A Study on Applying Data Mining in Airline Industry for Demand Forecasting by predicting Item Criticality

المطلوبة بقطع للتنبؤ الطيران مجال في البيانات تعدين تطبيق حول دراسة القطع بأهمية التنبؤ خلال من للصيانة

by

ZAINAB HAMDAN KARAM

A dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS
(KNOWLEDGE AND DATA MANAGEMENT)

at

The British University in Dubai

Professor Sherief Abdallah
March 2018

DECLARATION

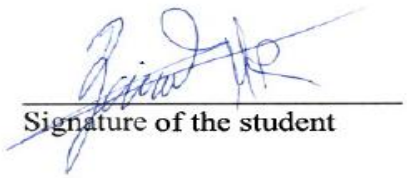
I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.



Signature of the student

COPYRIGHT AND INFORMATION TO USERS

Abstract

Organizations inventory forecasting plays an important role for supply chain management. It is very important for an organization to be able to identify the inventory demand required in future and this can be achieved by using the data stored in the company's data warehouse and with the help of data mining, future inventory demand can be predicted using specific data mining techniques. Several forecasting techniques have been developed for different businesses and each has its own advantages and disadvantages. In this research, the focus is in applying data mining technique to predict the item criticality for Expandable items (E-Class) which will support the organization to plan future demand.

This research is highlighting the use of data mining – predictive analysis using specific data mining classification methodologies to predict item criticality. This report is structured as following: introduction, Literature Review, Experimentation, Data Understanding, Data Preparation, Methodology, Results & Finding, Discussion and a conclusion.

Keywords: Inventory Forecasting, Data Mining, Classification Algorithm, Supply Chain

خلاصة

تلعب توقعات المخزون لدى المنظمات دورا هاما في إدارة سلسلة التوريد. من المهم جدا للمنظمة أن تكون قادرة على تحديد متطلبات المخزون في المستقبل وهذا يمكن تحقيقه باستخدام البيانات المخزنة في مستودع بيانات الشركة ومع مساعدة من تقنية استخراج البيانات، يمكن توقع طلب المخزون في المستقبل. وقد تم تطوير العديد من تقنيات التنبؤ لشركات مختلفة ولكل منها مزاياها وعيوبها. في هذا البحث، يتم التركيز في تطبيق تقنية استخراج البيانات للتنبؤ عن طريق دراسة بيانات مخزون شركة طيران لسنة 2017 التي ستدعم المنظمة لتخطيط الطلب في المستقبل.

هذا البحث يسلط الضوء على استخدام تقنية استخراج البيانات - تحليل التنبؤية باستخدام منهجيات تصنيف بيانات محددة لمعرفة أهميتها وبالتالي التنبؤ لمعرفة متطلبات المخزون في المستقبل. ويتمحور هذا التقرير على النحو التالي: مقدمة، مراجعة الأدب، التجريب، فهم البيانات، إعداد البيانات، المنهجية، النتائج و البحث، مناقشة و الاستنتاج.

DEDICATION

I would like to dedicate this paper to my father and mother, who always encouraged me and supported me to move forward.

Contents

List of Tables

List of Figures

Chapter 1

This chapter introduce an overview about the importance of aviation industry and it is contribution to the United Arab Emirates economy. The chapter also covers the usage of data mining in aviation as well as the goal and objectives of this research paper.

1. Introduction

Airline industry is estimated to have supported a $21.1 billion gross value added contribution to GDP in the United Arab Emirates in 2014 (IATA, Oxford Economics, Dec 2014).The service provider of the airlines, airport on-site enterprises, airport operators and air navigation has employed around 140,000 people in the United Arab Emirates in 2014. (IATA, Oxford Economics, Dec 2014).

Based on studies conducted by IATA in 2014, it was recognized that air-transport sector has supported around 99,000 jobs by buying goods and services from local suppliers. (IATA, Oxford Economics, Dec 2014).

Aircraft maintenance is considered one of the important aspects that every airline company have to look at. Based on so many studies conducted by the International Air Transport Association (IATA) in 2012, aircraft maintenance cost takes up to 13% of total operating cost of an airline company and this cost can be reduced by a good planning. (Jingyao Gu, Guoqing Zhang, Kevin Li, 2015)

Having an excellent maintenance program will help to avoid flight delays and cancellation which on the other hand will improve customer satisfaction. Since airline industry involves with a large number of parts and some of these parts are quite expensive, it is important to find an appropriate inventory forecasting model to look at the history of the inventory and predict the future demand based on the criticality of these items to operations.

Data mining is considered an important subject for research to help airline for better decision making. (Jingyao Gu, Guoqing Zhang, Kevin Li, 2015). A massive

part of the industry in today's world is based on the data. The amount of data is increasing every day and the process of extracting the data from a large set of data is called data mining. Due to huge data all around the world, the complexity of understanding the date related organisation is increasing (Moin and Ahmed, 2012). If an organisation does not have a proper knowledge and information about the market and customer preference, a remarkable turnover and greater customer dissatisfaction may happen.

Data mining is divided into two main aspects which are the descriptive data and the predictive data. However, in order to explain the predictive data mining, there are further discussions in this research which illustrates the importance of using data mining for predicting future organization's demands. Data mining has different methodologies that can be used to predict future demands. Some of these methods are: Classification, clustering, association rules, decision tree and others.

In this research, the focus is on predicting future demands of parts that will be required by an airline to perform aircraft maintenance. The dataset used is a real data from an airline system called Business Objects. The dataset includes 65533 records for different types of parts which are classified into 4 main categories: Rotables – T class (7585 records), Repairable – R class (24390 records), Expendables – E class (32403 records) and Consumables – C class (1155 records).

Below is the proposed inventory forecasting prediction framework that is used in this research

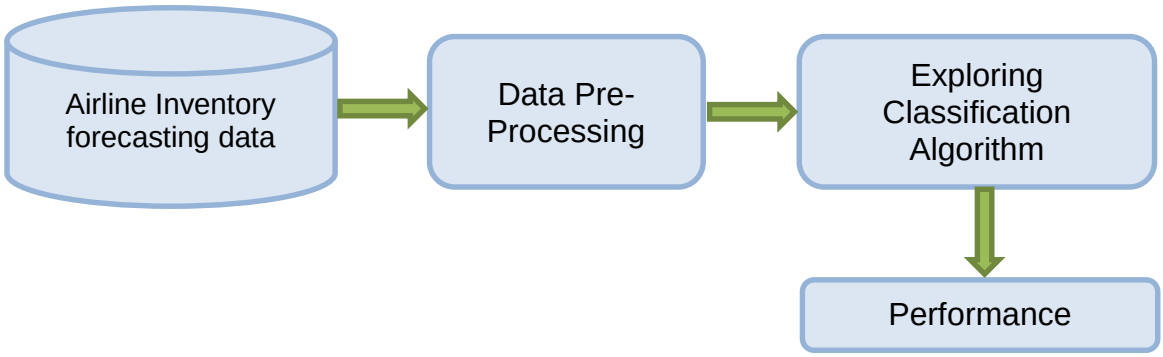


Figure 1: Inventory Forecasting Prediction Framework

1.1 Background

To predict the future demand for a company, the analysis of the state of inventory is very important for an organisation. Inventory means the stock of production of a company has not transacted in the present financial year but may be transacted in the next financial year (Bansal, Vadhavkar and Gupta, 1998). An organisation can determine the potential market for the following year by analysing the present stock of product of the organisation. A business organisation can overcome the fluctuation of demand by forecasting the future demand for the organisation (Provost and Fawcett, 2013). If an organisation has enough product or service capability for fulfilling the future demand for the product or service, it can reduce the rate of production.

The consumer forecasting and demand forecasting are the same things. By forecasting the demand for the product, the organisation can forecast the preference of the consumer. The organisation has to collect the data from the appropriate information about the rationale consumer of the company and the primary competitors of the company and their strategies. The predictive data mining process is the appropriate process to gather the information about the consumers and competitors (Sandborn, 2007). If the data mining is misused, the company cannot predict the proper demand of the consumers. For this cause, the company may have to face a huge loss in future.

The data mining process is significant to predict the future demand for the organisation and inventory forecasting process. Hidden and useful knowledge about the company can be found out by the data mining process. The marketing companies can attain the advantage to build the model on the marketing strategies based on the historical information (Vercellis, 2011). An appropriate production arrangement can have a retail company by conducting the data mining process. On the other hand, any manufacturer defect can find out by a manufacturing company by applying the data mining process in the operational engineering data.

1.2 Research Aim

To form the research aims of the research is an essential task of a research study. The aim of the research determines the primary objectives of the study. Therefore, the aim of this study is as follows.

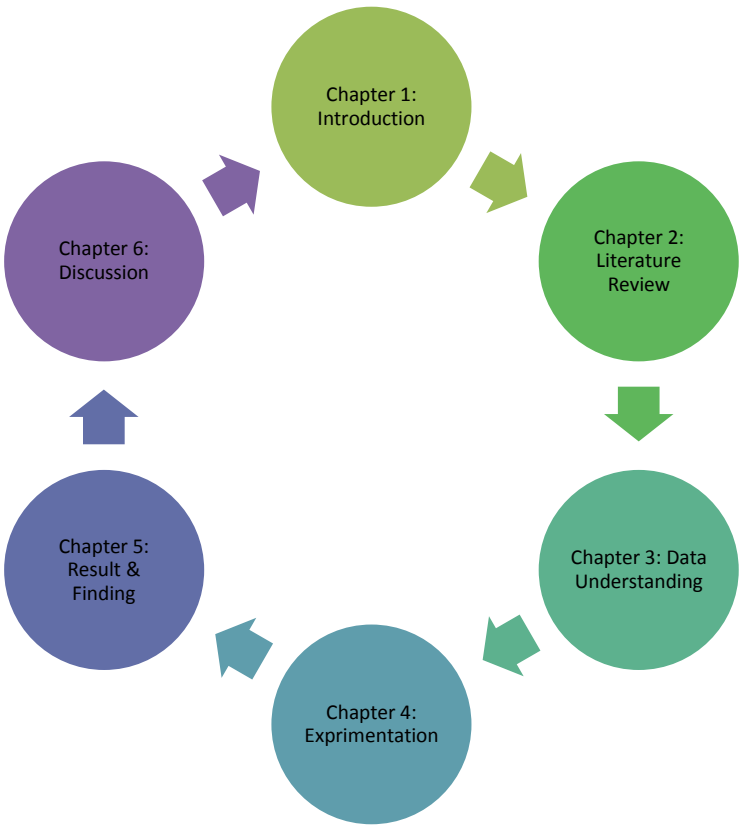
"To identify the process of data mining to predict the criticality of future inventory items which will help the airline organization in decision making and planning the future inventory demands"

1.3 Research Objectives

The objectives of a research study must be goal oriented. The objectives of this research study are as follows.

- To explore the data mining process for forecasting
- To explore the state of inventory for future demand prediction
- To identify the advantages of data mining in inventory forecasting

1.4 Structure of the research paper



Chapter 2

This section is the discussion of the research topic based on the components of the research problem according to other researches. The journal articles, reports and case study of previous scholars based on the data mining process and its importance in case of future demand prediction are discussed in this section.

2. Literature Review

As per the previous researchers, the data mining is a systematic process by which appropriate information about a company can be found out from a large set of data. The source of the data must reliable for appropriate future prediction. A gap in the literature will also be identified from the present literature and the previous literature.(Waller and Fawcett, 2013).

According to the research conducted by Dhond, Gupta, and Vadhavkar (2000), it is analyzed that there is much organization that is using the data mining techniques for the optimization of the information system. The use of the data mining techniques is getting very high in different fields of information technology. The use of these techniques also helps the user to understand the problem and that are the most efficient ways to solve that problem in the data information systems (Dhond, Gupta and Vadhavkar, 2000).

There are different kinds of the business benefits that associated with the data mining that eventually help the business to control its inventory. This paper also describes different principles of the data mining in the organization, and how these techniques can be used for the development of the business of the company. The outcome from data mining model in the big company of the medical distribution offers the underlying principle for policy to decrease the whole point of the inventory in the company by fifty percent (BANSAL, 1998).

According to the research conducted by Bansal, Vadhavkar, & Gupta, it is reviewed that how the decision-making skills of the company are important for the

agility, and for the more efficient inventory management, the decision-making skills are also needed to be polished. It is done with the experience of the particular field in which someone has to work. As an additional room to data mining neural network models, arithmetical actions and supposition are worn to supplement the model of a neural network are, explain in very detail (Bansal, Vadhavkar and Gupta, n.d.).

The large number of classes in the neural network, it is hard to recognize a specific model and class which propose the most excellent model of the inventory. This paper also explains the implementation of the customary techniques of the statistics to assist settle on the finest neural network sort for its use in the inventory control in the organization. A neural network-based system of the inventory management can allow the company to create effectual linkage with their associate organization in supply chain process (Bansal, Vadhavkar and Gupta, n.d.).

According to the research conducted by L, et al. (2014), it is analyzed that, how the use of the iMiner can bring the positive change to the inventory control and also the intelligent system of the organization. The data mining technologies in the inventory control can be used for distribution in the resources of the computer and it is also used to do a lot of task in the management of the inventory in the company (L et al., 2014).

One of the main reasons behind the use of the data mining in the inventory control of the company is that it helps to manage the whole inventory control in the organization and also prevent the overstock of the inventory. The iMiner is the very good software tool that helps the user in the future bases to inspect the aging process of the inventory and the timely analysis of the inventory.

So we can say that in this paper we have come to know about one of the best techniques that are been used by the iMiner tool for the forecasting of all of the inventory that is been present in the company (L et al., 2014).

According to the research conducted by Waller & Fawcett (2013), it is analyzed that, the field of the data need the very wide range of knowledge for the better analyzing and the manipulation, it required the deep knowledge of the domain and also some of the qualitative skills. The more someone has the knowledge of the domain the more he is likely to effectively manage all of the data mining in the company (Waller and Fawcett, 2013).

For the SCM data analytics, it is very important to get the basic knowledge of the field because it would help him a lot on the future basis. As we all know that the bid data now becomes one of the buzzwords and it is very important to have better manipulation of the data to get enough knowledge of the domain work with other skills related to the fields. By reviewing this article, we have come to know more clearly about the concept of the big data and its significance in the business of the company.

The big data manipulation, analysis, and the mining have the great impact on the supply chain management of the organization and eventually on the business of the company (Waller and Fawcett, 2013).

According to the research conducted by Liu, Bhattacharyya, Sclove, & Chen, (2001), it is reviewed that, the field of data mining is very vast and it is been used in the different kind s of companies, it can be a software company, food company or bank. In the case of the food company, the data mining needs to be done in the store and corporate level. We reviewed this appears we have to know about different policies of the data mining that can help the food company.

As it is mention in the paper can it for better forecasting of its inventory and the stock that is aging, and it also helps the company to get an estimate about the overstock (Liu et al., 2001).

For the forecasting in the food store data, mining can be used as the efficient tool, and it also assures more effective inventory management in the company. There are different procedures that have to be done in the time series of the data

mining. This paper also provides the complete review about the concepts and the procedures of the data mining in the company and how it may help the food company ton management and control all of its inventory (Liu et al., 2001).

According to the research conducted by Sandborn, Mauro, & Knox, (2007), it is reviewed that, the use of data mining in the organization for the lesser the risk in the database management and also use to improve the overall capabilities of the system. The data mining is very powerful techniques for the forecasting of the business electronic or inventory related details so that the risk in the organization can be minimized, by using the historical records of the business.

The data mining can be done one both of the individual parts of the system and it can be applied on the whole module but in both case, it very helps for the prediction of the future results and decision of the business of the company (Sandborn, Mauro and Knox, 2007).

Moreover, in the company, the forecasting of the data last order can also be combined with the inventory and the requirements of the business of the particular data, and further on what would be the best obsolescence data for the part of the system that is to forecasts (Sandborn, Mauro and Knox, 2007).

According to the research conducted by Kirkos, Spathis, & Manolopoulos, (2007), it is analyzed that, the role of the data mining techniques in predicting about the financial state of the firm and it also very helpful in the identification of all of the factors that are connected with the FFS. In this paper, we have also come to know about the significance of the decision trees in the data mining. The neural network can also be used in this approach for the identification of the financial statement and future condition of all of the finance in the organization's business. All of the things like assets, debut and the inventory of the company are forecast by using the data mining approach (Kirkos, Spathis and Manolopoulos, 2007).

In the data mining, the inventory of the organization can be check and examine by the comparing the current inventory with the rate of the sales in the company and by the total assets and the inventory. All of that information about the company is put together and data is manipulated in the manner that helps them to get an estimate about the inventory of the company and what would be the stock of the company in the future (Kirkos, Spathis and Manolopoulos, 2007).

According to the research conducted by BANSAL, (1998) it is reviewed about the significance of the inventory control in the company and that is the many procedures that are to do in order to get the forecast about the inventory of the company. it is obvious that the inventory of the company should be according to the demand of the customers otherwise the business have to go through loss. Data mining is the best approach for the management and the forecasting of the of the inventory in the company.

By reviewing, the article we also come to know about the concept of the neural network based forecasting techniques for the implementation of the inventory control in the company (BANSAL, 1998).

According to the research conducted by Rygielski, Wang, and Yen, (2002), it is reviewed about the forecasting of the trends in the Market by the use of the data mining is also the very effective technique to be used by the company. There are many companies that are now doing for the just in time inventory programs for more efficiency. The data mining also help the company to aware of all of the inventory is that it hold and what is the demand of them in the market so that is the future they may able to make the better decision about their inventory that is aging (Rygielski, Wang and Yen, 2002).

By reviewing this article, we have come to know about the use of the data mining for the relationship management in the company and it is also helpful in more efficient forecasting of the inventory that is in the company. For the better customer

relationship management, it is also very important to do efficient data analysis and data mining (Rygielski, Wang and Yen, 2002).

According to the research conducted by Chen & Liu, (2004), it is reviewed, many of the company are getting many challenges when it comes to the information technology. Because as we know that the technology is developing day by day and the employees in the company should know about all of the new trends that are been adopted by the market according to the domain of the company. The use of data mining in the information technology domain is also very vast. In this paper, we come to review the effects of the data mining in the information system of the company (Chen and Liu, 2004).

In the field of information science, the use of the data mining is progressing because the techniques that are been used in the data mining. All of the techniques that are discussed in this paper would be very helpful for the future application of the data mining in the information system. The more efficient the information system is the more business is likely to manage its logistics that may be related to the supply chain management or the inventory control (Chen and Liu, 2004).

According to the research conducted by Battistini, Segoni, Manzo, & Catani, (2013), it is investigated that, the ways to use the data mining for the online news analysis. It is also very important to upgrade the database of the geohazard. This approach related to the online news also enables the system to get nonstop feedback from different events.

All of the information that is to be given through the data mining is through the historical records of the inventories and it also helps to forecast about the future inventories of the geohazard in the company (Battistini et al., 2013).

This paper has the deep information related to the online data analyzing and the approaches that can be used for the data mining related to the online news. The maps of the inventory about the floods, landslides, earthquakes all of them can be

updated by the use of the filters. By using this approach, it is also become easier to automate all of the inventories of the geohazard at the national level by using the perfect and more efficient approach of the data mining in the system (Battistini et al., 2013).

According to the research conducted by Hsu, (2009) it is investigated that, the use of the data mining in the heavy industries and how the data mining approach can be used to improve the overall standard of the industry in the market. By using the data mining in the industry, it gives more efficient information about the current inventory, material, and other equipment so that business may likely to make more effective decision about its product and marketing (Hsu, 2009).

In this paper, they discussed a lot about the industrial use of the data mining, what techniques they are likely to use in this matter. In the large industries, the management of the production and the inventories and also been one of the huge problems that are too faced by the business.

By using, the data mining approach the risk of the failure in the production and the inventory management become very low and the industry is likely to make the more good decision about the future product and inventory control based on the demands in the market. Therefore, we can say that this paper is very beneficial when it comes to preventing the problems related to the production and inventory of the industry (Hsu, 2009).

According to the research conducted by Liao, Chu, & Hsiao, (2012), it is analyzed about the implementation of the data mining techniques, and the changing that has been occurring during the past decade in these techniques. The paper also discusses some of the future development in the data mining applications expected.

There are various kinds of the system that is been used in the for the data mining related to the specific field of concern. For the inventory control, there are different approaches in the data mining, many kinds of software of the data mining are used for controlling of the inventory of the business (Liao, Chu and Hsiao, 2012).

This paper has discussed in detail about the data mining techniques that are very efficient and the ways to implement them in the software. There are also huge differences in the data mining tools that are used from the year 2001 to the year 2011.

By the increment in the requirements of the marketing and technology the techniques of the data mining are also been changed and modified to the most efficient to be used by the companies and help them in varies procedures related to the production, marketing , information technology and inventory control (Liao, Chu and Hsiao, 2012).

Chapter 3

This chapter includes the details about the inventory forecasting dataset that was exported from airline system. The details includes: the collection of the initial data, description of the data selected, types of the attributes within the dataset and cleaning of the data.

3. Data Collection

3.1 Initial Data collection

The inventory related data was collected from reporting software that is used in Emirates Airline Engineering and is called as “Business Intelligence”. Business Intelligence is a solution that provides timely and accurate information to better understand the business and to make more informed, real-time business decisions. Full utilization of BI solutions can optimize business processes and resources, improve proactive decision making and maximize profits/minimize costs.

This tool is linked to another real-time application like UM – Ultramian which is linked directly to the aircraft’s system. If any changes performed physically on the aircraft (Install / Remove) of items, the system captures the transactions. The first step toward getting the data is to login to the tool, then select a universe which is considered as a folder that is linked to the main database which is ULTRAMAIN. The Universe that was selected for this project is UM_Materials as shown below

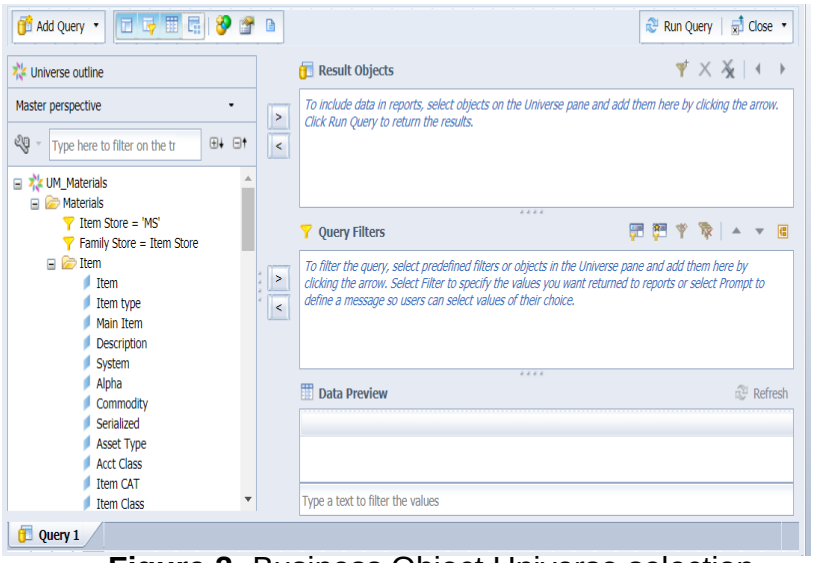


Figure 2: Business Object Universe selection

UM_Material: is a folder that includes the details related to all the materials / items used in the airline organization to perform the maintenance.

After selecting the universe, the tool open a “Query Panel” which shows different folders that includes set of objects and classes which can be filtered based on the need as below.

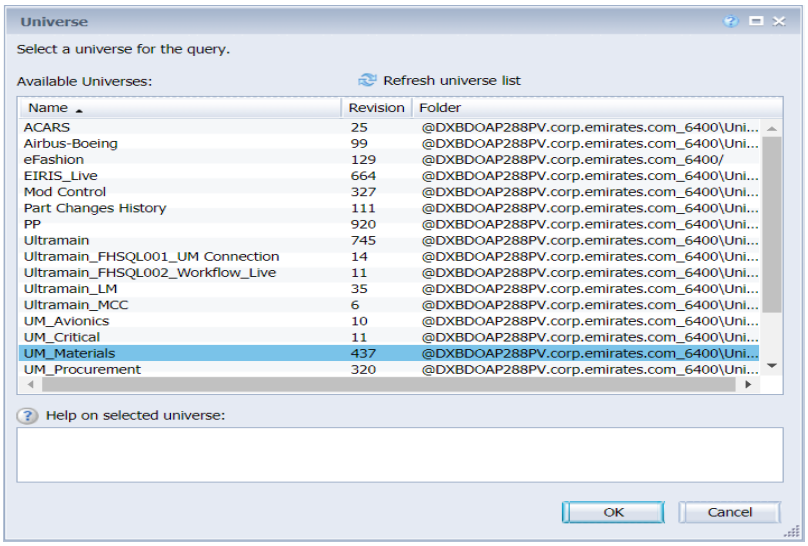


Figure3: Business Intelligence Fields selection

3.2 Description of the data collected

The data that was selected from the database are related to the inventory items from different categories for the year 2017. The items are located in different stores within the engineering centre. The below Table 1 describes the attributes / fields of the dataset. The dataset includes 14 attributes that describes each item listed in the dataset.

Field Name	Type	Definition
Item	Nominal	A unique identifier for item. Items may include a wide variety of entities such as assets, components, tools and equipment
Item type	Nominal	A string
Main Item	Nominal	A string
Description	Nominal	Description / Name of the item
System	Nominal	The system / ATA this item belongs to
Serialized	Binary	Yes indicates that item is tracked using a unique serial number (YES/NO)
Item CAT	Nominal	Item categories classify items by functional groupings of components such as avionics, mechanical, hydraulic, or lubricants for example.
Item Class	Nominal	Item with account class will defaults the costing method onto the item. Item can fall in one of these types (C, E, R, T)
Tool	Nominal	Indicated if the item is a tool or not (Yes or No)
Item Status	Nominal	Item status can be either (SETUP, ACTIVE, INACTIVE, TEMPORARY/ PROVISIONAL)
Stock Unit	Nominal	Unit of Measure
Store	Nominal	Identify the location of an item
Year	Nominal	The year in which usage has happened
Iss Total	Numeric	Total Issue quantity in a year when an issue, which typically recurs, was generated to meet a requirement caused by normal maintenance and part replacement activities
Item Criticality	Ordinal	It indicates whether the item is “Hight , Medium or Low” criticality to business operations.

Table 1: Inventory Forecasting Data Fields

3.3 Dataset Type & Attributes

Data Definition

A data is a collection of attributes and objects. The attributes are normally considered as the features of the objects. For example: the hair color of a person, the weather temperature, type of a car, etc. A collection of attributes describes an object which is also called as a record, sample or an entity.

There are different types of attributes as described in the below table

	Nominal	Ordinal	Interval	Ratio
Example	ID, Eye Color, Zip Code	Ranking, Grade, weigh, height (tall, medium, short)	Calendar Date, Temperature in Celsius	Temperature in kelvin, length, time, count

Attributes can be either “Discrete Attributes” or “Continuous Attributes”. Discrete attributes meaning when the values belonging to the dataset are distinct and separated which means the data can be counted. Despite the continuous attributes which means that the values belonging to the dataset are continuous and can take any value within a finite or infinite interval.

Dataset Types

A dataset can be from different types; it can be record dataset, graph or ordered dataset. The type of the inventory dataset is a Record Dataset as it includes a collection of records that is aligned with a fixed set of attributes. The dataset was exported from Business Intelligence tool and it includes different types of attributes as following:

- **Nominal:** Item, Item type, main item, description, system, serialized, item CAT, item class, tool, item status, stock unit and store.
- **Interval:** Year
- **Ratio:** Iss Total

Item class is considered an important attribute in the dataset as the data mining prediction will be based only on item class which is the Expendables items (E Class). Below is the description of each item class with the records (in value) available for each item class in the dataset.

Item Class	Description
Rotables – T class (7585 records)	Complex items Unlimited repairs no scrap is expected Controlled by serial number
Repairable – R class (24390 records)	components which can be repaired: Under normal conditions, a follow up of each individual serial number is not necessary. Have limited number of repairs and also have a possibility of scrap
Expendables – E class (32403 records)	Cannot be repaired and will be scrapped after removal and inspection result is unserviceable 100% replacement items Items which cannot be repaired
Consumables – C class (1155 records)	Any Item used once only Raw material Chemical material Items which merge on production with new product and cannot be removed

Table 2: Inventory Item Class Description

3.4 Data Preparation

3.4.1 Data Cleaning and Selection

The original dataset includes 65533 objects and 14 attributes. These objects are related to different items for the 2017. The report shows a total of 3.4 million of issue transactions performed for these items in 2017.

While looking at the data I have realized that so many information need to be removed as they are either not relevant or not meaningful. For example: “item type”, main item”, “system”and“stock unit records includes non-relevant information and need to be removed to

In addition, since the focus is on specific item class which is considered as very important/ critical for business operations which is the E Class – Expandable items category, other items classes R, T and C have been removed from the report. Iss total field is describing the number of transactions performed for each particular item in the report and is divided into the following

- Items with “0” issue transaction: which means that these items were never issued.
- Items with (- minus): This means that the items has been issued, not used and returned back to store.
- Items with normal transaction 1 issue transaction and above.

To get better and accurate results using classification technique, items with 0 and minus transactions have also been removed from the dataset as well as considering only the records related to 4 stores and they are: “MS – Main Store”, “LMS1 – Line Maintenance Store 1”, “LMS2 – Line Maintenance Store 2” and “LMS3 – Line Maintenance Store 3” as these stores are the biggest stores in the Engineering building and most of the part requests are orders through them.

While working with any dataset, it is very important and critical to insure that the right attributes are selected as it impacts the data mining modelling process. The below are some factors that may reduce the accuracy of a data mining model.

- **Misleading:** This means that the dataset includes redundant attributes that will impact the K-NN modelling algorithm as K-NN use small neighborhoods to determine classification prediction in the attribute space. These predictions can be highly impacted by redundant attributes.

- **Overfitting**: Keeping irrelevant attributes in your dataset can result in overfitting. Decision tree algorithms seek to create the best spits in attribute values. Those attributes that are more connected with the prediction are split on first. Deeper in the tree less relevant and irrelevant attributes are used to make prediction decisions that may only be beneficial by chance in the training dataset. This overfitting of the training data can affect the modelling accuracy negatively.

Chapter 4

This chapter includes detailed information about the data mining methodology used in this research paper. Classification has been selected as the methodology with K-nearest neighbors and Decision Tree algorithms. In addition, the section also includes the experimentation of the dataset on building both the models using Rapid Miner tool. The advantages and disadvantages of the algorithms are also explained.

4 Methodology

4.1 Data Mining Classification

Classification is a data mining methodology which is called as Supervised Learning which means that we know the criteria and features of the attributes within the dataset. It is as a function is data mining that assigns the items in the dataset to a target class or category and the goal of the classification is to predict the target label or class for each record in the dataset.

Classification is the data mining methodology used in this project. Classification is used to classify each item in the inventory dataset into one of predefined set of classes or attributes. The data analysis task classification is where a classifier or a model is constructed to predict categorical labels (the class label attributes). The goal of classification is to provide an accurate prediction based on the target class for record on the dataset. (G. Kesavaraj & S. Sukumaran, July 2013)

Classification process is divided into two main steps as following:

1- **Building the classifier model:** this step is considered as learning phase were classification algorithm builds the classifier from the training dataset which is made up of attributes and their associated label or class. As showing in **Table 3** is a sample of training data from the inventory forecasting dataset.

Item	Description	Serialized	Item Class	Item Status	Store	Year	Iss Total	Item_Criticality
00000584	BLOWER ASSY	NO	E	ACTIVE	MS	2017	6	Low
718-1592-001	MEDIA SET	YES	E	ACTIVE	LMS1	2017	3	Low
093118C156	COW LEATHER	NO	E	ACTIVE	MS	2017	17867	Medium
604015	FILTER BAG - VACUUM	NO	E	ACTIVE	MS	2017	12070	Medium
004955-018296-05Y	CARPET	NO	E	ACTIVE	MS	2017	124643.44	High

Table 3: Inventory Forecasting Sample Training Dataset

In the process of building the classification model, “Item_Criticality” has been set as the label with the below condition to help in predicting the criticality of an item to business operations:

- If “Iss_Total” value falls between “0.25” and “5000” then item criticality is “Low”
- If “Iss_Total” value falls between “5001” and “50000” then item criticality is “Medium”
- If “Iss_Total” value falls between “50001” or “above” then item criticality is “High”

Understanding the criticality of an item based on its usage for last one year will help the company to make better decision for the next year demand planning. This also means that the items falls under “High” criticality will be considered as important and the company will start planning the future demand first for this type of items compared to the items with “Medium” and “Low” criticality. **Table 3** is a sample that includes predefined classes which mean that each item in the list is labelled with the class it belongs to.

2- **Using the classifier model for classification:** this phase is considered as a testing phase to test the classifier model that was built in phase 1. To perform this phase a test dataset is required which is independent of the training set. In this

phase, the testing of the model on unknown objects as showing in the below **Table 4**:

Item	Description	Serialized	Item Class	Item Status	Store	Year	Iss Total	Item_Criticality
00-1428	CAP	NO	E	ACTIVE	MS	2017	483	Low
3M5N11N95	PARTICULATE FILTER	NO	E	ACTIVE	MS	2017	10694	Medium
005786RED	BAG, LAUNDRY-RED POLYTHENE	NO	E	ACTIVE	MS	2017	156200	High
005786GREEN	BAG, LAUNDRY-GREEN POLYTHENE	NO	E	ACTIVE	MS	2017	196200	?

Table 4: Inventory Forecasting Sample Test Dataset

Item criticality for item “005786GREEN” - BAG, LAUNDRY-GREEN POLYTHENE is unknown. For the classification model to be able to predict the item criticaly for the unknown item, it will look at the objects that has the nearest classes or attributes. In this case, item “005786RED” - BAG, LAUNDRY-RED POLYTHENE is considered as the nearest when looking at the “Iss_Total” for the last year “156200” is too close to “196200” compared with other items in the list. Hence, classification model predict that the missing value for item “005786GREEN” - BAG, LAUNDRY-GREEN POLYTHENE is “High”.

4.2 Classification Algorithms

4.2.1 K- nearest neighbour

K-Nearest Neighbor is considered one of the most popular and simpler algorithms that stores all available cases and classifies new classes based on measuring the similarity. Hence, KNN is being used by so many researches as it produces a very good performance when experimenting in different datasets. (N. Suguna1, and Dr. K. Thanushkodi, 2010)

K-Nearest Neighbor is a supervised learning algorithm which uses the training dataset to generate classification. It predicts the test samples according to the K training samples that are the nearest neighbors to the test sample. So if K=1, then it assigns the test sample to the nearest class from the training sample.

For testing of the KNN algorithm, first I have measured the performance based on 10000 records. The below demonstrates the operators used in rapid miner to build and classification K-NN model and to get the accuracy of the dataset

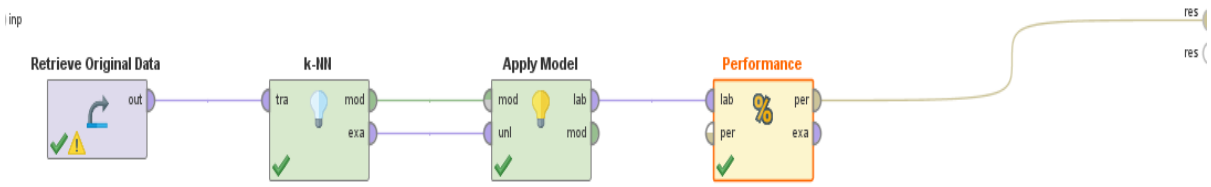


Figure 4: Rapidminer K-NN Model

Model Accuracy

The below figure 5 shows the confusion matrix for the K-NN model that was built. **Confusion Matrix** is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

accuracy: 100.00% +/- 0.00% (mikro: 100.00%)

	true High	true Low	true Medium	class precision
pred. High	8406	0	0	100.00%
pred. Low	0	1579	0	100.00%
pred. Medium	0	0	15	100.00%
class recall	100.00%	100.00%	100.00%	

Figure 5: K-NN Model Accuracy

Using the classification model for Prediction

In order to test the performance of the classification model, I have divided the original data into: Training Set with 7000 records and Test Set with 3990 records in order to verify the classification model built using the training dataset.

In Rapid miner the below operators have been used with the parameter details

- Inventory Training Dataset with 7000 records.
- Inventory Test Dataset with 3990 records and missing values “Item_Criticality”.
- K-NN with K=3, Measure Type= Mixed Measures, Mixed Measure = Mixed Euclidean Distance.
- Apply Model and Performance (Classification) operators.

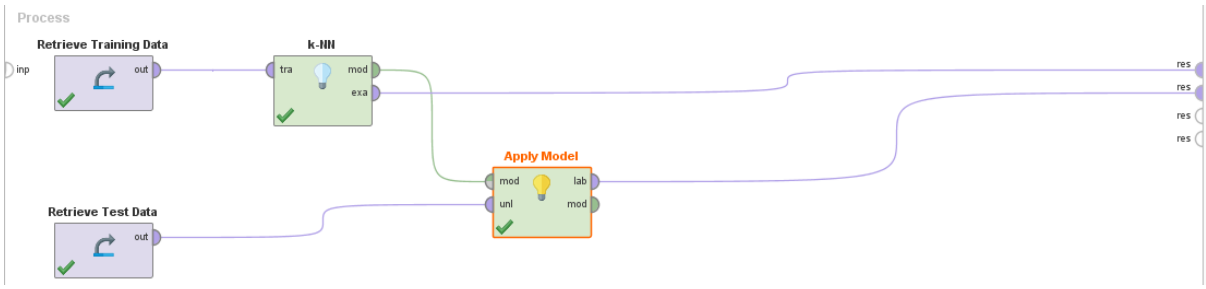


Figure 6: Building Classification Model for prediction

Any object in a dataset has different attributes and assigned to specific classes. For KNN to choose the best neighbors for the above record, it normally use distance functions to measure the similarity between the attributes in the datasets comparing to the target attribute. There are different types of distance function but **Euclidean Distance** is the one used in this project. Below is the Euclidean Distance formula:

$$d = \sqrt{\sum_{i=1}^{N} (Xi - Yi)^2}$$

Euclidean Distance: is a distance function that is used to find the similarity among objects within a dataset.

The training dataset was used to classify the unknown class with the help of Euclidean Distance and the below is the result from rapid miner considering K=3. As showing in figure 7, the KNN classification model prediction for the all un-known labels. As an example of the prediction below: item “3B3016-3R” Row No 1 was predicted as “High” item criticality which means that this item’s features or attributes

are only similar to other items labelled as “High” in the dataset. However, item “115684-001” Row No 5 with a confidence “High” with probability of 0.333 and confidence “Low” with probability of 0.667, the item has been predicted has “Low” Item criticality. This means that item 115684-001 has similar attributes to other items in the dataset with High and Low item criticality but because the probability of Low is higher than High, the model has predicted it as “Low” item criticality.

Row No.	Item_Critica...	prediction(Item_Criticality)	confidence(High)	confidence(Low)	confidence(Medium)	Item
1	?	High	1	0	0	3B3016-3R
2	?	High	1	0	0	810E102C5
3	?	Low	0	1	0	113W8111-17
4	?	High	1	0	0	113W8111-47
5	?	Low	0.333	0.667	0	115684-001
6	?	High	0.667	0.333	0	115684-001
7	?	Low	0.333	0.667	0	3210W605
8	?	High	1	0	0	52HS01
9	?	High	0.667	0.333	0	857787-001A
10	?	High	0.667	0.333	0	857787-001A
11	?	Low	0.333	0.667	0	3530-0279-09
12	?	High	0.667	0.333	0	62-0250-7
13	?	High	1	0	0	3530-0279-07
14	?	High	0.667	0	0.333	62-0250-11
15	?	High	0.667	0.333	0	77FV301-2
16	?	High	0.667	0.333	0	3215W504
17	?	High	0.667	0.333	0	7325225-891
18	?	High	1	0	0	332W5361-1
19	?	Low	0.333	0.667	0	0575YS
20	?	High	0.667	0.333	0	0575YS
21	?	High	0.667	0.333	0	0575YS
22	?	High	1	0	0	5021-5CPMA
23	?	High	1	0	0	435W1300-1A
24	?	High	1	0	0	00392-015
25	?	High	1	0	0	0020-6675
26	?	High	0.667	0.333	0	5155006001

Figure 7: Classification Prediction Result

4.2.2 Decision Tree

Decision trees are algorithms that build the classification in the form of tree structure. The logic of decision works in breaking the dataset into smaller and smaller subsets by having decision nodes (which can have two or more branches)

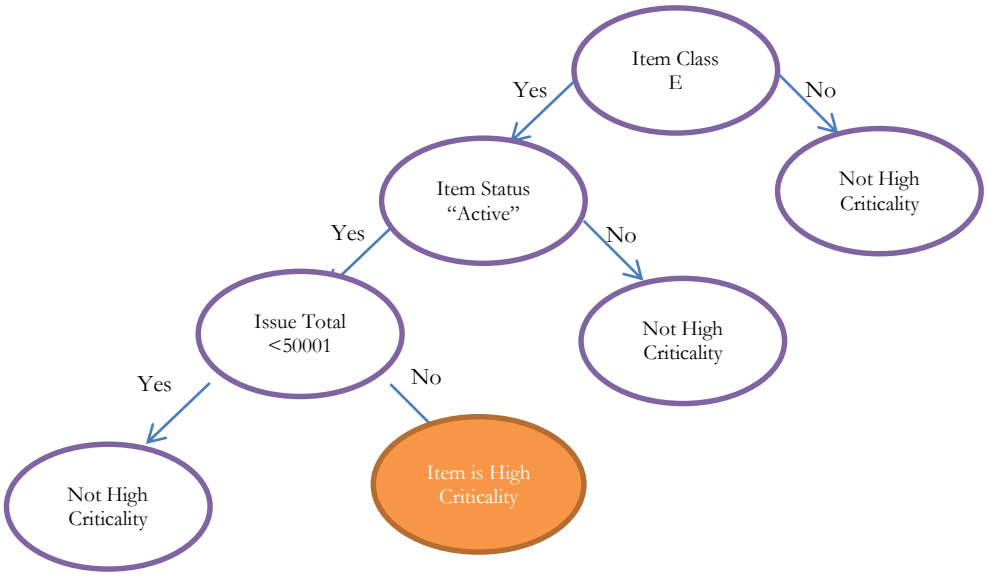
and leaf nodes which represents the prediction or the decision. Decision trees are very simple to use compared to other classification algorithms like K-NN as it works with different types of data including categorical and numerical data despite the K-NN algorithm which does not work with numerical attribute. The general idea of using Decision Tree is to build a training model which can be used to predict a class or value of target variables by learning decision rules inferred from prior data (training data).

Decision Tree Implementation process

The below steps are considered as the basic rule that need to be followed while building a decision tree model:

- Select the best attribute of the dataset to be at the root of the tree.
- Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
- Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

I have applied the above decision tree rule on the inventory forecasting dataset to make a decision whether the new item “00-1339 - PROFILE, SEAL” will be with High Criticality or not considering the attributes in the dataset as following:



The primary challenge in the decision tree implementation is to identify which attributes need to selected as the root node in each level. There are different attributes selection measures to identify the attribute which can be considered as the root node and information gain is one the measures.

- **Information gain**

  The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain. Information gain feature selection is computed using the below equation:

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Building Decision Tree Model in Rapid Miner

The inventory dataset has been imported to rapid miner to build the basic decision tree model as showing in the below Figure 8 using the Decision Tree operator. After running the module, rapid miner has given the result of the tree structure a showing in figure 9, the tree structure consist of “Iss Total” the top node and decision nodes at the bottom are High, Medium or Low.

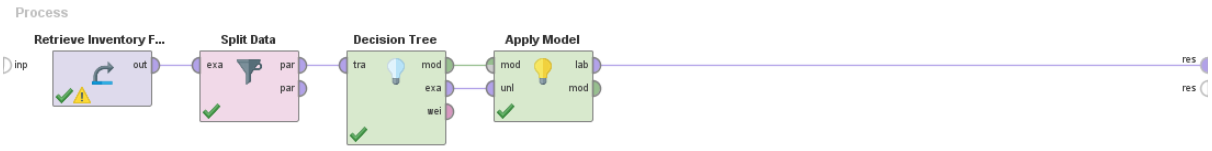


Figure 8: Decision Tree Basic Model

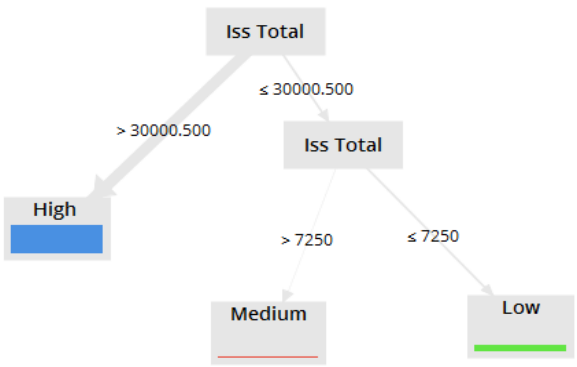


Figure 9: Decision Tree Model Result

Chapter 5

This chapter include the result and finding of the 2 classification algorithms: K-Nearest Neighbour and Decision Tree by using cross validation and multiple evaluation matrixes. This chapter also includes the differences between both the algorithms.

5 Result & Finding

K-NN Using Cross Validation

Cross Validation is a technique that is normally used to evaluate the predictive model by partitioning the dataset into Training and Test sets. The below table shows the KNN evaluation matrix based on different K values using cross validation operator in Rapid Miner.

Evaluation Matrix	K=1	K=200	K=800	K=1000
Accuracy	100.00%+/-0.00%	100.00%+/-0.00%	99.85% +/- 0.53%	99.85% +/- 0.05%
RMSE	0.000 +/- 0.000	0.000 +/- 0.000	0.010 +/- 0.037	0.038 +/- 0.006
Recall	100.00%+/-0.00%	69.17% +/- 8.78%	66.67% +/- 0.00%	66.67% +/- 0.00%
Precision	100.00%+/-0.00%	69.17% +/- 8.78%	66.35% +/- 0.10%	66.35% +/- 0.10%

Table 5: KNN Evaluation Matrix using Cross Validation

As showing in table 7 that the accuracy of the classification model decrease when the KNN value increase, hence the Error (RMSE) value also increase.

The KNN algorithm defines a category (label) in the test dataset for observation and compares it to the observations in the training dataset. Since rapidminer know the actual category of observations in the test dataset, the performance of the KNN model can be evaluated. The below equation is one of the

$$Average\ Accuracy = \sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i} / l$$

most commonly used parameter to compute model average accuracy:

Where TP is the true positive, TN is the true negative, FP is the false positive and FN is the false negative. The subscript i indicates category, and l refers to the total category. (Zhongheng Zhang, 2016)

There are different types of evaluation matrix that was used in this project to evaluate the KNN classification model such as: RMSE – Root Mean Squared Error, RRSE – Root Relative Square Error, Recall and Precision.

RMSE is mostly used with regression models; it demonstrates the sample standard deviation of the differences between observed values and predicted values. The RMSE values in Table 7 for different K values look not realistic as the classification model includes finite and countable class labels which do not correspond to numbers.

Recall is normally answering the question “"How many of all of the relevant results did the search show me?”. Looking at the values in Table 7, for K=1 Recall was 39.26% which indicates the positive prediction probabilities from the dataset. It is also observed that when K value increase, the prediction probability reduces. Recall is normally calculated using the equation TP/ (TP + FN). Similarly with Precision that is indicating “How many of the results is actually relevant?” and when looking at the prediction probability of precision when K=1 is 39.93% which also reduces when K value increase. Precision equation is = TP/ (TP + FP)

Decision Tree Using Cross Validation

Comparison between K-Nearest Neighbour and Decision Tree

The below figure shows the difference in evaluation matrix between K-NN and Decision Trees when K=1 and Decision Tree Depth = 10. It is obvious from the below figures that K-NN model has better performance and accuracy compared to the decision tree model.

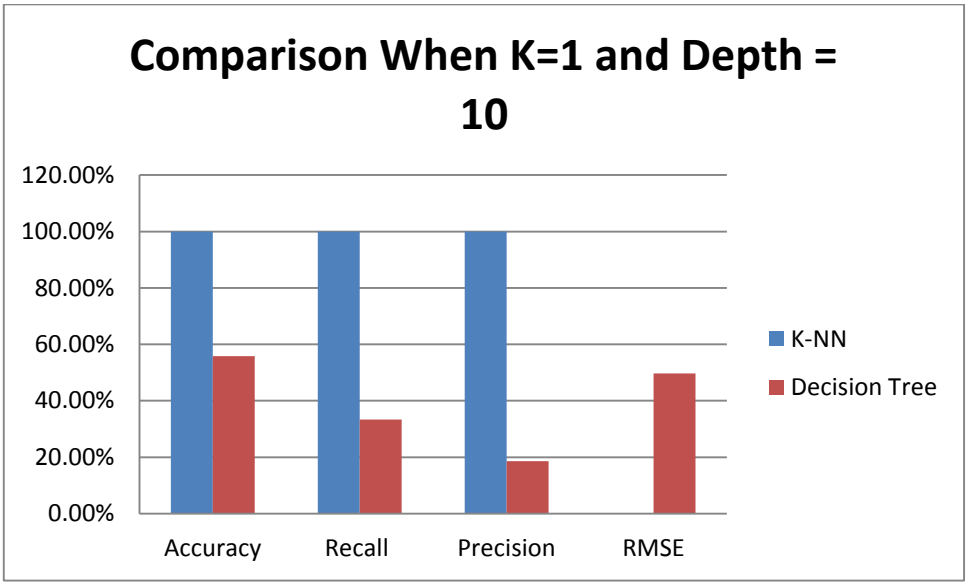


Figure 10: Comparison Chart

Advantages and Disadvantages of both models

Evaluation Matrix	Depth= 10	Depth= 1000	Depth= 3000	Depth= 5000
Accuracy	55.82%	55.82%	55.82%	55.82%
RMSE	0.497 +/- 0.000	0.497 +/- 0.000	0.497 +/- 0.000	0.497 +/- 0.000
Recall	33.33% +/- 0.00%	33.33% +/- 0.00%	33.33% +/- 0.00%	33.33% +/- 0.00%
Precision	18.61% +/- 0.01%	18.61% +/- 0.01%	18.61% +/- 0.01%	18.61% +/- 0.01%

Table 6: Decision Tree Evaluation Matrix using Cross Validation

	Advantages	Disadvantages
K-NN	1. Simple to implement 2. Flexible to distance choices	1. Require large search problem to find nearest neighbour 2. Data storage
Decision Tree	1. Leading toward a meaningful outcome. 2. Simple logic	1. Selecting irrelevant attribute will reduce the accuracy. 2. Complexity while preparing large tree model.

6 Discussion

Successful use of data mining depends mostly on the quality of the historical data available in any organization's data warehouse. The real inventory forecasting value is depending on the organization's ability to associate business operations continuity in case of any part failure. Having enough item demands in the organization's store will help act immediate in case of critical maintenance requirements. Being able to identify item's criticality to business operations play a major role in planning ahead for future inventory demands.

The data mining methodology presented in this paper is one of the mostly used methods and it provides an accurate result for the classification KNN model. Different evaluation matrix has been used to clearly evaluate the probability of the classification model to predict items with unknown label. However, the work presented does not represent a standalone solution. The approach requires to be combined with clustering algorithm to get better and improved accuracy.

The study included data mining in the forecasting of inventory of airline industry, the analysis were then compared to determine the methods and appropriate consideration for the supply chain management and enables the researchers to find and predict the item criticality which will help to proper plan future demands.

7 Conclusion

It is concluded that the research study can be made after discussing the result of the research. Data mining in inventory forecasting is not a new topic, however not very popular in airline industry as less researches found online compared with the implementing the same study on other industries. Based on the survey conducted on, supply chain management consider the item with Class E the most important, hence predicting the criticality of these items will help to make better decision for future demand planning.

Predicting future demands help in making better decisions and excellent planning ahead and hence better results of the inventory management. There are also expectations to increase the monetary growth rates through the better data analysis in order to hold up the statement. Moreover, it is analyzed through the research that data mining and its algorithms are very helpful while performing data analysis on small or large datasets.

It is concluded that data mining is very important for any airline organization as it helps to save cost on demand forecasting. Moreover, it is noticed that data mining processes are eventually helping the business to control its inventory and there are relevant data mining technologies in the inventory control. It is also analyzed that airline companies and supply chain manager are using the financial statement and future condition of all of the finance in the organization's business through the data mining process as they have recognized the values of the data mining process. The data mining also help the company to aware of the entire inventory so that company could able to make the better decision about their inventory.

Appendix

Questions	Response
Have you heard about Data Mining – Predictive Analytics?	Yes: 60% No: 40%
What is the current process of managing Inventory demands in EK stores?	Cabin Mobility First In First Out Historical Data Ultramain No Idea
Do you believe that data mining can help in managing inventory forecasting?	Yes: 90% No: 10%
How many items are available in our EK store?	100K 200K More than 200K – 100%
What is the most important item class from your perspective and why?	Rotables – T class 10% Repairables – R class 22.22% 2 Expendables – E class 66.67% Consumables – C class 1%

Reference

Aburto, L. and Weber, R. (2007) 'Improved supply chain management based on hybrid demand forecasts', *Applied Soft Computing*, vol. 7, no. 1, pp. 136-144.

Auramo, J., Kauremaa, J. and Tanskanen, K. (2005) 'Benefits of IT in supply chain management: an explorative study of progressive companies.', *International Journal of Physical Distribution & Logistics Management*, vol. 35, no. 2, pp. 82-100.

BANSAL, K. (1998) 'Brief Application Description Neural Networks Based Forecasting Techniques for Inventory Control Applications', *Data Mining and Knowledge Discovery* , pp. 97–102.

Bansal, K., Vadhavkar, S. and Gupta, A. (1998) 'Brief Application Description. Neural Networks Based Forecasting Techniques for Inventory Control Applications', *Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 97–102.

Battistini, A., Segoni, S., Manzo, G. and Catani, F. (2013) 'Web data mining for automatic inventory of geohazards at national scale;Nicola Casagli;', *Applied Geography*, vol. 43, p. 147e158.

Bhardwaj, B.K. and Pal, S. (2012) 'Data Mining: A prediction for performance improvement using classification.', *arXiv preprint arXiv*, vol. 1, no. 1, pp. 1201-3418.

Brandenburg, M., Govindan, K., Sarkis, J. and Seuring, S. (2014) 'Quantitative models for sustainable supply chain management: Developments and directions.', *European Journal of Operational Research*, vol. 233, no. 2, pp. 299-312.

Chen, S.Y. and Liu, X. (2004) 'The Contribution of Data Mining in Information Science ', *Journal of Information Science*.

Dhond, A., Gupta, A. and Vadhavkar, S. (2000) 'Data Mining Techniques for Optimizing Inventories for Electronic Commerce '.

Ellis, R., Allen, T. and Petridis, M. (2007) 'Applications and Innovations in Intelligent Systems XV', *Springer*, vol. 1, no. 1, pp. 120-135.

Elmaghraby, W. and Keskinocak., P. (2003) 'ynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions.', *Management science* , vol. 49, no. 10, pp. 1287-1309.

Ferrero, I. and Sison, A.J.G. (2014) 'A quantitative analysis of authors, schools and themes in virtue ethics articles in business ethics and management journals (1980–2011)', *Business Ethics: A European Review*, vol. 23, no. 4, pp. 375-400.

Hsu, C.-H. (2009) 'Data mining to improve industrial standards and enhance production and marketing: An empirical study in apparel industry', *Expert Systems with Applications* , pp. 4185–4191.

Jacobs, F.R. and Bendoly, E. (2003) 'Enterprise resource planning: developments and directions for operations management research.', *European Journal of Operational Research*, vol. 146, no. 2, pp. 233-240.

Kirkos, E., Spathis, C. and Manolopoulos, Y. (2007) 'Data mining techniques for the detection of fraudulent financial statements', *Expert systems with applications* , vol. 32, no. 4, pp. 995-1003.

Kohavi, R., Rothleder, N.J. and Simoudis, E. (2002) 'Emerging trends in business analytics.', *Communications of the ACM*, vol. 45, no. 8, pp. 45-48.

Koh, H.C. and Tan, G. (2011) 'Data mining applications in healthcare.', *Journal of healthcare information management*, vol. 19, no. 2, p. 65.

Kusiak, A. and Smith, M. (2007) 'Data mining in design of products and production systems.', *Annual Reviews in Control*, vol. 3, no. 1, pp. 147-156.

Lavrač, N., Motoda, H., Fawcett, T., Holte, R., Langley, P. and Adriaans., P. (2001) 'Introduction: Lessons learned from data mining applications and collaborative problem solving', *Machine learning*, vol. 57, no. 1, pp. 13-34.

Liao, S.-H., Chu, P.-H. and Hsiao, P.-Y. (2012) 'Data mining techniques and applications – A decade review from 2000 to 2011', *Expert Systems with Applications*, vol. 39, no. 12, pp. 11303-11311.

Liu, L.-M., Bhattacharyya, S., Sclove, S.L. and Chen, R. (2001) 'DATA MINING ON TIME SERIES: AN ILLUSTRATION USING FAST-FOOD RESTAURANT FRANCHISE DATA'.

L, L., Shen, C., Wang, L., Zheng, L., Jiang, Y., Tang, L., Li, H., Zhang, L., Zeng, C., Li, T., Tang, J. and Liu, D. (2014) 'iMiner: Mining Inventory Data for Intelligent Management'.

Min, H. and Zhou, G. (2002) 'Supply chain modeling: past, present and future', *Computers & industrial engineering*, vol. 43, no. 1, pp. 231-249.

Moin, K.I. and Ahmed, Q.B. (2012) 'Use of Data Mining in Banking', *International Journal of Engineering Research and Applications (IJERA)*, vol. 2, no. 2, pp. 738-742.

Provost, F. and Fawcett, T. (2013) 'Data Science and its Relationship to Big Data and Data-Driven Decision Making', *Big Data*, vol. 1, no. 1, pp. 51-59.

Renesse, R.V., Birman, K.P. and Vogels, W. (2003) 'Astrolabe: A robust and scalable technology for distributed system monitoring, management, and data mining.', *ACM transactions on computer systems (TOCS)*, vol. 21, no. 2, pp. 164-206.

Rossel, R.A.V. and Behrens, T. (2010) 'Using data mining to model and interpret soil diffuse reflectance spectra.', *Geoderma*, vol. 158, no. 1, pp. 46-54.

Rygielski, C., Wang, J.-C. and Yen, D.C. (2002) 'Data mining techniques for customer relationship management', *Technology in Society*, pp. 483–502.

Sánchez, A.M. and Pérez, M.P. (2005) 'Supply chain flexibility and firm performance: a conceptual model and empirical study in the automotive industry', *International Journal of Operations & Production Management* , vol. 25, no. 7, pp. 681-700.

Sandborn, P.A. (2007) 'A Data Mining Based Approach to Electronic Part Obsolescence Forecasting', *IEEE Transactions on Components and Packaging Technologies*, vol. 30, no. 3, pp. 397 - 401.

Vadhavkar, K.B.S. and Gupta, A. (1998) 'Neural networks based data mining applications for medical inventory problems', *International Journal of Agile Manufacturing*, vol. 1, no. 2, pp. 187-200.

Vercellis, C. (2011) *Business Intelligence: Data Mining and Optimization for Decision Making*, New Jersey: John Wiley & Sons.

Waller, M.A. and Fawcett, S.E. (2013) 'ata science, predictive analytics, and big data: a revolution that will transform supply chain design and management.', *Journal of Business Logistics*, vol. 34, no. 2, pp. 77-84.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J. (2016) *Data Mining: Practical Machine Learning Tools and Techniques*, Massachusetts: Morgan Kaufmann.

Wu, D.D. and Olson, D.L. (2013) 'Computational simulation and risk analysis: An introduction of state of the art research.', *Elsevier*, vol. 1, no. 1, pp. 1581-1587.