The Effects of Data Mining on Small Businesses in Dubai

آثار استخراج البيانات علْى الشركات الصغيرة في دبي

By Rasha AlMutawa
60013

Dissertation submitted in partial fulfillment of
MSc Information Technology

Faculty of Engineering & IT

Dissertation Supervisor
Dr. Sherief Abdullah

September-2011

DISSERTATION RELEASE FORM

Student Name	Student ID	Programme	Date
Rasha AlMutawa	60013	Msc IT	September 2011

Title
The effects of Data Mining on Small Businesses in Dubai

I warrant that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.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 one copy of my dissertation 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 that copy available in digital format if appropriate.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my dissertation 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

ACKNOWLEDGEMNENT

I would like to thank International House for their assistance in providing the databases to be used in this study as well as the time of the employees and their cooperation and input on the applications tested. I would also like to thank them for allowing the use of their hardware for testing the applications. I would also like to thank Sara El Rifaee, IT senior consultant for her IT support.

Declaration.

I declare that the thesis embodies the results of my own work and has been composed by myself. Where appropriate within the thesis I have made full acknowledgement to the work and ideas of others or have made reference to work carried out in collaboration with other persons. I understand that as an examination candidate I am required to abide by the Regulations of the University and to conform to its discipline and ethical policy.

Signature of student …… **Date** :…….

Word count: 15,756

Contents

Abstract

This study was conducted with the aim of finding the benefits of data mining on small businesses and specifically on the real estate sector in Dubai, and the criteria to be used when evaluating and selecting the best tools to consider for such businesses. While there are numerous studies on the best data mining models and their uses, even on certain industries, this study focuses on the applications more than the algorithms and models and their usefulness for small businesses specifically. Qualitative data was gathered from information found in relatively similar studies and from a real estate company in Dubai that allowed this study to be conducted on its databases with the help of its management and staff through interview, discussions, and direct implementations of the tools for testing.

The benefits of using data mining on such businesses was confirmed, a different criteria from the traditional kind for evaluating the available tools was used, and an existing cross industry standard for implementing data mining projects (CRISP-DM) was used to evaluate five selected tools for the business based on availability of models, price, platform, and popularity. The five selected tools that were evaluated for the real estate company are RapidMiner, Weka 3.6, NeuroXL, Knowledge Miner, and DMSK with RapidMiner leading the tools in advanatges for a small business in the real estate sector in the city of Dubai.

لقد أجريت هذه الدراسة بهدف العثور على فوائد استخراج البيانات على الشركات الصغيرة وتحديدا على قطاع العقارات في دبي ، و المعايير التي قد تستخدم عند تقييم واختيار أفضل الوسائل لمثل هذه الشركات. في حين أن هناك العديد من الدراسات حول أفضل نماذج استخراج البيانات واستخداماتها، وحتى في بعض المجالات تحديدا، فإن هذه الدراسة تركز على برامج استخراج البيانات أكثر من خوارزميات ونماذج و تركز ايضا على جدوة هذه البرامج بالنسبة للشركات الصغيرة على وجه التحديد. وتم جمع البيانات النوعية من المعلومات الموجودة في دراسات مماثلة نسبيا ومن شركة عقارية في دبي سمحت باجراء هذه الدراسة على قواعد البيانات الخاصة بها مع مساعدة من الإدارة والموظفين عن طريق المقابلات، والمناقشات ، والتطبيقات المباشرة للبرامج بهدف لاختبار .

وقد تم تأكيد فوائد استخدام برامج استخراج البيانات على هذه الشركات، ايضا تم اسنخدام معايير مختلفة عن النوع التقليدي لتقييم واستخدام الأدوات المتاحة. وعبر معايير موحدة موجودة لتنفيذ مشاريع التنقيب عن البيانات تعرف باسم كريسب-دي. ام. تم تقييم خمس أدوات مختارة لهذه الشركات على أساس توافر النماذج ، والأسعار ، وتوافر المساعده التقنية ، و مدى شعبية البرامج. الأدوات الخمسة المختارة التي تم تقييمها لشركة العقارات في دبي هية : رابيدماينر, ويكا, نيورو اكسل , نوليج ماينر ,و دي. ام. اس. كي.

Chapter One: Introduction

This chapter will present the background of the problem, the problem statement, rationale and significance of the study. The chapter aims at ensuring that a proper understanding of the need for the study is developed and communicated.

1.1. Background

Significant developments have been recorded in the protocols and techniques used in computing. The developments have been accompanied by increased usage of computer technology by small and large businesses. Small businesses which play an important role in overall GDP development and employ over 90% of the global workforce are as important as large businesses in supporting financial and economic development (Hölzl, 2009). Most businesses however use basic computing services such as communication within internal and external networks, processing word documents and spreadsheets and other basic office computer applications. Beyond the field of computer science, data mining and pattern matching are often considered technical specialist areas even though they have been practiced unknowingly by humans for decades a long time before computers could.

Small businesses face stiff competition from other small businesses. The flexibility offered by the small size and the large number of small and medium sized businesses all contribute to the high intensity of competition in small and medium sized businesses (Boden, Nett and Wulf, 2010). Market dynamics also determine the levels of competition and competitive pressures felt by players within different industries. The real estate industry in Dubai was affected and is still being affected by the 2008-2009 economic recession. Compared to large businesses that have easy access to external lending and highly experienced financial consultants and analysts within their ranks, small businesses are gullible to negative environmental changes (Boden, Nett and Wulf, 2010). This increases the need for better data management and analysis strategies among SMEs in the Dubai real estate industry. Pattern (data) mining as an advanced data analysis strategy can detect deep lying trends that can help businesses harness opportunities that are not visible to other firms. Determining challenges that businesses face in using pattern mining, the availability of the technology and its evaluation and its perceived importance within the real estate industry is vital in ensuring that the pattern and data mining are effectively harnessed by small businesses in the Dubai real estate industry.

1.2. Problem Statement

Changes in financial dynamics have resulted in operational environments that are demanding on small businesses. In competitive environments, the ability to determine the opportunities and threats that exist in the market and develop strategies aimed at harnessing or circumventing them plays an important role in determining the entities that will be successful and those that will fail (Hölzl, 2009). Small businesses have to effectively utilize the existing information on the market, industry and customer preferences to be able to effectively operate in competitive markets. Importantly, organizational ability to maximize on the enabling ability of technology and new capabilities may determine the short term competitive advantage that they gain (Boden, Nett and Wulf, 2010). Data and pattern mining are relatively new areas of computer application that are yet to gain extensive use as other areas of computer application in small businesses. This implies that the organizations that choose the technology are likely to gain advantages that are beyond the reach of other established businesses. Thus, pattern matching and data mining techniques offer a unique opportunity for small firms to gain competitive advantage and improve their position in the market with respect to industry trends and market needs. However, being a new technology, there are a number of challenges that hamper its utilization by small businesses in the Dubai real estate industry.

Since the emergence of data mining tools people have been trying to compare them and their effectiveness on different industries. There are many discussions on this issue dating back from the nineties to the current date from big seminars like the one done in Rhode Island in 1998, which was a workshop on artificial intelligence approaches to fraud detection and risk management that involved 50 participants from universities and industry researcher to online forums and discussion boards by new data mining tool users that are trying to implement these tools in their companies to make their jobs a little easier and more efficient (Fawcett, Haimowitz, Provost & Stolfo, 1998). While the problem just a few years ago was that there was little information and research on the challenges faced by small businesses specifically and the options they have when choosing a data mining or pattern matching tool, it has now become the abundance of tools and their numerous

evaluations that makes it difficult. Lack of clear criteria that can be used by small organizations to evaluate the pattern matching tools so as to determine the most effective is another issue that has yet to be comprehensively addressed in research. These are some of the issues that will be addressed by the study with the aim of not only developing greater appreciation of pattern mining applications but also improving their use by small businesses in the Dubai real estate industry.

1.2.1. Research Aim

The aim of the study is to investigate the potential gains in using pattern matching tools that small businesses stand to gain in Dubai and the challenges that they face in choosing and using existing pattern matching tools. Based on the definition of the aim, the study will focus on the challenges faced by small businesses in the Dubai real estate industry.

1.2.2. Research Objectives

The following research objectives helped in the attainment of the research aim:

a) To determine the pattern /data mining tools have been optimized for real estate industry and small businesses.
b) To determine the challenges that small businesses in the real estate industry face in evaluating and using the pattern /data mining tools.
c) To develop a criterion that can be used by small businesses in evaluating pattern/data mining tools.
d) To determine the utility of pattern/data mining tools to small businesses.

1.2.3. Research Questions

The following research questions were used in the pursuit of the research objectives

a) Which pattern /data mining tools have been optimized for real estate industry and small businesses?
b) What are the challenges that small businesses in the real estate industry face in evaluating and using the pattern/data mining tools?
c) What criterion can be used by small businesses in evaluating pattern/data mining tools?
d) What are the benefits of pattern/data mining tools to small businesses?

1.3. Rationale and Significance

There are several factors that have motivated this study. First, I have a special interest in the Dubai real estate industry and specifically the modalities that small businesses can use to effectively operate in the highly competitive segment. Having

information on patterns and trends that are invisible to others is a potentially effective avenue to successful operation in the industry. Pattern and data mining tools could therefore assist in maximizing the potential gains by individual small businesses and the overall performance of the real estate industry. However, there is limited research and information on key considerations when choosing data and pattern mining tools for small businesses specifically. Technical evaluations of expensive tools and their functionality is abundant, however, research on the best tools to use for small businesses is mostly limited to discussions by individuals interested in implementing cheaper tools in their businesses and comparisons by the developers of those tools themselves which is usually biased to their own developed tool. Additionally, there are few studies focusing on the utilities and challenges associated with the use of data or pattern mining tools by small businesses and in the real estate industry. This is the main motivation for carrying out this study.

The study findings will be of critical importance to development in the Dubai real estate industry and by small businesses. The potential benefits associated with the use of pattern mining by small businesses have yet to be realized due to limited research. This study will highlight the potential gains and develop a criterion that will set small businesses in the right track to maximizing the use of pattern matching tools. Next, the study will analyze the existing pattern matching tools with the aim of determining the best tools for small businesses in the real estate industry. This will help minimize the complexities associated with choosing the right data mining tool. In a nutshell, the study will help minimize the time spent on choosing the wrong data mining tool and therefore help small businesses effectively utilize and gain from the benefits associated with data mining.

Data mining is a fairly new area of computer application that has potentially infinite areas of use. It is an area of artificial intelligence that is extensively applied to decision support systems and can significantly aid business decision making at both the executive and operational levels (Boden, Nett and Wulf, 2010). The study will help create awareness on the practical challenges faced by small businesses in the real estate industry in evaluating and using data mining tools. In this way, the study will help propagate

research into the use of data mining and pattern mining tools by small businesses.

Chapter Two: Literature review

2.1. Data Mining Evolution

Despite the developments in technology that have led to huge databases and rapid development in computer application across industries, there are still challenges in analysis and utilization of the available data (Guoyin, & Yan, 2009). A large database is of little use to an organization if it is unable to use the data to effectively change its mode of operations and gain competitive advantage. Developments in information and technology have generally been more oriented towards the centralization of data for ease of access and management. As the importance of information and data became clear to businesses, the management and use of data has gained critical importance. This is one of the forces driving the rapid developments that have been experienced in computing in the recent past. It is important to note that the high demand forces manufacturers to develop better and highly intensive database management systems. However, the development and subsequent use of data warehouse and data warehousing techniques played important role in the development of pattern mining as one of the core areas in data mining.

Data mining in short is "the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform this data into information (Mehmed, 2003).” There are many uses for data mining. It is used in the form of pattern detection as face recognition or fraud detection; it is used medically for illness diagnosis, it is even used in gaming by some of the biggest gaming companies in the world such as Microsoft’s XBOX. In a single company, data mining can be applied in many cases and in almost all levels of a company. Data mining is used extensively in big scale projects such as search engines such as Google but is just as useful on smaller projects.

Data mining is a platform upon which the hidden value within large volumes of data can be unlocked. Through the use of algorithms and tools, data mining has remained an art rather than a process that is well understood within industries (Rupnik, Kukar, & Krisper, 2007). This is a major problem that hinders the use of data mining in large projects and is a barrier to the adoption of data mining by large corporate users. However, small businesses enjoy the flexibility associated with their small size and are therefore better placed to effectively utilize the immense benefits associated with data and

pattern mining. Decision making in small businesses is less bureaucratic and the risk in innovation or trying out new technologies is less pronounced than in large corporations. This places small businesses in a better position to experiment with new technologies and possibly gain benefits that can result in improved competitiveness. On the other hand, small businesses have limited resources and usually a lower number of technical personell than found in bigger organizations and therefore have little time to spend on evaluations and experimentations on different tools.

2.2. Data Mining Techniques

Before looking at the previous works done on the subject, a few clarifications on the subject are mentioned as a general refreshment or introduction of data mining concepts in this and the following sections.

Data mining could be done by humans, however it is much faster and much more efficient to use data mining tools. These tools use artificial intelligence. Artificial intelligent software are different from normal software by the fact that they continuously adjust themselves to the data that is being used on them. They have the ability to "learn". Data mining tools model, compute, graph, analyze data and much more.

There are too many data mining techniques to be discussed in this paper, therefore, only some of the most widely used techniques will be mentioned briefly in this section.

- Neural networks: also known as artificial neural networks to differentiate it from the biological neural networks, has interconnecting neurons or nodes to connect data to each other. The technology is derived from the basis within which the human brain works. . When simply put it sounds simple, but it is one of the most complex data mining techniques. One of the major reasons as to why the computer programmers have been slow in advancing neural networks for wider applicability is because neural networks are ambigious to comprehend (Witten & Frank, 2005).
- Naïve Bayes classifiers: based on Bayes Theorem. It assumes independency, that is, the existence of one feature of a class is in no way related to the existence of another feature of a different class (Mitchel, 1997). Naïve Bayes is dependent on probabilities. And even

though naïve bayes models are very simple they are very useful and widely used (Mitchel, 1997)

- KNN (K-nearest neighbor): classifies data according to the class that contains data with the nearest likelihood. This is another widely used technique which is highly effective. K is the number of nearest neighbors that are considered when classifying (Bremner, Demaine, Erickson, Iacono, Langerman, Morin & Toussaint, 2005).
- Decision Trees: are as the name implies tree-like models or graphs that make up decision support tools. They model decisions, their possible consequences, their probability, their cost and utility (Yang, 2006).

There are many more data mining techniques with various functions and models, and various ways of working. There is no way of deciding on a best technique as all of them have their strengths and weaknesses, however, there are ways to determine which techniques work better for certain types of datasets.

2.3. Standards of Data Mining

Ever since data mining increased in popularity, institutes as well as some organizations and individuals formed groups to set standards for the process of applying data mining on data to generate knowledge.

The Cross industry Standard Process for Data Mining (CRISP-DM) is a move started in 1996 and aimed at shifting the focus of data mining from technology to addressing needs at user levels (Sumathi and Sivanandam, 2006). The aim of the process is to ensure that businesses and non-technology specialist are included in the process of data mining with the aim of ensuring that business problems are comprehensively addressed. Data mining as defined under the CRISP-DM is applicable to virtually every sector and to different firms irrespective of their sizes and area of operation. CRISP-DM also makes large data mining projects run faster whereas improving the management, reliability and cost associated with the data mining process. Studies show that CRISP-DM has potential benefits to even small scale data mining initiatives and is an important step towards standardizing the processes involved in data mining (Sumathi and Sivanandam, 2006).

CRISP-DM addresses two issues namely: mapping business issues to data mining problems and capturing and understanding data. These are key issues in ensuring that any technology is of use to a business and is important in determining the actual steps that are taken in data mining operations. CRISP-DM has six stages starting with business understanding during which the business objectives are determined and better understood to have a better grip on what data mining is supposed to achieve. the data understanding stage follows in which all available data is explored, the quality of it assessed and outliers are found. Then follows the data preperation stage which is the most time consuming stage, during which the data related to the business is selected, the data is cleaned and prepared for the next stage. in the modeling stage, the modeling technique is selected, a model is built and assessed. Next, in the evaluation stage the model is evaluated on how well it performed on the data and the results generated. Finally in the deployment stage, it is determined how the results of the data mining will be used and by who andhow often.

Another data mining standard developed is KDD, knowledge discovery in databases, started in 1989 which referds to the process of finding knowledge in data and to focus on high level application of certain data mining methods. There are five stages associated with KDD which are: Selection, pre-processing, transformation, data mining and interpretation/evaluation. In the selection stage, the data to be used either as a sample or a complete data set is selected, after which the data is made consistant by cleaning the data and pre-processing it in the pre-processing stage. The transformation stage includes transforming the data into a more manageable set of data such as reducing the amount of data. The data mining stage is then applied in which a pattern is detected or a classification is made oe whatever the objective of the data mining is. Finally, the interpretation or evaluation stage evaluated the pattern or the knowledge that was mined or found as a result of the previous steps.

SEMMA is another data mining standard developed by the SAS institute. It stands for Sample, Explore, Modify, Model and Assess. As he first stage of KDD, the sample stage extracts a data sample from a large dataset that includes all the important information but is small enough to be controlled and tested in a short amount of time. Then comes the Explore stage in which the data is explored to find unexpected trends and discrepencies in order to understand how to deal with the data and modify it in the next stage, the modify stage. In the model stage, the data is modeled to obtain

knowledge from the data which is then evaluated in the Assess stage to determine the usefulness and reliability.

While these standards help a business perform data mining correctly, they do not set criteria for selecting the best data mining tools for certain industries or businesses.

2.4. Criteria for Choosing Data Mining Tools

Choice of the best tool is a major issue in the use of data mining applications to support business operations. Though developments in data mining have primarily focused on the development of strong analytic engines and algorithms, data mining is a process that is primarily dependent on the steps taken to collect and prepare the data to be analyzed. However, the development of tools that support or aid data preparation has been awarded little importance in the evolution of data mining tools. A proper evaluation of the models and functionalities offered by existing data mining tools is thus of critical importance when making decisions on the exact data mining tool to use.

Data preparation often includes the assessment and determination of missing values, outliers, collinearity and frequencies of multiple codes (Shmueli, Patel and Bruce, 2010). Other aspects that are included in the data preparation phase when mining for patterns include merging multiple datasets, mapping metadata, transformation of contents of similar variables, changing data types and derivation of new variables form existing variables. These tasks are estimated to constitute up to 80% of all the tasks involved in data mining (Shmueli, Patel and Bruce, 2010). However, most data mining tools provide little support for these important tasks. These initial tasks are demanding and businesses should look for tools that provide adequate support for data preparation.

Good data mining software should provide tools that allow for the evaluation of models. In theory, the best model offers the greatest levels of accuracy when predicting the classification state of target variables and is robust when working on validation data sets (Rao, Wegman and Solka, 2005). This implies that in using data mining applications it is important to consider the combined accuracy when the application is used on non-participants and participants. In practical settings this implies that the specific performance of

an application in the setting and its known levels of performance with respect to validity and accuracy must be guaranteed. This approach which is referred to the global accuracy method is used extensively by most data mining models with few users aware of the underlying assumption. Underlying the global accuracy theory is the assumption that the costs associated with all types of classification errors are the same and equal (Rao, Wegman and Solka, 2005). This assumption is well-intentioned in theory though it has vast weaknesses in practice. The global accuracy method has several flaws when used in CRM data mining operations and specifically in operations that drive direct mail campaigns. This is due to the failures in evaluation where focus is primarily on aspects that marketers consider important notably maximizing positive client responses and cost of attaining positive responses. Using a data mining tool that offers these capabilities improves the speed and accuracy of data mining processes and is thus an important consideration when choosing data mining applications.

Modeling data in data mining is complicated by lack of adequate knowledge on the nature of the modeling medium. The nature of the modeling medium is often known midway the modeling process thus the need for an iterative process (Shmueli, Patel and Bruce, 2010). This is an observation that has been noted by several researchers in data mining who concede that the data modeling process is circular rather than linear. The circular process also shows that once data mining has been instigated within an organization it should become part of their operation. This may require in-house data mining experts; a requirement that many small businesses cannot meet. Additionally, the circular nature of data mining has yet to dawn on small businesses. Most small businesses that carry out data mining take a research or project approach with clear deadlines.

A study done on the methodology for evaluating and selecting data mining software done in Northern Arizona University by Ph.D holders Ken Collier and Bernard Carey and fellow colleagues from the Center for Data Insight, Donald Sautter and Curt Marjaniemi gives details of the criteria to be used when selecting data mining tools. The four criteria mentioned are computational performance criteria, functionality criteria, usability criteria, ancillary task support criteria. The following tables taken from the same study show the details of each criteria.

Criteria	Description
Platform Variety	Does the software run on a wide-variety of computer platforms? More importantly, does it run on typical business user platforms?
Software Architecture	Does the software use client-server architecture or a stand-alone architecture? Does the user have a choice of architectures?
Heterogeneous Data Access	How well does the software interface with a variety of data sources (RDBMS, ODBC, CORBA, etc)? Does it require any auxiliary software to do so? Is the interface seamless?
Data Size	How well does the software scale to large data sets? Is performance linear or exponential?
Efficiency	Does the software produce results in a reasonable amount of time relative to the data size, the limitations of the algorithm, and other variables?
Interoperability	Does the tool interface with other KDD support tools easily? If so, does it use a standard architecture such as CORBA or some other proprietary API?
Robustness	Does the tool run consistently without crashing? If the tool cannot handle a data mining analysis, does it fail early or when the analysis appears to be nearly complete? Does the tool require monitoring and intervention or can it be left to run on its own?

Table1: Computational Performance Criteria

Criteria	Description
Algorithmic Variety	Does the software provide an adequate variety of mining techniques and algorithms including neural networks, rule induction, decision trees, clustering, etc.?
Prescribed Methodology	Does the software aid the user by presenting a sound, step-by-step mining methodology to help avoid spurious results?
Model Validation	Does the tool support model validation in addition to model creation? Does the tool encourage validation as part of the methodology?
Data Type Flexibility	Does the implementation of the supported algorithms handle a wide-variety of data types, continuous data without binning, etc.?
Algorithm Modifiability	Does the user have the ability to modify and fine-tune the modeling algorithms?
Data Sampling	Does the tool allow random sampling of data for predictive modeling?
Reporting	Are the results of a mining analysis reported in a variety of ways? Does the tool provide summary results as well as detailed results? Does the tool select actual data records that fit a target profile?
Model Exporting	After a model is validated does the tool provide a variety of ways to export the tool for ongoing use (e.g., C program, SQL, etc.)?

Table 2: Functionality Criteria

Criteria	Description
User Interface	Is the user interface easy to navigate and uncomplicated? Does the interface present results in a meaningful way?
Learning Curve	Is the tool easy to learn? Is the tool easy to use correctly?
User Types	Is the tool designed for beginning, intermediate, advanced users or a combination of user types? How well suited is the tool for its target user type? How easy is the tool for analysts to use? How easy is the tool for business (end) users to use?
Data Visualization	How well does the tool present the data? How well does the tool present the modeling results? Are there a variety of graphical methods used to communicate information?
Error Reporting	How meaningful is the error reporting? How well do error messages help the user debug problems? How well does the tool accommodate errors or spurious model building?
Action History	Does the tool maintain a history of actions taken in the mining process? Can the user modify parts of this history and re-execute the script?
Domain Variety	Can the tool be used in a variety of different industries to help solve a variety of different kinds of business problems? How well does the tool focus on one problem domain? How well does it focus on a variety of domains?

Table 3: Usability Criteria

Criteria	Description
Data Cleansing	How well does the tool allow the user to modify spurious values in the data set or perform other data cleansing operations?
Value Substitution	Does the tool allow global substitution of one data value with another (e.g., replacing 'M' or 'F' with 1 or 0 for uniformity)?
Data Filtering	Does the tool allow the selection of subsets of the data based on user-defined selection criteria?
Binning	Does the tool allow the binning of continuous data to improve modeling efficiency? Does the tool require continuous data to be binned or is this decision left to user discretion?
Deriving Attributes	Does the tool allow the creation of derived attributes based on the inherent attributes? Is there a wide-variety of methods available for deriving attributes (e.g. statistical functions, mathematical functions, boolean functions, etc.)?
Randomization	Does the tool allow randomization of data prior to model building? How effective is the randomization? How efficient is the randomization?
Record Deletion	Does the tool allow the deletion of entire records that may be incomplete or may bias the modeling results in some way? Does the tool allow the deletion of records from entire segments of the population? If so, does the tool allow these records to be easily reintroduced later if necessary?
Handling Blanks	Does the tool handle blanks well? Does the tool allow blanks to be substituted with a variety of derived values (e.g., mean, median, etc.)? Does the tool allow blanks to be substituted with a user-defined value? If so, can this be done globally as well as value-by-value?
Metadata Manipulation	Does the tool present the user with data descriptions, types, categorical codes, formulae for deriving attributes, etc.? If so, does the tool allow the user to manipulate this metadata?
Result Feedback	Does the tool allow the results from a mining analysis to be fed back into another analysis for further model building?

Table 4: Ancillary task support criteria

According to Eric A. king, Many companies make the mistake of purchasing their data mining tools by collecting product literature from vendors then inviting the vendors with the acceptable prices to a visit on site, gaining some useful information from that vendor and then purchasing the data mining tool from that vendor. The usual steps after that according to him would be putting some data in the tool and expecting magic results and being disappointed that the results that came out weren't the same as what were expected.

According to King, for a data mining tool to be purchased the right way, five steps would help in this process (King, 2005):

- Basic training on data mining would be very helpful for a company as most of the available software are fairly simple but require some knowledge in data mining and its techniques. This does not in any way mean that the company needs to invest vast amounts of money to train the employees on the software. In a small sized company, one or two employees can get the basic training and in turn pass it on to the other employees. This is much cheaper than getting extensive training on certain software and gives a bigger chance of adopting different technologies and applications.

- Assessment of the available resources, the knowledge acquired from the training, the level of expertise in the company, and the results that are to be attained from purchasing the tool should be clearly identified. The company needs to know exactly what kind of information is needed and who will use the tools before it purchases the software. The evaluation of the software discussed in this paper are mostly related to this step.

- Having a strategy for data mining is important. Having clear objectives and a clear plan would greatly help in the selection of the right tool and its successful implementation. As in any project, a clear strategy helps ensure that no steps are overlooked and everything can go smoothly with backup plans.

- Implementation should be an easy part of the process after the above steps. It is usually a straight forward task, but even in the case of it being slightly complicated, an IT knowledgeable employee can do the implementation as it would be a one time job.

- Finally, just like any successful project, the work does not end with the implementation. Monitoring the effects of the applied tools on the business must be assessed to ensure that all the benefits that could be realized from such a project are attained. Finding problems in this stage is not entirely a negative thing as it means the monitoring is working and a solution can be found.

By applying the 5 above steps, a business could be on the right path to selecting the right data mining tools.

Further research did not show any studies that attempted to find specific criteria for selecting data mining tools for small businesses or the real estate industry specifically. This can be attributed to the assumption that the criteria that is used in most cases is the general criteria that has been set to work for all industries and businesses.

2.5. Use of Data Mining in the Real Estate Industryand Dubai

Studies on data mining specifically for real estate is scarce and has only started being explored in the previous few years. One such study done in 2007 studies tenant behavior related to late payments by using data mining to

predict the likelihood that an existing commercial tenant will have at least one late payment within the next six months (Gschwind, 2007). The study follows CRISP-DM and the use of SAS Enterprise Miner as a data minng application to evaluate different models and to see if data mining does indeed give a business an edge which it was found to do. For their specific data that included tenant, accounts receivables and government data, neural networks worked best. However, that cannot be generalized for a real estate company as data found in each organization is different and even if it is similar, the quality might still deffer which might lead to a better result from a different model.

Another study done in December of 2010 by Carlos del Cacho on the real estate industry was conducted to compare pricing models in Madrid, Spain related to the valuation of houses in the city. There were several challenges in data mining because of the presence of different market segments. However the study found that M5 model trees constantly performed better on the data used for this purpose over linear regression and neural networks (del Cacho, 2010).

Further research showed data mining applications specifically created for real estate businesses that show the state of the market and the prediction for the coming months but for specific regions only such as UrbanDigs and StreetEasy for Manhattan, New York. No such industry specific tools were found for Dubai however some companies that provide data mining services were found in Dubai such as 4Fronts Consultants. These kinds of services are usually too expensive for small businesses to utilize.

2.6. Available Data Mining Applications

There are hundreds of data mining tools available. This section mentions briefly a few of these tools. R is a common statistical analysis and data mining tool. R is used by nearly 43% of all data miners and can be used with other leading data mining applications (Chambers, 2008). R provides a range of statistical and graphical techniques that allow for both linear and non-linear modeling and data mining. R is extensible and has an active community of developers who have introduced different package to enhance the functionality of R. The application is highly extensible and can be integrated to other open source data mining application such as Weka (Chambers, 2008). R unlike a number of statistical and data mining applications can produce publication quality graphs which improves its usefulness in data mining processes aimed at aiding executive decision making.

The Weka workbench (currently Pentaho) is a machine learning application that has been developed for professionals in different areas rather than machine learning experts. Domain specialists can effectively utilize Weka through the use of a graphical user interface (Holmes, Donkin andWitten, 2011). Weka offers easy access to a variety of machine learning technique via an intuitive and interactive interface. Additionally, Weka incorporates preprocessing and post processing tools known to be essential when working with real world data (Holmes, Donkin andWitten, 2011).

RapidMiner is the most commonly used open source data mining software. RapidMiner which was formerly known as YALE is an environment that supports machine learning, predictive analysis, data and text mining and business analytics (Rapid-I, 2011). The software is owned by Rapid-I and can be downloaded and used under an A-GPL open source license. RapidMiner has two major editions namely the community and the enterprise edition. The enterprise edition is for sale and offer more functionalities than the community edition.

Available commercial data mining applications include NeuroXL which has five different versions which include NeuroXL Predictor, NeuroXL Cluserizer, OLSOFT Neural Network Library, Predictor CRND, and Predictor CRND and OLSOFT Neural Network Library, KnowledgeMiner and DMSK; all of which are not open source applications but can be purchased for a low price.

Other recent versions of statistical analysis and data mining applications like SPSS have additional extensions for data mining for instance clementine and Amos. The main challenge that small businesses face in using commercial data mining software is the associated high cost of acquisition. SAS Enterprise Miner can cost up to $9000 per year for a single user license. When compared to open source software, SQL server data miner, Oracle Suite, STATISTICA data miner and SAS enterprise miner are quite expensive and offer more capabilities especially for large and multiple databases.

Chapter Three: Methodology

The research methodology is important as it determines the validity and reliability of the findings. This chapter presents a discussion of the research philosophy, research design and the research approach used to determine the benefits of data mining for small businesses and the criteria for selecting a tool for a real estate small business in Dubai. Other issues discussed in the chapter include ethical measures and the limitations of the research approach.

3.1. Research Philosophy

Constructivism, which is to generate knowledge and meaning from an interaction between experiences and ideas, is the main research philosophy that was adopted in this study. It is important to note that the study seeks to determine the best data mining tool for small businesses in the real estate industry in Dubai and the challenges and benefits associated with the use of data mining applications. The suitability of any application can be assessed from multiple perspectives thus the need for a criterion. Constructivism is a post-positivist research philosophy that acknowledges the need of adopting multiple perspectives in analyzing research problems (Kothari, 2008). Under constructivism, knowledge is assumed to be constructed via interaction between the researcher and the participants. Additionally, the research philosophy ascribes to the view that there are multiple interpretations for data. The use of constructivism in the research ensures that existing knowledge and knowledge specific to the small businesses in the Dubai real estate industry are all captured and used in addressing the research problem. Thus the choice of constructivism was due to its suitability to the research problem as knowledge of existing data mining applications and criteria for choice are critical to meeting the research objectives.

3.2. Research Design

A mostly qualitative research design was used in the study. The study focused on the collection of qualitative data on the performance of selected data mining applications in a database owned by a small business in Dubai’s real estate industry as time limitations restricted the ability to gather enough quantitative data for this study. Quantitative data for the performance of the applications would be very useful for the study as well as for the company, however, that kind of data requires vasts amounts of time dedicated by the resources of the company for this purpose which was not possible in the time period provided for this study as it fell during the summer season

during which the resources of the compary are already reduced from their normal numbers.The qualitative dimension of the study focused on determining the challenges and benefits associated with the use of data mining applications by small businesses in Dubai's real estate industry. A qualitative research design is thus suited for the study due to time constraints as well as the type of objectives to achieve. Additionally, the qualitative study is important in the developments of a criteria that was used in assessing the best tools for data mining for small businesses in the real estate industry.

3.3. Research Approach

The research approach is vital in minimizing biases and determines the overall quality of the data collected. Due to the importance of the research approach, various measures were included to ensure that the data collected was accurate. A proper definition of the population and the sample is the first step to improving the accuracy and reliability of the research exercise (Jackson, 2008). The next step in the study involved the collection of secondary data. This is a process that was aimed at assisting in developing suitable criteria for analyzing the data mining software. The last phase involved the collection of primary data from the participants. Emphasis in this phase was on the challenges that they face in using data mining software and the associated benefits. It is important to note that this step comes after a structured and formal process aimed at determining small businesses that have and use data mining software in Dubai's real estate industry. Permission was sought from the small businesses that met this criterion. In seeking permission, the aim and the methods to be used were relayed to the company. One of the available standards of applying a data mining project, CRISP-DM, was used to ensure the reliability of the process undertaken.

3.3.1. Sample and Population

The population in this study is defined as small businesses in the Dubai real estate industry that use data mining. The sample is however a single organization and at least five employees from the small business. An important consideration in the choice of the participants in the study is that the organization must implement and run a data mining application. Convenience sampling was used because the population is rather small and few organizations were expected to grant access to their databases. Also, the use of

such tools in comparison to normal real estate management software was found to be rare in this part of the world. The five participants in the study were interviewed to determine their views on the challenges and benefits associated with the use of data mining applications. In choosing the participants, different departments and levels of the small business were represented. As for the data mining tools selected for this study, five tools based on the factors that were deemed to be of high importance for the company after discussions with employees of the company

3.3.2. Data Collection

Secondary data collection involved the extensive use of online electronic databases. The IEEE database, EBSCO databases and other electronic databases were used as sources of peer reviewed journal articles. To ensure relevance of the sources used, the study set the date of retrieval to be after 2005 for most of the sources. The keywords data mining, pattern detection, small businesses and real estate industry and Dubai were used to search for articles. The abstract for each article was skimmed through to determine the articles that were more relevant to the research problems. Books were also used to determine the available information on the criteria used in choosing data mining applications and assessing their performance.

Primary data collection involved the use of unstructured in-depth discussions. The focus of the discussions was determining the benefits the real estate company was hoping to aquire from the data mining application as well as the challenges associated with the use of these applications by small businesses. In-depth unstructured interview is used as the mode of data collection due to the flexibility it offers researchers in seeking clarification. The last phase in data collection involved the use and assessment of different data mining applications on the small business's database. The different data mining applications were used and their performance assessed in terms of quality of their output, speed and usability of different utilities, the user interface provided and the level of support available. To improve the validity and reliability of the experiments carried out, each data mining application was used or tested at least three times and a detailed record of the steps tried and the issues faced was kept. The applications that posed problems were tried on different machines to ensure that the problem was in fact related to the application itself. It is important to note that at least one employee from the company with knowledge of the databses being tested was always present during the process in order to ensure the validity of the results and their meanings and provide input on the ease of use and graphical user interface.

3.3.3. Data Analysis

Data analysis involved critical analysis of literature reviewed to develop a criterion for assessing different data mining applications. Other issues of concern in the critical examination of the literature were the benefits and challenges associated with the use of data mining software by small businesses in the real estate industry. A critical examination of the data collected from the in-depth interviews and discussions and integration and comparison with findings from the literature review was also important in the data analysis. The last phase in the analysis involved the use of measures of central tendency in accessing the overall performance of different data mining tools. Additionally, even though there is a usual criteria that is used for the evaluation of such tools, several new criterias were used in this study as this study is directly related to a company and the way that the applications needed to be evaluated for their specific use seemed to need a different approach from the norm.

3.4. Ethical Considerations

The main ethical consideration in the study was seeking permission from the small business and the participants in the interviews so as to ensure informed consent. The study involved active steps aimed at ensuring that the organizations involved in the study was consenting and individual participants understood the aim and the method used in the study beforehand. The study did not involve the use of secrecy, force or coercion.

The other ethical consideration is the privacy and confidentiality of data stored in the databases. The study did not involve copying data on business performance and output of the experiments was kept safe by employees attached to the study. This is a measure that is aimed at ensuring that organizational data on performance and emerging patterns are kept confidential and that the study does not jeopardize the integrity and safety of organizational data.

3.5. Limitations

The main limitation of the research approach is that it is highly dependent on the researcher and the capability of the business. The dependence increases the risk of researcher bias which is a key issue. The use of critical analysis in the research requires considerable objectivity on the part

of the researcher. However, past experiences and preferences may affect researcher objectivity thereby resulting in biases. An external entity (the employees involved in the research) was included in the research to ensure that the analysis and findings are objective and supported by the study. The lack of full fledged IT team left the IT problems up to the reacher and one IT employee to figure out. Any IT problems that could not be resolved by the two and by the support from the application teams were mentioned as weaknesses of an application and a disadvantage to a small business.

Furthermore, the time restriction posed a big limitation. As a small business has limited resources who are always kept busy by given more than one role in the company to cust costs, it was difficult to have the time with the employees to continuously test out the applications in their presence. Also, as this study was only supposed to be completed in a certain time period and a big chunk of that time was spent on literature review and secondary data, there was not efficient time to try out all the options on each application or resolve issues faced. More time would increase the level of expertise with such applications and a better understanding and a bigger chance of resolving their issues would be possible.

Chapter Four: Findings, Results & Data Analysis

The company involved in this study is a real estate company based in Dubai with 2 offices in different locations in dubai and is made up of 18 employees many of which perform several roles in addition to their primary ones. The only IT is the same person that does most of the marketing tasks, the rental manager is the one that comes up with the marketing ideas for implementation. The company secretary is also part of the rental team. Therefore, every employee of the company is extremely busy all the time. Even though the number of employees is small, there are big projects for the company and clients being managed by the company. As the company does everything from designing, constructing, renting, maintaining and managing these properties, as well as brokering for other properties, selling and buying existing and new project, deploying a data mining tool might take up some valuable time away from the employees and increases their work load for a while, but the same tool should make their work more efficient and yielding better results in the future if used correctly. The following sections mention the benefits of data mining specifically in the real estate sector in dubai, the criteria for evaluating the data mining tool which was used for this sector and can be used for other small businesses, and finally the findings of the evaluation of the five selected tools for testing done in cooperation with the employees.

4.1. Benefits of data mining on small businesses

Discussing with the management of the company the benefits were expected to be obtained from applying data mining on all parts of the business resulted in the following:

- Better understanding of the clients and the target market for each season (by detecting patterns in the amounts of requests for certain types of units during certain parts of the year)
- More efficent advertising for the clients (by detecting which marketing means result in more paying customers in order to concentrate on those means instead of spending time on using all means some of which do not lead to any profit)
- Reduction of expenses (by looking at the expenses trends in the company to see if there is a chance of reduction in that area)

- Getting a better or further insight into the financials of the company (detecting unusual financial activities that are either harmful or beneficial but were unknown before the mining of the data)
- More efficient procurement of company supplies (detecting purchasing patterns to make the procurement process more efficient and cut procurement costs if possible)
- Increasing customer satisfaction (by detecting patterns in thhe issues faced by clients in order to anticipate them better and resolve them faster)

Small businesses have the option of outsourcing their data mining processes to data management companies. In Dubai, there are a number of companies that offer data mining and consultation services. Dataline FZ, GeoBird Management, Interactive limited, Seabird Exploration, SEVdotcom are examples of companies in Dubai that offer data mining services. When the idea was discussed with the employees of the company, all responded negatively about it some stated the reason to be that those services are usually too expensive and the company has no budget for something like that, others stated that they need a tool that is available for them all the time without restrictions in order to use it in every part of the business.

Real estate is to a large degree concerned with the landscape. In addition to support for classification, clustering, association and numeric predictions, real estate firms often seek data mining applications that support spatial data mining. These applications are generally characterized by functionalities that support data visualization and algorithms aimed at findings patterns in geography. Most open source software has additional packages that can be installed to improve the applications' support for spatial data mining. When this was discussed with the employees, again, the response was negative. In Dubai this is not very popular and doesn't seem to be as useful for the company as it might be for other companies since the company focuses on management and renting rather than buying and selling which is what that software is usually associated with. The benefits anticipated from the data mining tool were not related to this option at the moment but could be in the future.

An important area of application of data mining is customer relationship management. Data mining software can detect patterns in customer behavior for instance correlations between the products that they buy that cannot be detected through casual observation. This information can help small businesses in the real estate industry to offer products that meet specific customer demands. For instance knowledge showing

that customer seeking a particular house design like a specific grass carpet can be used to improve the house and compound design to meet the requirements. This ensures that the products offered by the small businesses are close to the client expectations. This falls in line with the benefits expected for the company being worked with and the employees working in the departments related to this area were all in agreement about this point.

When the idea that better knowledge of the customer and the market as well as expenses and costs could lead to better advantage and having more chance of competing with the bigger names in the business such as Better Homes, most employees were hopeful about the matter. Some were less optimistic and stated that it would take years to get to the same level as Better Homes or similar companies not only because they already dominate big chunks of the market but also because they have the finances to finance a data mining project as well or even better.

In spite of all the benefits that the management and the employees of the company were hoping to obtain, concerns were discussed. Some of the concerns were the time needed to evaluate and implement a data mining tool, the difficulty that would be faced when implementing the tool and working with it, and the possibility that it would not work out correctly or as expected.

4.2. Standards Followed and Tools Selected

To select the data mining applications to evaluate for this study, CRISP-DM was generally used. First, the business objectives were determined, which in this case is to increase profit by acquiring more customers by targeting the right markets at the right time, by identifying the most useful advertising means to eliminate spending on useless tools and concentrating on more profitable ones as well as reducing expenses. To achieve this, databases from different departments were requested. The marketing department provided a database of all the callers and viewers that detailed the date of call or visist, the requested type of unit, the means of which they have heared about the company from. The rental department provided a simplified database of the rentals tha happened during each month for 2 years. The finance department provided several reports on the expenses of the company. As there was limited time and the data from the finance department needed a lot of work to be used effectively in a data mining tool

after studying the data, only the databases from the marketing department and the rental department were used. And since the rental department dataset was already simplified, it was used the most. The rental department data was the cleanest in the sense that it was simple, straight to the point and could easily be converted to numerical figures if required for certain data models.

After understanding and preparing the data to a certain extent, the next step would have been to build a model or test it. That would have included determining which data mining method or model was most suitable. Since there was different datasets for the same company and since the tool selected is supposed to help in all departments rather than only one part of the business, selecting data mining applications (such as weka, rapidminer, …etc) to evaluate rather than a technique (such as neural networks, decision trees, …etc) was done. Since the traditional way of selecting a data mining application as found by the literature review was to evaluate the models, the speed and other technical aspects, this could not be done in this case as there would most likely be more than one method to work best on each of the datasets. However, a small business does not have the means to purchase more than one tool or spend more time on training for several tools, therefore the five applications that were selected for evaluation were selected on the following criteria: variety of methods included in the tool, popularity, and price. RapidMiner and Weka were selected because they are currently popular amongst data miners, are free, and have a diversity of models. DMSK was selected for its cheap price as a commercial tool and to evaluate against free open source applications such as RapidMiner and Weka. NeuroXL was selected for its specialization to one method in order to see if this specialization yields better results. Knowledge Miner was selected because it only runs on one platform which is different from all the other previously mentioned applications. By selecting these tools for evaluation, a range of different applications with different aspects and different characteristics yet all affordable by a small business that does not have the means to spend big amounts on such tools. Other commercial tools that are popular but too expensive for a small business such as SAS Enterprise Miner and SPSS and Oracle Analyzer were not selected for evaluation because in reality even if they do perform better and have additional extensions they would not be purchased by a small business because of their high prices.

4.3. Criteria Used for Dubai's Small Real Estate Business

The criteria for selecting a data mining tool as found in the literature review were general to all sectors and industries, but that does not necessarily mean that they work as well for small businesses as they do for larger ones. A lot of the research conducted was done by either institutes or organizations that had the time and the money to spend on testing commercial tools developed by big brand names. That is not to say that the mentioned criteria is not valid and highly important for small businesses, however, almost all of them require more expertise in the field of data mining than usually found in small businesses and especially the one this study is conducted in cooperation with. Therefore other criteria that were deemed fit after discussion with the different departments of the company, were given more focus and weight in the evaluation. One of the most influencial criteria on the selection of an application is its ease of use and user interface which is one of the criterias mentioned in the literature review but is one of the less technical evaluation criteria and therefore was given priority. If the application is found to be too difficult to use or even too difficult to install, small businsses are likely to go with another option. Another criteria with a big weight on decision is the price of the software, there are many applications that might have a good user interface, good training programs, and good support but are too expensive to be even considered by a small businesses. Support and popularity are other criteria to consider, an application that has a lot of users is more likely to have more user guides, user discussions and information especially online where it is accessable for everyone. A more popular application usually has better support by the development team itself as well as users than a less popular tool.

While these are all very important aspects to consider for small businesses when selecting a data mining tool, the evaluation of the performance of the applications in general as well as specifically on the businesses's data is the most important criteria to consider since a cheap, easy to use application with incorrect results is a lot more costly for a business than a more expensive but accurate option. Table 5 summerizes the criteria focused on in this study.

Criteria	Desciption

Platform	runs on windows, mac, linux, …etc
Methodology Available	available models to use: neural networks, decision tress, clustering,…etc and their accuracy
Price	how much does the application cost
Popularity	how popular is the tool in the data mining world
Evaluation	how accessable are the evaluation copies and how useful they are
Support	how much support is available for the tool

Table 5: Criteria used for this study

In the following sections, a comparison of the five selected tools in above mentioned criteria is conducted in cooperation with the employees of the company in order to determine which applications should be adopted or studied more before implementing in the company.

4.3.1. Methodology and Platform

4.3.1.1. RapidMiner

RapidMiner can be ported to an R interface and can work on multiple databases. RapidMiner offers data loading and transformation, visualization, modeling, visualization and deployment utilities (Rapid-I, 2011). The application is written in Java programming language which allows it to run on multiple platforms. Moreover, RapidMiner integrates learning schemes and attribute evaluators from Weka and statistical schemes from R (Rapid-I, 2011). This implies that RapidMiner has comprehensive statistical and data mining utilities and can be used in various areas.

RapidMiner functions can also be called from programs written in other languages and systems for instance Perl. However, the major strength of RapidMiner is the broad array of data mining algorithms ranging from decision trees to self-organizations maps it offers. It includes all kinds of data mining methods from K-NN, Bayesian classification, neural networks, to linear regression and many more.

Additionally, plugins can be used to offer extra functionalities for instance text analysis and mining. It has plug-in as well as extension mechanisms and knowledge discovery processes that have been modeled as operator trees. The program has an internal XML module that maintains a standardized data interchange format which leads to less data conversion for the purpose of processing. In addition it has a multilayer data viwer which has efficient as well as transparent data handling.

RapidMiner also has a plotting facility which offers models and data visualization schemes (Clausen, 2009). The application can be used for text mining, feature engineering, tracking drifting ideas and distributed data mining.

4.3.1.2. Weka

Weka has systems for machine learning experimentation intended for researchers focusing on the comparative efficiency of algorithms. Weka incorporates *consultant* which is an expert system that allow domain experts to choose learning algorithms that meet their needs. *Consultant* works by assuming that a machine learning algorithm that can directly be applied to problems exist in every domain. This is a development that is a result of the realization that domain experts need environments where they can manipulate data and run experiments by themselves. Clustering (by the use of K-Mean, Expectation Maximization, and Cobweb), classification, and visualization are all functionalities of Weka.

The Java based applications that can run on different platforms focuses on ease of use of the application functionalities by end users in different domains rather than machine learning experts (Holmes, Donkin andWitten, 2011). The system levels implementation details of the application including the input formats are hidden from the user. This makes it a suitable application for small businesses that cannot access the input of data mining specialists. Furthermore, Weka uses Java database connectivity to allow access to SQL databases and can also process results from database queries (Holmes, Donkin andWitten, 2011). Though Weka does not support multi-relational data mining, there are collections of software that can be used to transform multiple databases to a single table that can be processed using the application.

4.3.1.3. NeuroXL

Just as the name implies, NeuroXL uses neural networks for data mining and pattern recognition. This software plugs right into MS Excel and has both a classifier and a predictor for data. It hides the complexity that usually comes with neural networks and produces results in forms of graphs and simplified data. This software forecasts and estimate problems for all kinds of fields. NeuroXL has increased accuracy and precision to various tasks that include sales forecasting, cluster analysis, stock price

prediction and sports prediction among other predictions in the business and entertainment fields.

Out of the five available versions, NeuroXL predictor and clusterizer are the most widely used versions among the five. The predictor is a neural network forecasting software which accurately resolves estimation and forecasting problems. It is developed in order to assist the experts to solve real world forecasting problems. The new version of the application includes 5 transmission functions which are threshold, zero-based log-sigmoid, log-sigmoid, bipolar sigmoid and hyperbolic tangent. As for NeuroXL Clusterizer, in finance, business and science analysts are regularly faced with the task of cluster analysis on the basis of measured and historical data. The neural networks technology has proven to solve such complex classifications. However, despite the fact that neural networks are effective they are not widely applicable in cluster analysis since they are complex (Veart, 2008). Nevertheless, the NeuroXL clusterizer eliminates the complexity of the advanced neural network based methods by taking the advantage of the analysts experience in using excel spreadsheet application. As no evaluation from a valid source was found for this application, mostly because of its newness, there are no figures on its accuracy but the reviews about it from software downloading sites proved to be positive ones.

4.3.1.4. Knowledge Miner

Knowledge Miner has many functions, which include classification, modeling, forecasting, clustering and sequential patterns. The techniques used for these functions are: Self-organizing Networks of Active Neurons (SONAN) which is a kind of neural networks, Analog Complexing (AC), Fuzzy Rule Induction (Fuzzy) (KnowledgeMiner.com). According to the "Evalaution of Fourteeen Desktop Data Mining Tools" study, Knowledge Miner has an above average of 82.75% accuracy in datasets in different fields ranging from medical to business (King & Elder IV, 1998). In the same study, the software was evaluated on 20 criteria but was given a score of average on several of them, and needs improvement and poor on most of them. However, this study was done in 1998 and there have been several newer versions of the application in the past few years. The fact that this application is growing in popularity compared to other bigger applications implies that the scoring would be much higher for this application if it was done today. There are now two new different editions of KnowledgeMiner for Excel, the Silver edition for solving smaller and

simpler modeling problems, and our Gold edition for high-performance, high-end professional knowledge mining needs.

4.3.1.5. DMSK

As for DMSK, The data miner software package has 4 main applications that have different functions:

- the rule induction kit for texts (RIKTEXT) which learns rules from document collections that are simple and are generated from the data. This application does not include the complexity that usually goes hand in hand with numerical data mining and modeling.

- The text-miner software kit (TMSK) that predicts texts. This is useful for XML documents to enable their processing. It uses Naïve Bayes, linear models, K-nearest neighbors, and document matching and clustering.

- The Rule Induction Kit (RIK) which discovers decision rules that are highly compact from data. Just like RIKTEXT, the rules are simple and highly predictive, that is, the main goal is to find the best set of tules for prediction and classification to keep errors to a minimum. Such applications are useful in the medical field.

- Enterprise Data-Miner is used big data by using classical and more computationally expensive methods such as data reduction or sampling, prediction, and data preparation.

Just like Knowledge Miner, DMSK had an above average accuracy in the low 80s (81.25), and got poor evaluations on most of the 20 criteria in the evaluation of 14 tools study (King & Elder IV, 1998). However, it has an excellent and average scoring in allowing data transformation, testing options, linking to other applications, model adjusting flexibility and exporting data.

4.3.2. Popularity of the Applications

4.2.2.1. RapidMiner

In May 2011, an annual poll done by KDnuggets, a data mining community with more than 1,100 participants showed that Rapidminer was the

most used data mining application in 2011 with 27.7% of the votes which is a drop from the 37.8% in 2010. R was the second most used with 23.3%.

4.3.2.2. Weka

In the same KDNuggets poll in which RapidMiner came first, Weka came in at seventh place with 11.8% of the votes. The number might not seem very encouraging, however compared to hundreds of other data mining applications, the placement is a good one.

4.3.2.3. NeuroXL

Neuroxl is relatively new to the data mining market compared to both knowledge miner and DMSK. However, even though it is newer, it seems to be much more popular the DMSK as there are many returns when a search on NeuroXL is done. At a random software purchasing website, DownloadPipe.com, there were hundreds of downloads of predictor application alone. The neuroxl package seems to be getting increased popularity, but no specific companies have been mentioned as users of the application.

A current look into what users were saying about the application showed that many did not find the results of the application satisfactory. However, some users argued for the application by stating that the reason the results were not accurate is because of the incorrect selection and input of parameters.

4.3.2.4. Knowledge Miner

NASA, Boeing, MIT, Columbia, Notre Dame, Mobil Oil, Pfizer, Merck, Dean & Company. According to Ware Adams from a strategy consulting firm in the USA called Dean & Company, "Knowledge Miner is the only product that I have found that makes it easy to try non-standard equation formats on a data set. Many standard regression tools are as easy, but they limit you to a small set of potential relationships. Knowledge Miner combines spreadsheet-like set up with an algorithm that doesn't "over fit" the model. Also, the output is in a readily usable format (e.g. not C++ code) (KnowledgeMiner.com)."

Furthermore, from the Ukraine, Gregory Ivakhnenko, a specialist at the National Institute for Strategic Studies of the Ukraine states that "It is really easy-to-use tool. It helps me to find laws which acts in my object directly from the data sample only. (KnowledgeMiner.com)"

4.3.2.5. DMSK

While DMSK was as popular and had approximately the same results as Knowledge Miner at the end of the 1990s, DMSK did not have the same success. The software that was once written about, evaluated, and reviewed in different languages seems to have lost popularity and failed to achieve the same advances as Knowledge Miner. The software is still sold to businesses but not as a single user license to individuals anymore and might be used still but no data could be found on such users or customers.

4.3.3. Prices

RapidMiner	free
Weka	free
DMSK	$75
NeuroXL	$200
Knowledge Miner (Silver Edition)	$300
Knowledge Miner (Gold Edition)	$1500

Table 6: Prices of Data Mining Tools

4.3.4. Interface and Ease of Use

4.3.4.1. RapidMiner

At a first glance, rapidminer looks like an easy and straight to the point application. The graphical user interface provides an interactive platform for user to navigate through the program. Most parts of the application has wizards that simply the process for the user. Everything on the application has instructions and information popping up by simply having the mouse pointer pointed at it. Rapidminer has the best consideration for users from all the evaluated software. Besides controlling the program through the graphical user interface the program can also be controlled through the command line client which gives more experienced users more options and control. Despite all of that, simply loading the data into the program posed a problem and gave constant errors. Understanding how to run any data mining process on this application though was a challenge and reading the results of any successfully run process was an even bigger challenge.

4.3.4.2. Weka

While this tool had fairly average looking user interface, it was much less complicated than the other tools, in addition to that, the results generated upon testing were the easiest to comprehend from all the tools. Also the application allowed easy importing of data even though it did not accept Microsoft access or excel formats, it did accept the CSV format which was easy to convert into from an excel format. The data was slightly changed for the purpose of getting more accurate results but that was no real disadvantage as the same had to be done on different parts of the dataset in most of the other tools. The major disadvantage of this tool however was its inability to run on normal computers with new specifications and processors. A not enough memory error kept coming up as soon as the application was run until it was installed and tried on a much more powerful computer which has specifications that are not usually found on normal desktop computers in any business either small or large. And even on that computer, the same error came up and the application had to be re-installed before it worked seamlessly again. Even though the application could be installed on a powerful server with employee client computers connected to it to utilize the application, it is less convenient than a tool that can be easily installed on all employee computers that will be using this tool. Having to set up a server means also having to set up a proper network in case one is not available which could be costly especially if there are more than one locations for the company, which is not unusual in a small business.

4.3.4.3. NeuroXL

NeuroXL gives the impression of a simple interface but for a non-IT or more specifically a non-AI knowledgeable person, the software is not as easy to use as it is being advertised. The help section of the software states: "NeuroXL Predictor requires no prior knowledge of neural networks, and is extremely easy-to-use. Being integrated into Microsoft Excel, PredictorXL eliminates the need to export data and import the results, and leverages your existing knowledge of Excel [NeuroXL.com]."
Even though the examples section is helpful with good examples, it is still unclear for a person with no prior knowledge of how to select the correct parameters to use this program. Selecting the wrong parameters could lead to very different results which could lead a company in the wrong direction. However, most users of clustering software have raised concern on the speed at which the applications run on and the

difficulty in use as well as the specific formats of data required in order to be processed. However, the NeuroXL Clusterizer applies the neural network technology which is fast enough in speed and easy to use thereby reducing complexity. A given cluster analysis can be performed within minutes. A user is only required to specify input as well as output references and then perform some few mouse clicks and the cluster will be achieved.

As it is a plug-in into excel directly, it does not require an extra application to be open which makes for a very practical function for the users of MS Excel. However, in the case that the business has its own big database other than excel, this tool would not be very practical as data would have to be copied into Excel and that might not work due to the size of the data or other technical details. However, for most small businesses, this is not a big issue as the data is usually not very large.

The data set from the real estate company was used to test the application. After several tries that resulted in errors, a graph did come up, however, it was unclear what that graph meant according to the data as there were no indications.

This application only accepts numeric figures and therefore not any company's data can be used without conversion first or special storing. Textual data can be saved as numerical data for these purposes however, that adds more work as information has to be continually converted. Also it makes data harder to read for the employees.

One of the biggest strengths of the application though is that it can be used for all sorts of data. It is not limited to a certain category. It can be applied in finance, medicine, business, and others.

4.3.4.4. Knowledge Miner

As this application is made for MAC pc users and since the computers both personal and belonging to the companies were all windows based computers, this application could not be tested directly with the test data set for ease of use. While there is supposed to be a windows version of this application, it could not be found anywhere, even through a link from the main website. Even in the case of finding that version of the application, it would not be a stand-alone application, but one that runs on an emulator that runs MAC applications, which would use up extra resources and makes the installation and set up much more complicated compared to the other researched applications.

Also, as many small businesses do not require big databases and especially individual departments when they save their data, tend to use MS Excel. This is not viewed as a big setback though because Knowledge Miner has its own spreadsheet that data can be inputted to and therefore data could be taken from most databases and put into Knowledge Miner.

After some research, details of the user interface of the application were found, which looked much simpler than the other applications studied in this paper. Some understanding of data mining techniques are required to select some of the choices but nothing as complicated as the previously discussed tools. The options are fewer and the user interface as a whole seems less intimidating. Also, with the tutorial provided on the knowledge Miner website, it all seems fairly easy for an average user to learn with minimum self-taught training without much difficulty. However, we stress that some researching on the topics and some practice and tutorials would be needed to operate this application unlike what is mentioned on the website which states no training is required at all [KnowledgeMiner.com].

4.3.4.5. DMSK

DMSK which stands for Data Miner Software Kit is a collection of stand-alone Java application that run without having to be plugged into any other application. The interface looked relatively simple, however it was still confusing to use as it is unclear what should go in the required boxes.

The data can be analyzed in 3 different ways in this application: Data Preparation Methods, Data Reduction Methods and Prediction Methods. Once one of them is selected, further options of sub-methods will be displayed. For an average company user in a department other than IT, it is unlikely that the user will know what to select and what further parameters to input whenever one method is selected.

Documentation in the form of a book is included with the original copy of the application called “Predictive Data Mining: A Practical Guide”, which might be helpful to the user.A demo application is available on the main DMSK website There are also DMSK courses given by the same developer company, but in Australia only. However, in a study done by Michel A. King and John F. Elder IV, Ph.D., the application was evaluated as having poor learn-ability and usability (King & Elder IV, 1998).

4.3.5. Support and the Availability for Evaluation

4.3.5.1. RapidMiner

Support for rapidminer is available on the enterprise version which isn't a free version such as the community version but support for non-technical users and newcomers to data mining might be essential for small businesses. Training can also be provided for the enterprise edition which is a great addion for a small business

4.3.5.2. Weka

The help section of the program was very useful. In addition to that, the application's website is full of information and details related to the application as well as useful examples. There is detailed documentation of the application found on the website as well. There is a lot of information and discussions about this application found online. There are forums, discussion groups, questions and answers all found in big amounts for this application

4.3.5.3. NeuroXL

Support for NeuroXL was average. Three options are listed under support on the website. The first option is to send an email to the support group which was done more than once. A quick response was provided and there was an apparent eagerness to support and help the user. The second option is to talk to data mining experts through a website in which any user can sign up as a professional and charge the person for each question or a certain amount of support time; payment must be done for the expert before the support is provided so there is no guarantee that proper support would be provided for the amount charged per issue. The third option under support listed is "books" which just refers the user to books on neural networks and have no relation to neuroXL in any way. Discussion groups and other user interaction about this software is not found easily on the internet. The evaluation version is a complete version of the application and allows its use for 10 days, afterwhich the NeuroXL support team advises you to purchase the application and try it out further within 30 days as there is a 30 day money back guarantee offer that comes with the software.

4.3.5.4. Knowledge Miner

Knowledge Miner has a lot of options when it comes to support including a tutorial, free book on self-organizing data mining with a purchase, various examples with data and models. Users can contact the development team, the research team, the service

team, as well as a data mining consulting service. Additionally, there is a frequently asked questions section. Therefore, from all the applications, Knowledge Miner seems to have the most support from the application's team. However, the support is not limited to that; discussion boards for Knowledge Miner were also available and easily accessable. An evaluation version of the application is easily accessable as well.

4.3.5.5. DSMK

DMSK provides support for the software since the software sold to organizations is customized specifically for each organization. Training is also provided for the oraganizations. As the software was not purchased and there are no evaluation copies of this application, the level of support could not be evaluated properly.

Chapter Five: Discussion

At the beginning of this study, the objectives were to determine if data mining is beneficial for a small real estate business in Dubai, to determine the criteria for the evaluation of the tools for such a business, and to determine some of the best tools for this purpose. The following sections discuss each objective's result.

5.1. Benefits of data mining on a small real estate business in Dubai

Use of data mining can bolster organizational efforts to adopt knowledge oriented decision making. Information driven operations is key to success in highly competitive industries. The changes in norms and ultimately culture can help improve firms' chances of survival considering the complex and unstable operational environment. The uncertainty and shocks characteristic of the global economy are taunted as key drivers to greater appreciation of information by businesses. Use of data mining by small businesses is an important first step to ensuring that they meet the demands of the modern operational environment.

There are no doubts about the benefits that a data mining tool can supply to a business. For any business, it could retain more customers by improving customer satisfaction; it could increase business by identifying the best and most profitable clients in order to concentrate on them and reduce wasted time; it could enhance company sales and marketing practices; and it could reduce costs significantly.

The real estate industry in Dubai is competitive due to the existence of established firms and the high demand for units. Information and knowledge of the market and industry is critical in such industries. Knowledge of patterns in the market, industry and internal operations gained through the use of data mining software can be used to reengineer operations to address issues that others have yet to realize. This can give small businesses an edge over established players who on the other hand have vast working capital. For a company as small as the one being worked with to have such knowledge would give it a leading edge even against huge companies in Dubai such as Better Homes who are dominating a big chunk of the real estate market in the city. Even though positive response to this idea was not unanimous, it should be pointed out that smaller businesses have smaller datasets and therefore it is

a lot easier to manage their data and get better and accurate results from data mining than bigger businesses can.

In spite of the great demand for data mining and the clear benefits, challenges in data mining persist. Data mining is a highly time intensive endeavor that requires organizational appreciation and support for its potential benefits. Most small businesses are however owned and run by entrepreneurs who are mostly concerned with the maximization of profits and minimization of cost. Moreover, most small businesses do not plan to evaluate their strategies and operations. As a result, data mining in most small firms barely gets the support required. Another problem that small businesses face is lack of working capital and clout to attract professionals in specialist areas like data mining. However, this problem is offset by availability of comprehensive open source data mining software that can easily be used by domain specialists. The applications can be used by sales executives and operations managers thereby minimizing the role and need for specialist data miners.

However, if the data mining tool is used incorrectly, the negative effect might be greater than not investing in it at all. A data mining tool is almost dangerous in that it might lead to a wrong path if used incorrectly which would lead to business losses instead of increased profits. Simply purchasing a tool and failing in its implementation has a big negative impact on the company morally and financially. Is this a risk worth taking? As the benefits seem to overweigh the risks, I would recommend taking the risk for the company being worked with. However, the risks might be a lot bigger for other small companies especially in Dubai as there are many factors that affect the calculation of risk specific to each business.

Small businesses have the option of outsourcing their data mining processes to data management companies which is an option that could reduce the risks. However, the third party providers of data mining services encourage a project approach to data mining and reduce the likelihood of adopting data mining as part of daily organizational operations. This minimizes the gains that small businesses can make through data mining. It could also be more costly as such services from companies are not cheap.

5.2. Availability of Data Mining in Dubai

Although forms of data mining are commonly found in some of the bigger organizations in Dubai, such as Etisalat, RTA Authority and others, data mining has still not gotten popular enough to be diffrenciated from business intelligence in this

part of the world. This poses the problem of not enough support or training facilities. Data Mining Tools with the training and support options are usually present and provide those services in the countries in which they operate and rarely in the Middle East region. Therefore, the lack of training and support is a serious problem that faces companies in Dubai that want to implement data mining. The best way to deal with that issue is to select a tool that has good support online, good documentation and good tutorials online in order to be able to utilize the tool to its fullest. However, the fact that data mining is not very popular in this region gives any company that ventures into the data mining field a competitive advantage. Even just considering the idea and reading into it can have a company look at its data in a different way and have the mining of knowledge as an aim to be achieved either through human work if not by data mining tools.

5.3. Criteria for data mining tool evaluation for small businesses

The criteria that was set based on the discussions and the influencing factors of the small business are not the traditional way of going about evaluating data mining tools. Some might argue that these are not as important as the traditional criterias but from the company's point of view and possibly from many other small businesses' point of view, these are some of the most important factors to consider as long as the performance of the tools are reliable and the results are verified and deemed to have a high accuracy rate, then there is no reason to determine these criterias as unimportant as previously set general ones. Using this criteria reduces the time spent on bigger more expensive tools that get evaluated with no chance of being purchased or used by a business of this size.

5.4. Evaluation of Tools

From the previous chapter's evaluations on each category the following was deducted:

- The advantages associated with Weka include a freeware license which guarantees the availability of the software, high degree of portability due to its full implementation in Java language, a collection of pre and post processing modeling techniques and an easy to use graphical interface with high documentation and support. Also, the most important advantage found on this application is the usefulness of the results. Users with no prior knowledge of data mining could easily interpret the information displayed after using the

clustering, classification and especially the visualization functions. Its main disadvantage is its requirement for high specifications in machines that normal employee machines cannot handle.

- The strong points of RapidMiner as a data mining application are derived from its key features which include being written in Java and support for different technical aspects of presentation and layouts. The application is open source software and this reduces the overall costs associated with producing and using the software. It has a lot of additional plug ins including Weka, R and other useful add-ins. It is popular with a lot of support and easy to install on normal employee machines.

- While NeuroXL is relatively new, it is apparent that it is gaining popularity in the data mining world. It has acceptable support, good documentation, and an acceptably easy user interface. With a little bit knowledge in the ideas of data mining this application would be as easy to use as mentioned in the documentation. The fact that this application plugs right into excel serves as a strength especially with the new version that can be pluged into other applications. As there are no academic or professional research on this specific application, the reviews had to serve as indications of its performance. Unlike Knowledge Miner, RapidMiner and Weka which use several data mining techniques, NeuroXL focuses on mainly neural networks, which might be a slight disadvantage. However, this application seems to have good uses for the present and potential in the future and for various fields and sectors.

- While Knowledge Miner is made for MAC computers, it looks like it has great potential as it withstood many other more expensive applications and managed to keep advancing. It has valid customers and good reviews. The fact that there are ongoing projects using this application as well as the customization aspect for each business makes it a good option for small businesses that use MAC computers. The better user interface also increases the chances of selecting this application as one of the most suitable options for a small sized company but not one that uses Windows or Linux based platforms which is what most small businesses use. It is therefore not applicable for the company being worked with in this study as all the computers used are Windows based.

- DMSK had below average evaluations when it was in its most popular days in 1998 when compared to other commercial versions. The lack of an evaluation version is another disadvantage to it. however, even though it could be argued that the reason a single user license version is no longer sold is that the software was not good enough to grow in popularity for that sector of the market which could be worrying for potential users, the same point can be used to argue that this means the focus is now on improving the application for small businesses. The application is now being sold as a customizable application specifically for each organization with training which might mean the availability of a good option for a small business. However, as there is no proof of any obvious advancements that kept this tool in the running with newer ones, it was fairly acceptable in the past to give this application a chance when compared with the more expensive ones, but since there is barely any data about this product, it seems rational that this software is not the right way to go for a business as it is risky to go with an un-advanced technology which has no real support or future from the developers. It is also not a good option for the real estate company being worked with as it is not existing in Dubai and therefore no in-house training can be done for the employees which is its most appealing feature.

While each application had its own functionality, used its own techniques, and had its own strengths and weaknesses, from the commercial applications, NeuroXL and Knowledge Management both seemed to have an advantage over DMSK. It is hard to pinpoint one of the two applications as the best one according to the above evaluation as each one has been made to a certain kind of customers specifically and at the same time provided enough functionality and possibility to be for any type of customer in general. One of the best features of both the applications is the ability to work for all types of fields. However for this study's purpose, Knwoledge Miner will not be considered as a viable option since it runs on MAC computers and to run on Windows a lot of work and additional cost would be associated with it.

As for the open source applications, Weka had an advantage over Rapidminer in this case since it generated results that were found to be useful for the company and were used to immediately consider some strategies for the year 2012. The

following are some of the visuals that were obtained from Weka when a dataset that includes information on the rentals of each month was used. The information used for the visualization tool were the type of unit, the month it was rented in, and the year it was rented in. based on that, valid results were displayed in the following figures. Figure 1 shows some useful information generated as soon as the dataset got loaded into the application

Figure 3

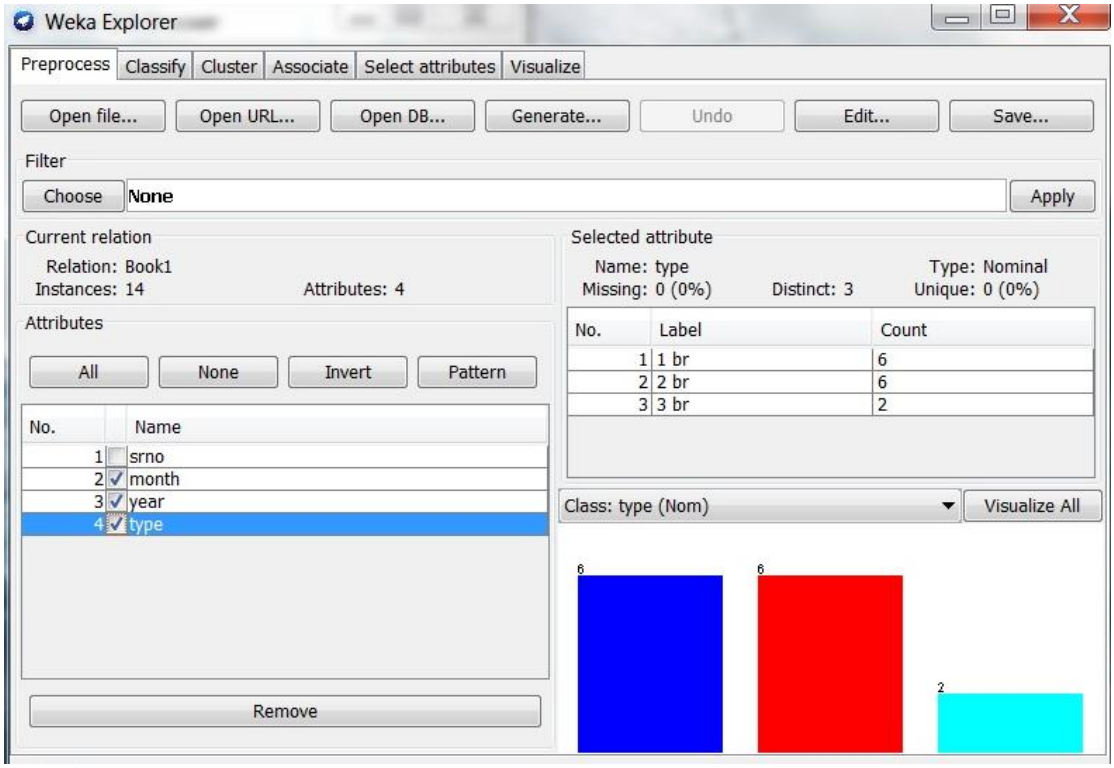

Figure 1: Information Displayed Upon Loading

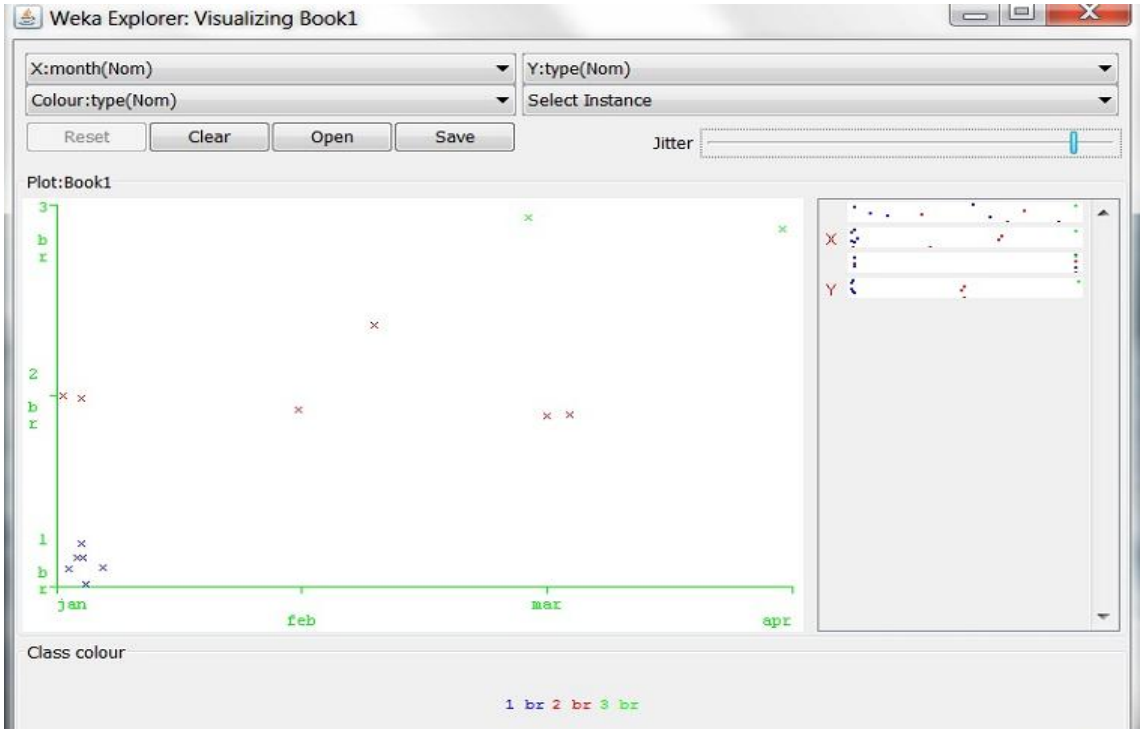

Figure 2: Type vs. Month

Figure 2 shows what type of units were rented in certain months. The way the data is displayed clearly gives the company an idea of what market to target in certain months.

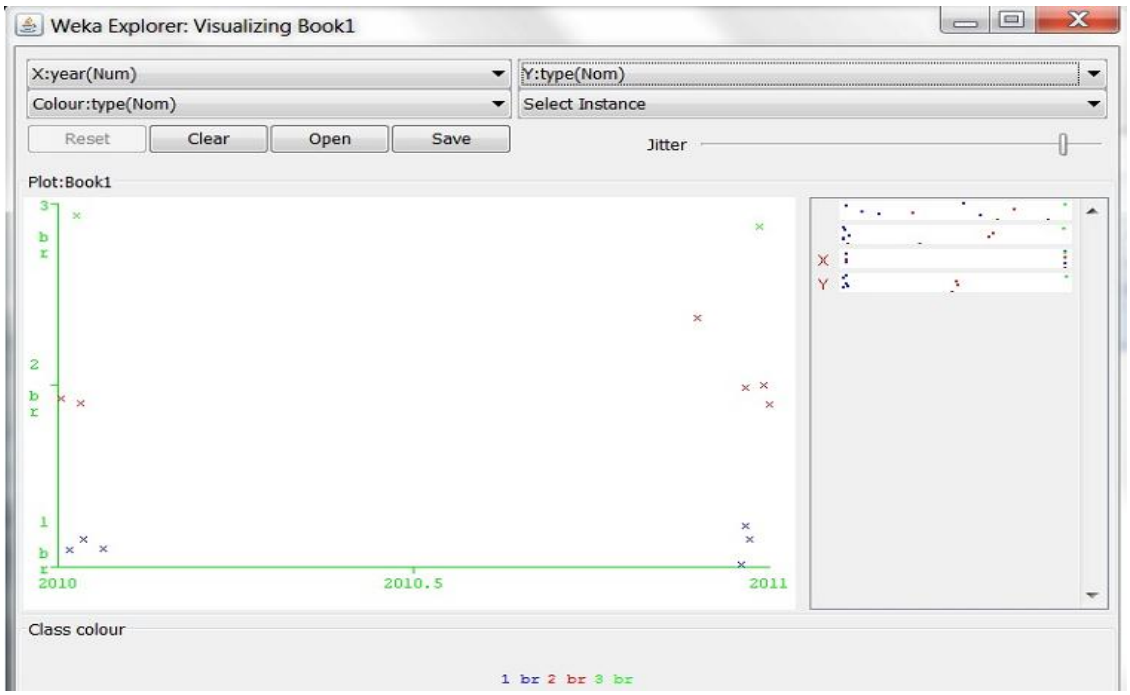

Figure 3: Type vs. Year

Figure 3 shows what type of units were rented the most in each year. The way the data is displayed clearly gives the company an idea on the performance of the company as well as the change in the market if any.

Therefore, further research could go into Weka and NeuroXL before determining which is best for the company. However as an initial result and based on the usefulness gained during the trial, Weka is the obvious choice. It came out to be easier to use, generated easier to understand results, had more models, is free to use, had more support online, and proved to be useful right away to non-technical users. Weka is also suitable for this specific company because a network is already setup between its two locations, which means the disadvantage of having to run the program on a powerful server that clients can use to access the application is not a big one in this case.

Even though these tools are fairly cheap for a small business to acquire, the cost of the resources time to be spent on more evaluation, implementiation and training might prove too costly for some businesses. Therefore, King's five basic steps to a better chance of choosing the right data mining tool: training, assessment, strategy, implementation and iteration would provide a good reference when attempting such a task(King, 2005).

All of the researched applications have claimed to be easy to use for the average user who has no prior knowledge of the data mining techniques. This however, is not what was concluded from this research. While these applications are in fact very simple for advanced users or people with some technical background and knowledge of the artificial intelligence techniques, they are very complicated for an average computer user found in every department in almost all small businesses. A person must have the right knowledge to be able to know what part of the data must be used in which part of the application as well as to know what the results mean and how the result could be changed according to the data inputted. Even after going through the help sections of the programs, the way to get a result was clearer but what the result meant and which options were the right ones to be picked were still unclear even in Weka to an extent,.

Now that the options have been narrowed down to a single tool for the company, extensive testing of the data must be undertaken in order to make a final

decision. More time could also allow for the selection of more popular open source applications for evaluation. Even though data was tested on all five applications, it was done very briefly due to the lack of time and resources and only with the aim of testing the graphical user interface and assessing the level of difficulty of reading the results. This however, is not a good judge of the consistency of the accuracy of the tools and their level of performance which is key when selecting a data mining tool.

This study however, serves as a good basis for a small business in Dubai to start a data mining project as it provides a more simplified and more customized way of dealing with such a project from a small business perspective rather than dealing with the traditional technical way.

Chapter Six: Conclusion and Project Extensions

6.1. Conclusion

In conclusion, the benefits of data mining go far beyond any doubt of their usefulness for any business, however, the high cost of acquiring and using commercial data mining software reduces their suitability to small businesses. Therefore, open source software and cheap commercial software are viable options that can be used by small businesses. The criteria for evaluating what data mining tools to use for a small business is not set in stone and does not have to be followed without consideration for the factors that have heavier weights on the decisions made by a small business. Finally, whichever data mining tool is selected, if used correctly, could lead to a big advantage in the Dubai market, and especially the real estate sector as there are barely any company's that use this technology at the current time in Dubai.

6.2. Future Research and Project Extensions

As there are numerous tools online, and many being added every day, researching more tools must be done to get a more comprehensive idea of the available applications. Also, a more detailed evaluation of the results of the applications must be conducted to get an updated as well a confirmed decision about the potential and benefits of the applications. Furthermore, working in collaboration with more companies from different sectors could provide further insight on the subject and further requirements.

Given more time, an extension of this research could lead to the development of a friendlier user interface for the average user to be incorporated to some of these applications, making the applications more useable as well as more useful as they would be used correctly.

Another extension to this research would be the development of a new tool that tries to incorporate most if not all the strengths of the researched applications in one tool as a better understanding of what a small business needs and expects have been attained.

References

Abbott. D., Matkovsky, I., Elder IV, J., 1998. 'An Evaluation of High-end Data Mining Tools for Fraud Detection.' Viewed 25 May 2011 <http://www.datamininglab.com/pubs/smc98_abbott_mat_eld.pdf>

Azevedo, A., Santos, M.'KDD, SEMMA and CRISP-DM: A Parallel Overview', viewed 05 Sep 2011 <http://www.iadis.net/dl/final_uploads/200812P033.pdf>

Boden, A., Nett, B., & Wulf, V. (2010). 'Operational and Strategic Learning in Global Software Development', *IEEE Software*, vol. 27, no. 6, pp. 58-65

Bremner, D., Demaine, E., Erickson, J., Iacono, J., Langerman, S., Morin, P., Toussaint, G., (2005). 'Output-sensitive algorithms for computing nearest-neighbor decision boundaries'. *Discrete and Computational Geometry*. 33 (4), 2005, pp. 593-604

Carey, B., Collier, K., Marjaniemi, C., Sautter, D. (1999).'A Methodology for Evaluating and Selecting Data Mining Software'. *Proceedings of the 32nd Hawaii International Conference on System Sciences*,1999.Viewed 25 Aug 2011 <citeseerx.ist.psu.edu/>

Chambers, J.M. (2008). *Software for data analysis: programming with R*, Springer, New York, NY

del Cacho, C. (2010). 'A comparison of data mining methods for mass real estate appraisal' *University Library of Munich, Germany*. MPRA no. 27378. Viewed 02 Sep 2011
< http://ideas.repec.org/p/pra/mprapa/27378.html>

Duda, R., Hart, P., Stork, D. (2001) Pattern classification (2nd edition), *Wiley*, New York, ISBN 0-471-05669-3

Fawcett, T., Haimowitz, I., Provost, F., Stolfo, S. (1998). 'AI Approaches to Fraud Detection and Risk Management'. 1998, 19 (2) *American Association for Artificial Intelligence* (1998). Viewed 17 May 2011
< http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1372>

Guoyin, W., & Yan, W. (2009). '3DM: Domain-oriented Data-driven Data Mining', *Fundamenta Informaticae*, vol. 90, no. 4, pp. 395-426

Gschwind, M (2007). 'Predicting Late Payments: A Study in Tenant Behavior Using Data Mining Techniques' *Journal of Real Estate Portfolio Management*. Vol.13, No.3, 2007
Viewed 01 Sep 2011.

<http://business.fullerton.edu/finance/jrepm/pdf/vol13n3/07.269_288.pdf>

Holmes, G., Donkin, A., & Witten, H.I. (2011). *WEKA: A Machine Learning Workbench*, viewed 2 July 2011 <http://www.cs.waikato.ac.nz/~ml/publications/1994/Holmes-ANZIIS-WEKA.pdf>

Hölzl, W. (2009). 'Is the R&D behavior of fast-growing SMEs different? Evidence from CIS III data for 16 countries', *Small Business Economics*, vol. 33, no. 1, pp. 59-75

Jackson, S. (2008). *Research Methods and Statistics: A Critical Thinking Approach*, 3rd edn, Cengage Learning, Hampshire

Kantardzic, M. (2003). 'Data Mining: Concepts, Models, Methods, and Algorithms'. *John Wiley & Sons*. ISBN 0471228524. OCLC 50055336.

KDNugget(2011). *Poll Results: analytics/data mining tools used for a real project*, viewed 6 August 2011 < http://www.kdnuggets.com/2011/05/tools-used-analytics-data-mining.html>

King, E., (2005). 'How to Buy Data Mining: A Framework for Avoiding Costly Project Pitfalls in Predictive Analytics'. *Information Management*. Viewed 27 Aug 2011 <http://www.information-management.com/issues/20051001/1038094-1.html?pg=1>

King, M., Elder IV, J., 1998. 'Evaluation of Fourteen Desktop Data Mining Tools'. *IEEE International Conference on Systems, Man, and Cybernetics.* San Diego, CA: 12 Oct 1998. Viewed 20 March 2011 <http://www.slideshare.net/pierluca.lanzi/machine-learning-and-data-mining-14-evaluation-and-credibility>

Kothari, C. (2008). *Research methodology: methods and techniques*, 2nd edn, New Age International, London

Mitchel, T., 1997. Chapter 1:Generative and Discriminative Classifiers: Naïve Bayes and Logistic Regression. *Machine Learning Inc.* Viewed 25 Aug 2011 <http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf>

NeuroXL. 2010. Software Help Documents. Viewed 21 March 2011. <http://www.neuroxl.com/>

Rao, C.R., Wegman, E.J., & Solka, J. (2005). *Data mining and data visualization*, Elsevier, London

Rapid-I (2011). *RapidMiner 5.0 Manual-English*, viewed 5 July 2011 <http://sourceforge.net/projects/rapidminer/files/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf/download>

Rupnik, R., Kukar, M., & Krisper, M. (2007). 'Integrating Data Mining and Decision Support through Data Mining Based Decision Support System', J*ournal of Computer Information Systems*, vol. 47, no. 3, pp. 89-104

Script Software Intl' 2001. *Knowledge Miner*. Viewed 23 May 2011 <http://www.knowledgeminer.com/>

Shmueli, G., Patel, N.T., Bruce, P. (2010). *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*, 2nd edn, John Wiley and Sons, New York, NY

Sumathi, S., & Sivanandam, N. (2006). *Introduction to data mining and its applications*, Springer, New York, NY

Yang, T. (2006). Computational Verb Decision Trees. *International Journal of Computational Cognition* (Yang's Scientific Press) **4** (4): 34-46.