia, Díaz, and Correa, 2016); choosing the best among the following: autoregressive integrated moving average (ARIMA), exponential smoothing (ES), a Bayesian regression model (BRM), and a Bayesian dynamic linear model (BDLM). The paper results have shown that the model can provide a better solution for inventory management than what was achieved in the real case.

(Tanizaki, Hoshino, Shimmura, and Takenaka, 2019) did a comparison between Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression and Stepwise method as a demand forecasting method for Restaurants PoS System extracted data. The dataset included internal data such as PoS data and external data in the ubiquitous environment such as weather, events, etc. The results of the paper resolved that there was no big difference in the forecasting rate using Bayesian, Decision, and Stepwise, however the forecasting rate of Boosted was a little low.

A case study was conducted by (Merkuryeva, Valberga, and Smirnov, 2019) on the demand forecasting in the pharmaceutical field, and specifically for one pharmaceutical product. The paper included a comparison between three forecasting methods, that are: simple moving average method, multiple linear regression, and symbolic regression with genetic programming. The paper concluded that symbolic regression-based forecasting model provides the best fitting curve to history demand

data, lower error estimates across all scenarios and performed experiments, and the ability to enhance the accuracy in the prediction of demand peak sales in the study.

(Guanghui, 2012) created a time-series forecasting based on Support Vector Regression (SVR) by optimizing the parameters of SVR using the Genetic Algorithm (GA). (Guanghui, 2012) compared the results to the RBF neural network method. The result showed that SVR was in fact superior to RBF in prediction performance.

2.5.2 Supply Chain Management Forecasting and Analysis in ERP Systems

The utilization of clustering techniques for detecting deviation in product sales and also to identify and compare sales over a particular period of time shows in (Hanumanth and Babu, 2013). The paper shows how the annual sales data of a steel major were also utilized to analyze Sales Volume and Value with respect to dependent attributes like products, customers and quantities sold. In their paper, (Hanumanth and Babu, 2013) have analyzed sales data with clustering algorithms like K-Means andEM which revealed many interesting patterns useful for improving sales revenue and achieving higher sales volume. The study confirms that partition methods like K-Means and EM algorithms are better suited to analyze our sales data in comparison to Density based methods like DBSCAN and OPTICS or Hierarchical methods like COBWEB.

Table (2.1) provides a summary for the related work discussed in Chapter 2, both Framework papers and Forecasting and Analysis Papers.

Table (2.2) Key Related Work Summary

#	Paper and Author	Algorithms Used	Data	Results	Purpose of the Study
	Framework Papers				
1	**Integrated Data Mining and Knowledge Discovery Techniques in ERP** (Amirthalingam et al., 2014)	-	-	Framework for integrating ERP with the multi-dimensional Data warehouse concept.	A framework for implementation of Data warehouse from ERP
	Forecasting and Analysis Papers				
2	**Support Vector Regression for Newspaper/Magazine Sales Forecasting** (Yu et al., 2013)	Support Vector Regression	Dataset of 1000 records	Prediction of sales	Overcome overfitting problem in traditional regression models
3	**Data Mining Techniques for Sales Forecasting** (Ozsaglam, 2015)	Regression Analysis and Naïve Bayes	Not Available	Prediction of sales	Combine Regression Analysis and Naïve Bayes classifier
4	**Analysis and Prediction of Sales Data in SAP- ERP System Using Clustering Algorithms** (Hanumanth and Babu, 2013)	Compared numerous clustering algorithms to analyze sales data	Not Available	Partition methods (K-Means and EM) are better suited to analyze sales data in comparison to Density based methods (DBSCAN and OPTICS) or Hierarchical methods (COBWEB).	utilized clustering techniques for detecting deviation in product sales
5	**Inventory Prediction Research Based on the Improved BP Neural Network Algorithm** (Pan, 2016)	BP Neural Network	Dataset of 150	An improved BP neural network method	Inventory Prediction Based on the Improved BP Neural Network
6	**Multi-product inventory modeling with demand forecasting and Bayesian optimization**	Bayesian optimization to Bayesian dynamic linear model	Dataset of 92 records	Bayesian dynamic linear model proved to be the best in comparison to	a comparison of model forecasts of demand for multiple products

	(Valencia et al., 2016)				
7	**Demand Forecasting of Supply Chain Based on Support Vector Regression Method** (Guanghui, 2012)	Support Vector Regression and Neural Networks	Dataset of 52 records	Support vector regression was better than neural networks in the accuracy of prediction	Demand forecasting using Support Vector Regression (SVR) by optimizing the parameters of SVR using the Genetic Algorithm (GA) and comparison with BP neural networks
8	**Demand Forecasting in Restaurants Using Machine Learning and Statistical Analysis** (Tanizaki et al., 2019)	Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression and Stepwise method	Not Available	No big difference in the forecasting rate using Bayesian, Decision, and Stepwise. Low forecasting rate of Boosted	comparison between Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression and Stepwise method as a demand forecasting method for Restaurants PoS System data
9	**Demand Forecasting in Pharmaceutical Supply Chains: A Case Study** (Merkuryeva et al., 2019)	simple moving average method, multiple linear regression, and symbolic regression with genetic programming	Dataset of 41 Points	Symbolic regression-based forecasting model provides the best forecasting for values	Comparison between three forecasting methods for a single pharmaceutical item

2.6 Chapter Summary

Supply Chain Demand Forecasting is a field of importance to organizations. Academic society has provided several research papers, along with its relation to Data Mining techniques, which this chapter has discussed. The research hence adopts (Amirthalingam et al., 2014) approach in integrating ERP with the multi-dimensional Data warehouse concept in order to have a solid ground for the data mining processes. The next chapter discusses the methodology thoroughly.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter will thoroughly deliberate the methodology used in this thesis. In order to achieve research objective, a framework combining the data mining CRISP-DM process in conjunction with Ralph Kimball's data warehouse dimensional modelling methodology was used. As previously shown in Figure (1.3) of this research study, CRSIP-DM consists of 6 main interrelated phases, each consisting of a number of tasks and respective outputs. These phases are: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

The conjunction between CRISP-DM process in conjunction with Ralph Kimball's data warehouse dimensional modelling methodology happens in the third phase "Data Preparation", where the Kimball methodology is used to create the data warehouse. Kimball methodology is a bottom-up methodology, where a datawarehouse is created as a series of one data mart at a time. (Kimball and Ross, 2013).

3.2 CRSIP-DM Process

The application of CRSIP-DM went through two clear iterations. These iterations are discussed in the following sections of this chapter. Figure 3.1 shows the sequence of steps used in this research.

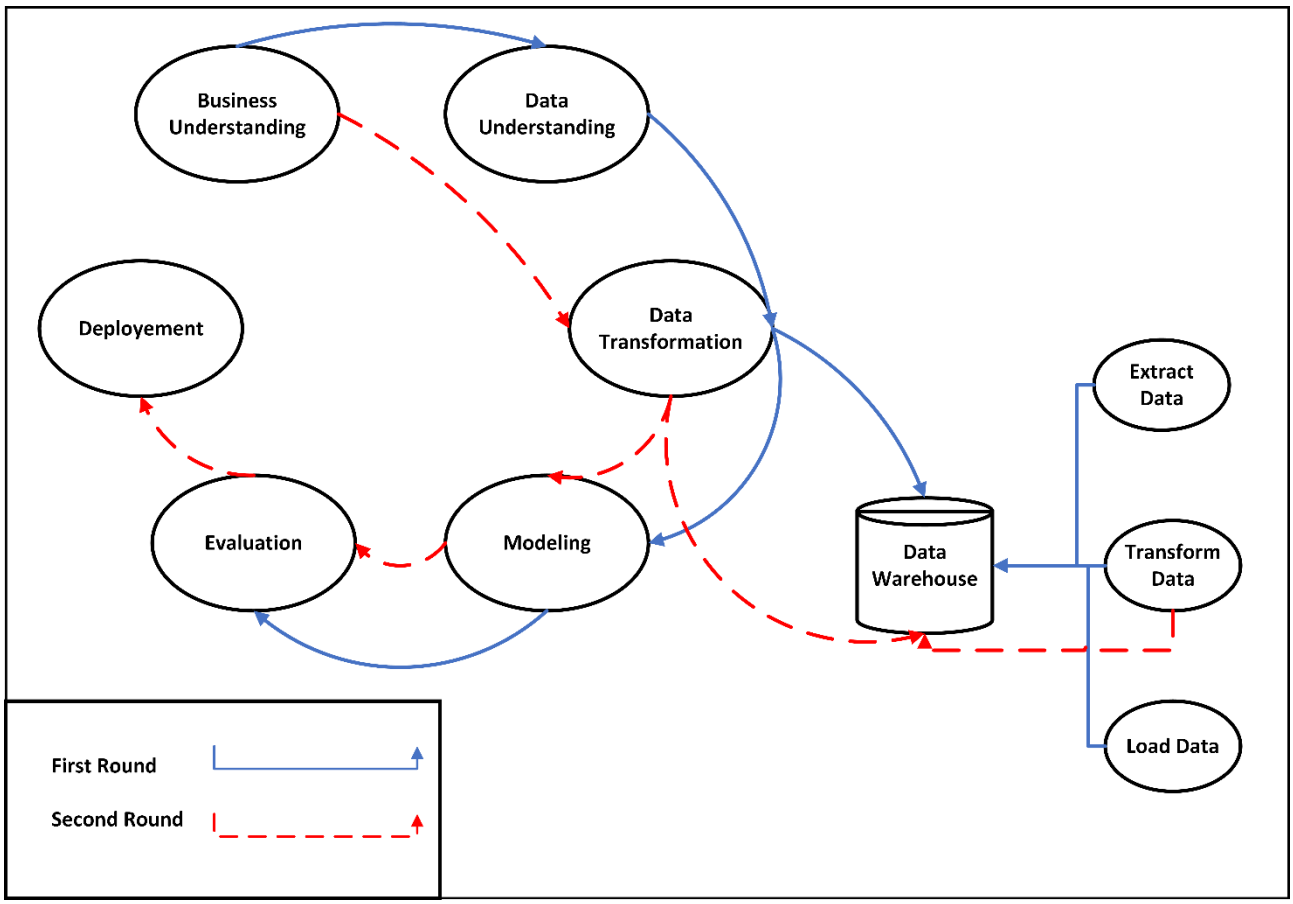

Figure (3.1) Research Methodology

3.2.1 First Round of CRISP-DM Cycle

As illustrated in Figure (3.1), the first round went through 5 of the 6 CRISP-DM Steps. They are: Business Understanding, Data Understanding, Data Transformation, Modeling and Evaluation. The details of the steps are described below.

3.2.1.1 Business Understanding

The first step in business understanding is to determine business objectives. The most apparent business objective is to be able to shift the inventory management strategy from push-based to pull-based. In order to do so, a forecast of customer demand must be provided.

Setting a Business Success Criteria is essential in business understanding step. A successful implementation for this data mining project would be in accurately predicting sales demand in a given strategic time frame.

The second step in business understanding is assessing the situation. That can be done firstly by studying the Inventory of Resources, Requirements, Assumptions, and Constraints Inventory of Resources.

In this research, the Employees are experts in either Sales or Supply chain processes or both. This provide a solid business point of view guidance. A more in-house technical expertise is provided through the company's ERP system administrator. They provide knowledge on the sales transaction logs and its relation to the business's workflows, in addition to a transactions issue log from a functional point of view.

However, it is clear that there is a lack in basic datawarehousing concepts and therefore no previous knowledge in data cleaning for analysis tasks. In addition to that, an access to the company's operational database was provided. This insured access to both sales and inventory data. Also, the hardware for the project needs to have a good storage space for the datawarehouse, and good RAM for ETL processes.

Some assumptions provided in the research state that it is assumed that the data quality to be high. This means that data entry issues are supervised (controlled) from the beginning, not like in flat files. Also, data types are governed, this means less time spent on data cleaning tasks.

Another assumption is that there will be quite a number of sale data, both vertically and horizontally, putting into consideration two factors for each aspect. The first considered point which leads to vertical expansion - in the number of features - is the comprehensive nature of Odoo ERP system. Being an ERP means that there will be an integration between modules and therefore their data. In this case it's the sales and the inventory module. The other aspect is the reported nature of sales given by the sales department. According to the sales department, there's an average of 500 sales order per month, and this leads to the horizontal expansion of data – meaning a lot of records.

Also, a definition of all the risks and contingencies associated to this research was a necessity. If the quality of the data is too bad, this might lead to a lengthy data cleaning and transformation process, and this in itself might cause a delay in the project.

The third step is determining Data Mining goals. Data mining goal of this particular research is to predict the demanded quantity of each product given their sales history for the past two years.

The success criteria for this data mining project is the produced model's ability to accurately predict the demanded quantity of each product.

The last step in business understanding is producing a project plan. A preliminary project plan was created with the management in the company. The plan consisted of the business objective of the project, a schedule with detailed tasks and durations. The iterative nature of the data mining process was put into consideration.

3.2.1.2 Data Understanding

The first step in data understating is data collection. Data was collected from the company's database server, then was loaded into the researcher's machine using an ETL tool. This step created the main datawarehouse.

Database Structure was a relational database, that is PostgreSQL, version 9.3. The existing data consisted of Time series operational data. These operational data belonged to either Sales, procurement, inventory, or accounting and Finance departments. Transactions were all in the period of May 2017 - July 2018.

The second step in data understanding is data description. Since all tables are coming from a single ERP's transactional database, the coding scheme is completely unified. Being a transactional, enterprise-based database, the OLTP consisted of a large number of tables that is 473 table. In these tables, data types ranged from numeric, categorial, nominal, date and time and boolean.

3.2.1.3 Data Transformation

In this part, the conjunction between the CRISP-DM and data warehouse dimensional modeling took place. Ralph Kimball's data warehouse dimensional modelling methodology consists of creating data marts that build up to become a datawarehouse.

Snowflake schema was chosen to create the data mart. Snowflake schema was selected because it provides better data quality. Nature of data in a PostgreSQL is more structured, because it's an Object-relational database. Hence, the aim of the data mart schema was to reduce data integrity problems, and use less disk space is used then in a denormalized model. This is what a snowflake schema provides. (Drkusic, 2016)

Datamart Creation

1) Extraction

The process of creating a data mart started by connecting the Talend ETL tool to the OLTP to extract the relevant tables (later in the process called dimensions).

2) Transformation

The Sales fact table was created, from the accompanying dimensions. These dimensions are: Company, Users, Currency, Sale Layout category, Sale Oder, Partner, Product, Product Packaging, Product Template, Stock Location, Users, Tax, Invoice.

The produced Sales Fact table contain 18,348 records, and a total of 34 attributes. Table 3.1 shows The attributes of the produced Fact table.

Table (3.1) Generated Sales Fact Table

Sales Fact Table					
#	Attribute	Key Type	Table	Data Type	Comments
1	id			Integer	
2	order id	F	Sale Order	Integer	
3	product_uom	F	Product	Integer	
4	write_uid	F	Users	Integer	
5	currency_id	F	Currency	Integer	
6	create_uid	F	Users	Integer	
7	price_tax	F	Tax	Float	
8	customer_lead	F	Customer	Integer	
9	company_id	F	Company	Integer	
10	order_partner_id	F	Customer	Integer	
11	product_id	F	Product	Integer	

12	name	F	Product	Text	
13	salesman_id	F	Partner	Integer	
14	product_packaging	F	Product Packaging	Integer	
15	route_id	F	Stock Location	Integer	
16	cost	F	Product	Float	
17	price_unit	F	Product	Float	
18	invoice_status	F	Invoice	Text	
19	create_date			Date/Time	
20	quantity			Integer	
21	price_subtotal			Float	
22	price_reduce_taxexcl			Float	
23	qty_to_invoice			Integer	
24	layout_category_sequence			Integer	
25	state			Text	
26	qty_invoiced			Integer	
27	sequence			Integer	
28	discount			Float	
29	write_date			Date/Time	
30	price_reduce			Float	
31	qty_delivered			Integer	
32	layout_category_id			Integer	
33	price_reduce_taxinc			Float	
34	price_total			Float	

The next step in data transformation is to select relevant data. The selection versus exclusion can happen on both a vertical and a horizonal manner.

A horizontal selection was based on the attribute **state**, where the value of the attribute was equal to "Sale". Other values "draft, cancel" were ignored because they do not serve the purpose of the study.

Another horizontal selection was based on the quantity attribute, where any instance with quantity = Zero was excluded.

On a vertical exclusion manner, some attributes were excluded. The table below provides a rationale for attribute exclusion. Table 3.2 shows the excluded attributes and the rationale of exclusion.

Table (3.2) Excluded Attributes

Excluded Attributes		
#	Attribute	Rationale for Exclusion
1	write_uid	Salesman id is more relevent in this case
2	currency_id	The value is = 1 for all records
3	create_uid	Salesman id is more relevent in this case
4	price_reduce_taxexcl	Similar to price_unit field
5	price_tax	The value is = 0 for all records
6	customer_lead	The value is = 0 for all records
7	company_id	The value is = 1 for all records
8	name	Similar to product_id field
9	product_packaging	The value is = 0 for all records
10	route_id	The value is = 0 for all records
11	price_subtotal	Is a computed field and won't affect
12	qty_to_invoice	Similar to product_uom_qty field
13	layout_category_sequence	The value is = 0 for all records
14	qty_invoiced	Similar to product_uom_qty field

15	sequence	The value is = 0 for all records
16	write_date	Similar to create_date for all records
17	price_reduce	The value is = 0 for all records
18	qty_delivered	Similar to product_uom_qty field
19	layout_category_id	The value is = 0 for all records
20	price_reduce_taxinc	The value is = 0 for all records
21	State	The value is “Sale” for all records
22	Id	Row identifier
23	Order_id	Order identifier

After exclusion a list of selected attributes shown in the table 3.3 below.

Table (3.3) Selected Attributes

Selected Data Set Attributes	
#	Attribute
1	product_uom
2	order_partner_id
3	product_id
4	salesman_id
5	cost
6	price_unit
7	invoice_status
8	create_date
9	quantity
10	discount
11	price_total

After the selection process, the sales fact table was down-sized to 17,985 records and a total of 13 attributes.

The third step in data transformation is the creation of derived attributes or generated records. Here, an addition of one derived attribute “month” was made, creating a new dimension called Date. This categorial attribute is to provide an organization view on the time-series records in the fact table.

3) Loading

The final stage of the ETL process is loading the produced fact table and accompanying dimensions to an environment where we can apply data mining techniques to achieve research objectives.

Figure 3.2 shows the final Sales Fact Table and Dimensions.

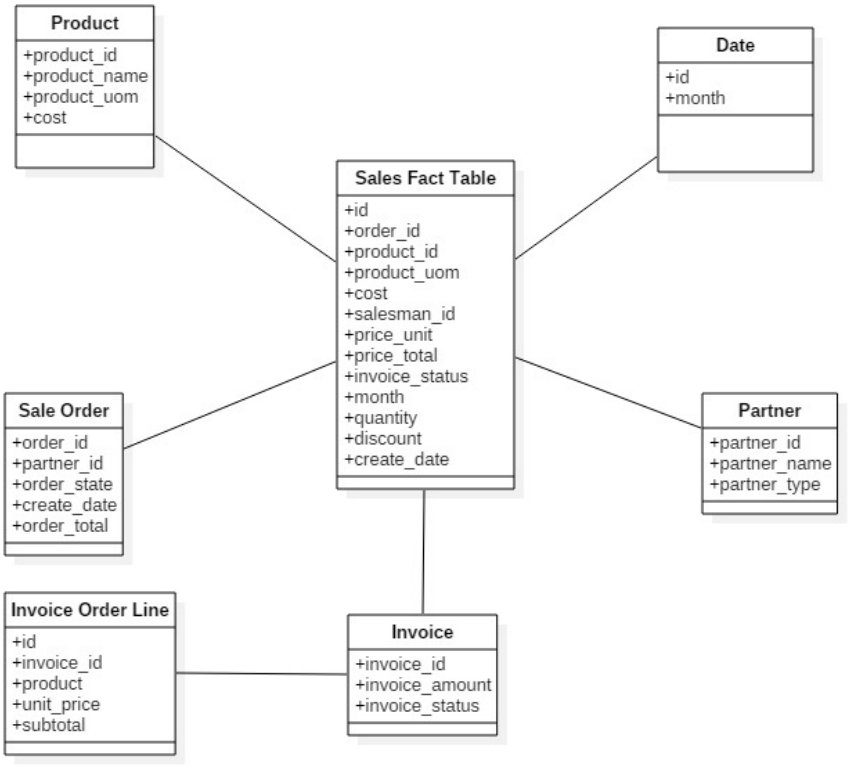

Figure (3.2) Sales Fact Table and Dimensions

Table 3.4 shows the final form of the Sales Fact Table attributes.

Table (3.4) Final Sales Fact Table

Sales Fact Table		
#	Attribute	Key
1	id	
2	order id	F
3	product_uom	F
4	order_partner_id	F
5	product_id	F
6	salesman_id	F
7	cost	F
8	price_unit	F
9	invoice_status	F
10	create_date	
11	quantity	
12	discount	
13	price_total	
14	month	

3.2.1.4 Modeling

The first step in the modeling phase is the selection of a suitable forecasting technique. (Chambers et. al, 1971) argue in their article that the selection of a method depends on many factors—the context of the forecast, the relevance and availability of historical data, the degree of accuracy

desirable, the time period to be forecast, the cost/ benefit (or value) of the forecast to the company, and the time available for making the analysis.

They also argue that there are three basic types—qualitative techniques, time series analysis and projection, and causal models. The first uses qualitative data (expert opinion, for example) and information about special events of the kind already mentioned, and may or may not take the past into consideration. The second, on the other hand, focuses entirely on patterns and pattern changes, and thus relies entirely on historical data. The third type uses highly refined and specific information about relationships between system elements, and is powerful enough to take special events formally into account. (Chambers et. al, 1971).

Studying the data available, it is clear that the research should use the third type of forecasting methods. These methods include: Regression modeling, Econometric modeling, Leading Indicator modeling.

a) Regression Model: Regression is one of the most common techniques used to understand a variable relationship in a dataset. In this method, a function is estimated using the least square technique between the dependent and independent variables which defines the interaction among them. A simple example would be forecasting the margin of a business (dependent variable) based on factors like cost of goods sold, inventory holding etc. (independent variables). b) Econometric Model: The econometric modeling technique uses economic variables to forecast future developments. It relies on the interaction between the economic variables and the internal sales data. Some of the economic variables are CPI, Exchange rates, inflation, employment rate etc. Econometric models are a system of interdependent regression equations and it is this nature of the model that gives better results in explaining causalities as compared to ordinary regression. c) Leading Indicator Models: The leading indicator technique uses a combination of regression models and willingness to buy survey results to identify causation between movement of two time-series variables. One of the variables here is an economic activity and the other is the dependent variable. A good example of Lead Indicator would be to find if the time series of an economic activity (say CPI) precedes the movement of times series of the dependent variable (say Sales of a company) in the same direction.

Causal forecasting can be used to forecast at a granular level. For sales, it can be used to forecast by product, product category, subclass etc. It can also be used for any forecast where there are multiple forces at play which impact the dependent variable.

The research objective is to predict demand numbers, which directly translates to data type ‘integer’. With the absence of economical factors and time limitations restricting the ability of willingness-to-buy surveys, regression modeling was the optimum choice.

Linear regression is the process of computing an expression that predicts a numeric quantity. According to (Jiawei and Kamber, 2006), Multiple Linear Regression is an extension of straight-line regression so as to involve more than one predictor variable.

It allows response variable y to be modeled as a linear function of, n predictor attributes A_1, A_2, …, A_n, describing a tuple, X. (That is, X = (x_1, x_2, …,x_n).) In a training data set, D, containing data of the form with associated class labels, y_i. An example of a multiple linear regression model based ((x_1, y_1.), (x_2, y_2.),. . . , ($x_{|D|}$, $y_{|D|}$.),, where the x_i. are the n-dimensional training tuples on two predictor attributes or variables, A_1 and A_2 is:

$$y = w_0 + w_1 x_1 + w_2 x_2$$

Where x_1 and x_2 are the values of attributes A_1 and A_2 respectively, in X.

Linear regression was selected due to its ability to predict numeric values. The attribute to be predicted is "Quantity".

The second step is to set the test design for the model. Cross Validation (10-k fold) was selected to test the model. According to (Khandelwal, 2018), Cross-validation is a statistical technique which involves partitioning the data into subsets, training the data on a subset and use the other subset to evaluate the model's performance. Whereas the K fold cross validation is a technique that involves randomly dividing the dataset into k groups or folds of approximately equal size. The first fold is kept for testing and the model is trained on k-1 folds. (Gupta, 2017) argues that this significantly reduces bias, as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation set.

The success of the model was to be measured using Mean Absolute Error (MAE), which is the average of the difference between the Original Values and the Predicted Values. (Mishra, 2018) says "MAE is absolutely robust to outliers. It gives us the measure of how far the predictions were from the actual output".

The model was created and executed using WEKA (The Waikato Environment for Knowledge Analysis) in the explorer window.

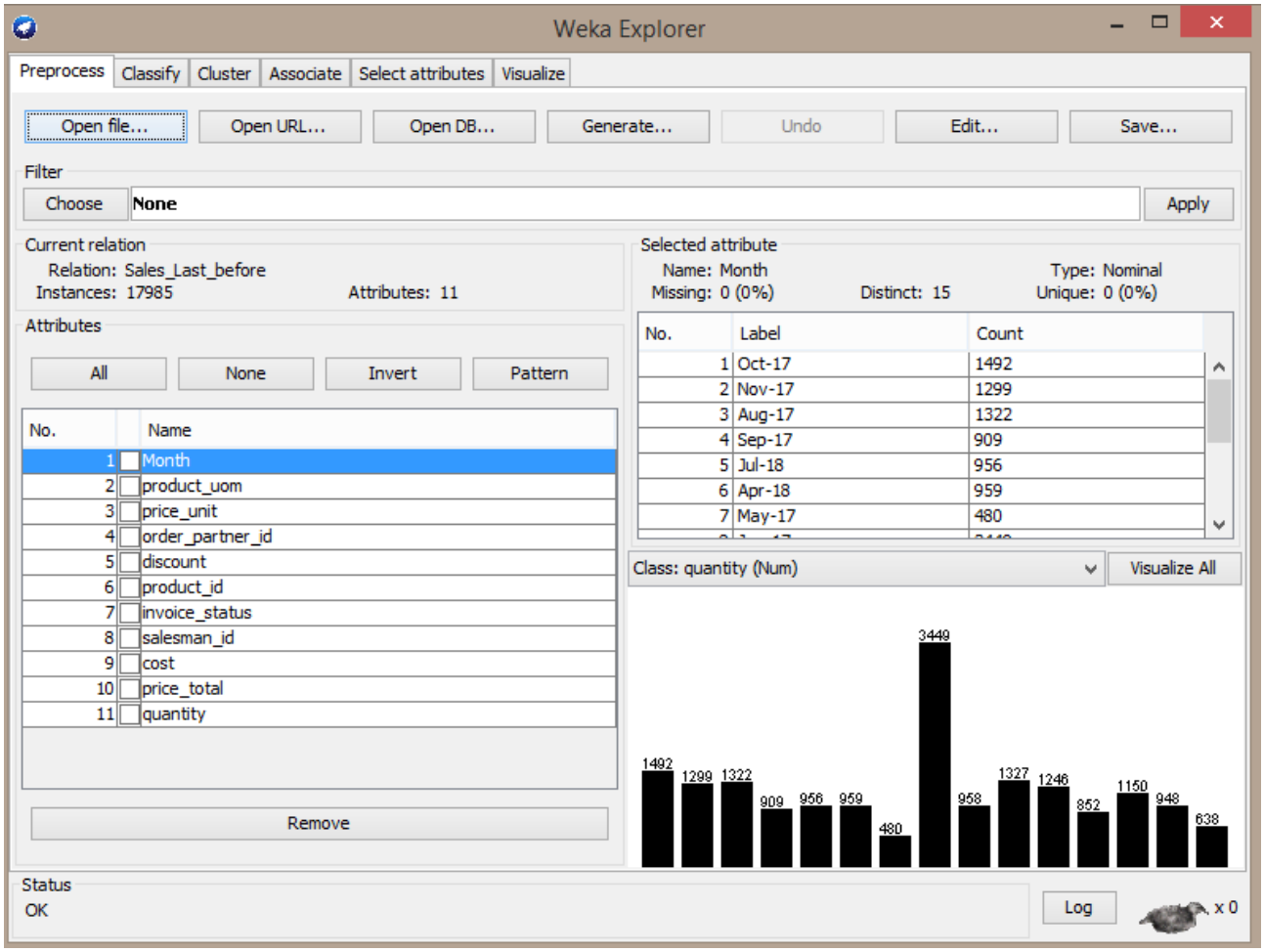

Figure (3.3) Attributes description for CRISP-DM Round 1

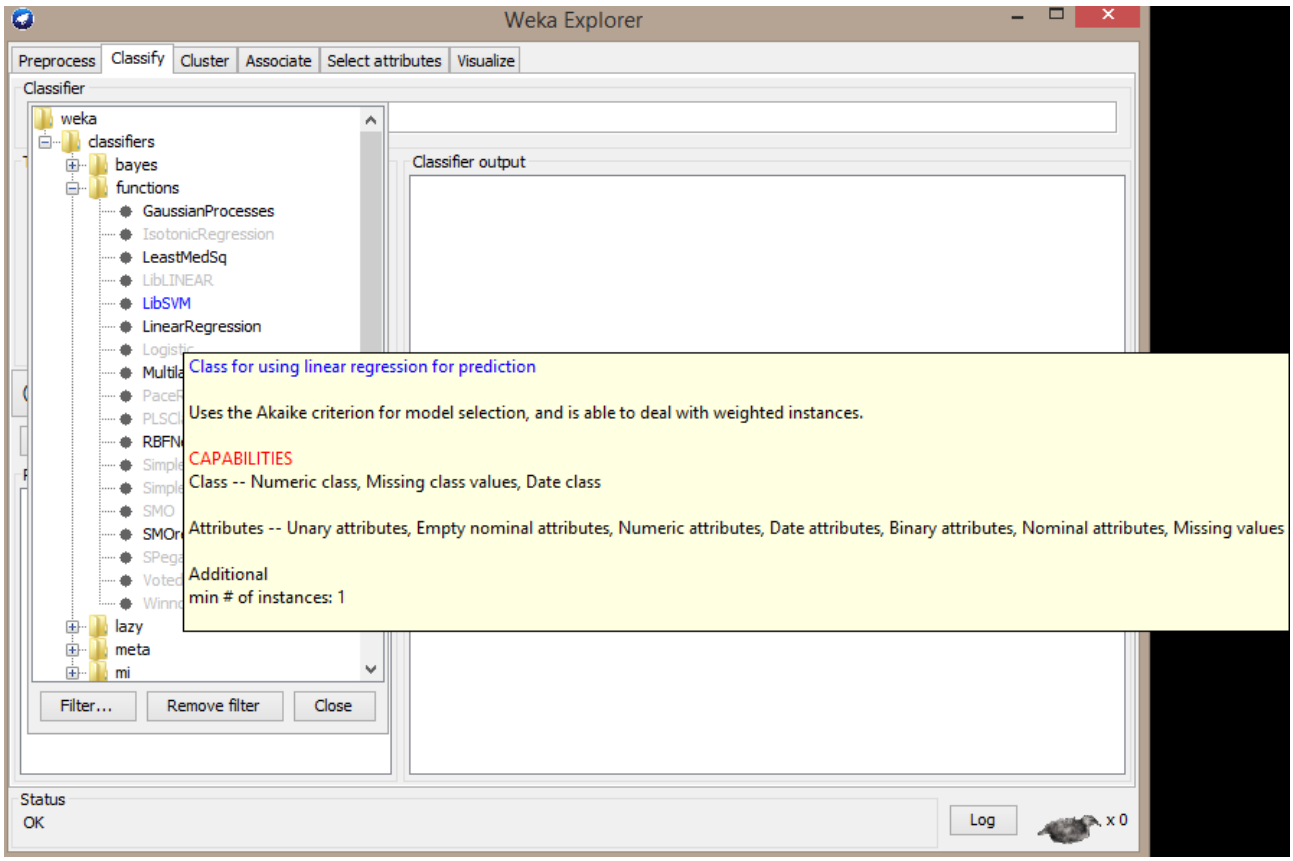

Figure (3.4) Linear Regression in WEKA explorer

3.2.2 Second Round of CRISP-DM Cycle

As illustrated in Figure (3.1), the second round went through 5 of the 6 CRISP-DM Steps. They are: Business Understanding, Data Transformation, Modeling, Evaluation and deployment. The details of the steps are described below.

3.2.2.1 Business Understanding

The results of evaluation in the first round of the CRSIP-DM process were shown to the stakeholders of the project, and a preposition of changing the format of the prediction was made. It was agreed that the purpose of the study was to predict quantities of products in a given period of time, however from a functional point of view, the operation of purchasing these items from abroad comes in the form of packets or known-numbered batches. Upon that statement, a change to the predictor was to be made, putting the outcomes into consideration.

3.2.2.2 Data Transformation

Instead of numbers to predict, a list of classes was presented after thorough discussion with stakeholders. The data type of the attribute to be predicted was changed from numeric to categorial. Table 3.5 shows the classes newly created.

Table (3.5) Iteration Created Classes

Class name
Q < 10
10 <= Q < 20
20 <= Q < 50
50 <= Q < 100
100 <= Q < 250
250 <= Q < 500
500 < Q

3.2.2.3 Modeling

In model selection, a change of model was necessary because the predicted value was changed from a number to a category (class).

(Foxworthy, 2020) argues that there are four models to forecast a category. These are, Logistic Regression, Decision Trees, Random Forest Classifier, Support Vector Classifier. Table (3.6) shows a comparison between the four types on their main attributes.

Table (3.6) Models for predicting categories

Model	Main Attributes
Logistic Regression	• Base model for forecasting a classification • Gateway to neural networks/deep learning
Decision Trees	• No data normalization or scaling • Less work • Missing data will not affect processing
Random Forest Classifier	• Captures non-deterministic processes • Non-linearity • Captures outliers
Support Vector Classifier	• Captures data outliers accurately • Linear separability

Analyzing Table (3.6), it's found that Decision Trees will better match the type of existing data since it's capable of handling denormalized data.

According to (Gavrilov, 2016), Decision trees are beneficial, since they are: Interpretable at a glance, Suitable for handling both Universal for solving both classification and regression problems, Capable of handling missing values in attributes and filling them in with the most probable value, High-performing with regard to searching down a built tree, because the tree traversal algorithm is efficient even for massive data sets.

A decision tree is a flowchart like data structure where each non-leaf node specifies a test of some attribute. The algorithm starts from the root node—according to test outcomes— and moves out to from branches until the final level of leaves are created. These nodes or leaves are given class label, or are values of target attributes. (Gavrilov, 2016)

Decision Tree C4.5 was chosen because, as pointed out by (Saha, 2018), the algorithm inherently employs Single Pass Pruning Process to Mitigate overfitting, It can work with both Discrete and Continuous Data, It can handle the issue of incomplete data very well. The implementation of the C4.5 tree is by using the newly created classes as shown in table 3.5 and substituting the numerical values with them as classes.

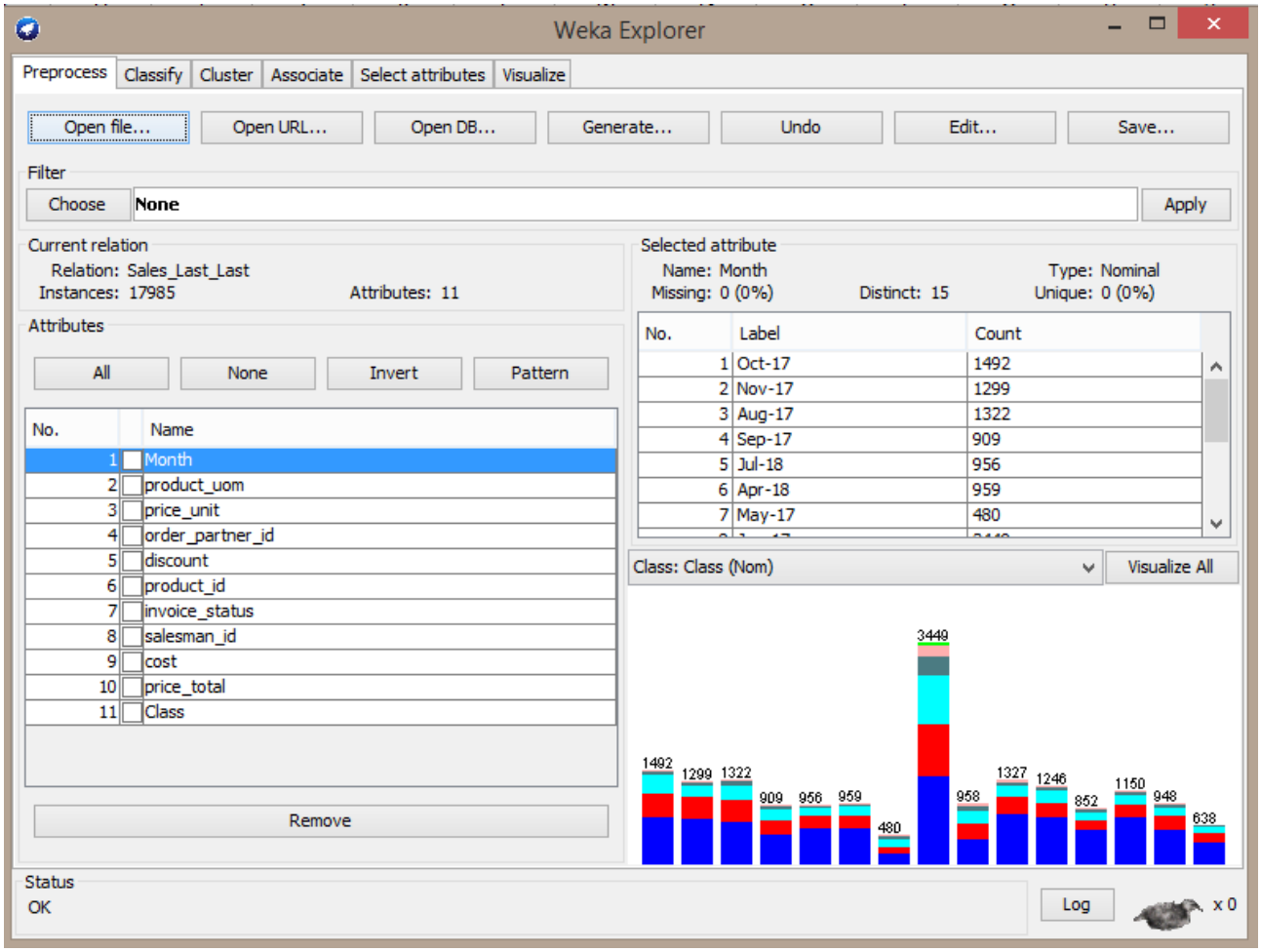

Figure (3.5) Attributes description for CRISP-DM Round 2

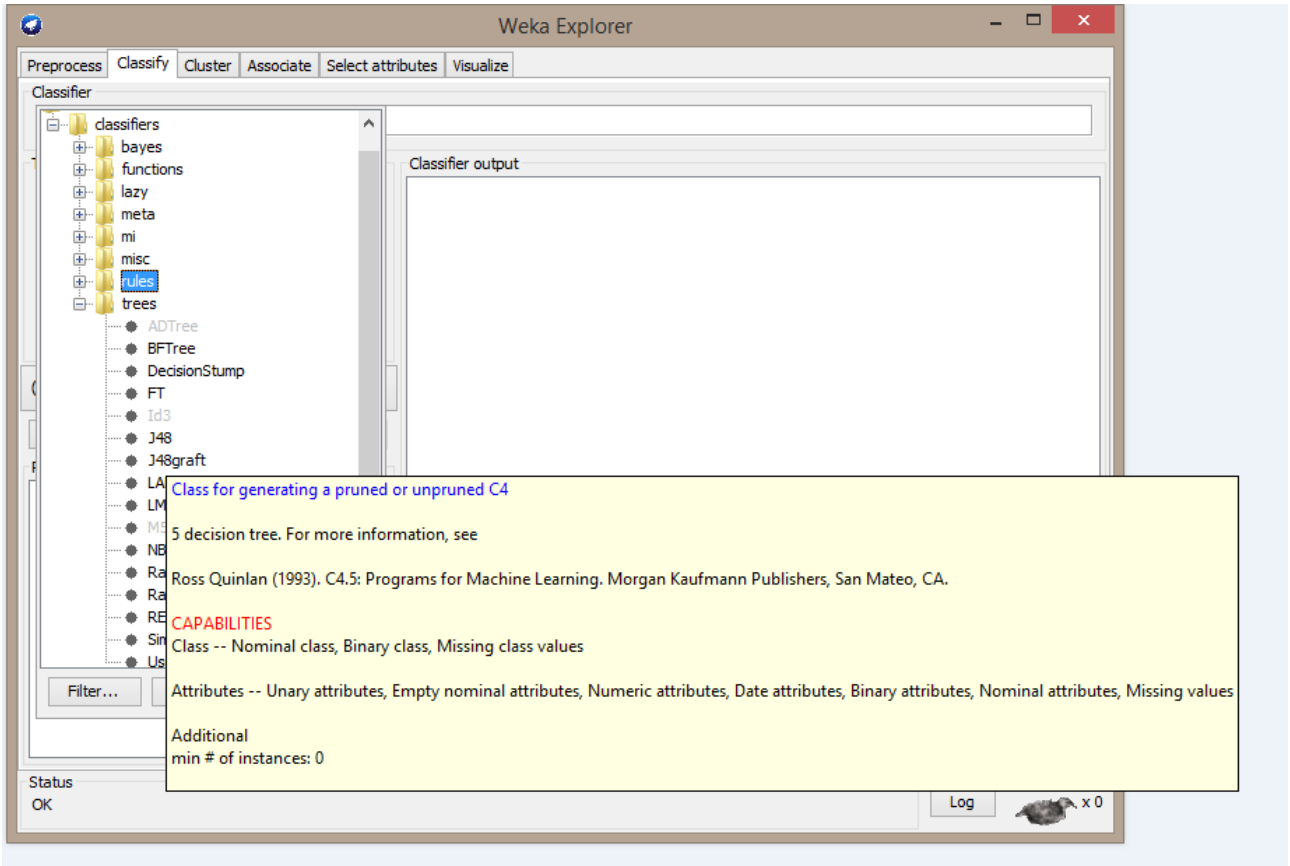

Figure (3.6) Decision Tree C4.5 in WEKA explorer

The second step is to set the test design for the model. Cross Validation was selected to test the model, 10 K fold technique, just like in iteration 1 previously mentioned in section 3.2.1.4.

3.3 Chapter Summary

The CRISP-DM methodology was implemented in two cycles. In the first cycle, a data mart was created for the sales part using Kimball's bottom-up approach in creating a datawarehouse. A snowflake schema was chosen to implement the data mart. The model chosen (Linear Regression) showed limitations in the run, which led to a second cycle of the CRISP-DM.

The second cycle of the CRISP-DM process, a change in the predicted class data type was made. From continuous it was changed to categorial, and the C4.5 decision trees were used as a predictor.

CHAPTER 4

DISCUSSION AND FINDINGS

4.1 Introduction

This chapter provides a thorough discussion and interpretation of the results found in the previous chapter. The greater part of the results was extracted from the CRISP-DM's step "Evaluation".

4.2 Results and Evaluation

As mentioned in Chapter 3, the methodology of CRISP-DM has gone through two consecutive iterations. This section thoroughly details the results and finding of each iteration.

4.2.1 First Round of CRISP-DM Cycle Results

The built model results after the 10k fold cross validation showed the following: The correlation coefficient showed a high positive result, indicating that there is indeed a strong linear correlation between the variables. However, a high mean absolute error (MAE) indicated that there was an obvious variance between the actual versus the predicted values.

Table 4.1 demonstrates the outcomes of the linear regression model run. The figure shows a high positive result correlation coefficient of 0.7589, indicating that there is indeed a strong correlation between the variables. However, a high mean absolute error (MAE) indicated that there was an obvious variance between the actual versus the predicted values.

Table (4.1) CRISP-DM First Cycle Model Build Results

Run attribute	**Value**
Correlation Coefficient	0.7589
Mean Absolute Error	18.574

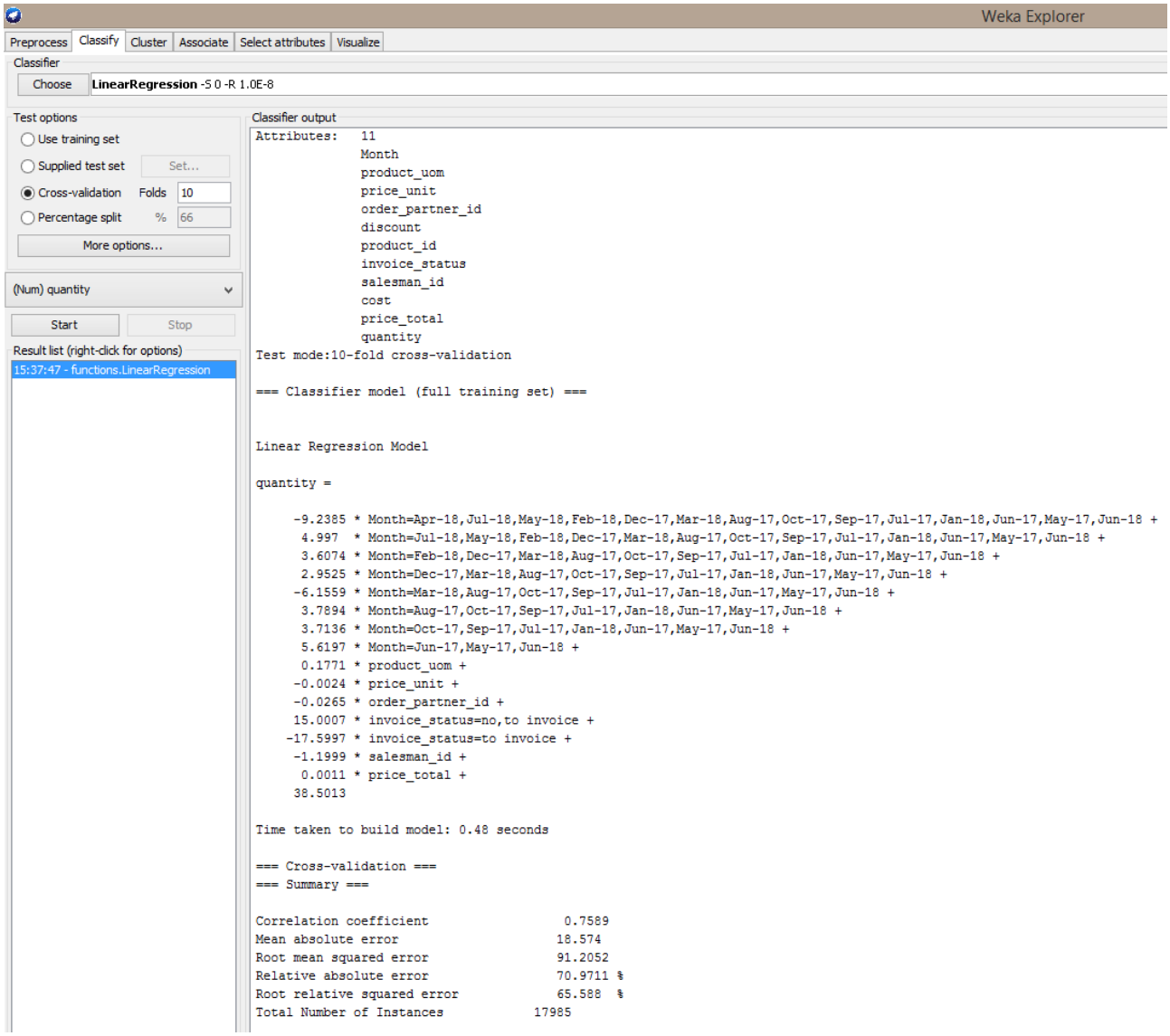

Figure (4.1) Deployed dataset on Linear Regression

Figure (4.1) shows the deployed dataset using linear regression model and the outcome in the summary.

Evaluation of Model

(Wu, 2020) states that there are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square

   R Square measures how much of variability in dependent variable can be explained by the model. It is square of Correlation Coefficient(R) and that is why it is called R Square
2. Mean Square Error (MSE)/Root Mean Square Error (RMSE)

   Mean Square Error is an absolute measure of the goodness for the fit.

3. Mean Absolute Error (MAE)

   Mean Absolute Error (MAE) is similar to Mean Square Error (MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of absolute value of error

According to (Yarnold and Soltysik, 2013), in a time-series data containing two or more different groups (or mixed data), it is bound to have a paradoxical results. Therefore, it was resolved that the variance shown by the high MAE was due to the used linear regression's inability to incubate mixed types of independent variables to predict a numeric dependent variable without a statistical intervention. Trying to have a "one-size-fits-all" linear model is restrictive, and therefore a form of conformity is a necessary action.

According to table (3.3), there are 6 categorial variables, while the rest were continuous (or numeric). There were two clear choices to take from. Either to use a statistical remedy, which is the use of dummy variables, or to change the model to a more convenient one in accordance to research problem. A Dummy variable or Indicator Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. (Skrivanek, 2009). The use of dummy variables would have affected the time frame of the project drastically, due to the further statistical steps to be taken. Therefore, the second solution was sought to be proposed, which is changing the model. This has led to a second iteration of the CRISP-DM process, moving back the Business Understanding step.

4.2.2 Second Round of CRISP-DM Cycle Results

The built model results after the 10k fold cross validation showed noticeably accurate predicted values. These results are thoroughly discussed in Chapter 4.

```
=== Summary ===

Correctly Classified Instances       17558               97.6258 %
Incorrectly Classified Instances       427                2.3742 %
Kappa statistic                          0.9637
Mean absolute error                      0.0093
Root mean squared error                  0.0783
Relative absolute error                  4.9931 %
Root relative squared error             25.6301 %
Total Number of Instances            17985

=== Confusion Matrix ===

    a    b    c    d    e    f    g   <-- classified as
 9163   82    6    0    0    0    0 |    a = Q < 10
   59 3886   53    0    0    0    0 |    b = 10 <= Q < 20
   12   39 2938   48    0    0    0 |    c = 20 <= Q < 50
    2    0   36  910   28    0    0 |    d = 50 <= Q < 100
    1    0    1    9  523   13    0 |    e = 100 <= Q < 250
    0    0    0    1   16  100    8 |    f = 250 <= Q < 500
    0    0    1    0    1   11   38 |    g = 500 < Q

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.99      0.008      0.992     0.99      0.991      0.997    Q < 10
                 0.972     0.009      0.97      0.972     0.971      0.992    10 <= Q < 20
                 0.967     0.006      0.968     0.967     0.968      0.993    20 <= Q < 50
                 0.932     0.003      0.94      0.932     0.936      0.986    50 <= Q < 100
                 0.956     0.003      0.921     0.956     0.938      0.987    100 <= Q < 250
                 0.8       0.001      0.806     0.8       0.803      0.959    250 <= Q < 500
                 0.745     0          0.826     0.745     0.784      0.911    500 < Q
Weighted Avg.    0.976     0.008      0.976     0.976     0.976      0.994
```

Figure (4.2) Deployed dataset on Decision Tree C4.5

Figure (4.2) shows the deployed dataset using decision tree c4.5 and the outcome in the summary.

In the second iteration, the evaluation step of the CRISP-DM has shown the following results: In Table (4.2), the model build summary for the Decision Tree C4.5 showed a noticeably low MAE of 0.0093, and a correctly classified Instances of 97.6%. Low MAE indicates that the variances between the actual results for the classes and the predicted values is very low.

Table (4.2) CRISP-DM Second Cycle Model Build Results

Run attribute	**Value**
Correctly classified instances	97.626%
Mean Absolute Error	0.0093

A more detailed set of readings is shown in Table (4.3) for the given seven classes. The figure shows a weighted average 0.976 F-measure. The F-measure is the weighted harmonic mean of the precision and recall.

Table (4.3) CRISP-DM Second Cycle Model Detailed Accuracy

Reading	Value
Weighted Average True Positive (TP) Rate	0.97
Weighted Average False Positive (FP) Rate	0.008
Weighted Average Precision	0.976
Weighted Average Recall	0.976
Weighted Average F-Measure	0.976

Evaluation of Model

According to (Shchutskaya, 2018), a classification problem is about predicting what category something falls into.

Metrics that can be used for evaluation a classification model:

- Percent correction classification (PCC): measures overall accuracy. Every error has the same weight.
- Confusion matrix: also measures accuracy but distinguished between errors, i.e false positives, false negatives and correct predictions.
- Area Under the ROC Curve (AUC – ROC): is one of the most widely used metrics for evaluation. Popular because it ranks the positive predictions higher than the negative. Also, ROC curve it is independent of the change in proportion of responders.

The obtained results in the second cycle of the CRISP-DM process shows a high accuracy of the C4.5 decision tree predictor. These results were sufficient to stop the CRISP-DM process at the second cycle and accept the outcomes for deployment.

4.3 Discussion

The results of the first iteration of the CRISP-DM process shows in no doubt that the choice of Linear Regression – or in fact any form of linear modeling – would lead to paradoxical results. As previously pointed to in section 3.2.1.5 of this research study, variances in the predicted values will occur due to Linear Regression's inability to incubate mixed types of independent variables to predict a numeric dependent variable without a statistical intervention.

Observing the structure of ERP system's OLTP, it's clear that such mixed data existence is inevitable, due to the integration of its modules at the database level. The solution used in this research study, depending on the business case - for example, solely depended on the stakeholder's business needs flexibility. This is a plus provided through using the CRISP-DM as a methodology for completing the data mining task. It allowed the mapping of model selection according to the expected form business results.

Upon studying the technical documentation of the used ERP system 'Odoo', and later on during the Data Cleaning process, it was found that the quality of the data extracted from the ERP's transactional database was high, hence only a handful of transformation tools were used. This stresses on the validation rules used to build the ERP system itself.

The results obtained in the second iteration of the CRISP-DM process indicates that the decision tree C4.5 provides an accurate set of predictions to the business problem, which is Supply Chain demand forecast for the sales. This has led to a successful data mining project, and adoption of the methodology suggested by the research in the stakeholder's strategic planning, as detailed in section 1.3 of this research study.

4.4 Summary

This chapter has discussed the results obtained in Chapter 3. It covered the knowledge obtained throughout the methodology implementation process. It first discussed the results of the first iteration of the CRISP-DM process applied to the research, then it also discussed the second iteration of the CRISP-DM process, which came as a result of changing some elements in the process such as changing the model used for forecasting.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

Demand Forecasting is undoubtedly an essential strategic tool to any profit-seeking organization. It is yet however an understudied field when it comes to the application of data mining. The research studied the case of a medium-sized company based in Khartoum, Sudan. It has analyzed the company's supply chain data in order to provide an accurate model for forecasting the company's sales demand in a given time frame. The CRISP-DM was used as a methodology of implementation in order to achieve the research's objectives stated in section 1.3 of this thesis.

The methodology applied two different data mining forecasting models in two consecutive iterations (or phases). The two models are Linear Regression and Decision Trees. The discussion in Chapter 4 demonstrates that the Decision Trees presented a better forecasting model for the particular research problem, whereas Linear Regression was eye-opening to more research-related questions. The research objectives underlying in creating a datawarehouse from a transactional database and using the data to evaluate the most suitable prediction model of sales data, were successfully achieved after two cycles of CRISP-DM.

5. 2 Recommendation and Future Work

Upon the finishing of this research study, and based on the results found and points discussed in Chapter 4, the research has exerted some recommendations in order to add more to the field of Demand Forecasting:

- Enhance Linear Regression model for Demand Forecasting using Dummy variables.
    - It will provide a numeric prediction of dependent variable using the source of categorial or mixed independent variables.
- Customer segmentation prior to classification task's effect on Demand Forecasting
    - In order to give access to more variance in the data, and therefore improve the bias-variance trade-off.

REFERENCES

Murray, P. W., Agard, B., & Barajas, M. A. (2015). Forecasting supply chain demand by clustering customers. *IFAC-PapersOnLine*, *28*(3), 1834–1839. https://doi.org/10.1016/j.ifacol.2015.06.353

Archer, B. (1987). Demand forecasting and estimation, 77–85. Retrieved from https://www.cabdirect.org/cabdirect/abstract/19871847153

Khandelwal, R. (2018). K fold and other cross-validation techniques – Data Driven Investor – Medium. Retrieved January 24, 2019, from https://medium.com/datadriveninvestor/k-fold-and-other-cross-validation-techniques-6c03a2563f1e

Mishra, A. (2018). Metrics to Evaluate your Machine Learning Algorithm. Retrieved January 24, 2019, from https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Te Grotenhuis, M., & Thijs, P. (n.d.). *Dummy variables and their interactions in regression analysis: examples from research on body mass index*. Retrieved from https://www.ru.nl/sociology/mt/bmi/downloads/

Saha, S. (2018). What is the C4.5 algorithm and how does it work? – Towards Data Science. Retrieved January 24, 2019, from https://towardsdatascience.com/what-is-the-c4-5-algorithm-and-how-does-it-work-2b971a9e7db0

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: the complete guide to dimensional modelling*. *John Wiley & Sons, Inc., Indianapolis, Indiana*. https://doi.org/10.1145/945721.945741

Rey, T. D., Dow, T., Company, C., Wells, C., Kauhl, J., & Services, T. C. (2013). SAS Global Forum 2013 Data Mining and Text Analytics Using Data Mining in Forecasting Problems Defining the Need SAS Global Forum 2013 Data Mining and Text Analytics, 1–17.

Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating Forecasting Methods by Considering Different Accuracy Measures. *Procedia - Procedia Computer Science*, *95*, 264–271. https://doi.org/10.1016/j.procs.2016.09.332

Mullick, S. K., & Smith, D. D. (1971). How to Choose the Right Forecasting Technique, (July).

Kramer, S., & Widmer, G. (2001). Prediction of Ordinal Classes Using Regression Trees ∗, (August 2013). https://doi.org/10.1007/3-540-39963-1

Armstrong, J. S. (2009). Selecting Forecasting Methods.

Sukkerd, R., Beschastnikh, I., Wuttke, J., Zhang, S., & Brun, Y. (2013). Understanding Regression Failures through Test-Passing and Test-Failing Code Changes, 1177–1180.

Relich, M., Witkowski, K., Saniuk, S., & Šujanová, J. (2014). Material Demand Forecasting : an ERP System Perspective, *527*, 311–314. https://doi.org/10.4028/www.scientific.net/AMM.527.311

Berry, Michael J. A. ; Linoff, G. S. (1996). *Data mining techniques*. *SIGMOD Record* (Vol. 25). https://doi.org/http://doi.acm.org/10.1145/235968.280351

Pan, F. (2016). Inventory Prediction Research Based on the Improved BP Neural Network Algorithm. *International Journal of Grid and Distributed Computing*, *9*(9), 307–316. https://doi.org/10.14257/ijgdc.2016.9.9.26

Velcu, O. (2007). Exploring the effects of ERP systems on organizational performance: Evidence from Finnish companies. *Industrial Management & Data Systems*, *107*(9), 1316–1334.

Gupta, P. (2017). Cross-Validation in Machine Learning – Towards Data Science. Retrieved January 24, 2019, from https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

Janvier-James, A. M. (2011). A New Introduction to Supply Chains and Supply Chain Management: Definitions and Theories Perspective. *International Business Research*, *5*(1), 194–208. https://doi.org/10.5539/ibr.v5n1p194

Drkusic, E. (2016). Data Warehouse Modeling: The Snowflake Schema. Retrieved January 24, 2019, from https://www.vertabelo.com/blog/technical-articles/data-warehouse-modeling-the-snowflake-schema

Pontius, N. (2019). What is Inventory Management? Definition, Best Practices – Camcode. Retrieved January 19, 2019, from https://www.camcode.com/asset-tags/what-is-inventory-management/

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. In *The Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*.

Gneiting, T. (2010). Making and Evaluating Point Forecasts, 1–38.

Ncr, P. C., Spss, J. C., Ncr, R. K., Spss, T. K., Daimlerchrysler, T. R., Spss, C. S., & Daimlerchrysler, R. W. (n.d.). Step-by-step data mining guide.

Lupetin, M. (2003). A Data Warehouse Implementation Using the Star Schema, 1–5.

Yu, X., Qi, Z., & Zhao, Y. (2013). Support vector regression for newspaper/magazine sales forecasting. *Procedia Computer Science*, *17*, 1055–1062. https://doi.org/10.1016/j.procs.2013.05.134

Adaileh, M. J., & Abu-alganam, K. M. (2010). The Role of ERP in Supply Chain Integration. *Journal of Computer Science*, *10*(5), 274–279.

Gavrilov, V. (2016). Benefits of decision trees in solving predictive analytics problems | Prognoz blog. Retrieved January 24, 2019, from http://www.prognoz.com/blog/platform/benefits-of-decision-trees-in-solving-predictive-analytics-problems/

Kauffman, M. (2009). Machine Learning for Numeric Prediction, *4*, 223–233.

Hand, D., Hand, D., Mannila, H., Mannila, H., Smyth, P., & Smyth, P. (2001). *Principles of data mining*. *Drug safety : an international journal of medical toxicology and drug experience* (Vol. 30). https://doi.org/10.2165/00002018-200730070-00010

Rajan, C. A., & Baral, R. (2015). Adoption of ERP system: An empirical study of factors influencing the usage of ERP and its impact on end user. *IIMB Management Review*, *27*(2), 105–117. https://doi.org/10.1016/j.iimb.2015.04.008

Hicham, A., Mohammed, B., & Abdellah, E. F. (2012). Sales forecasting based on ERP system through Delphi, fuzzy clustering and back-propagation neural networks with adaptive learning rate. *International Journal of Computer Science Issues*, *9*(6–3), 24–34.

Moorman, M., & Roi, K. (n.d.). Data Warehousing Design Issues for ERP Systems Single Business Template, (Dm).

In, S. (2012). EVALUATION OF THE EFFICIENCY OF INTEGRATED ERP, *1*.

Bhargava, N., & Sharma, G. (2013). International Journal of Advanced Research in Decision Tree Analysis on J48 Algorithm for Data Mining, *3*(6), 1114–1119.

Ozsaglam, M. (2015). Data Mining Techniques for Sales Forecasting. *International Journal of Technical Research and Applications*, *34*(34), 6–9. https://doi.org/10.1002/9780470685815.ch3

Jain, J., Dangayach, G. S., Agarwal, G., & Banejree, S. (2010). Supply Chain Management: Literature Review and Some Issues. *Journal of Studies on Manufacturing*, *1*(1), 11–25. https://doi.org/10.1108/JAMR-09-2017-0090

Simitsis, A., & Polytechniou, I. (n.d.). Modeling and managing ETL processes.

ADVANTAGES OF IMPLEMENTING A DATA WAREHOUSE DURING AN ERP UPGRADE Advantages of Implementing a Data Warehouse During an ERP Upgrade. (n.d.).

Jain, N., & Srivastava, V. (2013). Data Mining Techniques: a Survey Paper. *IJRET: International Journal of Research in Engineering and Technology*, *2*(11), 116–119. Retrieved from http://esatjournals.net/ijret/2013v02/i11/IJRET20130211019.pdf%0Ahttp://ijret.org/Volumes/V02/I11/IJRET_110211019.pdf

Guanghui, W. (2012). Demand Forecasting of Supply Chain Based on Support Vector Regression Method, *29*, 280–284. https://doi.org/10.1016/j.proeng.2011.12.707

Jiawei, H., & Kamber, M. (2006). *Data mining: concepts and techniques*. *San Francisco, CA, itd: Morgan Kaufmann*. https://doi.org/10.1007/978-3-642-19721-5

Moriya, G., & Gosawi, D. (2015). Data mining intelligent system for decision making based on ERP. *Binary Journal of Data Mining*, *5*, 8–12.

Elragal, A. A., & Al-serafi, A. M. (2014). The Effect of ERP System Implementation on Business Performance : An The Effect of ERP System Implementation on Business Performance : An

Exploratory Case-Study. *Communications of the IBIMA*, *2011*. https://doi.org/10.5171/2011.670212

Saleem, F. (2017). The effects of data mining in ERP-CRM model : a case study of MADAR THE EFFECTS OF DATA MINING IN ERP-CRM MODEL –, (April).

Os, S. K. Ł., Patalas, J., & Niak, W. W. O. Ź. (n.d.). a Data Analysis Based Methodology, 55–72.

Matende, S., & Ogao, P. (2013). Enterprise Resource Planning (ERP) System Implementation : A case for User participation. *Procedia Technology*, *9*, 518–526. https://doi.org/10.1016/j.protcy.2013.12.058

Umble, E. J., Haft, R. R., & Umble, M. M. (2003). Enterprise resource planning: Implementation procedures and critical success factors. *European Journal of Operational Research*, *146*, 241–257. https://doi.org/10.1016/S0377-2217(02)00547-7

Nazemi, E., & Tarokh, M. J. (2012). ERP : A literature survey ERP : a literature survey. *International Journal of Advanced Manufacturing Technology*, (August). https://doi.org/10.1007/s00170-011-3756-x

Moon, Y. (2007). Enterprise Resource Planning (ERP): a review of the literature. *Int. J. Management and Enterprise Development*, *4*(3), pp.235-264. https://doi.org/10.1504/IJMED.2007.012679

Tanizaki, T., Hoshino, T., Shimmura, T., & Takenaka, T. (2019). Demand forecasting in restaurants using machine learning and statistical analysis. *Procedia CIRP*, *79*(ii), 679–683. https://doi.org/10.1016/j.procir.2019.02.042

Merkuryeva, G., Valberga, A., & Smirnov, A. (2019). Demand forecasting in pharmaceutical supply chains: A case study. *Procedia Computer Science*, *149*, 3–10. https://doi.org/10.1016/j.procs.2019.01.100

Kot, S., Grondys, K., & Szopa, R. (2011). Theory of Inventory Management Based on Demand Forecasting. *Polish Journal of Management Studies*, *3*, 148–156.

Yarnold, P. R., & Soltysik, R. C. (2013). Ipsative Standardization is Essential in the Analysis of Serial Data. *Optimal Data Analysis*, *2*, 94–97.

Skrivanek, S. (2009). *The Use of Dummy Variables in Regression Analysis*. Retrieved from http://www.moresteam.com

Al-Mudimigh, A. S., Ullah, Z., & Saleem, F. (2009). A Framework of an Automated Data Mining Systems Using ERP Model. *International Journal of Computer and Electrical Engineering (IJCEE)*, *1*(5), 651–655.

Hanumanth, S., & Babu, P. (2013). ANALYSIS & PREDICTION OF SALES DATA IN SAP-ERP SYSTEM USING CLUSTERING ALGORITHMS. *International Journal of Computational Science and Information Technology*, *1*(4), 387–393.

Valencia, M., Díaz, F., & Correa, J. (2016). Multi-product inventory modeling with demand forecasting and Bayesian optimization. *DYNA*, *83*(198), 236–244. https://doi.org/10.15446/dyna.v83n198.51310

Kalchschmidt, M. (2008). Demand Forecasting Practices and Performance: Evidence From the Gmrg Database. *Unibg.It*, (May), 1–30. Retrieved from http://medcontent.metapress.com/index/A65RM03P4874243N.pdf%5Cnhttp://www.unibg.it/dati/bacheca/530/42054.pdf

Akkermans, H., Bogerd, P., & Van Wassenhove, L. (2003). The Impact of ERP on Supply Chain Management : Exploratory Findings From a European Delphi Study THE IMPACT OF ERP ON SUPPLY CHAIN MANAGEMENT : EXPLORATORY FINDINGS FROM A EUROPEAN DELPHI STUDY. *European Journal of Operational Research*, *2217*(April). https://doi.org/10.1016/S0377-2217(02)00550-7

Murray, P. W., Agard, B., & Barajas, M. A. (2015). Forecasting Supply Chain Demand by Clustering Customers. *IFAC-PapersOnLine*, *48*(3), 1834–1839. https://doi.org/10.1016/j.ifacol.2015.06.353

Amirthalingam, G., Shaheen, R., Kousar, M., & Bilfaqih, S. M. (2014). Integrated Data Mining and Knowledge Discovery Techniques in ERP, *2*(4), 210–214.

Kolkas, M. K., El-bakry, H. M., & Saleh, A. A. (2015). Integrated Data Mining Techniques in Enterprise Resource Planning (ERP) Systems Integrated Data Mining Techniques in Enterprise Resource Planning (ERP) Systems, (January 2014).