19-October-2015

LAPPEENRANTA UNIVERSITY OF TECHNOLOGY

SCHOOL OF BUSINESS AND MANAGEMENT

INDUSTRIAL MANAGEMENT

Tea Suuronen

Utilizing analytic modelling process in industrial retail context

Master's Thesis

Examiners:	Post-doctoral researcher, D.Sc. (Tech) Samuli Kortelainen
	Researcher, D.Sc. (Tech) Olli Pekkarinen

ABSTRACT

Author: Tea Suuronen

Title: Utilizing analytic modelling process in industrial retail context

Year: 2015 **Place:** Espoo

Master's Thesis. Lappeenranta University of Technology, Industrial Management.
97+20 pages, 26 figures, 9 tables and 1 appendix
Examiners: Post-doctoral researcher, D.Sc. (Tech)) Samuli Kortelainen
Researcher, D.Sc. (Tech) Olli Pekkarinen

Keywords: analytical modelling, data-driven modelling, customer network

In today's complex business world companies that can transform their operational data into information assets could gain competitive advantage. The use of advanced analytics methods is on rise because companies need to identify factors to gain competitive edge in order to forecast upcoming trends more efficiently. This in turn provides essential support for accurate, real-time decision making that can be seen as basic requirement to survive in this highly competitive and volatile business environment.

The emphasis on this thesis was to design a theoretical framework based on existing literature from the business user point-of-view. Theoretical model was deployed in empirical part with diagnostic customer network modelling and to identify predictive factors related to the sales activities. The study is executed in a Finnish industrial filter technology wholesaler with business operations in Finland, Russia and Baltic states.

Research methods used in this thesis were derived from the case study strategy approach. The main data collection technique for primary data in this quantitative research was operational data and transaction data from the case company's ERP-system.

TIIVISTELMÄ

Tekijä: Tea Suuronen

Työn nimi: Utilizing analytic modelling process in industrial retail context

Vuosi: 2015 **Paikka:** Espoo

Diplomityö. Lappeenrannan teknillinen yliopisto, tuotantotalous.

97+20 sivua, 26 kuvaa, 9 taulukkoa ja 1 liite

Tarkastajat: Post-doctoral researcher, D.Sc. (Tech) Samuli Kortelainen
Researcher, D.Sc. (Tech) Olli Pekkarinen

Hakusanat: Analyyttinen mallinnus, Data-pohjainen mallinnus, asiakasverkostot

Nykypäivän monimutkaisessa ja epävakaassa liiketoimintaympäristössä yritykset, jotka kykenevät muuttamaan tuottamansa operatiivisen datan tietovarastoiksi, voivat saavuttaa merkittävää kilpailuetua. Ennustavan analytiikan hyödyntäminen tulevien trendien ennakointiin mahdollistaa yritysten tunnistavan avaintekijöitä, joiden avulla he pystyvät erottumaan kilpailijoistaan. Ennustavan analytiikan hyödyntäminen osana päätöksentekoprosessia mahdollistaa ketterämmän, reaaliaikaisen päätöksenteon.

Tämän diplomityön tarkoituksena on koota teoreettinen viitekehys analytiikan mallintamisesta liike-elämän loppukäyttäjän näkökulmasta ja hyödyntää tätä mallinnusprosessia diplomityön tapaustutkimuksen yritykseen. Teoreettista mallia hyödynnettiin asiakkuuksien mallintamisessa sekä tunnistamalla ennakoivia tekijöitä myynnin ennustamiseen. Työ suoritettiin suomalaiseen teollisten suodattimien tukkukauppaan, jolla on liiketoimintaa Suomessa, Venäjällä ja Balteissa.

Tämä tutkimus on määrällinen tapaustutkimus, jossa tärkeimpänä tiedonkeruumenetelmänä käytettiin tapausyrityksen transaktiodataa. Data työhön saatiin yrityksen toiminnanohjausjärjestelmästä.

FOREWORDS

This Master thesis has been the most intriguing project I have ever worked with. For me the driving force throughout the project was enthusiasm about my topic. I was easy to stay motivated when I saw also people around me to get excited about my work. The delivery of the thesis has been pretty straightforward since the beginning; steady uphill by learning new skills on topics I wasn’t really studying for, as well as self-management skills to deliver this kind of project from start to the finish line. I can truly relate myself now with the ideology that Industrial Management students are dynamic multi-talents who adapt into new challenges with smile on the face and bring-it-on attitude.

I wish to express my gratitude to my first examiner Samuli Kortelainen who shared his knowledge, supported and guided me through this project with a professional manner and helped me to see the light at the end of the tunnel in times I couldn’t see it. Secondly, I would like to thank all the people from the case company that were involved in this project, thank you all for supporting me.

The past years in the university have been the time of my life and since this also means an end of an era, I would like especially thank Industrial Management guild Kaplaaki and all my dear friends for their priceless support during this thesis project as well as during my studies. Special thanks goes to Kimulit who I am honoured to have been part of and keep on rocking with. Finally, and most importantly, the biggest thanks goes to my family who have encouraged and cheered for me always.

Tea Suuronen
Espoo, October 19th 2015.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBENDIXES

LIST OF ABREVIATIONS

AMP	ANALYTIC MODELLING PROCESS
B2B	BUSINESS-TO-BUSINESS
ERP	ENTERPRISE RESOURCE PLANNING
SME	SMALL- AND MEDIUM-SIZE ENTERPRISE

1 INTRODUCTION

This master thesis focuses on studying analytics opportunities in a Finnish B2B company operating in air and hydraulic filter industry serving both as an on premise and e-commerce platform. The purpose of this thesis is to study analytic modelling process in literature and how these could be utilized in small size business in retail context in order to enhance current business processes with the information gained from the internal data sources currently unemployed.

The outcome of the study will be two data processing models designed to study the current customer network position and to find relevant predictive indicators that drive sales. The models employed in this thesis can be used as support tools for more accurate, real-time decision-making in different levels of organization. The study is executed in a Finnish industrial filter technology wholesaler with business operations in Finland, Russia and Baltic states.

1.1 Background

Traditionally, organizations have used data to evaluate what has already happened and to find justifications and patterns for their choices of action. (Alstete & Cannarozzi, 2014) However, just recently leading organizations have harnessed their operational data into usage prospectively. The aim of this behavior is to anticipate potential behavior of target units and automate prescriptive decision-making that allocates the resources specifically, and ultimately grows the business while decreasing the risk and investments needed to perform these actions. (Siegel, 2013, p. 60) Data mining provides information assets from company’s data to be utilized later on with predictive indicators that can be leveraged to set and achieve precise strategic targets. (Accenture, 2015) Recently in Harvard Business Review (McAfee & Brynjolfsson, 2012) it was stated that companies that are placed among the top third in their industry for data-driven decision-making gain on average 5% higher levels with productivity and are 6% more profitable compared to their competitors.

In today's complex business world companies that can transform their operational data into information assets could gain competitive advantage. (Barton & Court, 2012) The use of advanced analytics methods is on rise because companies need to identify factors to gain competitive edge in order to forecast upcoming trends more efficiently. (Croon Fors, 2010) This in turn provides essential support for accurate, real-time decision making that can be seen as basic requirement to survive in this highly competitive, global and volatile business environment. (eMetrics Summit, 2013)

Predictive analytics is a data-driven support tool to be used with company's decision-making support system. (Ranjit, 2009) With current systems on the market, predictive analytics programs complement and engage with plenty of other IT and big data capabilities, for example OLAP's, query and reporting as well as with traditional statistical analysis tools. (Berry & Linoff, 2000) While the data mining provides the data used for the analysis, the predictive modelling with algorithms and machine learning provides a new way of mastering the data with ability to model underlying patterns that can be discovered from the large volumes of data. (Deloitte University Press, 2014) The data focus differentiates advanced analytics from the retrospective data technologies since it produces an actual model that acquires and represents hidden patterns, and highlights the interactions in the data to be utilized.

The produced models are at the same time prospective and highly descriptive since they address why certain things or actions happened as well as imply what is likely to happen next. The more data the model is fed with, the more accurate prediction the model can provide. This also underlines the biggest difference with traditional statistical analysis, where the model is the king. (Provost & Fawcett, 2013)

Statistical analysis provides more accurate forecasting depending on the quality of the model while with predictive analytics, the model is merely a mean to an end and the actual data from the organization forms the base for these operations. This results that a user can pose traditional 'what-if' questions with different scenarios to the predictive model that cannot be posed directly to the raw data. (Hardesty, 2015) These scenarios can vary from expected lifetime value of certain customers, popularity of different product categories into warehouse operations management directly from the predicted sales figures. (McAfee & Brynjolfsson, 2012)

Companies that can transform their operational data into information assets and effectively automate the decision-making process with this knowledge can realize considerable return of analytics investments, higher incomes as well as lower risks with their business operations. (Alstete & Cannarozzi, 2014) The outcome of using machine learning methods instead of traditional statistical methods does not have to be impeccable accurate, it just needs to be more accurate than the current methods in use in order to gain higher benefits. (Siegel, 2013, p. 18)

This study is executed to the Finnish SME operating in retail industry. Currently the case company's data-related activities are slowly evolving, and the managers have understood the potential value of an analytics function in order to support the decision-making. The case company was recently acquired by British investing group and the pressures to grow the business are present. The new management of the case company actively seeks opportunities for growth and due to these new growth goals also this thesis project was launched.

The company does not have systematic data processing or analysing processes, which lead to a situation where departments are: 1) making decisions based on their former actions and intuition, 2) wasting time by acquiring the same information that is already available internally, and 3) not having realistic knowledge about the current or future needs of customers. The current information flow can be described as reactive and it is unsystematic in all parts of the organization, whereas the ideal data function or at least data activities would possess a solid structure and a balance of pro-activeness and reactiveness.

The case company does not have a solid view about its customer base and the segmentation of the customers is executed based on the customer's business activities. With analytics it is possible to identify customer segments based on their actual purchasing behaviour, instead of old-fashion one-size-fits-for-all mentality that divides customers based on how they are perceived to have common needs, interests and priorities. For the sales, the data analytics in this thesis is mainly related to identify customers.

1.2 Research scope, objectives and limitations

Advanced data analytics is the discovery and interpretation of meaningful, hidden patterns that can be found in the company's operational and transactional data related to the company, products, channels and customers. (Davenport, 2013) Compared to the traditional statistical forecasting models, in advanced analytics the actual model or the value of single data entry is not the main focus, but the signals in and across the data clusters. (Barton & Court, 2012)

The ultimate goal of this research is to make the case company aware what kind of data analytics options they could benefit with their current data and ultimately unleash the data potential the case company holds with advanced analytics to make better, more informed decisions to grow the company. The thesis is a part of a larger project scope of the case company's top level strategy to seek growth options.

Analytics as a term is a huge playing field which can be applied to the organizations on several different levels depending on the skills and available resources to focus on. The main focus of this study lies in diagnostic analytics with mapping the current status of customers' network position and in predictive analytics on identifying variables that could be used when predicting future actions. The scope of the study is presented below in figure 1.

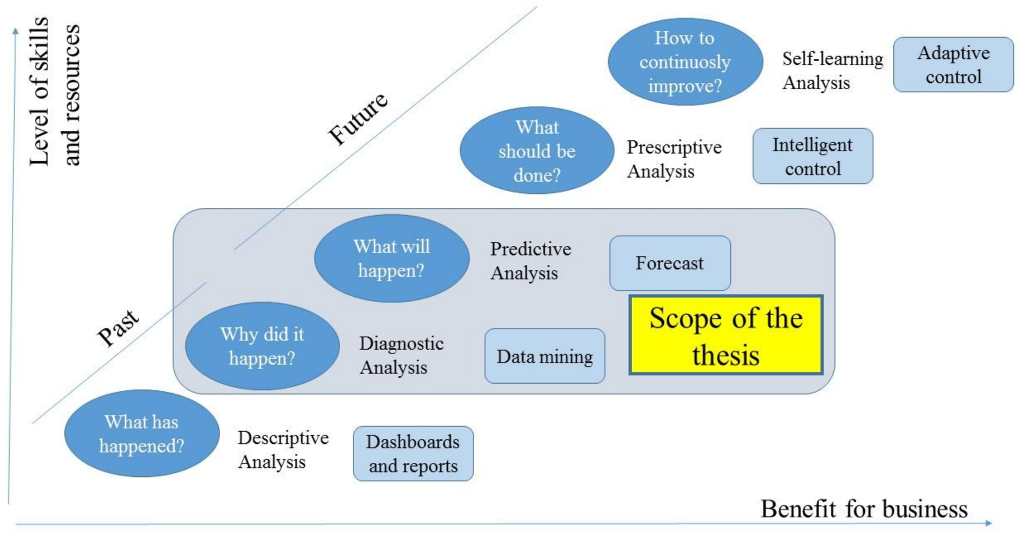


Figure 1. Scope of the research

For this research, the focus was decided to centralize in exploring the current customer situation. The case company has been traditionally segmented their customers based on their business purposes; Original equipment manufacturers (OEM’s), distributors and end-users. With new analytic modelling processes, it is possible to identify similarly behaving customers based on their purchasing behavior and perform the customer clustering and segmentation with valid and accurate knowledge about their actions.

The objective of this research was to design an analytic modelling process based on existing literature from the business user point of view and apply this compiled model into action with real case setting. In order to establish a solid foundation for defining objects for the modelling process in the case setting, it was necessary to familiarize with the current state and quality of the data usage activities in a Finnish wholesale business-to-business company, understand the requirements and opportunities produced from the current system that could be used for advanced analytics, and finally to implement gathered results into business initiatives to grow business further.

To attain the goals of the study, two research questions were formulated. The topic and the two research questions (RQ's) are shown below in figure 2.

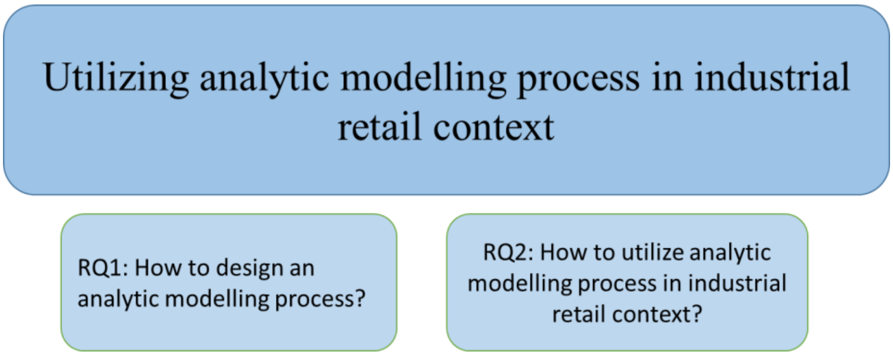

Utilizing analytic modelling process in industrial retail context
RQ1: How to design an analytic modelling process?
RQ2: How to utilize analytic modelling process in industrial retail context?

Figure 2. The topic of the thesis and research questions

The aim of the first research question is to create a theoretical framework based on existing literature and compile a modelling process that will be later used on the empirical part of the thesis with the second research question. The second research question is focused on adapting the planned model that is the results of the first research question, and deploy it with the real case company's data.

The limitations of the thesis were related to the limitation of the data sets. Since the case company operates in Finland, Russia and Baltic States it was necessary evaluate the data sources already in the early stage of the research. Since all the export activities from Finland to other countries are performed via subsidiaries and there was not available any end-customer data except from Finland, it was decided to limit the study consisting only Finnish customers. The collected, available data was operational and transaction data and it was used to accurately map the position of the company's end customers in network modelling.

The second limitations was related to the single-time entries in the transaction data. There were several single-time purchases from customers that were all marked under same customer number. This is the normal procedure in the case company; when the customer purchases at least a second time, the individual customer number is attached to the customer. These single-time purchases would create distortion to the data set by presenting these customers in larger scale than they are in reality, and in order to avoid these false positions, it was decided to rule this customer group completely out of the data sets.

The final, and most important limitation regards the data handling. Since the case company operates as a wholesaler in extremely competitive and volatile industry, the customer-related information is at the core essence of their business practises. Hence, all customer related data that would enable identifying customers has been encrypted and the gained results in this thesis remain on general level describing the progress of the process and what kind of outcome the models can produce instead of examining single results from the data set or from customer point of view.

1.3 Structure of thesis

This thesis has been divided into eight main chapters illustrated in figure 3 below. The chart presents how the thesis is structured chapter by chapter with the information flow. On the left side is the substance value and intake for each chapter and on the right side concrete deliverables for later use.

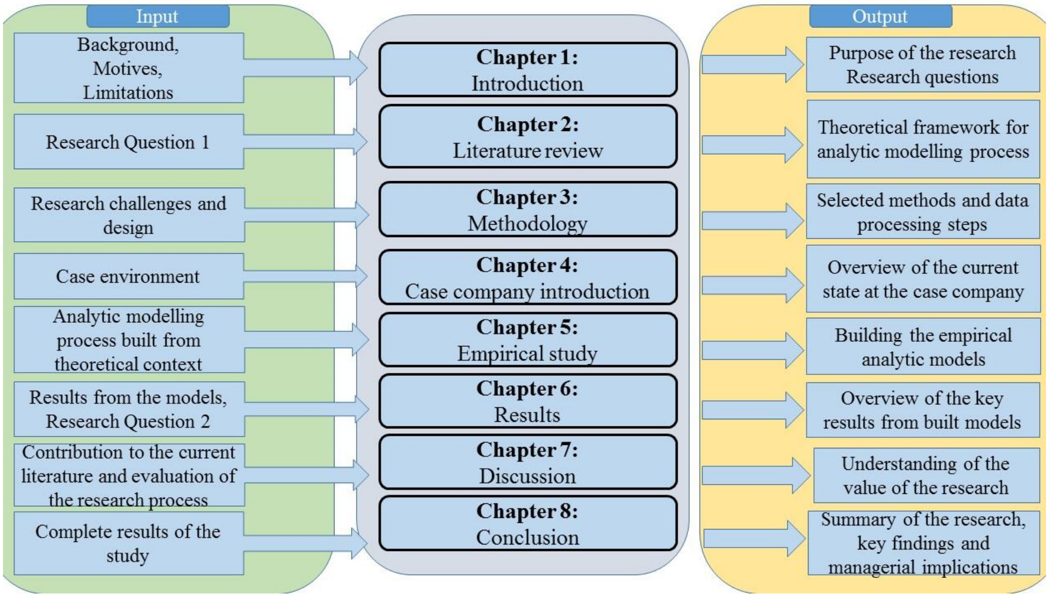

Input
Background, Motives, Limitations
Research Question 1
Research challenges and design
Case environment
Analytic modelling process built from theoretical context
Results from the models, Research Question 2
Contribution to the current literature and evaluation of the research process
Complete results of the study
Chapter 1: Introduction
Chapter 2: Literature review
Chapter 3: Methodology
Chapter 4: Case company introduction
Chapter 5: Empirical study
Chapter 6: Results
Chapter 7: Discussion
Chapter 8: Conclusion
Output
Purpose of the research Research questions
Theoretical framework for analytic modelling process
Selected methods and data processing steps
Overview of the current state at the case company
Building the empirical analytic models
Overview of the key results from built models
Understanding of the value of the research
Summary of the research, key findings and managerial implications

Figure 3. Outline of the thesis

Research begins with chapter one addressing the background and research problem for this study. The output of the first part is to identify objectives and limitations, and formulating the research questions as well as introducing the reader to the topic. The theoretical part of the research is presented on chapter 2 with descriptive literature review about the current knowledge of machine learning, predictive analytics and modelling processes. This part provides the theoretical foundation for information to be incorporated on answering the first research question on designing an analytic modelling process in the synthesis of chapter two. The different methods to be used in empirical part are introduced in the third chapter when outlining the methodology of the thesis, research design and data processing techniques.

The fourth chapter introduces the empirical case environment by describing the current data usage level and challenges the case company faces. The aim is to provide an overview of the present situation in the case company. In chapter five the theoretical model framework that was created based on existing literature has been applied into the case context with two modelling processes. The chapter presents the adaptation of the theoretical model into each situation. Following the modelling processes, chapter six presents the results gained by using the models. In the synthesis of the results both of these built modes are being compared by their usability from the business user point of view.

After wrapping up the results, chapter seven reflects the findings of the research to the current literature presented in the literature review and displays contributions of the research to the existing theoretical frameworks. Also the quality and reliability of the research are discussed in this part. Finally, key findings of the research and managerial implications are concluded at the end in chapter 8.

2 LITERATURE REVIEW

In today's dynamic and complex business world, organizations must find innovative ways to differentiate themselves from competitors in order to stay in business. Companies need to become more agile, adaptive, mobile, virtual and accurate in order to rapidly respond to the changing markets and customer needs. (eMetrics Summit, 2013) Many of these companies have noticed that they can differentiate themselves with the data they own and produce. (Demirkan & Delen, 2013) Data and information are becoming primary assets for organizations, which has resulted that companies and organizations are collecting as much data as possible. (Croon Fors, 2010)

Previously statisticians have built models to estimate and forecast the future actions. However, the accuracy (Edelstein, 1999) of these models have been highly related to the quality of the model. Nowadays, data analytics prefer to use the actual data to reveal hidden patterns and the model built is just a tool to support main purpose. (Barton & Court, 2012) While statisticians need to prove why certain underlying results are useful based on their model, analytics using predictive methods can simply discover how the results are useful. Advanced analytics is guide-lined with the idea that it is not necessary to know an explanation how certain model works, since this part is taken care by the machine learning, but how the results can be used to make better decisions. (Siegel, 2013, p. 67) This is also the main reason why advanced analytics is gaining more attention especially from the business user point of view, who can just reap the benefits the model produces without the need to dig deep on how the model actually functions. (Alstete & Cannarozzi, 2014)

The chapter has been divided into three sub-chapters. First part introduces the environment of modern advanced analytics and machine learning. Second part focuses on defining the processes behind analytics modelling and in the third sub-chapter learned knowledge is combined into theoretical analytic modelling process that will be deployed on the empirical part further on.

2.1 Advanced analytics

This chapter introduces the reader into the dynamic and exciting world of analytics and cognitive computing. Company examples are being used to illustrate the practical applications of successful business cases built on theoretical modelling.

The first part introduces the environment for predictive analytics and presents the mind-set of leveraging the existing in-house data of the companies into more efficient use. Since the roots of analytics and machine learning lies heavily in statistics, the second part presents the main differences between these two approaches; machine learning and traditional statistical methods. Finally in the third sub-chapter the context of cognitive computing and how it functions is being introduced.

2.1.1 Predictive analytics

Information about data transactions has existed for quite a while already. However this data was not seen as a resource to improve business operations or to gain competitive advantage but rather as evidence of past transactions (Joe F, 2007). According to Siegel, data describes a collection of events to learn from (Siegel, 2013, p. 23). Only recently this data has been given value and translated into knowledge. In today’s dynamic business world, the ability to convert information into knowledge and further on into business case is worth of gold (Siegel, 2013, p. 24). According to Ranjit (Ranjit, 2009) this means to deliver accurate, timely and efficient decisions on strategic, tactical and operational levels to face customer’s requirements and preferences.

To simply describe, predictive analytics, or advanced analytics, is a generic name to describe the process of answering questions and/or solving problems by applying various analytic methods and techniques to process existing data (Ranjit, 2009). On the other hand, Eric Siegel (2013, p. 31) presents his definition as "Technology that learns from the experience, data, to predict the future behavior of individuals in order to drive better decisions." As a process, predictive analytics uses confirmed relationship information, for example from transaction data, between interpretive and norm variables from past events to forecast upcoming behavior (Joe F, 2007). The more data is collected and integrated, the more comprehensive and accurate model can be build based on recognized patterns and underlying relationships (Ranjit, 2009). Operationally, predictive analytics transforms daily data into strategic information which can be then used on all levels of organization's decision-making units to gain competitive advantage (Ranjit, 2009)

From managerial point of view, predictive analytics enables right information available to the people who need it at the right time (Ranjit, 2009). In other words, to receive this information, relevant people can simply log into their system and fetch the valid knowledge to support their decision making without additional data analytic in the process, which were used in the recent past when the analytics processes were more complex by the nature (Klatt, et al., 2011). The ultimate goal of using predictive analytics on managerial level lies in comprehensive knowledge about existing data. This leads to focusing on relevant information, gaining actionable insights, making smarter decisions, and ultimately, securing better outcomes. (Accenture, 2015)

Data sets that companies handle are still extremely small amount from the total data provided by the whole industries. (Deloitte University Press, 2014) However, even though these data sets are relatively small, there is still enough material for model building and testing. The beauty in predictive analytics, as Siegel explained in his book (2013, p. 11) is that predictive analytics model does not have to be 100 % accurate to obtain results. It is enough to cover better percentage than the current system in use to gain better outcome.

While fraud detection and fast tracking claims are currently the most popular applications, the ideology of predictive analytics’ one-size-fits-for-all provides opportunities in entire range of business operations both on strategic, tactical and operational level. (Siegel, 2013, p. 28) According to Guardian (The Guardian, 2014), last year in UK fraudulent insurance claims were reported as worth of £1.3 billion in total. A cautious estimation reveals that in northern America alone the value of fraudulent insurance claims could be almost $80 billion dollars on a yearly level. Even a 10 % difference in these number would create significance difference. (Coalition Against Insurance Fraud, 2014)

The processes of handling the data set remain duplicable while the focus and posed angle to the data set and expected outcome change (McKnight, 2008). Direct marketing and digital online marketing with click-on ads and conversion rates are the areas profiting highly from predictive analytics. With information based on these customers’ web behaviour, rough customer clustering can be performed. (eMetrics Summit, 2013)

To illustrate previous situation with an example from direct postal marketing with a mailing list of 1 000 000 prospective buyers. These buyers are all ranked as equal and the cost of advertisement sheet and mailing is 2 € per receiver. It has been observed that out of these 1 000 000 prospective buyer only 1 % makes an actual purchase. Profit of these 1 % customers is 220 € per buyer, making the total profit based on these estimations

$$\begin{aligned}\text{Overall profit} &= \text{revenue} - \text{cost}\\ &= (220\text{ €}\ \text{x}\ (0{,}1\ \text{x}\ 1\ 000\ 000)) - (2\text{€}\ \text{x}\ 1\ 000\ 000)\\ &= 200\ 000\text{ €}\end{aligned}$$

When applying predictive methods into data sets, it was noticed that one quarter (250 000) of customers are three times more likely to buy the product compared to average rate. Three percentages from 250 000 is 7 500 respondents and while 3 % is still extremely inaccurate, it already has a huge impact to the results.

Overall profit = Revenue – cost

= (220 € x 7500) – (2 € x 250 000)

= 1 150 000 €

Overall profit increased by 5.75 times over by contacting fewer people. Predictive analytics enhances the possibilities to target the customer groups that are more likely to response to the marketing campaigns. In this case already a small and rather simple predictive model applied to existing data set created an increased results. (Siegel, 2013, pp. 19-20)

The usage of predictive analytics in the processes can be defined as causal-relationships. Predictive analytics applications in business processes use algorithms that can be applied to large volumes of data to reveal hidden patterns and predict the future behaviour (Liyakasa, 2013). Each application is defined by what kind of behaviour, for example action or event, will be predicted for each customer, customer segment, stock or other kind of element, and what are the actions taken by the organisation in response or after being informed by these predictions. (Wang & Chen, 2015)

The first phase, data processing, generates a predictive score for each examined unit or organizational element; individual consumer, customer segment, or even a warehouse unit, based on how likely they will act according to prediction. The second phase includes the decisions driven by the prediction in different decision-making levels. (Siegel, 2013, p. 39) For example on sales and marketing function this can be used to predict churn rates among customers and customer segments, and then taking action to retain the valuable ones. (Computerworld, 2013) Sales and marketing function can also gain valuable insight and optimization of resources with predictive analytics tools. For example when altering the focus of the selling resources among customer segments by estimating who are sound customers and will buy anyway, and which segments and customers require more attention (Gibbons, 2015).

Amazon (Huffington Post Tech, 2014) has taken this approach over the current used limits by creating and patenting a process called “anticipatory package shipping”, APS. This APS process is predictive modelling process that deploys Amazon goods to the specific geographical areas based on speculative shipment system. Amazon uses this APS approach to pack items before the actual orders are even made. According to APS these ready packaged items are then sent to the Amazon’s fulfilment warehouses located nearest to the customers who are, according to their predictive modelling, most likely to purchase these items. (Huffington Post Tech, 2014) Especially for vendors this is a time-saving procedure since by using these kinds of predictive models and big data, vendors can identify the consumers most likely to place in an order for certain items and have these accessible on stock at the time of consumer’s purchase order. These items are already packaged and ready to be shipped to the customer immediately after adding the buyer’s details. (Davenport, 2013)

Especially useful advanced analytics has become for marketers who, by integrating consumer information with data processes, create consumer habit profiles. (Davenport, 2006) According to Hsieh (2004), these profiles have proven to be more accurate by large and describing consumer behaviour from individualistic point of view based on person’s history, previous actions and choices that have been made and then comparing the findings with existing customer profiles, instead of basing marketing segmentation on external factors, that cannot be influenced, such as country of origin or gender. Hsieh (2004) proposes an integrated data mining and behavioural scoring model that is similar to the statisticians’ classification analysis by neural networks, which was presented already in 1995 by Lancer & al (1995).

Since the total volume of individual data sets can be enormous, business analytics use statistics and machine learning to capture, store and process this information instead of traditional method of manually exploring the information. (Barton & Court, 2012) With regularly purchasing customer, this is rather easy process, since for example Amazon captures the info of previous purchases and returns, browsing and clicking history, length of time spent on each page and Amazon Wish List items to get customized list of interests and, for example, when certain customer’s favourite author’s new book is released, it is already on the way to the nearest warehouse location of this customer. (Huffington Post Tech, 2014) The items won’t be shipped straight to the end-customer, but this kind of system would decrease the delivery time for these items. For customer this could mean the same day, or by best even only couple of hours, delivery time, which results in content customer enhancing the overall shopping experience, which can be then used in marketing purposes for upcoming purchases. (Davenport, 2013)

2.1.2 Differences between machine learning and statistics

Statistics, data mining and machine learning present the same branch of science, who have set as a common aim to uncover structure in the data. These disciplines have such similar aim and overlapping tools that sometimes data miners are seen as a sub-category of larger statistician even though this cannot be seen as realistic judgement since the object of using these differs significantly. (Hand, 1999).

Machine learning, computational learning theory, cognitive computing and terms alike are regularly used in the same context as data mining. (Chen, et al., 2004, p. 25) These terms are used to indicate the application of comprehensive model-fitting or classification algorithms for predictive data mining (Lin, 2002). While the stress in data mining and machine learning is usually on the accuracy and usability of the prediction, traditional statistic data analysis methods aim to find the estimation of population parameters by statistical inference. The accuracy of the prediction is seen as the key focus disregarding whether or not the models or techniques that are used to produce the prediction is understandable or open to uncomplicated explanation (Hand, 1999). Good examples of this type of techniques often applied to predictive data mining are neural networks or meta-learning techniques. (Dell, 2015) These methods usually involve the fitting of intricate universal models that are not related to any reasoning or theoretical understanding of underlying causal processes. Instead of classifying these methods under causal processes, these techniques can be shown to generate accurate predictions or classification in cross-validation samples (Provost & Fawcett, 2013). From historical point of view, machine learning and data science can be seen as spin-offs of statistics. This is rather interesting argument since earlier also statistics was seen as spin-off from mathematics. (Hand, 1999)

In the table 1 below are presented some key objects that separate these two divisions that have been collected from the existing literature. The first column presents the object in comparison between machine learning and statistics and in corresponding columns are presented the output from each district.

Table 1. Differences between machine learning and statistics

Object	Machine Learning	Statistic
Outcome	Network, graphs, predictive models	Model
Variables	Weights	Parameters
Method	Learning	Fitting
Model Fit	Generalization	Test set performance
Model Building	Supervised learning	Regression/ classification
Data Analyzing	Unsupervised learning	Density estimation, clustering
Time Concept	Impromptu, Ad hoc	Rigorous
User Attitude	Experimental	Fixed

Althought machine learning and statistics have plenty of common points and similarities, there are some key differencies to be addressed between these two. Firstly, while statistic provides a model as an outcome to be used for a dataset, the output from machine learning is networks, visual graphs and predictive models to describe the data sets fed into the system. Machine learning uses weights of individual entries of data in the modelling as variables to describe the appearance. Statistic uses parameters as variables for characterizing and summarizing data for larger sample sets. (Provost & Fawcett, 2013)

Statictians use fitting of the data samples as their method of testing the model. Sample sets of data are used to find generalisable models that can be fit into larger population. According to its name, machine learning learns from the data it is fed with. The more data is available, the more accurate the model built can be. Also in the future, when original model is fed with new data, the model learns from this and can produce even more accurate model by developing the original model. (Barton & Court, 2012)

Approach to the model building process separates these two movements. While machine learning is dependent on supervised learning and provided algorhytms, based on which the computer produces the model, statistics rely on regression and classification. Statistical approach reins out the outliners and null values to provide centralized view, which present the results from the majority of data set. Since machine learning is fed with large sets of data, it can produce generalizable models from the complete set of the data. (Luciano, et al., 2010) On the contrary, statistic can only produce models that are valid within their test set, and are thus limited with their test set performance.

Data analysing process is one of the most interesting differences between machine learning and statistics. While statistics rely on density estimation and clustering of the gained results, machine learning has been designed to find underlying patterns in the data set and can produce insights from the data set from wider range. (Hand, 1999)

Another distinguishing feature between these two orientations is related to their concept of time. This is related to the idea that statistics as a discipline has a tendency for conservativeness when addressing issues impromptu. Statistics tend to prefer the rigorous time span instead of ad hoc evaluations that could be more beneficial in some situations. This conservativeness for addressing issues spontaneously has its roots in mathematics, which operates strictly under set rules. Statisticians rely on rigorous, linear models when in real life there does not exist events that follow linear modelling since forecasting cannot be 100 % exact. (Hand, 1999)

While statisticians are mostly interested in about the mathematical argumentation from the validation of the model, machine learning includes also ideas, tools and methods from other areas of science, such as database technology and data mining from computational branch (Fayyad, et al., 1996). This resemblance in aims of both statistics and machine learning has caused bafflement with the traditional view for data sciences. This issue has evolved when analytics discipline concentrated on solving topics that statisticians had previously considered as their dominion. Furthermore, the new subject had particular relevance in business use, especially in marketing, additionally with its scientific and other applications. (Klatt, et al., 2011)

The roots in mathematics and stress for strictness and accuracy has encouraged the trend of demanding the proof of concept for the proposed methods and solid evidence of its functionality prior to the usage of the chosen method. In computer science and machine learning environment, user attitude is much more experimental. The possibility to pose “What-If” questions and creating scenarios with the data enables curiosity towards the results. (Berry & Linoff, 2000) This has resulted in situation, in which more experimental disciplines of science have faced the same challenge as statisticians but have managed to produce methods which according to test results work, even if they cannot proven these to work. (Chen, et al., 2004)

The mentality to share these findings tend to be difficult, since in general, statistical journals seem to avoid publishing research that utilize ad hoc methods. Instead these journals seem to favour of findings which have been proven by precise mathematics to work. Data mining and machine learning practitioners have received adventurous attitude towards data modelling process. This does not mean that data mining professionals do not value rigour or precise methods, but rather hints if the certain methods produce valid results, the emphasis is not on the correctness of the model, but in the value of the results. (Hand, 1999)

When researching through statistical literature, the results reveal the attention of statistics on mathematical strictness and formulas. There seem to exist heavy emphasis on inference. Any general statistics text indicates that a central concern is about how to make statements from a population when only a sample of the total population has been observed (Hand, 1999) even though there are some categories within statistics who are focused on description. (Provost & Fawcett, 2013)

Interference is often also a valid concern in data mining. A defining attribute of a data mining problem is that the data sets are large. This means that often one might want, for reasons of practicability for example, to work only with a sample and make statements about the larger data set from which the sample was drawn. (Berry & Linoff, 2000) However, data mining faces different problems, since the available data normally includes the data of the complete population. These data sets include details of the companies' employees, database of all customers or complete set of transactions made on previous years. In such cases notion of significance testing loses its meaning and the observed value of the statistic, for example the mean value of all the year's transactions, is also the value of the parameter since these sets were handled as complete. From the statistical point of view this means that if proposed features in the model do not differ significantly from zero, these features can be discarded from the model. Data mining is not concerned about these features if the entire population is involved. (Kamber &

Han, 2000) Instead of using mean value or features alike, there are other options that can be used, such as score functions. Score functions are designed to measure the adequacy of description a model provides for the data since the set includes the total data set instead of sample. The mind set of data mining by focusing to the results, rather than the correctness and theoretical hypothesis of the model, simplifies the model searching process. To illustrate this further, simplistic features of accuracy measures can be deployed in data mining when searching model algorithms, when by using probabilistic statements about generalisability of the results these simplistic features might not be noticed. (Berry & Linoff, 2000)

For data mining practitioners, the core of processes and excitement lies in the attitude of discovering valuable information unexpectedly. In practise this means that the process itself is highly exploratory. These data mining practitioners are not concerned about the best way to collect the data in the first place, or what is the most suitable method to handle the research further. (Edelstein, 1999) These professionals are focusing with the core aim of data mining, which is discovery, to find answers for specific questions. The starting point in data mining is based on assumption that the primary data has been collected and the aim of the process is simply how to discover the hidden patterns in the data set. Roughly summarizing the main difference being that statistics analysis is model-driven, while advanced analytics is data driven (Hand, 1999). According to Joe (2007) classic statisticians criticize the unscientific user attitude of data miners and predictive analytics as the techniques they use in their processes are data driven instead of theory driven.

Statisticians tend to have rather narrow set on how they see the data. According to Hand (1999) statisticians see the data as “convenient flat table, with cases cross-classified by variables, stored on the computer and simply waiting for analysis”. If the data set under research is small enough that it fits into the computer’s memory on premise, this might be acceptable way to see issues, but normally situation with data mining is quite in contrast. Actually, large data sets might even be dispersed across several machines and require multiple platforms to handle these properly. (Edelstein, 1999)

The invention and development of computer can be said to have its roots on statistics. Because of this connection, many of the tools used in statistics can still be applied manually. The link from past is still constantly present since for example many statisticians consider a data set of 1000 points as large. (Hand, 1999) The used term large can obviously be seen relative, since for example Visa, UK's largest credit card provider handles 350 million annual transactions (The UK Cards Association, 2014) and the largest long-distance carrier AT&T handles daily 200 million long distance phone calls (AT & T, 2015). It is rather obvious that manually applied techniques are not really practical when facing such rapid increase in numbers. New approaches are required and today, computers are fundamental in every business, no matter what industry the companies present. (Madison, et al., 2012) Computers are used for data manipulation and data analysis, and provide the necessary filter between the data and the analyst, since due to the large mass of data, it is not possible for the analyst to interact directly with the data without computational tools. This essential approach for data sets is another reason for the emphasis on algorithms in machine learning. (Hand, 1999)

2.1.3 Cognitive computing

Analytic computing programs are at the cutting edge of the new era of computing, cognitive computing. This is radically new kind of computing that differs from the current programmable systems as much as those systems were from tabulating machines a century ago. (IBM, 2015) Traditional computing solutions are based on mathematical principles with rules and logic intended to write mathematically precise answers, often following logical decision tree approach. But with today's dynamic business environment and enormous amount of data, the need for more complex, evidence-based decisions is needed, such a rigid approach of breaks or fails to keep up with available information. (Knapp, 2011)

Cognitive computing enables people to create a profound for new kind of value, finding answers and insights hidden away in volumes of data that has not yet been able to research. (Dobbs, et al., 2011) The applications for this kind of systems are enormous in all industries and company-levels. For example when considering a doctor diagnosing a patient or a marketing manager optimizing their customer portfolio, new approaches are needed to put into context the volume of information they deal with on a daily basis, so the value could be derived from this process. (IBM Research, 2015)

Cognitive computing's process aim to enhance human expertise. Computer and its cognitive capabilities mirror some of the key cognitive elements of human expertise; systems that reason about problems like a human does. (Hardesty, 2015) When human seeks to understand something and to make a decision, he goes through four key steps. These steps are presented below in figure 4.

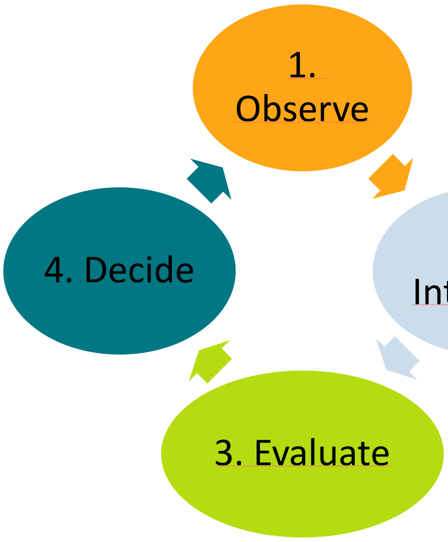


Figure 4. Decision-making process for human

In the first stage the person observes occurred visible phenomena that embodies evidence. On the second stage, he draws on what he knows to interpret what he is seeing to generate hypothesis about what it means. On the third stage he evaluates which hypothesis is right or wrong, and finally decides choosing the option that seems best in acting accordingly. (Montgomery & Svenson, 1992). For example when facing a decision whether to take stairs or elevator in a building, first step is to observe the issue at hand that one needs travel up. In the second stage the person interprets what this situation means to him, for example into which floor he needs to travel, and in this case that there are two possible choices; stairs and elevator. On the third stage the person evaluates the options based on possible factors that each choice carries. For example traveling with elevator might be faster but taking the stairs is better for overall health. Finally person makes the decision based on both prior experience and what outcome these produced as well as evaluated pros and cons precisely in this situation. Just as humans become experts by going through processes by observation, evaluation and decision-making, cognitive systems use similar processes to reason about the information they receive. Compared to humans, these systems can do this with massive speed and scale. (Forbes tech, 2011)

Unlike traditional approaches to computing, which can only handle neatly organized structure data, such as what is stored in databases and archives, cognitive computing systems can understand unstructured data, which according to Forbes (Barrenechea, 2013) is 80 % of the total data of today. All of this information is produced primarily by humans for other humans to consume. This includes everything from literature, articles and research reports to less organized forms of social media data, blogs, videos, posts, news and tweets. Well-structured data is governed by clearly defined fields that contain unambiguous information. (IBM Research, 2015)

Cognitive computing relies on natural language which is governed by rules of grammar, context and culture. As with natural language, it is implicit, complex and a challenge to understand when processing. Cognitive computing focuses to the root of causes instead of keyword matches and synonyms like search engines. (Accenture, 2015) These programs actually read and interpret text like a person, finding contact surfaces on similar used context before interpreting the meaning.

This takes place by breaking down a sentence grammatically, relationally and structurally discerning meaning from the semantics of the written material. (Deloitte University Press, 2014)

Cognitive computing programs understand context which separates these systems from simple speech recognition, which is how computer translates human speech into a set of words. Cognitive programs try to understand the real intent to the user’s language and uses that understanding to possibly extract logical response and draw interferences to potential answers to a broad array of linguistic models and algorithms. (Computerworld, 2013)

When cognitive computing systems work in particular field, these systems learn the language, jargon and the prevailing motive that domain the field. For example, when researching the term cancer from healthcare. There are many types of cancer that each has different symptoms and treatment. However, these same symptoms can also be associated with other diseases. Also treatments can have side-effects which affect people differently depending on many factors. Cognitive systems evaluate standards of care practices and existing literature that capture the science in the industry. From the gained results the system identifies the therapies that offer the best choices for the doctor to consider in their treatment of exactly that patient currently under study. (IBM Research, 2015) With the guidance of human experts, cognitive systems collects the knowledge required to have literacy in particular domain, which is called a corpus of knowledge. Collecting a corpus starts with loading the relevant substance of the literature onto cognitive system. Building of the corpus requires also human assistance in order to teach the computer what is relevant information. In this stage, it is the person interacting with the machine to discard anything that is out-of-date, poorly regarded or immaterial to the problem domain. This phase is called curation of the content and as by its name imply, aims to confirm the information used is accurate and valid. (Modha, et al., 2011)

In the following phase, the data is pre-processed by the cognitive system building indices and other metadata that makes working with certain data content more efficient. This is known as ingestion. At this point, cognitive systems may also create a knowledge graph to assist in answering more precise questions. In this phase the cognitive system has established a corpus and it needs to be trained by a human expert to learn how to interpret the information, to learn the best possible responses and acquire the ability to find patterns. (IBM, 2015) These cognitive systems are paired with experts who train it in using an approach called machine learning. An expert will upload data into system in the form of question-answer pairs that service as ground truth. This process does not give an explicit answer for every question the system will receive but rather teaches at the linguistic patterns that have meaning to the domain. (Davenport, 2013) Once the system has been trained with question-and-answer pairs, it can continue to learn through ongoing interaction. These interactions between users and cognitive systems are periodically reviewed by experts and fed back into the system to help the system better interpret information. Likewise, as new information is published, these systems are being updated so that the system is constantly adapting to the shifting knowledge and linguistic interpretation in any given field. (Forbes tech, 2011)

After these phases, cognitive computing system is ready to respond questions about highly complex situations and quickly provide a range of potential responses and recommendations that are backed up by evidence. It is also prepared to identify new insights or patterns hidden in the information. (Barrenechea, 2013)

Cognitive computing and modeling is not something that only aimed for certain tasks but it can be utilized in wide range of industries. For example choice modelling in engineering design looking for new alloys (Wang & Chen, 2015) to researchers looking to develop more efficient drugs, human experts are using cognitive computing to uncover new possibilities in data and make better, evidence-based decisions across all of these different industries. According to IEEE, big data and predictive analytics systems have been successfully utilized in neonatal intensive care both in UK and US (IEEE Life Sciences, 2013). By utilizing predictive analytics system the hospitals managed to identify key factors with complication that occur in emergency care of neonatal patients. By monitoring the brain waves and driving these results into advanced analytics systems, doctors were able to identify critical changes in brain activity up to 24 hour earlier that it was possible with older systems. (IBM, 2015) Just recently also Helsinki University Hospital acquired IBM Watson, highly predictive analytic tool to deploy cognitive computing with their neonatal care. (Varho, 2015)

2.2 Data modeling

This chapter presents the ideology behind the data modeling processes and framework on how models can be designed. The first part describes the ecosystem of data mining and the second part lays out a foundation for modeling process.

2.2.1 Data mining process

There is a growing trend for using data mining as a business information management tool. The reason behind this lies in the thought that data mining is expected to reveal hidden knowledge structures that can ultimately guide business decisions in more uncertain conditions. Lately, there has been an increase in interest to develop analytic techniques specifically designed to address the relevant business needs of data mining. For example using decision trees can be nowadays seen as data mining technique. (Davenport, 2013) However, data mining is still based on at least with conceptual principles on statistical modelling, which includes aslo for example traditional exploratory data analysis as well as modelling. Data mining also shares some components in its general approach as well as techniques with these more traditional approaches. (Provost & Fawcett, 2013)

When comparing data mining with traditional exploratory data analysis, an important general difference is in the focus of these two approaches. The traditional approaches are interested about the basic nature of underlying phenomena while data mining is clearly more oriented towards possible applications and implications of the revealed phenomena than the phenomena itself. In other words, data mining is less focused to identifying underlying reasons between involved variables than driving new initiatives with the results. (Linoff & Berry, 2000)

Data mining is an analytic process, which is designed to explore data sets. These sets contain usually large amounts of data, which, depending on the objects of the data collections, usually focus on business or market related topics. These kind of data sets are sometimes referred as “big data” although the term “big data” is rather confusing, since it can include anything from company level to industry and even global level on information assets. Data mining is used to search consistent patterns and systematic relationships between variables. These findings are then validated by applying the detected patterns and models into new sets of data. (Hand, 1999)

As it comes to using terms, predictive analytics and data mining are sometimes viewed as the same, while in reality they are individual interacting processes that can be used sequentially. The first process is to use data mining to search and identify patterns and interesting relationships. These results are based on searching data (numbers), text (words and phrases), web movements (click through and time-spent patterns), and visual image. (Chen, et al., 2004) Predictive analytics is then used to explore these gained results, identified patters and relationships from data mining to predict future behaviour. (Joe F, 2007)

Different combinations of data mining goals and methods are used to ensure flexibility and the greatest accuracy possible in the process. The major benefit of data mining is that it can be used as an aid on strategic, tactical and operational level decision making processes in situations where numerous variables can affect costs or benefits of actions that a company might decide to take and which cannot be modelled with traditional methods. (Berry & Linoff, 2000)

The results of data mining can be visualized in the form of a more traditional methods such as decision trees. The modelling following data mining phase absorbs the information on costs and benefits of alternative courses of action. (Kantardzic, 2011) Companies can use such information to find new opportunities for growth, choose more effective means to achieve their business goals and streamline business processes to lower their costs since with the decision tree the pinpoints of the processes can be evaluated. (Edelstein, 1999)

Data mining and visualization tools are used in combination to improve the usability of advanced analytics systems. The purpose of data visualization system is to give the user an understanding of what is happening as far as data mining models and their outputs are concerned. (Edelstein, 1999) Since data mining involves extracting hidden information from datasets, the understanding process for results can get complicated. As the results of the data mining cannot be predicted, the user does not know beforehand what the data mining process has discovered. Without clear goal setting and mind set on data modelling process, it is rather difficult to understand the output the model produces and translate it into an actionable solution to a business problem. (Fayyad, et al., 1996)

Data mining is primarily used for competitive advantages by companies with a strong consumer focus. (McKinsey & Company, 2012) The focus of data mining applications amongst the business leaders has been steadily evolving from customer analytics to relationship analytics. (Gibbons, 2015)

As the business and consumer marketplaces are turning into increasingly competitive environment, it becomes necessary for companies while attracting new customer, to also retain them. This is particular with normally small percentage of highly profitable customers that generate steady income. Retention strategies are used to promote loyalty for these valued customers. (Gibbons, 2015)

When seeking to maximize customer value, companies build loyalty through brand management and service differentiation as there are only few companies that can rely on economies of scale to sustain competitive differentiation on price alone. This approach highlights the interaction and quality of each individual customer since each interaction serves to either build brand or destroy it. (Harrison & Callan, 2013)

Business leaders use advanced analytics with data mining to optimize their customer relationships. Examples include: improving the effectiveness of marketing campaigns and attracting new customers, maximizing the value of sales to existing customers with cross-selling and up-selling possibilities, minimizing customer loss by predicting churn rates among customers, credit risk scoring, and lifetime value modelling. Data mining techniques are also used to analyse and monitor levels of customer satisfaction and loyalty, and diagnose the causes of changes in these levels. So far, it has been the most helpful in the areas of fraud detection, identity theft, and tax evasion. (Siegel, 2013)

2.2.2 Data modelling process

When designing new predictive models, raw data preparation is the most time-consuming and critical phase. As the models are extremely well thought out, it is rather easy to feed a set of data into the model and generate results if the data that is being fed into the system is good quality. After the data has been fed into the model and the model produces first results, it is essential to understand what the model says about the data in order to validate the results, and refine it. (Fayyad, et al., 1996) In the processing of the data, the data modellers encounter incorrect and null data, rein in outliers, and transfer datasets to identify business issues for evaluation. Another challenge arises when the data modeller have to interpret the results to ensure the output makes good business sense, which in the end is the terminal interest. (Berry & Linoff, 2000)

The conclusive goal of data mining is driving business initiatives, and predictive data mining can be seen as the most common type of data mining since it provides direct business applications from the results. Data mining process contains three stages: (1) the initial data exploration, (2) model building or pattern identification with validation, and (3) deployment of the achieved results. (Berry & Linoff, 2000) (Kamber & Han, 2000) Each of these steps are explained further on.

Step 1 includes the initial data exploration. This phase contains usually data preparation which could involve extracting the raw data, cleaning it, data transformations, and selection of suitable sets of records and finally, possibly carrying out preliminary feature selection. This preliminary feature selection takes place to reduce the number of variables to a manageable range. (Berry & Linoff, 2000) This reducing of variables is performed also in order to handle the data sets with normal computing devices instead of need for super computers to process the data sets. After the data sets have been prepped, depending on the nature of analytic problem, the first actual stage of the process of data mining involves choices of used methods, which can range from simple, straightforward predictors for regression model to more elaborate exploratory analyses using a wide variety of graphical and statistical methods in order to identify the most relevant and interesting variables and determine the complexity and general nature of the models that could be exploited in the next stage. (Edelstein, 1999)

The second step in the modelling process consists of the model building and model validation. This phase incorporates various different models and the decision on the process is to choose the best suitable model based on each model’s predictive performance. This means explaining the variability in question and producing reliable results from the total sample. After the data preparation this might seem simpler operation, but in reality this stage involves more elaborated processes. (Edelstein, 1999) Variety of techniques has been developed to apply the different models into data sets and then comparing the results and performance when choosing the optimal model. These techniques are based on so-called

“Competitive evaluation of models” (Hastie, et al., 2001) and are generally considered the core means in predictive data mining. (Provost & Fawcett, 2013)

The final phase of the modelling process is to deploy the gained results into action. This stage involves exploiting the model selected in the previous phase as the most suitable model and to apply it with new sets of data in order to generate new outcome and predictions. (Kamber & Han, 2000)

After all the stages of modelling process has been taken, the work is not complete, but continues evolving. After the data has been prepped, model developed and tested, and prosperously process deployed in business use, the final aspect of advanced analytics is incremental model revision when new sets of data are available to feed into the model. (Berry & Linoff, 2000) In this aspect also lies the genius of predictive modelling, since it continually learns from the new data. (Provost & Fawcett, 2013)

If the patterns in the data sets change, or more accurate data sources become available, the models are able to capture and adjust to these changes. Every company faces the challenge of being proactive to these changes by constantly monitoring their situation and validating their results instead of being reactive. If the models are not on track and the quality of the results drastically decreases, the model should be re-evaluated and information fed again to the system. After all, in order for the models to be useful, people must believe in them. (Davenport, 2013)

2.3 Synthesis

In this chapter, the model designing process is compiled in the light of the theoretical knowledge learned previously in this chapter and, by doing so, lays out the theoretical framework onto which the actual empirical model will be built on. The framework designed in this thesis is based on the three-step model processing presented earlier in this section, introduced by Han & Kamber among many other models (2000). The greatest difference for designing this modelling process lies in the starting point of designing this model into direct business purposes and keeping the modelling context simple enough for using it in business context by the business person, instead of requiring high-level data-expertise in processing.

This presented model will be the base for the empirical part of this thesis by introducing the stages required to build complete data handling model with this case study and thus answers the first research question:

RQ1: *How to design an analytic modelling process?*

Below in figure 5 the data modelling process has been built as combined from the theoretical framework. This model is called analytic modelling process, AMP, containing three stages.

The modelling process has been divided into three stages by chronological order. These stages are planning, building and implementation. The main phases of the model follow chronological order even though the stages inside the main phases can be viewed as continuous improvement cycle until reached demanded level of accuracy.

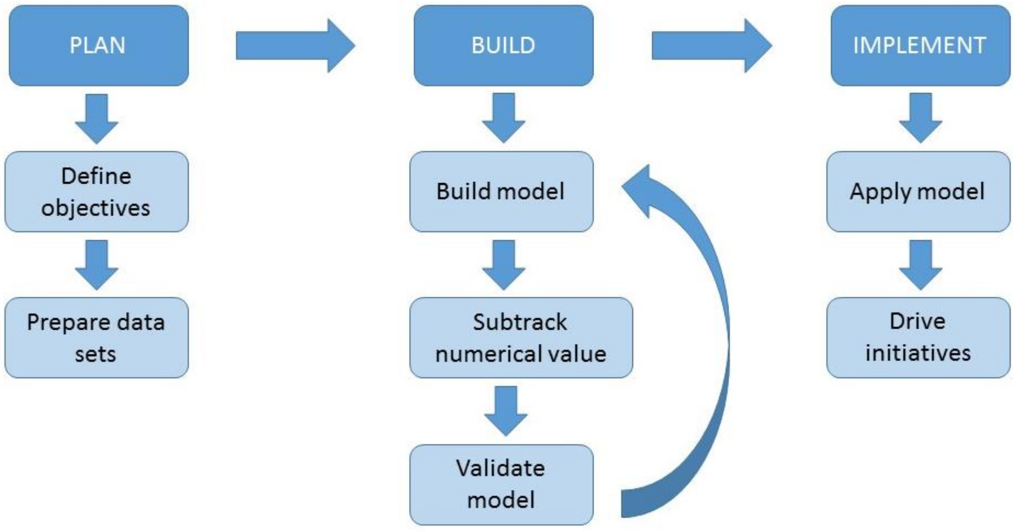


Figure 5. Analytics modelling process

Planning phase

The first phase includes the goal setting for data modelling process and preparing data sets. This is extremely important when handling large quantities of data since without proper focus, it is easy for one to get lost with all the information and different directions that data sets provide. Especially cognitive computing and machine learning automatically find several interesting samples and paths of information that can lead on from the original goal if it is not clearly defined.

Defining the objects includes the data-driven methodological objects as well as clear business objects why this data-driven information is needed, and how it enables better informed business practises. This stage includes also identifying the main sources of data for the next phase.

After the clear objects has been defined, the next step is to prepare the data. Defined objects helps in this phase since these will guide the data selection process as well. While the data- sources and -bases can be in numerous different formats, the goal setting stage determines the selection of main sources based on the expected outcome from the process point of view, not the actual results.

From the time management point of view, preparing the data sets might be the most time consuming phase. This is due to various sources and formats of data that needs to be unified and standardized in similar form to enable data processing and comparison in the later stages. Preparation of data sets includes data cleaning and the initial clustering of data sets based on the set objectives.

Building phase in model design

The second step in analytical modelling process is the actual model building phase. This means that the cleaned and initially clustered data sets are set to be tested based on the primary model version. As stated earlier, in advanced analytics the actual model acts as blue print how to interpret the data sets. The focus is strictly concentrated to the value of the extracted data from the sources instead of the accuracy of the individual features in the model. Like explained earlier, one of the biggest differences between machine learning and statistic lies in the focus. Statisticians aim to produce as accurate model as possible while data scientists using machine learning and cognitive computing programs rely on the accuracy and validity of the raw data.

After the model has been built, the results will be in numeric form. The output can be in visual forms of graphs and networks, but the numerical values can be extracted from the results. After the model has produced these test results, results need to be validated. The validation of the model is an iterative process that aims to refine and fine-tune the gained results with the ultimate goal of making good business sense in the end.

Implementation phase in model design

The final phase of designing and building analytic modelling process is to deploy the gained results into action. This step includes implementing the validated model into substance business aspect. Since the model has been tested in the previous phase and validated until accurate, it can be used in the empirical business environment.

Deploying the model into current, valid data sets generates on-time results that can be used to make better informed decisions and drive new business initiatives to grow business further. The value of clear goal setting in the first phase of the process cannot be highlighted enough, since without proper focus, the end results cover such a wide-area of information that the understanding of what the results even mean might take lots of resources.

3 RESEARCH METHODOLOGY

Every researcher needs to make certain choices on the basic principles of the study. This chapter presents an introduction into the relevant research methodological approaches in the scope of the thesis including research design and research strategy as well as data collection techniques.

Research methods used in this thesis were derived from the case study strategy approach. These methods were the descriptive literature review to form knowledge about existing literature and quantitative analysis of the operational and sales data to harness the theoretical knowledge into empirical context. The main data collection technique for primary data in this quantitative research was extracting quantitative operational data and quantitative transaction data from the case company's ERP-system.

Saunders et al. (2009, pp. 107-108) have created the so called "Research onion" to describe the different level of decisions that must be made when designing an academic research. In the figure 6 below is presented the complete scope of academic research approaches according to Saunders et al (2009, p. 108)

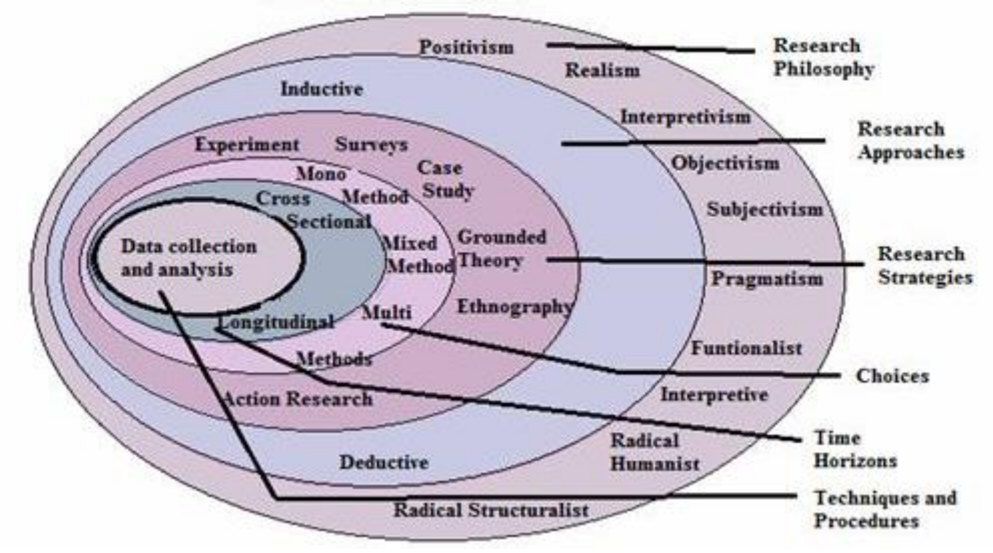


Figure 6. Research onion (Saunders, et al., 2009, p. 108)

Research onion is widely used in academic research since it visually presents the layers that every researcher must decide in clear manners. The outer edge of the onion presents more theoretical and higher level decisions that ultimately define the spirit of the research while levels closer to the core are more practical activities to be performed (Saunders, et al., 2009, p. 108).

3.1 Methodology

This chapter presents the methodological choices undertaken for this thesis following the approach of research onion by Saunders et al (2009, p. 108). Below in the figure 7 is presented the methodology of this research. These steps are presented in decreasing order from top-down.

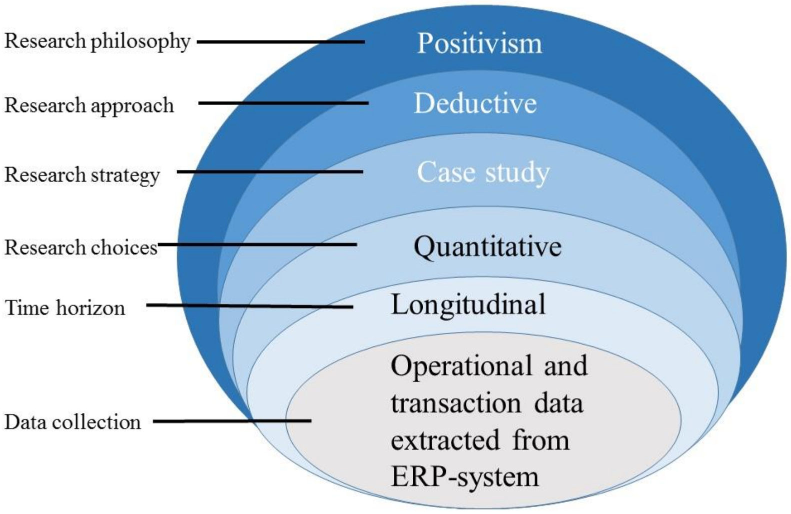


Figure 7.Methodological outline of the thesis

According to Saunders et al (2009, p. 110) and Malhotra (2009, p. 41), research philosophy refers to the development of the existing knowledge based on the increasing knowledge of the topic for the researcher. In other words this means the will that drives every researcher to develop the level of general knowledge in particular field that has been chosen to examine. This theory is used to direct the actions the researcher conducts when outlining the procedure for research design, strategy, data collection and analysis (Malhotra, 2009, p. 41). Philosophical approach for this topic was chosen to be positivism, since the emphasis of the topic is on applying the theoretical framework into empirical situation, testing the theory and, based on the results of testing, developing a broader view about the case situation.

While philosophical approach defines the nature of the thesis, it is not explicit when designing the research In order to indicate the research design more specifically, research philosophy is extended with research approach. (Saunders, et al., 2009, p. 124).

In this research the chosen approach was deductive. There were two main reasons for choosing this approach. Firstly, the intention of the thesis was to test the validity of a model that has been developed based on the existing academic knowledge. And secondly, the aim of the research was to test the validity of the build model, observing the results and, based on the outcome, reach a confirmation for the original hypothesis in a single case environment.

Research strategy is the general plan of executing the research and presents the process how the research is conducted. The strategy includes a systematic approach of answering the research questions (Malhotra, 2009, p. 42). An effective research strategy also helps the researcher to define the most suitable data collection methods to support the underlying arguments. (Saunders, et al., 2009, p. 126)

According to Yin (2002, p. 13) case study is an empirical research strategy focusing on concurrent phenomenon within real-life context. This approach is especially useful when the boundaries between theoretical framework and examined phenomenon are not evident. Case study is also comprehensive research method since it covers the areas from logic of research design, data collection to the data analysis in case related context (Yin, 2002, pp. 13-14). Additionally, Eisenhardt (Eisenhardt, 1989) suggests that case study as a research strategy focuses on comprehending certain dynamics and effects that are related only to the certain business environment or activities.

In this thesis, case study was chosen as the main research strategy since the theoretical framework compiled from literature was to be applied into single company environment. Case study is also featured as preferred approach when the research questions are formed in action-related form “How” or “Why” (Ghauri & Gronhaug, 2010, pp. 109-110). In this thesis both of the research questions are formed with question word “How”.

Research choices are used to define the form of data that will be utilized in the research process; data collection techniques and data analysis procedure. Terms quantitative and qualitative can be differentiated in simplest form by explaining whether the data in research includes numerical or non-numerical (words) data. Quantitative methods include numerical data handling, such as statistical analysis and thus can be measured and compared. In general, quantitative researches have more clarity on what is asked with research questions. Since quantitative approach is based on pure data and relies on scientific methods instead personal observation, subjective judgements or intuition, and when properly performed, it generates results that are objective and statistically valid. (Saunders, et al., 2009, pp. 151-153)

Quantitative methods were the choice of data form in this thesis, since the goal of the thesis was to enhance cognitive computing and machine learning algorithms into the existing data produced by the case company’s ERP-system. The product of the thesis should be network analysis of the case company’s current customer base based on solely occurred buying behavior.

Time horizon of research could be either longitudinal or cross-sectional studies. Both of these approaches are subjected to observations and researcher does not interfere with the subject. Benefits of cross-sectional study are related to the fact that it allows the comparison of multiple different variables at the same time while longitudinal is more focused on how certain aspects of study have evolved over time and can provide definite information about cause-and-effect relationships. (Saunders, et al., 2009, pp. 154-155)

Since customer networks evolve overtime and understanding the current situation properly, it is crucial to study how the current point has been reached in order to evaluate the existing customer relationship. With this in mind, the time horizon fell naturally into the longitudinal category.

The final, and most hands-on stage of the research is to define the data collection techniques with data sources. The data source for this research was based on the case setting environment and was extracted from the case company's operating system. The raw data extracted from the system was then compiled and manipulated into processable form to derive results through machine learning algorithms. The data processing and taken steps are described more thoroughly later on in chapter 3.3 Data processing techniques.

3.2 Research design

In order to replicate the results of the study, research design acts as a blueprint of the study and describes actions taken to perform the research. Smith & Albaum (2012, pp. 16-18) suggest, that the research design is used as a framework and guidelines on how to execute a study and the data collection plan. The research design includes all the practices and procedures for data collection in order to study being repeatable and the results validated. (Smith & Albaum, 2012, pp. 16-18).

The focus of this thesis is to utilize analytic modelling process in retail context. Research design applied in this thesis is presented below in figure 8. The left column describes the stage of the research and it proceeds in progression downwards. Middle column reflects the actions taken in each stage and the outcome of the performed actions is described on the right column.

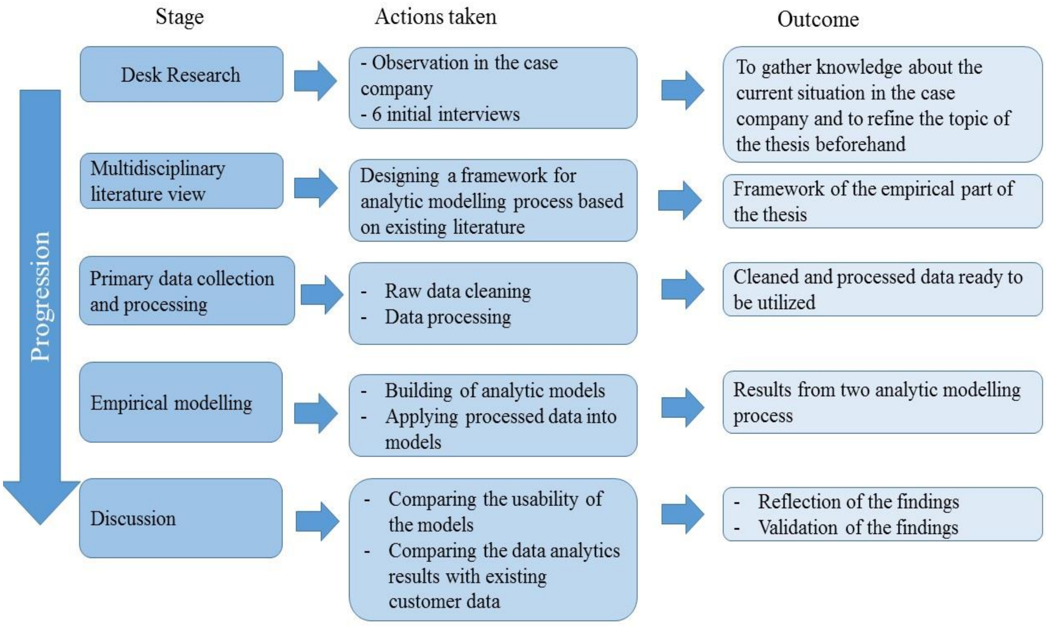
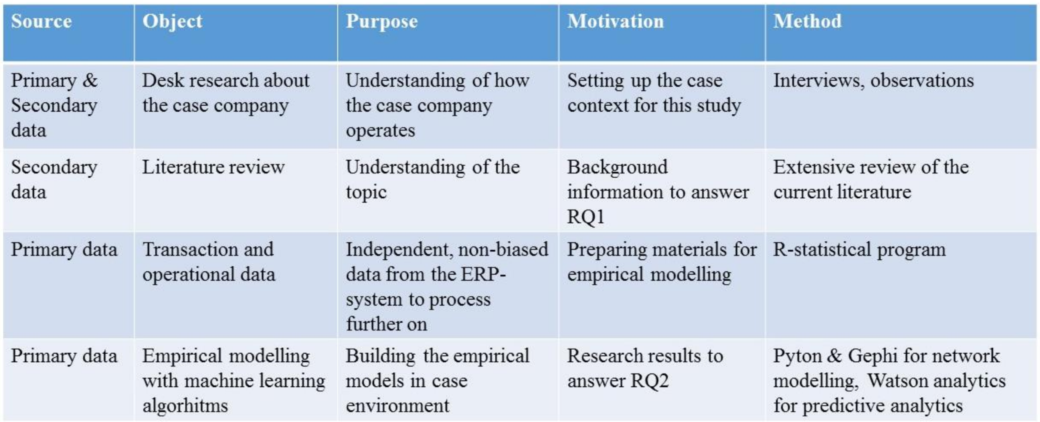

Stage
Actions taken
Outcome
Progression
Desk Research
- Observation in the case company
- 6 initial interviews
To gather knowledge about the current situation in the case company and to refine the topic of the thesis beforehand
Multidisciplinary literature view
Designing a framework for analytic modelling process based on existing literature
Framework of the empirical part of the thesis
Primary data collection and processing
- Raw data cleaning
- Data processing
Cleaned and processed data ready to be utilized
Empirical modelling
- Building of analytic models
- Applying processed data into models
Results from two analytic modelling process
Discussion
- Comparing the usability of the models
- Comparing the data analytics results with existing customer data
- Reflection of the findings
- Validation of the findings

Figure 8. Research design applied in the thesis

The research practices and procedures used for this research include desk study, literature review, primary data collection and processing, empirical model building based on theoretical framework defined earlier in the synthesis of the literature review, and finally discussion of the usability of the built models. Validation of the findings is used to ensure the used data is valid and accurate and that the gained results are reliable.

Data sources used for this thesis are presented below in table 2. The first column defines the type of data collected and second column the object of this used data source. The objects are in progressive order according to appearance in this study. The last three columns gather the purpose of the data objective and why it is valid for the research, what is the output and underlying motivation to perform each stage, and in final column is presented the chosen method to perform these actions.

Table 2. Collected data for the research



Source	Object	Purpose	Motivation	Method
Primary & Secondary data	Desk research about the case company	Understanding of how the case company operates	Setting up the case context for this study	Interviews, observations
Secondary data	Literature review	Understanding of the topic	Background information to answer RQ1	Extensive review of the current literature
Primary data	Transaction and operational data	Independent, non-biased data from the ERP-system to process further on	Preparing materials for empirical modelling	R-statistical program
Primary data	Empirical modelling with machine learning algorhitms	Building the empirical models in case environment	Research results to answer RQ2	Pyton & Gephi for network modelling, Watson analytics for predictive analytics

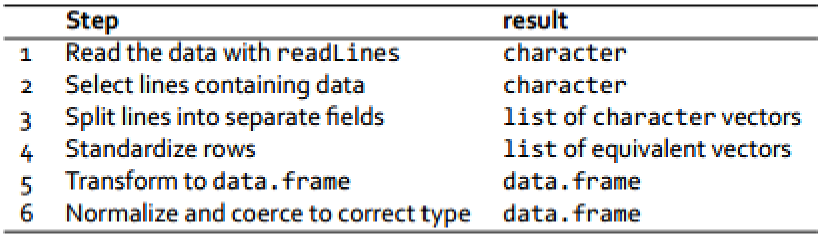
The secondary data source related to this to study were attached to the theoretical part when gathering general understanding of the case company's processes and when reviewing material for building analytic modelling process based on current literature. The desk research utilized both primary and secondary data. The primary data collection in this phase included multiple interviews to create understanding the case company's operating manners and processes. The results of these interviews are not part of the outcome of this thesis but provide general knowledge about the case setting for the researcher. The interviewees were in different operational units in the case company; sales, marketing, operations and purchasing. The extensive literature review covered more than 100 pieces of academic researches, journal articles, studies and other valid literature. The purpose of this review was to gather wide understanding about the current state of literature and act as a context when creating theoretical framework used in the case setting based on academic literature.

Primary data sets used in this study were extracted straight from the case company's ERP-system in several listings. Primary data used in this research included data sets both from operational side as well as monetary transactions. The data was collected between years 2010 and 2014. Before 2010 company undertook large changes in strategy that still took place in 2008 and 2009. 2011 was the first year with state of normality from the business point of view.

Operational and transaction data includes every single sell and purchase between the set periods of time. The extracted raw data included 332 622 entries on the transaction side as well as corresponding amount from operational side. The data was limited to include only Finland between years 2010-2014 so the data set was additionally filtered to line out export data sets. With these in mind, the valid data set included 245 516 data entries.

3.3 Data processing

Data cleaning is the transformation process that turns raw data into consistent data that can be analyzed. (De Jonge & Van Der Loo, 2013) The raw data sets were cleaned by using program R, which is both an environment and programming language applied especially in statistical computing. In this case, the used feature was data manipulation with indexing in order to classify the attributes under correct variables. The aim of this cleaning process was to transform raw data into technically correct, consistent data, where each value can be recognized as belonging under certain variable and that the numeric format is constant across the data set. The used program was R version 3.1.3 (Smooth sidewalk), which has been released in March 9th 2015. Below in figure 9 has been presented the data cleaning steps.



	Step	result
1	Read the data with readLines	character
2	Select lines containing data	character
3	Split lines into separate fields	list of character vectors
4	Standardize rows	list of equivalent vectors
5	Transform to data.frame	data.frame
6	Normalize and coerce to correct type	data.frame

Figure 9. Raw data processing steps. (De Jonge & Van Der Loo, 2013, p. 16)

First phase of the cleaning was to import data sets into the program. This was done in csv.-file format. Second step was to exclude all the export data from the data set as well as remove single-time cash purchasing customers that were marked under general customer group 9999. Finally, the data lines were disconnected into separate fields. After these steps, the customer base for the data set was valid.

The next phase, standardize rows, included adding headers and labels to missing variables. The raw data set was missing some of the headings, and some of the headings were named incorrectly. After this step was taken by correcting incorrect labels and headers, the classification of data entries became simpler since the end locations, under which variable data entries belonged, was rather easy to notice. This phase included also checking the data type and confirming that every row has same amount of fields and that these fields are in right order. It was noticed, that the data types were correctly defined in all sections. After this phase, the data can be said to be technically correct, but this does not mean that the values inside the data sets were error-free or complete.

When the data sets were technically correct, the variables that were needed into this thesis were extracted into separate data set. The variables that remained were: list of items, invoicing date, sales value, cost value, invoiced quantity and customer ID. The complete data set included plenty more variables but when defining the scope of the thesis, these were the variables that were needed to perform the actions that have been taken in this thesis in order to reach set goals.

After the final data set was brought together, the elements were first copied into matrix, and then transformed into data frame. In the matrix form, the revision of the data take turn since the abnormal values were rather easy to identify for the final step.

The last step of the cleaning process was to ensure the values in the data set were constant and in correct form. The first run produced the outcome of data sets that were not consistent and fit with the rest of the data. The data set included some *null* values, *not available* –values and *not a number* –values. After reviewing the data, it was noticed that most *not a number* –values were date entries that were added in incorrect form, mostly typed with comma “,” instead of dot “.”. This was corrected by simply replacing the incorrect values with correct ones. Even though *null* value suggests that there was something wrong with the data, in this case most of the *null* values were by their context correct. In the sales data this meant under *Sales value* and *invoiced quantity* variables that customer either received compensation or sample items with resulted in negative and zero values. These were confirmed from bookkeeping to be accurate. *Not available* values were related under category C*ost value* which is linked to the purchasing in operations. These values could not be tracked to the confirmed purchasing event which produced invalid *Cost value* price for the item. There was one main reason for these values: If item had been returned from customer and re-sold, the original purchase action cannot be traced back to the single item. Since the purchasing values are highly dependent on the buying quantities, the prices were slightly different for each purchase actions, which was especially remarkable to certain suppliers. After these actions, the data set was accurate, complete and ready to be further processed.

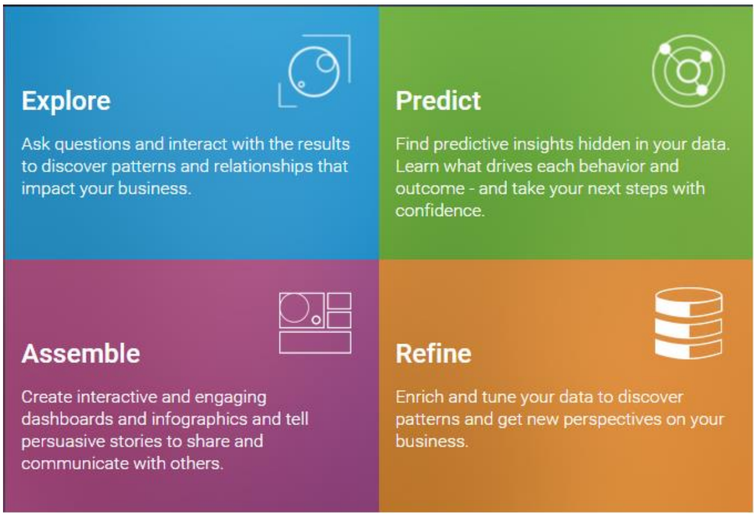
For customer network position modelling computer program Gephi was used. Gephi is an open-source software designed especially for visualization and network graphs. Used version for Gephi was Gephi-0.8.2-beta and it was released in December 30th 2012.

Gephi's modelling is based on algorithms that can be either created by oneself or using ready imported algorithms provided by the program. In this thesis it was decided to use Gephi's own Force Atlas 2 algorithm since it is a force directed layout that simulates physical systems, like it is meant to in this thesis. Force Atlas 2 algorithm has a particularity for degree-dependent repulsion forces that cause less virtual cluttering. In general this means that results can be shown in clear three dimensional network model even though the clutters drift away from each other if the distance is too short. However in this case the repulsion does not matter for the end results since the focus is on what kind of clusters are forming, not on what distance those are to each other.

The network in Gephi consist of nodes and edges. Nodes repulse each other in a similar way as same end of the magnets, and edges function as connecting links between nodes. These forces create a movement that ultimately lead to a balanced state in simulated system. In this research customers and products are defined as nodes and purchased quantity in each transaction act as edges. The bigger the purchased quantity, the thicker the edge appears in the model. Force vectors in this study were defined as undirected in order to examine the results from both product and customer point of view.

The script for the data processing in Gephi was executed by using Python programming language. Used version of Python was Python-2.7.10.amd64 released in May 23rd 2015.

The second analytic model designed in this research was executed by using IBM Watson. Watson is cognitive computing system that has been developed to simulate human mind in business context. This enables both scenario planning with posing "What if" –questions as well as comprehensive data analytics without the need of being a data analytics. Below in figure 10 has been presented the starting screen on Watson with its four main features.



Explore

Ask questions and interact with the results to discover patterns and relationships that impact your business.

Predict

Find predictive insights hidden in your data. Learn what drives each behavior and outcome - and take your next steps with confidence.

Assemble

Create interactive and engaging dashboards and infographics and tell persuasive stories to share and communicate with others.

Refine

Enrich and tune your data to discover patterns and get new perspectives on your business.

Figure 10. Starting screen in Watson

Exploration enables researching the data set that has been imported to Watson. Immediately Watson highlights interesting facts and pinpoints paths that could be beneficial for business to investigate further on. Prediction-section is used to find relevant factors that have influenced the outcome in the imported data set. With Prediction-section key drivers for behavior and outcome can be identified. Watson provides the outcome of the prediction as percentage number about how likely certain variables would occur.

Watson makes it possible for business person in need of support for decision making to make the conclusions from the data by using plain English for asking questions from the data set. The data processing, analyzation and modelling has been targeted especially for business people who can by themselves explore their data sets and find insights into their businesses, without the data analytics or data scientists processing the data as an additional step in the process between. Watson has been designed especially for end-user with easy plug-and-play mindset, which enables the customer, in this case the businessperson in need for data-driven support for decision-making, to find the relevant information from the data with only few clicks by posing natural language question to the data set.

4 CASE COMPANY INTRODUCTION

This chapter introduces the case company and the environment it operates in. In this research the case setting is focused on a Finnish SME operating in industrial markets in retail context. The case company is one of the leading suppliers of industrial filters in Europe, and market leader in Russia.

The first sub-chapter presents the general company information. Second sub-chapter introduces company's product portfolio, customer segmentation basis as well as current data activitites. The last sub-chapter exposes the issues in current situation.

4.1 General information about the case company

Case company is an importer of heavy equipment components and engineering service provider, specializing in heavy filters used in industrial and heavy machinery context. Customers of the case company include original equipment manufacturers (OEMs), machine manufacturers, industrial maintenance companies, machine contractors, machinery rental companies and operators.

The head office of the case company is placed in Finland. Addition to this, the company has subsidiaries in Moscow, Saint Petersburg, and in all Baltic states. Through subsidiaries the case company serves more than 5000 industrial customers additional to Finland's inland customers. The company's most important and fastest growing operating area is export to Russia, which in total covers more than 70 % of the total net sales. In 2013, the company's net sales totaled approximately 21 million euro.

Last years have been tough due to the economic crisis and radical changes in Russia's export opportunities, the effects of which still can be seen. During the last two years the case company's consolidated net sales contracted sharply from previous years. Net sales for the financial year of 2013 amounted to only 15,4 million, a decline from the previous year of 5,2 million euros, in total of -25,2 percent.

While the case company’s net sales were in brisk decline, so was the situation with the overall profit. Total earnings fell down to 1,5 million euros, compared to previous years result of 1,8 million euros. The profit deteriorated from the previous year’s 290 000 euros, in total 15,9 %. According to leading Finnish financial newspaper Kauppalehti, despite the clear decline in profit, case company’s profitability measured by return on equity was excellent in financial year 2014.

From 2014 onwards, the case company was acquisited by british corporation listed in the London stock. The acquiring group is an international group of businesses supplying specialized technical products and services. The acquisition was worth of 80 % of the total shares of the case company including Finland, Russia and Baltic states. With this acquisition the case company was able to expand its product portfolio into industrial seals along with filters. Both of these product categories in the end are fitted with same heavy machinery so these categories actively support each other. The group is now actively seeking further options to accelerate growth and to build substantial, broader based businesses in its core sectors and thus pressures for growing business activities are also highly related to the case company.

4.2 Company activities

The company engages in importing, exporting, manufacturing and wholesale of its complete product portfolio. The product portfolio covers heavy-duty vehicle filters, industrial filters, hydraulic filters, exhaust and fuel filters, hydraulic controllers and machine building components.

In addition to the technical products referred to above, case company provides logistics services and filtering services for use in manufacturing, contracting and resale to companies home and abroad. Increasingly, the case company also engages in consulting and performing design of the above mentioned sectors in order to expand its services.

Case company identifies currently three different customer segments based on their purchasing quantities and purchased items. So far the segmentation has been done only based on the purpose of the purchasing. These three customer groups are original equipment manufacturers, distributors and end users.

Original equipment manufacturers (OEM) are customers who order their own branded products from the case company. The purchased items are identical to normal items on product catalogues but these OEM items are branded with the customer name already from the manufacturer and produced especially for these customers. The OEM-customers also order from rather narrow scope of items compared to other customer groups.

Distributors act as lower level retailers for smaller end customers. These can be seen as partners in business process value chain between the case company and the end customers. The product portfolio in purchasing is wider than with OEM's but much narrower than with the end-users. Distributors segment in buying frequency is weekly active.

End users embrace the largest use of case company's product portfolio. However the average value of single purchase is much smaller than with OEMs and distributors. The purchasing frequency is also more seldom compared to previous groups but still in bimonthly level on average.

Case company's data usage activities are currently extremely low. The data is collected from every point of operations but only a small fracture of the data is used later on. On operational level purchaser uses sales data to decide the purchasing amounts based on the warehouse balance levels. Warehouse balances has been decided by operational manager based on earlier years sales figures.

Case company’s product portfolio covers about 12 000 products, out of which around 3000 products are on constantly available from the case company’s warehouse and the rest of the items are by order. The delivery time of ordering items is highly related to the suppliers but varies normally between 3 working days and 1.5 weeks.

The ERP-system of the case company is Matilda, targeted especially for small and medium-size companies with export activities. Reports exported from Matilda are in numerical form and do not provide sufficiently complete picture of the upcoming business trends or an overall view of the current situation.

4.3 Complications in current situation

There are two main issues identified as complications in the current situation. These key issues were used as the starting point when developing the goals for the empirical part of the thesis. The key issues are related to 1.) Inaccurate customer segmentation and 2.) Inefficient data usage in decision-making.

Case company currently identifies only three customer segments and these segments are clustered based on solely their business purposes. When acquiring new customers, the classification of the customer class is decided based on to the activities of the customer company. Customer segmentation approach that emphasizes the value of each customer and appoints the value of individual customer in the customer network position enables case company to individualize their customer segmentation process based on actual, realized purchase actions.

Data usage in the case company is currently re-active. Data is being collected but it is not used actively to support decision-making. With the acquisition of the case company and new growth strategy, analytics has been appointed as one of the areas of interest. With data analytics company seeks internal growth as well as identifying new business opportunities.

The current ERP-system produces valid data, but this data is raw data that needs further processing before it can be efficiently utilized in decision-making processes. The data is also stored in separate location that requires further access rights.

The tools that could be used to utilize this data should be easy to use since as an SME, and by company profile, the case company does not have separate data analysts and at least currently it is not relevant to have one. The used solution must be easy to use from business user point of view and provide sufficient information to make more informed decisions.

5 EMPIRICAL STUDY

This chapter presents two analytics modelling processes based on the theoretical framework designed and presented in the chapter 2.3 Synthesis earlier. The designed framework is used to establish solid basic structure for designing an analytic modelling process for the needs of this exact case company.

The scope of the thesis, which is again presented in the figure 11 below, includes both diagnostic analysis about the current situation of the case company as well as testing the extracted data sets into predictive, advanced analytics computing systems in order to identify what are the key drivers influencing the outcome. Since the aim of these two outcomes differs greatly from several aspects, it was decided to design and build a model for each of these objectives. Both of these modelling processes are presented in the following chapters 5.1 and 5.2. The results gained by using these models are then presented later on in the chapter 6. Results.

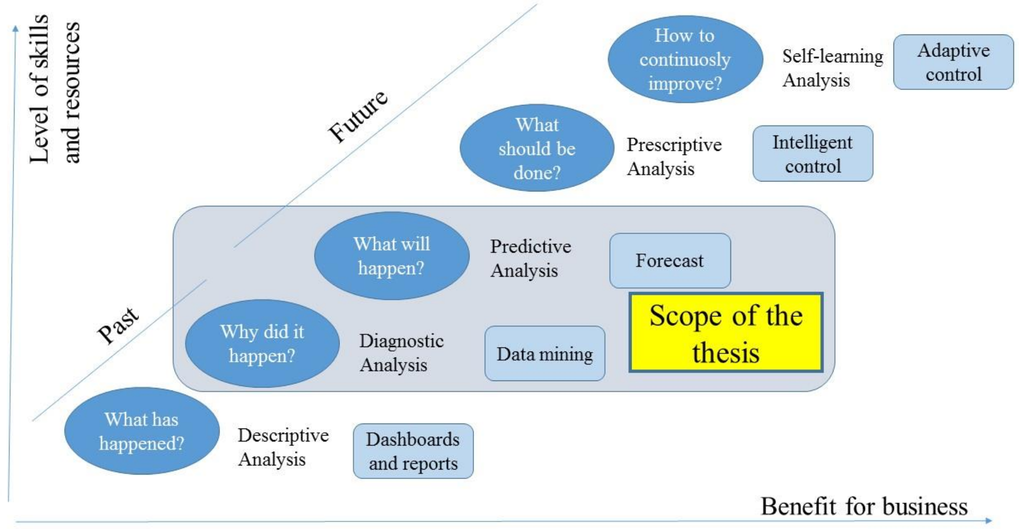


Figure 11. Scope of the study

5.1 Designing diagnostic model

The first model is used to examine the network position of case company's customers based on their transactions both from monetary point of view, meaning the value of their purchases, as well as quantitatively, based on the amount of items customers purchase products. The model processing steps are presented below in figure 12.

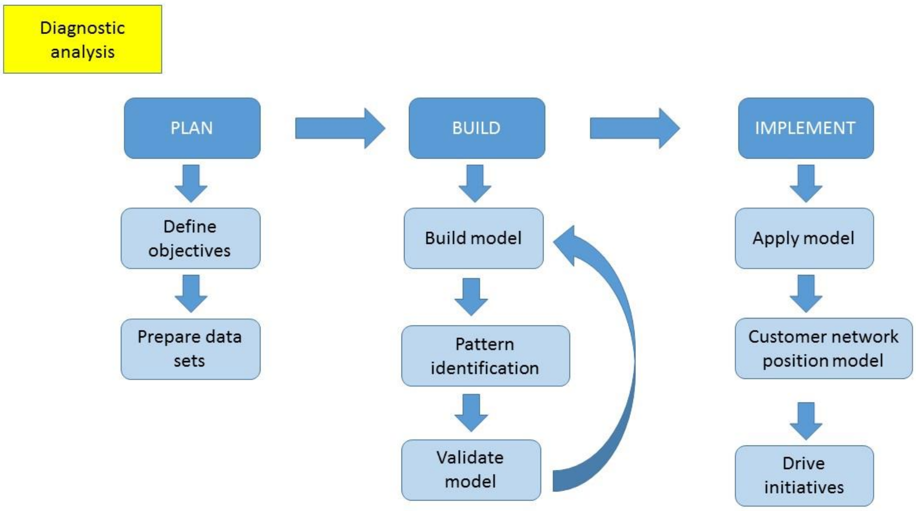


Figure 12. Modelling process for customers network position

Planning

In the first stage, planning, the object of this model was decided to identify the current customer segments based on solely their buying behavior, since the case company did not hold valid or accurate knowledge about their customers' individual buying behavior. The empathy was given for purchased quantity of the items as well as the total value of the purchases.

The data sets that were cleaned earlier in the pre-processing phase from incorrect entries, outliers, and other errors were modified to suit for this purpose. For building a customer network position model, the important variables needed to replicate the model from the transaction data were customer ID’s, products, price of the items, quantity of purchased items and finally, if the results were to be demonstrated with simulation as how the network has developed over the time, the date of purchase. All these variables were listed into one csv.-file from the transaction data in chronological order. Since customers also have different price classifications based on their value from business relationship point of view, the csv.-listing was carried out item by item in chronological order to match the used price of the item with correct customer.

Building

Building phase includes initial building of the model as well as testing the preliminary results by validating the model and then adjusting the model if needed. Having the data sets being prepared, the model building stage started with incorporating the data sets into the program used for modelling.

The modelling for creating a script for the actual model was done by using programming language Python and the actual outcome of the modelling was extracted by using computer program Gephi. In order for Gephi to read the csv.-file and create the connections between each data entry both with customers and items, Python programming language was used, in cognitive computing terms, to teach Gephi-software how to read the csv.-listing and what actions to perform in different situations. Basically Python was used to create the framework for the network and then Gephi was used to visualize the results in easier to understandable form and actually identifying the relations between nodes, which in the network were both customers and items, while the connections, edges, were the purchasing quantities for each item. Colors in the model are defined in the network script for the Gephi in order to separate the clusters more effectively. The starting point colors from the outer edges of the framework were red, green and blue.

Since Gephi uses machine learning algorithms, purpose to use it was to discover hidden patterns in transactional relationships between customers through the case company’s product portfolio. The outcome of Gephi-model was three-dimensional network model presenting the position of each customer. The customers were identified based on their purchasing activities and numerical clustering was performed to map similar customer profiles.

The validation of the model was tested manually in order to determine if the identified customer clusters were expected to behave similarly. Since the data material used for this thesis is at the core of the case company’s business, the manual handling was a prerequisite in order not to reveal any compromising customer information. During the validation the model was updated slightly and all these changes took place only to the data sets, since the model is based on machine learning algorithms that learn from the fed data sets. After the model was concluded valid, it was ready to be used for the implementation step.

Implementation

The final stage of analytical model designing process is to implement the built model into business use. After the customer network position model has been successfully tested and validated, the complete operational data can be fed into the system.

The total cleaned data set included all the Finland’s transaction data from 2006 and since the company went through major changes in year 2008 and 2009, it was decided to define the starting point for this phase onto year 2010. Especially year 2009 involved significant changes in company’s strategy resulting in alteration also in company’s product portfolio as well as in customer base, and in year 2010 the changes in business approach were still in developing phase. Year 2011 was the first year when the case company functioned with almost its current product portfolio and similar strategy for customer portfolio management. With these starting points, there were 245 516 valid data entries to regard for this modelling process.

Results of Gephi’s network modelling can be extracted from the program in numerical file, which in this case was done using excel. Since the network model was designed with the object of examining the current customer positions based on their purchase actions that have actually taken place, it was expected that the algorithm used in this model would provide certain classifications and thus clustering and segmentation would be based on given variables; product items, quantities, value of purchases and purchasing dates.

The beauty in these machine learning algorithms, which are also used in Gephi, is that the model can be developed further based on newly fed data into the system. For example in this case, it was possible to extract network development model in video-format to study the development of the customer network positions. This was added into the program by including the exact dates of purchases into the item specific data to enable sequential assessment. Also when new data is available, the transformation and development of the network can be easily reviewed.

In this stage, the results of this modelling process were ready to be analyzed and used to drive further business initiatives to grow the business. The network position modelling provides solid stand and overview for the current stage of business as well as how that point has been reached. After gaining genuine understanding about the present situation and where the company stands currently, can the viewing point be changed into the predicting future actions.

5.2 Designing predictive model

The second model is used to examine the evolution of gained results from network position of case company's customers, as well as finding indicators that can be used for predicting future events. Predictive analytics aims to find the indicators that influenced the appeared actions and behavior.

The result in the network model produced customer segmentation model, which is further analyzed using advanced analytics program IBM Watson to find predictors from the existing data set. The model processing steps are presented below in figure 13. The stages of the predictive modelling are identical to diagnostic modelling, however the input inside the main phases differs. When network analysis focused on current situation and how it had been reached, this model aims to find important patterns and indicators upon which future behavior can be estimated.

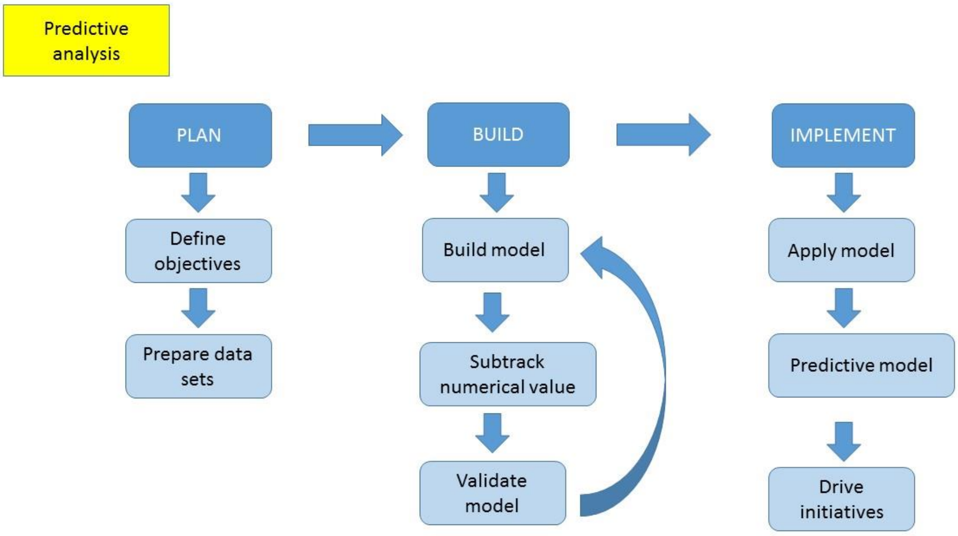


Figure 13. Modelling process for predictive analytics

Planning

In the first stage, planning, the object for this model came from previous modelling process and the focus was continuing the process by making the next step. The aim of this predictive, or advanced analysis was to identify meaningful, hidden patterns and features that rely heavily on the current and past behavior of the case company's customers. The target feature with special interest was naturally recognizing features that enable predicting customer behavior, since in the end that is the root of all actions taking place in these model, and in business in general. Target for this model was set into identifying predictors for sales.

The data sets that were used earlier were again being used in this phase. Since the original transaction data that was cleaned in the pre-processing phase, includes plenty of more information than just the entries used in the network analysis, it might provide additional information about the underlying causes for the sales and thus these data sets were used additionally. However since these listings are extensive in file size, it is not feasible, or even possible with normally operating computers, to process or fed this complete data into Watson. It was decided that the complete operational data was focused to the entries covering complete year of 2014. Watson understands both ready excel-files as well as csv.-listings so in order to decrease one extra data handling and transformation step, it was decided to use the existing csv.-files from which unnecessary columns and entries outside the defined target timeline, meaning all the entries before year 2014 and all the entries after 2014, were simply deleted by using normal csv.-editor.

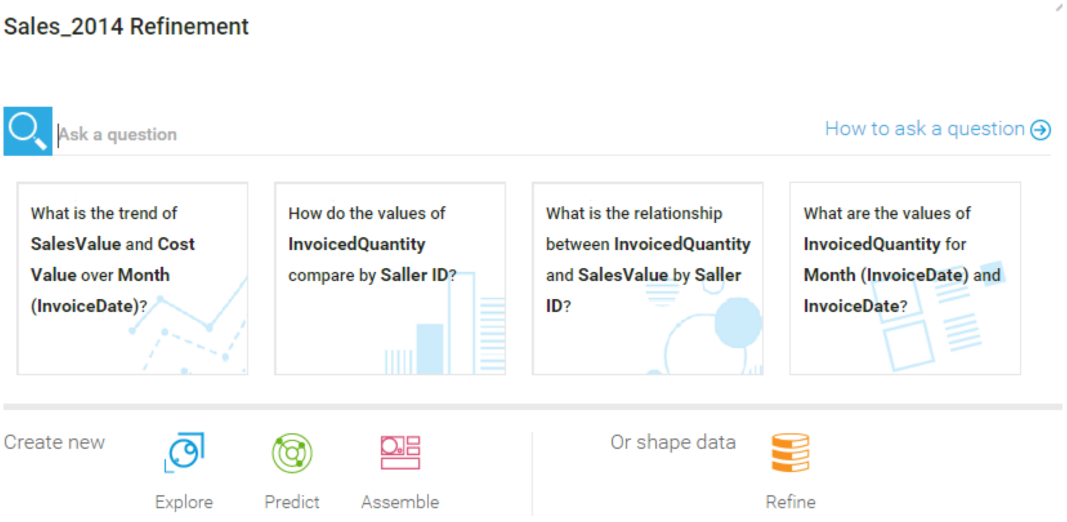
Building

Building phase includes initial building of the model as well as testing the preliminary results by validating the model and then adjusting the model if needed. Having the data sets being prepared, the model building stage started with incorporating the data sets into the program used for modelling.

As well as with Gephi in the previous model, predictive model is also machine learning, and even taken further, cognitive computing system. This means that even though the data will be fed into the system and the results need to be validated, the cognitive computing program produces the actual model. Compared to Gephi, Watson is already such a civilized program that it does not require special data prepping, excluding the data formatting into a file type that Watson can automatically process.

When data was fed into Watson, it automatically evaluates the overall quality of the data sets. This is based on the sheer amount of data in order to find patterns as well as evaluating the quality of the data entries in the file, for example how much outliners and null values the data set includes. The data quality score measures the degree to which the data is suitable for predictive analysis. It is an average of the data quality score for every field in the data set, as determined by missing and constant values, influential categories, outliers, imbalance and skewness. The data set is ranked in percentage number from 1 to 100 %.

After uploading the data set into Watson, it immediately starts processing the data. For example from the sales data of 2014 it pinpointed automatically interesting selection of paths to discover as presented in below figure 14.



Sales_2014 Refinement

Ask a question

How to ask a question

What is the trend of **SalesValue** and **Cost Value** over **Month (InvoiceDate)**?

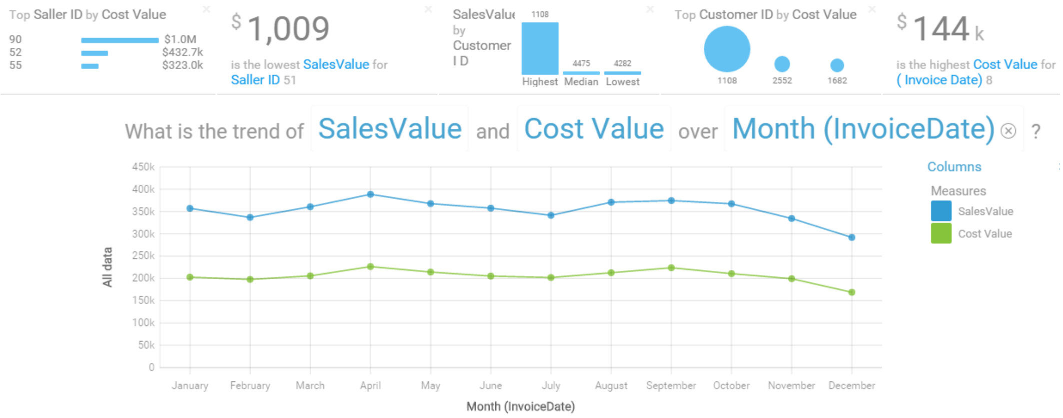
How do the values of **InvoicedQuantity** compare by **Saller ID**?

What is the relationship between **InvoicedQuantity** and **SalesValue** by **Saller ID**?

What are the values of **InvoicedQuantity** for **Month (InvoiceDate)** and **InvoiceDate**?

Figure 14. View from Watson after data processing.

When choosing any of these ready-made paths, or asking other questions in plain English, Watson provides visually easy to understand figures by extracting numerical value. From the provided reports about the current state, it was rather easy to check from the case company's internal reports how well the model built matches with the actual event. Below in figure 15 has been presented the trend between sales value of certain item and its comparison to the purchasing value over the year 2014. Watson also proposes additional interesting details related to the presented figure on the top. In this case it showed top seller of this chosen item, lowest sales value that has been invoiced from certain customer as well as the biggest buyers for this chosen item in this time period.



$ 1,009 is the lowest SalesValue for Saller ID 51

$ 144 k is the highest Cost Value for (Invoice Date) 8

What is the trend of SalesValue and Cost Value over Month (InvoiceDate) ?

Figure 15. Trend between sales and purchasing value for chosen item

Validation of the data was again done manually since all the identifying numbers used in this data set have been encoded in order to prevent any identifications. The validation resulted into extremely similar, almost identical conclusion as the numerical report provided by the case company's ERP system.

Implementation

As with the diagnostic model, the final stage of predictive analytical model designing process is to implement the built model into business use. After the operational sales data and the clustering results of the customer networking analysis has been successfully tested and validated, the model can be harnessed into real business use to search hidden patterns and insights.

The original data sets included all the sales transaction data between years 2011-2014. With these starting points, there were 245 516 valid data entries to regard for this modelling process. The number is same as in the first modelling but these entries hold more information since the number of columns is bigger. As it was explained earlier, it was not feasible to feed all this data into the system since, firstly, the program would slow down considerably, and secondly, normal PC or laptop does not hold enough processing power to handle this without the fear of machine breakage. Also when looking from the data mining point of view, when searching for meaningful insights and patterns, it was decided to focus on the data from year 2014 since the development of the customer network from the previous years was already provided by the first model. In the end, the first data set fed into Watson included 97 025 data entries.

The second data set was the result data set from the network clustering. This data set was considerably smaller, since single purchasing customers were lined out. This was due to the fact that the single purchase customers cannot be reliably modeled as there simply does not exist enough valid data to create behavior patterns. The second data set including the clustering information included 8666 data entries.

Results provided by Watson can viewed in several different forms. Program itself provides clear visual results which can be also extracted in numerical file form from the program for example in excel-file if further data examination is needed. Program also includes data presentation tools by providing possibility to create own dashboards and presentations from the data sets.

Since the predictive model was designed with the object of finding variables that would influence the future behavior and which could be, after further analysis be identified as indicators, expectation of the results was some kind of predictive score as it was presented in the literature view as the outcome. The predictive score presents the likelihood of certain indicator to have influence on the determined variables.

In this stage, the results of this modelling process are ready to be analyzed and used to drive further business initiatives to grow the business. This level of data processing produces already quite extensive results on the company’s competition abilities and market position by pointing out the strong areas as well as the weaker ones if the data sets are comprehensive.

Similar with the results from Gephi, these machine learning algorithms can be developed further and more accurate based on the available data. These cognitive computing systems are constantly developing to bring the fore-front of data processing to the hands who really need it in a form that does not require special data-scientist skills to handle the data or to understand the results but rather interested mind and excited user attitude to gain complete usage of the machine learning programs.

6 RESULTS

This chapter presents results from the models designed in the previous chapter. The chapter has been divided in three sub-chapters presenting results separately from both models; customer network and predictive analytic model, and finally a synthesis about the usability of these models from the business user perspective. To understand the value of each used model, both models were compared to each other to find the relevance for using these analytic modelling methods in retail context from the business user point of view.

6.1 Customer network position

Aim of the customer network position was to identify case company's customer positions in clustering model that was build based on customer's realized purchasing activities. Colors in the model are defined in the network script for the Gephi in order to separate the clusters more effectively. Same color clusters are similarly behaving customer groups and the closer the clusters are together, the more similar products the customers in these clusters purchase.

The connecting lines present the purchasing quantities. The thicker the line is, the more of these products the customer in the line purchases. Used algorithm for this model was Force Atlas 2 and due to the attributes of this algorithm, the spread of network clusters in this picture is not relevant. On the other hand, the more clusters are connected to each other, the closer the customer profiles for these clusters are. Network model for customer's evolvement over time resulted in network graph presented below in figure 16.

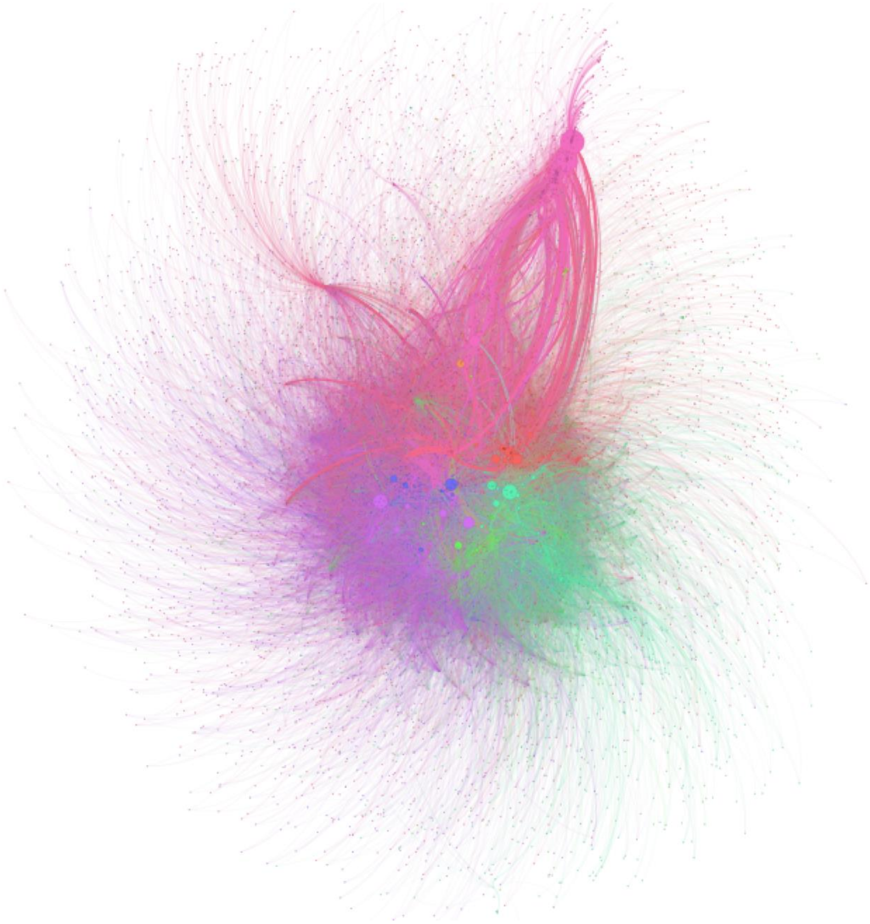


Figure 16. Customer network position based on purchased quantities

Each individual spot reflects one customer or one product. After the test run and model validation, in the customer network analysis 16 differently behaving customer clusters were identified. From the figure 16 it can be seen that these customer groups are un-evenly spread and clearly one group differentiates from the rest of the network to the upper-right in pink colour with large purchasing quantities since the connecting lines are a lot thicker than the average in visible picture. On the center of the main body several customer clusters can be observed, differentiated with the color codes. These clusters are marked with red, blue, purple and different shades of green.

The clusters in the figure 16 are formed solely around the customers, instead of products in this model since the main purpose of this model was to examine customer's evolvement. As it can be seen, there is a concentration on the middle of the figure 16 but that does not have a round circle shape around it since this spot presents a single item. Same situation applies to the concentration on upper-left from the main body of the formation. Both of these items seem to act as important access points in the network and should be seen as items that draw customers into the network. However when manually confirming the sales amount for these items, neither of these items is ranked among top 30 most sold items or classified as level A items with ABC-classification that is used with stock balance monitoring in the case company.

The first outcome of the network produced the customer position based on purchased quantities. The second run was performed with adding the monetary value of each purchase into the model. The results of this run can be seen below in figure 17.

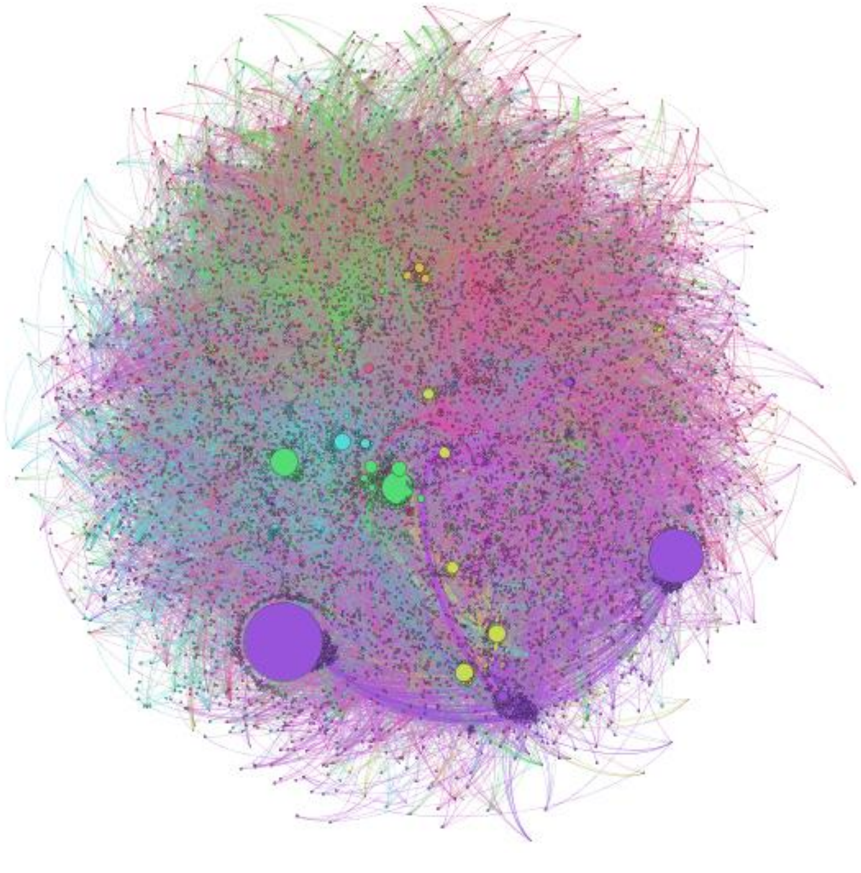


Figure 17. Customer network based on total purchased value

This figure was taken from slightly different angle than with the previous run with figure 16 in order to illustrate the network as accurately as possible. The network changed noticeably when adding the value of each purchase into the script. The main body of the formation is much tighter than with the previous run but the distances between individual customers remain.

The reason of the shift in network when adding the monetary value of transactions is related to the total value spent on single purchase. Because of the attributes of algorithm Force Atlas 2, each edges acts like a spring. The difference to the previous phase when only purchased amounts were examined, lies in the position that in this phase the edges, when buying bigger amounts on single purchase, become tighter, since the impact for the whole network is also bigger. These edges are pulling these concentrations closer to the central body of the network.

The colours used in the round circles are used to describe the similarity of individual customer's buying behaviour. The model using monetary value addition to the purchased quantity highlights the importance of individual customer while the previous version was more focused on the purchased items. Naturally since the value of each purchase is important variable to measure, this model presents the power inside the network of the customer positions.

These models provide an access to extremely deep level of customer relationships in the network positions. It is possible to observe and learn from a single customer point of view, what it the meaning for the network and how this situation has evolved over time. In figure 18 below, the customer network has been enlarged and applied with customer labels.

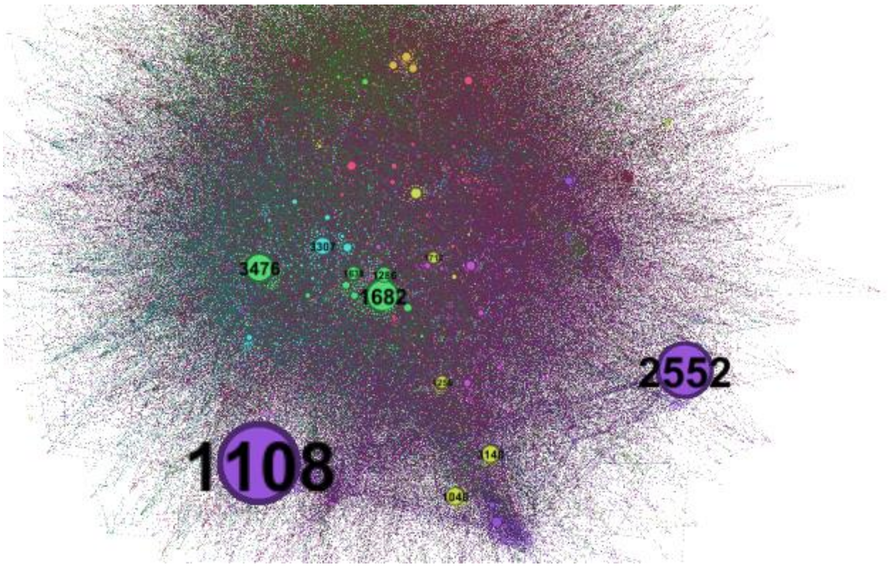
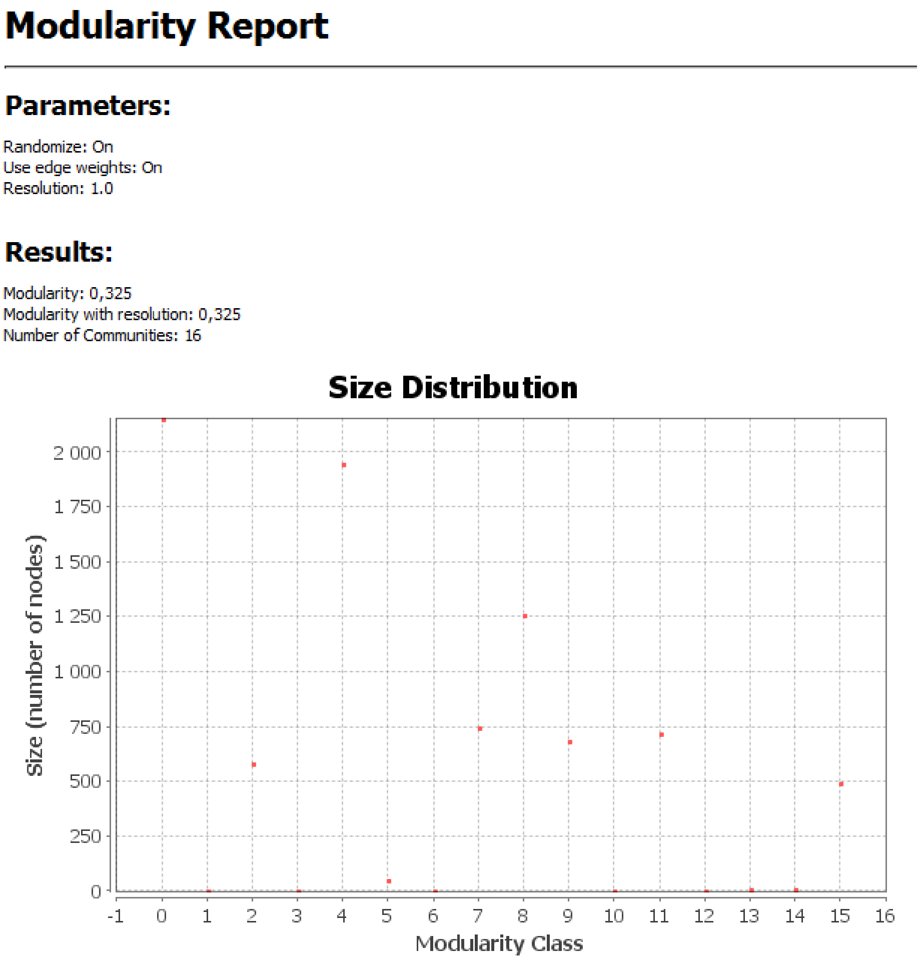


Figure 18. Enlargement of customer network with customer labels

Figure 18 includes the customers and products. As can be seen, the closer to the network the viewer approaches, the clearer the connecting edges between each customer and items appear. This can be used to observe and identify customers behaving similarly in same product environment as well as identifying the key items for each customer. When simulating the customer relationships evolvement over time, changes in buying behaviour can be noticed and further analyzed for example if the machinery base of the customer has changed.

Figure 16 presented the network model run through Gephi-software solely focusing on quantity of the purchased items and figure 17 presented the same model added with the value of each purchase. Like stated earlier, in total of 16 customer cluster were identified. These 16 customer clusters are spread over in the network model, as can be confirmed from the modularity report, which is presented below in figure 19.



Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0,325
Modularity with resolution: 0,325
Number of Communities: 16

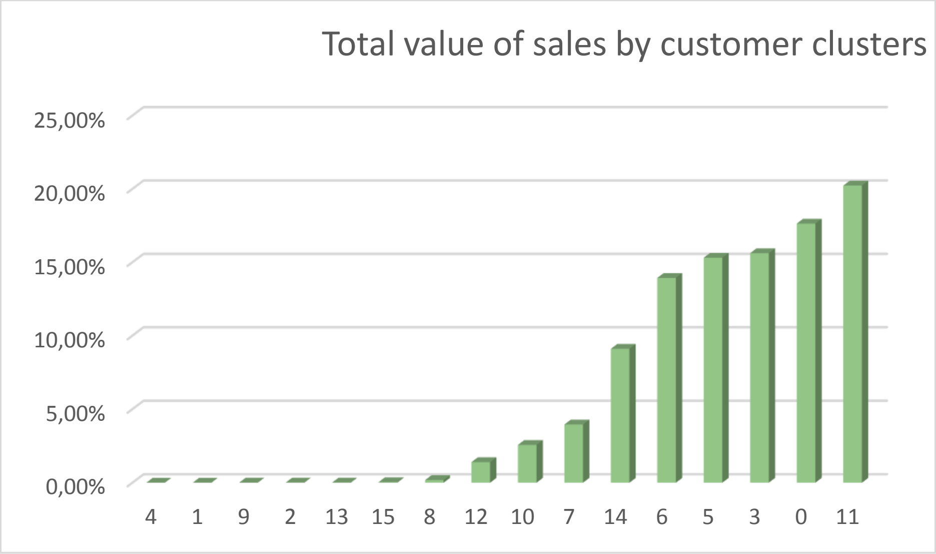
Figure 19. Results from modularity run

The x-axis describes the identified modularity class, in total of 16, and the y-axis presents the total number of nodes connecting this customer to the products. Parameters define how the modularity was defined. Since it was expected that there would be some larger customer clusters as well as smaller ones, modularity class would need to provide equally accurate, and realistic outcome describing all the network clusters adequately. Due to these reasons, this network model was produced by using modularity resolution 1.0.

In this case the distribution of the customer clusters around the table is random, meaning the order from one to 16 is based on the formation of the cluster over time and when it was observed. For example, according to this report, modularity class seven, meaning the customer cluster number seven, evolved over time and was observed before class eight. As can be seen from the figure 19, some of the modularity classes are placed on extremely low level compared to some other classes. For example classes 3, 10, 12, 13 and 14 are at the lowest level on y-axis while on the other hand groups 0, 4 and 8 stand out from the figure.

The customer clustering listing was extracted from Gephi and the value of each customer and cluster was compared to the total value of sales. The results of this comparison can be seen below in table 3.

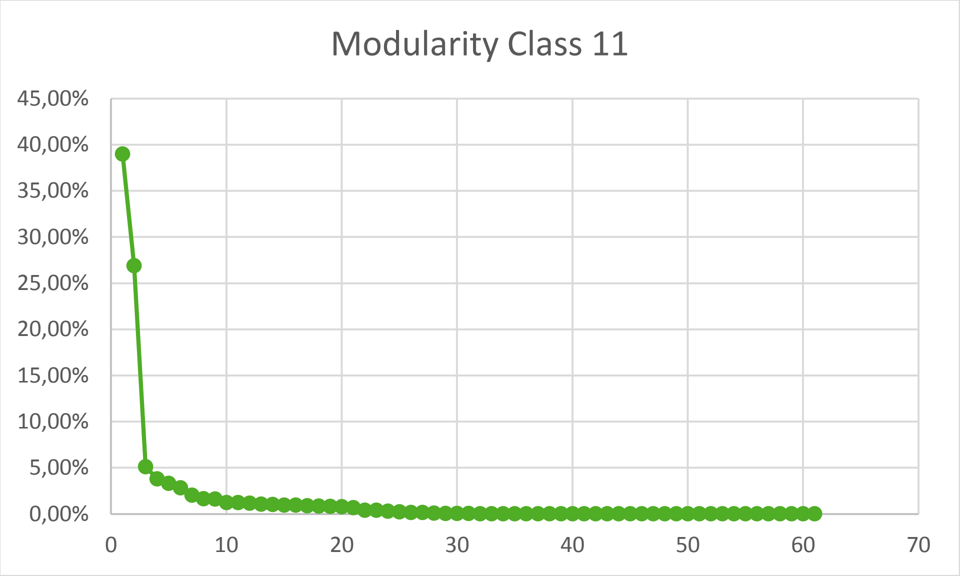
Table 3. Total value of sales by identified customer clusters



The customer clusters were sorted from smallest to largest in order to adequately compare the gained results. From the table it can be observed that the clusters 5, 3, 0 and 11 cover around 67 % of total sales alone. The most valuable cluster from monetary point of view is cluster number 1 covering almost 20 % of the total sales. Comparing to these, the first clusters; 4, 1, 9, 2, 13, 15 and 8 play only a small role in significance. In order to fully harness the results of this segmentation, the decision must be made what are the valid customer clusters to further investigate in short and long term.

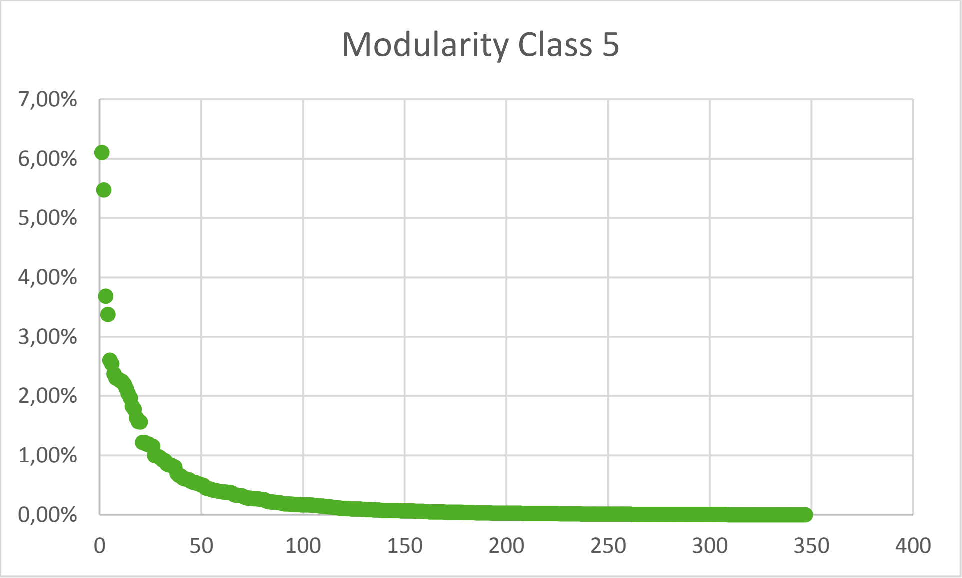
The general customer clustering reveals how many customer concentration can be found in the network. However this is not yet adequate to define the customer behavior and to gain actionable insights. When exploring deeper into the customer clusters rather interesting trend can be seen with the distribution of customers inside the segments. Below in table 4 is presented the distribution of customers from modularity class 11, which in the previous phase resulted as the most value deriving cluster, over purchased amount during the year 2014.

Table 4. Customers in modularity class 11



Modularity class 11 includes 61 customers in total, which have been identified having similar buying behavior. As it can be seen from the graph, only first three customers generate almost 70 % of the total value produced by this cluster. This same patterns emerges also in other modularity classes. Below in table 5 are presented the results from cluster 5.

Table 5. Customers in modularity class 5



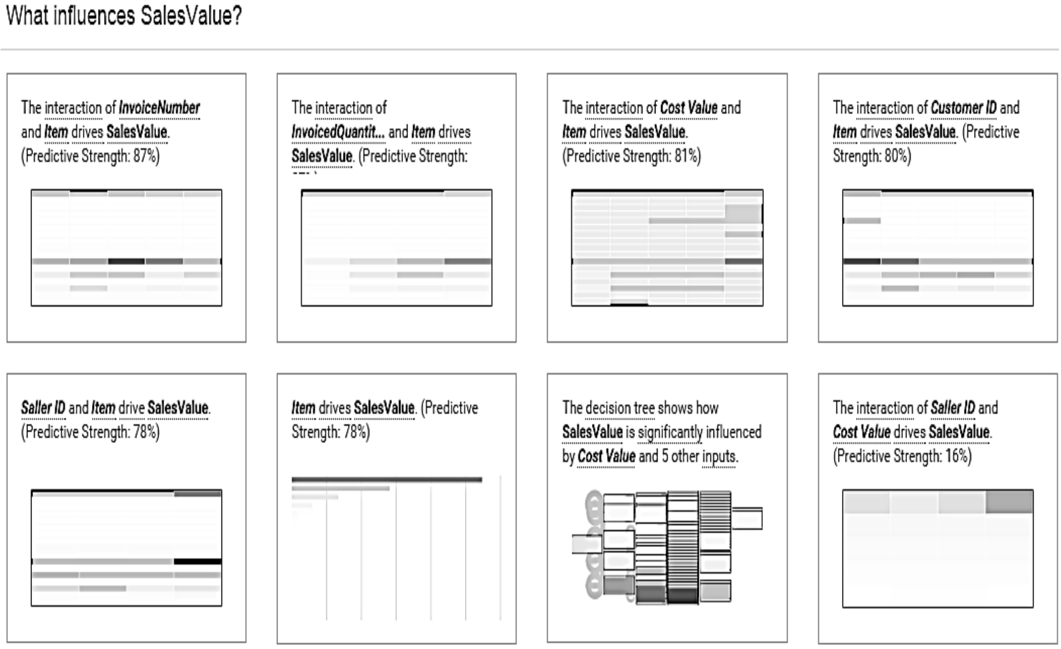
Customer cluster 5 includes total of 346 customers. Even though the distribution of customers over the table is more even than with the customer class 11, the same pattern repeats of having rather small part of the customers generating most of the revenue from this cluster as well.

In order for the case company to gain full understanding on their customer position, each of these modularity classes need further investigation. While naturally all customers are important for the business, the priority of targeting marketing and sales efforts between customer clusters and even individual customers differ radically.

6.2 Predictive analysis

The target of the predictive data analytics was to utilize and evaluate the usability of the predictive modelling. This was done by case framework set to find the key drivers for predicting sales value from the existing data set. The data set imported to Watson included 10 fields of variables; Date of purchase, item, invoiced quantity, sales value, cost value, customer ID, seller ID, invoice number, invoice line number and warehouse location.

When posing a question to the data “What influences the sales value?” the prediction model identified in total eight correlations on top predictors for sales value. The outcome correlations of posing the question to Watson is presented below in figure 20.



What influences SalesValue?

The interaction of ***InvoiceNumber*** and ***Item*** drives **SalesValue**. (Predictive Strength: 87%)

The interaction of ***InvoicedQuantit...*** and ***Item*** drives **SalesValue**. (Predictive Strength:

The interaction of ***Cost Value*** and ***Item*** drives **SalesValue**. (Predictive Strength: 81%)

The interaction of ***Customer ID*** and ***Item*** drives **SalesValue**. (Predictive Strength: 80%)

Saller ID and ***Item*** drive **SalesValue**. (Predictive Strength: 78%)

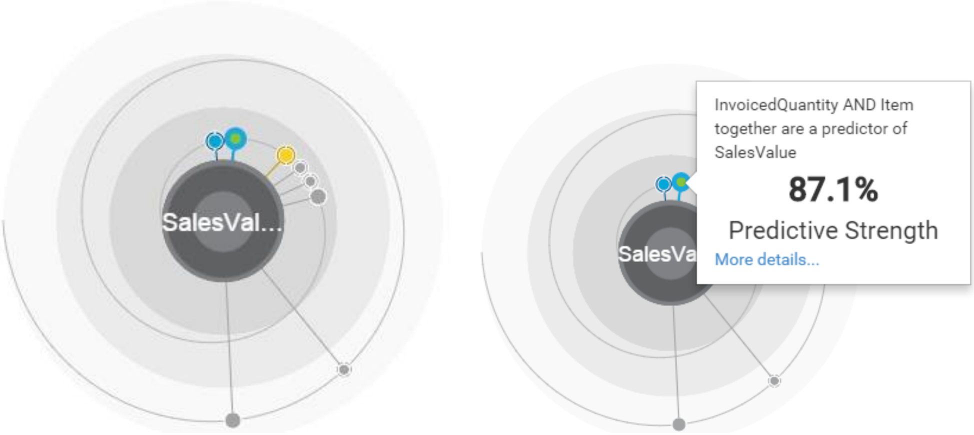
Item drives **SalesValue**. (Predictive Strength: 78%)

The decision tree shows how **SalesValue** is significantly influenced by ***Cost Value*** and 5 other inputs.

The interaction of ***Saller ID*** and ***Cost Value*** drives **SalesValue**. (Predictive Strength: 16%)

Figure 20. Correlations for sales value

The distribution of these eight factors is presented below in figure 21. The left side of the figure presents the overall distribution of the predictors and on the right side one field has been highlighted as an example how the model notifies about the results. In total, the model identified six predictive factors with predictive score over 75 %, while the remaining two identified factors were both with predictive score below 20 %.



InvoicedQuantity AND Item together are a predictor of SalesValue

87.1%

Predictive Strength

More details...

Figure 21. Top predictors for sales value

The first note on the figure 21 is that the interaction between invoice number and item drives sales value with predictive strength of 87.2%. The second interaction is the one in the example in the figure 21 proposing that interaction between invoiced quantity and item together drive the sales value with predictive strength of 87.1%. The third strongest prediction was between cost value and item together resulting in predictive strength of 80.6%. Interestingly, the customer ID and item present the fourth strongest predictive feature with level of 79.8 %. Following on the next place is the interaction between seller ID and item with predictive strength 78.1%. Item alone is the sales value predictor by 77.9%.

So far all the fields have had only one or two fields correlating each other. Watson also provided predictive combination of fields with significantly lower predictive strength. The seventh predicting combination of attributes included cost value, seller ID, customer ID, invoice number, item and invoiced quantity with predictive strength 17.9%. The eight and weakest recognized predictive feature was interaction between sales value and cost value with predictive strength of only 12.9%. This formed an interesting feature since this proves the statement, that while the purchasing are done at same time for the whole demand, the pricing for each customers varies.

From each of the predictions factors, further information can be investigated. Below in figure 22 is presented the interactions of the data for predictive strength between invoiced quantity and purchased item, which had predictive strength of 87.1%.

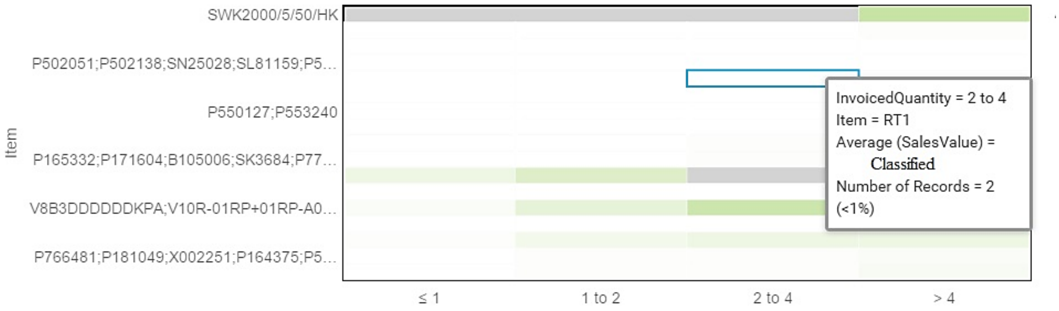


Figure 22. Interaction between invoiced quantity and item

Figure 22 maps the average invoiced quantities for all the items that have been purchased. In this case item RT1 was chosen as an example item. The information box notifies that with invoiced quantity 2 to 4 items, there has been only 2 records of purchase, which in total is less than 1 % of the total records. Normally, the average sales value would appear into the information box as well, but since that is classified information it has been removed. In figure 23 below can be seen the progression of sales value over invoiced quantity with the items included into the previous figure 22.

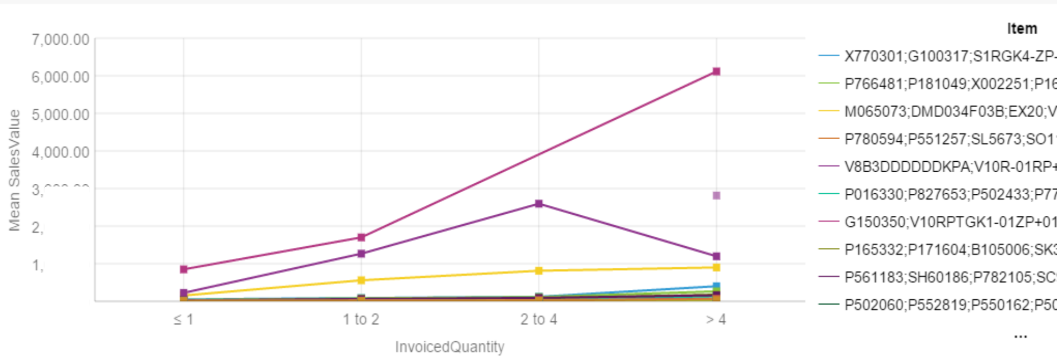
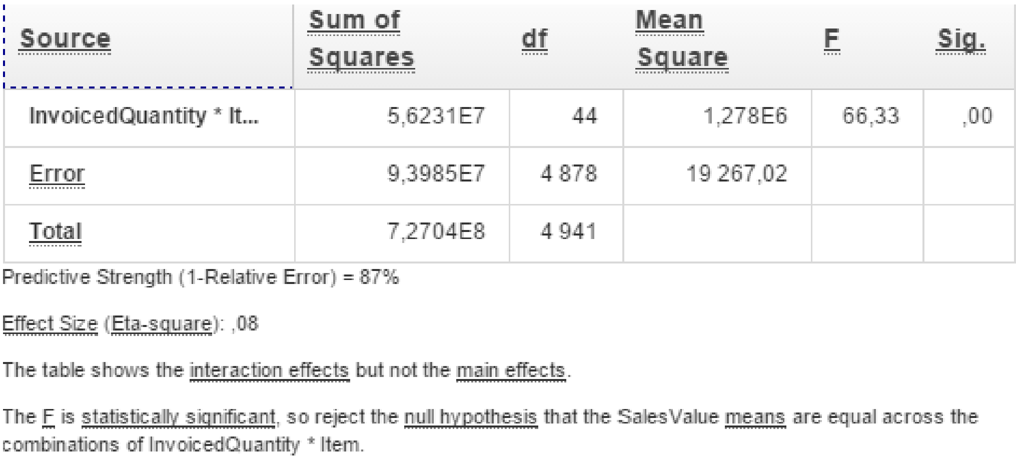


Fig 23. Additional information about predictive factor

These additional information charts provide statistics and frameworks for modelling if the data processing behind the models is in interest. Since Watson is designed for support in decision-making, this is information that can be looked if time is not the essence.

The chart also implies what approach has been used for the prediction target. In this case for sales value Watson informs that linear regression approach has been used. Analysis of variance, ANOVA, is a linear modelling method that is being used to evaluate relationships among fields. (Laerd statistics, 2013) For field association, it tests whether the mean target value differs across categories of input. If the variation is significant, there is a main effect. For key drivers, ANOVA tests whether the mean target value varies across combinations of categories of two inputs. If the variation is significant, there is an interaction effect. The results of ANOVA tests are seen below in figure 24.



Source	Sum of Squares	df	Mean Square	F	Sig.
InvoicedQuantity * It...	5,6231E7	44	1,278E6	66,33	,00
Error	9,3985E7	4 878	19 267,02		
Total	7,2704E8	4 941			

Predictive Strength (1-Relative Error) = 87%

Effect Size (Eta-square): ,08

The table shows the interaction effects but not the main effects.

The F is statistically significant, so reject the null hypothesis that the SalesValue means are equal across the combinations of InvoicedQuantity * Item.

Figure 24. Results of ANOVA test

As seen, on can dive to the roots of causes with the results that Watson provides. If even more detailed information is needed, it is possible to dig deeper to find the underlying connections.

From more practical level, the decision tree is extremely useful to identify line of variables linked in the prediction. In figure 25 below the decision three shows how predictors for highest sales value is significantly influenced by five factors in the current data set.

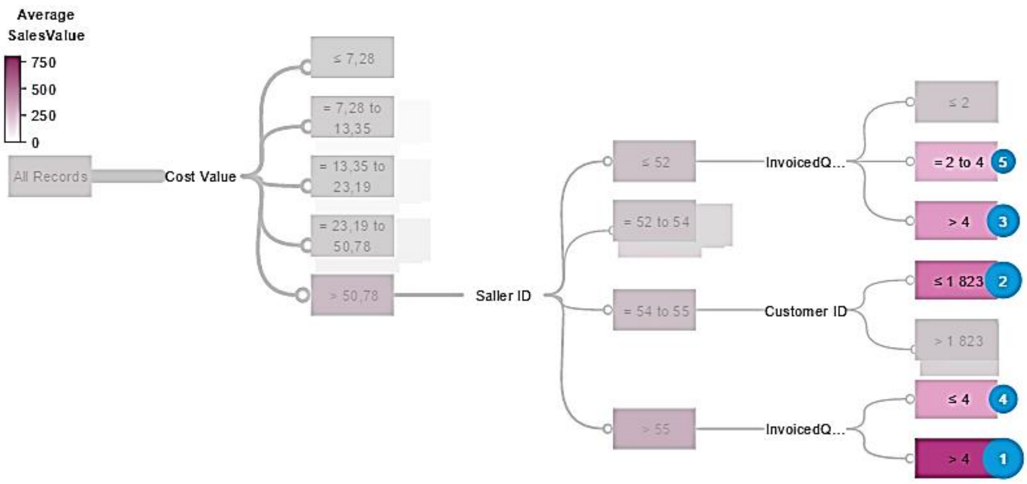


Figure 25. Decision tree for sales value

The marked five decision rules predict the highest sales values. For example the rule number 1 leads to the highest predicted sales value and is interpret as follows; The highest sales value is generated if the invoiced quantity is over four with seller ID over 55 with the cost value of item going over 50,78 euros. The second strongest prediction rules follows the path which starts with number 2; If the customer ID is below than 1823 and the seller ID is 54 or 55 with the cost value averaging over 50,78 euro. The complete decision tree for predicting sales value can be found in Appendix A.

When data was fed into Watson, it automatically evaluates the overall quality of the data sets. The data quality score measures the degree to which the data is suitable for predictive analysis. It is an average of the data quality score for every field in the data set. Figure 26 below describes the data quality report from Watson.

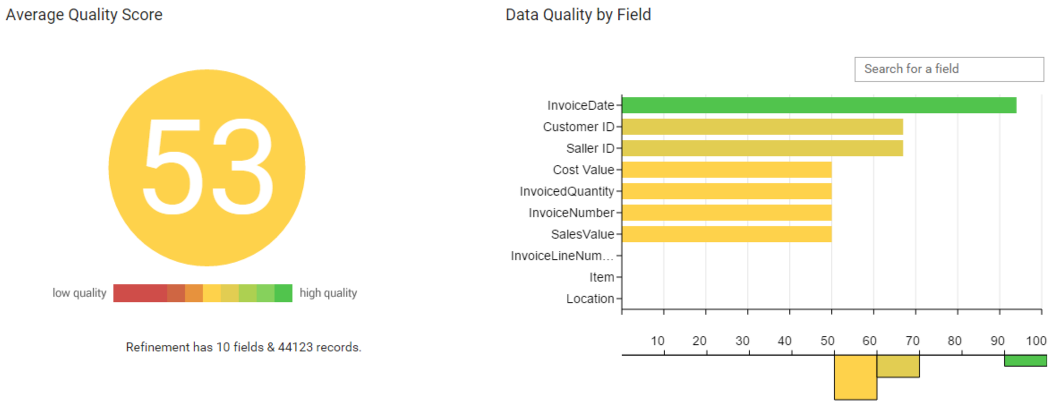


Figure 26. Data quality report

The data quality for these used files were on medium quality level, 60 % for the customer network data set and 53 % for the complete sales data from 2014. The level of data quality remained on medium level since the sales data includes plenty of individual purchasing operations from customers who only buy once or twice. All of these actions are defined as outliners since the weight of the data points lean on customer network nodes that have more transactions, and thus all the outliners were removed from the data set under study. The data set has total of 10 fields, and out of these 10, six field had both outliers and skewed distribution.

6.3 Synthesis

Case company for this research was a small Finnish retailer. It was essential to find data analytic solution that is user-friendly and can be utilized by the decision-maker. Small businesses do not have the luxury of having a full time in-house data analytics since the business focus lies heavily in operational level.

This thesis was executed in order to evaluate different options and how beneficial these options would be in the case company's context. The comparison of usability of designed analytic models is purposely set to the framework of business person point of view and how much additional guidance these analytic models could bring to the decision-making processes.

The used models were customer network position model created by network visualization software Gephi and identifying predictive attributes from given data set by using cognitive computing program IBM Watson. Below in table 6 are listed the main differences from business user point of view. The figure has been divided in three sections. First part presents the variables for general usage, the second part focuses on using the actual software, third part usability of the results and the final one reporting of the outcome.

Table 6. Comparison of usability

Comparison attribute	Gephi	Watson
Cost of use	Free	Free, but professional version costs 80 $ per month
On-premise or cloud	On-premise	Cloud
Scalability	No	Yes
Internet connection required	No	Yes
Data processing skills	Data needs cleaning and pre-processing	Cleaning needed. Pre-processing can be done in Watson
Coding skills	Script needs to be created	No
Software user-friendliness	Medium-level	Easy to use
Usability of results	Highly relevant	Medium-level
Reporting tools	Simulation model, visualization pictures, excel-listings	Daschboards and ready templates for reports

Both of the used programs are free or require only a small investment. Gephi is a free open-source software developed exactly for modelling of complex networks. Since it is an open-source software, nobody can guarantee its safety and data security in the future since it is literally open for anyone to develop further. It is on premise software that needs to be downloaded onto user's computer. After the software is downloaded to the end user's computer the software do not need internet connection to work. Depending on the amount of data, the data processing might take lots of processing power from computer. In this thesis, a normal laptop and PC were adequate for the data handling, but it was noticed already with this size of data package, around 64 MB that handling of larger file sizes require more efficient computer.

A free version of IBM Watson is available also and it is cloud-based service as with the professional version and thus require internet connection when working with. The invoicing method for Watson is related to the file sizes and how many rows and columns needs to be operated at once. While free version has certain limitation on the used data (100 000 rows, 50 columns and 500 MB of data storage), it was at least adequate in this thesis. The professional version of Watson operates by monthly fees and is cloud-based. The monthly fee for professional version is currently 80 $ per month and one can buy additional space when needed (IBM , 2015). This means that instead of heavily investing in new software frontloaded, the monthly fee enables changing these capital expenditures into operational expenditures which are easy to budget further on with the monthly fee.

The beauty in Watson lies in its scalability opportunities. If the business needs for using Watson grow higher and demand more space, since it is cloud-based it is simple to buy more storage without losing any of the gained benefits or results already in the software. The same argument applies for updates of the software, which makes sure the data is stored.

Another important aspect to bear in mind with Watson is that even though the data needs to be uploaded or imported to Watson, the rights for the data remain in the company as user. This is something to highlight since the process is not industry standard currently when using cloud-based application.

Software usability is on medium or easy level with both programs. IBM Watson has been designed for the end-user automatizing all the time-taking data processing steps in order to gain maximum benefits from the data. To use Watson on a regular basis and to dig into the roots of the issues, it enables the more traditional statistics data viewing in order to find causalities from the relations of the data but since the results of these connections are already available, these are more nice-to-know information and something to look further on if needed. The layout of the software is simplistic and after importing the data set, one can immediately start to explore the data.

Gephi requires at least medium-level data handling and processing skills. The data fed into the program need to be pre-cleaned and processed from the raw data in order to fully harness the data into the modelling. The data cleaning and manipulation took also in this research plenty of time to transform raw data into technically correct data.

There are not that many programs available yet on the market that can provide sufficient network modelling opportunities and are still rather easy to use. While Gephi is definitely not the most sophisticated from its user interference, it is still clear to navigate further on and execute the commands. The issue with using Gephi as end-used is that it need to be taught how to read a script for building the network. If the csv.-file that was gathered from case company’s ERP-system and which was cleaned and prepared by using program environment R, would have been imported to Gephi as such, it would have resulted as a bundle of same size dots spread around the screen. Instead, now it was defined with Python how the dots should be connected, how to differentiate the customs from the products and how to separate the purchasing profiles.

The results gained by using Gephi were highly relevant. While earlier the case company identified only three different customer groups, with network modelling in total of 16 similarly behaving customer clusters were able to be identified and confirmed from the company data. The result was a clear network model graph where these customer clusters were easily marked to be spotted.

It could be argued from the network model, that the closer the customer clusters are to the core of the network model, the closer the clusters indicates the core business activities of the case company. However, since one of the attributes of the used Force Atlas 2 algorithm is that it avoids over clustering and if the clusters seems to get too close to each other, the algorithm spreads the overall network into wider area. Also even though the large mass in quantity-based model was rather far away from the core of the network, it cannot be said that these customers would not be important customers for case company simply because they have radically different buying behavior than smaller customer clusters, especially since their monetary value as well was high.

The results gained from Watson in this thesis context were not as interesting as with network model, mainly due to the case company context of being new for the data analytics. While connections between variables were found and predictive factors identified, the data set itself did not hold enough information to make valid, customer-related predictions. The data set that would include more data about individual buying methods of each company would might have provided more useful results in order to target sales and marketing resources more accurately. However the gained results did provide comprehensive causal-relation from sales action with clear decision-tree backtrackings.

Each of the used modelling programs provide at least average reporting methods. While outcome of modelling from Gephi can be exported in standard excel-files and pictures of the network models, as an outstanding feature it also provides the simulation possibility in video format in order to examine how the network has developed over time. This enables also tracking the development of singular customer relationships over time.

Watson includes ready reporting templates and dashboards to build customized presentations already in the program, which were also used in this thesis to demonstrate results without adding additional layer in the process, such as creating a presentation of the results with MS Powerpoint. The report templates are interactive and enable to tell a compelling story from the results. As it is extremely sophisticated program, Watson provides in every stage of the process the most exploratory reporting base which the user can decide to use or then choose another format from the wide-selection. Even by using Watson with diagnostic approach, the data exploration opportunities provide interesting insights that could be deployed together with customer network modelling.

7 DISCUSSION

This chapter reflects the findings of the research to the existing literature and evaluates the research process. The quality of the research has been evaluated based on its reliability, validity and generalizability.

7.1 Reflection on the findings

For decades already, companies have dealt with information they own with logical ways. These ways include systematically exploring data sets to gain insights. The issues rising is dependent on the sheer amount of available data for the user to focus on accurate and valid data sets. (Barrenechea, 2013) Whether the exploring of the data sets have been done by using queries, reports or advanced analytical models, strict and explicit rules have been applied to provide assistance for decision-making processes. (Barton & Court, 2012) These logic rules are related to the roots of statistics and computational sciences in mathematics. The elemental technologies on how to store, visualize and model data have evolved and today's business intelligence solutions are as well derived from these logical models applied in clearly written business rules. (Deloitte University Press, 2014) However while currently used systems still work and we are still far from reaching the limits of these traditional used techniques, cognitive computing provides a new era on how to approach the decision-making process by automatizing it.

Since the current marketplaces are highly unstable, mostly due to the possibilities provided by the digitalization (Croon Fors, 2010) companies need to focus on product and customer related functions in order to adapt efficiently (Lynch, et al., 2012). The findings of the research are align with the idea that advanced analytics and cognitive computing provide means to deliver valuable insights about the current operations even on a single customer or single product level to dig in the roots of what is causing certain type of behavior (Siegel, 2013, p. 70). By identifying the similarly behaving customer clusters, the case company was able to make more accurate segmentation of their customer base and update their operational models into more efficient form.

Back in the old days, the production –centered business model was seen dominant, especially with industrial companies (Kotler, et al., 2006). With this approach, production capacity and efficiency were considered as the enabling differentiation factors among competitors, and companies mainly focused on these to improve their market position (Lynch, et al., 2012). Even today, most industrial companies still continue to compete with the scale of volume, low cost and low pricing (Hirata & Matsumura, 2011) as was also the starting situation with the case company, and the key drivers for change was to gain solid understanding about the current customer positioning with the operations.

With opportunities and challenges provided by digitalization, production-centered business model alone is simply not enough to survive and operate with growth mind-set, and analytics is used to redraw customer relations management lines (Krill, 2012). The genius and beauty in big data and analytic philosophy is that is provides a mean to target individual needs if the technological requirements in the company can support this.

However, the theoretical modelling processes of these new systems focusing on predictive analytics and cognitive computing have not evolved with the same speed as the ideology behind the technology. During the literature review it became evident, that while there are plenty of academic sources for the cognitive computing ideology listing possibilities, opportunities, threats and ethical questions about what these systems mean and what companies could gain or lose by using these methods, at least in this research only a few examples of model designed for business user, the real end-user of these methods, could be find. It became clear that before taking any further actions related to the data processing and data modelling, a framework model on how to utilize this data needed to be compiled.

Today, every company collects and possess data, even though they might not know what to do with it and what are the optimal ways to utilize this data. Data is being collected from various sources but without having a clear mind set on what to do with this data, it is merely a waste of resources to save and preserve. Structured, organized data is already a legal requirement for having detailed transaction data for accurate bookkeeping. (Croon Fors, 2010) Also as every company has some kind of ERP-system, whether it is a state-of-the-art business ecosystem via all business processes are handled or a simple excel-sheet to look over daily actions in a smaller scale, this system produces valuable data that can be used to find insights about the company’s operations. The case company’s data was extracted from the company’s ERP-system and when identifying the goals for the modeling process, it was clear that the data was adequate for the customer modeling as well as creating a simple model to identify predictive factors.

However, the results gained from the predictive modelling are valid if these are reviewed only from the company’s internal situation. But sales and marketing are highly customer-centric, two-way processes together with the client and the predictive indicators for customers should be derived from larger data set that includes also external data; customer-, competitor- and market environment-related data. (Kotler, et al., 2006)

As it was demonstrated in this thesis, the data format does not have to be very complex, or covering wide area of activities to find relevant insights and focus points to derive and pinpoint growth opportunities. With proper identification of system requirements and goal setting, the data activities can be incremental if the resources are limited and company is just starting to explore data usage opportunities. Both of the used modelling programs require none or little resources with basic level usage and the user-interface is rather simple. These programs, Gephi and Watson analytics, are designed for business person responsible for decision-making.

Machine learning, data mining, artificial intelligence and cognitive computing used to be experimental concepts that was mainly researched in computational sciences and brought to public via movies like Terminator and Matrix. Nowadays these concepts have transformed into potential business disruptors that can drastically change and unbalance the markets faster than ever. Attraction to instant gratification of information attracts decision-makers in all industries and company sizes, the case company not being an exception (Alstete & Cannarozzi, 2014).

By harnessing the internet speed, cloud opportunities and adapting business processes to provide and even drive data insights, companies can gain accurate, real time support for their decision-making. Companies that look for ways to bridge the gap between the intent of big data and the reality of their company-related decision making processes, cognitive analytics can be a powerful aid if the existing technique can support this vision. (Deloitte University Press, 2014)

Applications of cognitive computing range for almost every industry. Even though the targets are different for example in healthcare, commerce, finance, education and governmental services, in every industry humans engage in discussion, show curiosity by asking questions, test ideas and be a part of larger decision-making processes. (Computerworld, 2013)

The actual modelling in this thesis focused on two areas with different aims. As the case company did not have a comprehensive view about their current customer base, it was necessary to gather an understanding of this in order to form an accurate basis to continue. The network model is related to diagnostic analytics and it is meant to provide detailed information about the current situation with the customers. The second model using advanced analytics was deployed to identify patterns and variables that could be used as predictors for sales activities.

For a company that does not hold proper understanding of its customer structure, this is highly valid information. One cannot apply future forecasting models unless knowing where they stand now. The predictive models used with unclear picture of current situation end up in false predictions that cannot be properly interpreted. On the other hand, for a company that has solid understanding of their current situation, predictive analytics could provide interesting hidden gems of information that could benefit the whole business. (McAfee & Brynjolfsson, 2012)

The modelling process of the research went in line with theoretical foundation introduced for example by Han & Kamber (2000), Edelstein (1999) and Hand (1999). The modelling examples were all from statistic and science while the theoretical framework, analytic modelling process, was compiled based on existing literature on what the process would look like from the business user point of view. This also shows the shift in mind set in the literature of modeling processes. While most modeling articles before and in the early 2000's are derived from statistics and computational sciences, research after 2010 clearly focuses on business applications analytics can provide. This might also relate to the trend that IT and software providers have understood that the information need to be instantly available for the right person and the core of these application remains in easy user-interfaces, usability and reporting features provided by the program. (Alstete & Cannarozzi, 2014)

In overall, this thesis provides contribution to the existing literature about data analytics in small- and medium-size companies operating in industrial markets. The literate review revealed a gap in designing the modelling process from the business user side, in which this thesis aimed to provide tools. This research offers more pragmatic approach on how to design and applying analytic modelling into business practices and what kind of benefits could be expected. The concept of data analytics is highly relevant for companies but the topic has been researched relatively little on academic side in marketing and business development related topics. The topic is mainly covered in statistical and heavily IT and software development focused academic circles.

7.2 Quality of the research

Although the results of this study were based on transactional data, machine learning methods and the data did not include qualitative features that might have been misinterpret , the reliability and validity of the study need to be reviewed. The models were tested and validated manually on building phase but in order to identify if these models bring real business value for improving business processes, the results should be reviewed against the current operating models.

Miles & Huberman (2014, p. 276) argue that the quality of the research can be assessed by different criteria. The most common aspects include reliability, validity and generalizability. Additionally, according to Lincoln & Guba (1994, p. 106) these criteria are suitable especially for quantitative research. In order to research to be valid, in needs first to be reliable (Laerd Dissertation, 2012) According to Roberts (2006, p. 41) reliability and validity are presenting the ways to demonstrate and communicate the progression of research process, as well as evaluating the trustworthiness and usability of the gained results.

Reliability indicates the constancy of research and how likely the same results apply if the research is re-done and same conditions are used (Laerd Dissertation, 2012). Validity refers to the extent on how accurate, conclusive and credible the findings of the research are (Roberts, 2006). Generalizability of the research demonstrates to the extent on how likely the particular process would provide similar results in different circumstances assuming everything else remains similar in test setting (Given, 2008). The evaluation of the research based on these three criteria is presented below in table 7.

Table 7. Methods to confirm quality of the research

Criteria	Method to confirm quality	Action taken
Reliability	Engagement and Observation	Researcher was engaged to the research and observed the case organization persistently
	Versatility of the data	The data sources included extracted data from the ERP-system, which was confirmed from the records.
	Research design	Research design and all the phases are clearly written in order to re-do the research
Validity	Testing	The model was tested by using real data sets that were manually validated.
	Formalized practises for data handling	Data collection, analysis and interpretation of the data
Generalizability	Implementation	The designed model was used in case environment

The reliability of the research was confirmed by using three different methods; engagement and observation, versatility of the data and research design. During the whole research process the researcher engaged to the project and actively observing the case company throughout the thesis project.

The data collected for this study can be seen as a reliable source of information as it is based on the actual transactional data from the company's ERP-system. During the pre-processing phase of the data sets it came evident that there were missing values in the data set. As there is always a possibility for data distortion, all damaged and defective data that could not be confirmed reliable from other source was removed in order to have more accurate results.

The research was well structured and the study progressed as planned. The implementation was in line with test modelling and the phases are clearly written in order to replicate the research in case needed.

The testing of the planned models were carried out by using real data sets which were then manually confirmed. Slight alterations were done for the data sets but this was due to the research limitations of excluding one-time single-item purchases from the data. Formalized practices for data handling were undertaken as described in the research design in order to track back all the performed actions.

Finally when assessing the generalizability, the models were planned for certain case setting, and in the implementation phase these practices were carried out with real case company's data. However, the planned theoretical modelling process was compiled based on the existing literature about the topic. The built models are designed for the use of this case company, but with similar data sets could be applied to another company as well. One must remember though, that this model has been tested accurate only in this case environment, and the validation of the model should always be performed before utilizing the models.

8 CONCLUSIONS

Companies that can leverage the data they possess into information assets and streamline their decision making processes can gain significant benefits. Especially small- and medium-size companies that don’t possess that much resources, need simple and accurate tools on how to deploy their data power. These tools need to be simple to use in order to enable real-time decision making from the business user point of view.

The object of this thesis was to design an analytic modelling process based on current known theories and apply this model into case company’s needs, operating in retail industry. The case company was a Finnish industrial filter wholesaler. The data collection techniques used in this thesis included 1.) Desk research, 2). Literature review, 3) Processing of quantitative data and 4) Analytical modelling process.

The chapter 8 concludes this thesis by summarizing the results of this research. The results for both of the research questions are summarized in sub-chapter 8.1 and finally in sub-chapter 8.2 the managerial implications of this study are presented.

8.1 Summary of results

The first objective of the research was to map current practices for the modelling process. It was quite evident since the beginning of this research that almost all the literature focusing on the modelling of analytics was related to the modelling processes from the discipline of statistics instead of focusing on end-user actions as a business person. It was decided to plan a simple framework that could be exploited by the end-user when planning analytics activities. The results for first research question: *How to design an analytic modelling process?* are presented below in table 8.

Table 8. Results for RQ1

Research Question 1: How to design an analytic modelling process
The analytical modelling process includes three stages. The first stage focuses on the planning of the model by defining clear goals on what the intrinsic business value is and why the modelling will be conducted. The planning phase includes also creating the data sets by defining the data sources as well as cleaning the possible disorganized data sets based on the business objectives set earlier. The second phase of the modelling process is the actual model building process. Depending on the chosen method, the model building could be either teaching the program how to read the data or then simply feeding in the prepared data sets if the modelling program already has set parameters on reading data sets. After the data sets have been fed into the system, the model produces results which can be in text or in visual formats. The numerical value can be extracted from these results and this is needed in order to validate the model. If the primary results provided by the model are not accurate enough, or include some false values that interfere with the complete data sets, the model need to be refined. This incremental improvement cycle will be executed as many times with the test data set as necessary to reach the desired level of accuracy. The final stage of the modelling process is to implement the validated model into real business use. This will be performed by feeding the real company data into the model. The results gained from this phase should be use to drive further business initiatives in order to fulfill the underlying business purposes the modelling process was created in begin with.

After the theoretical modelling process was created based on existing literature, the modelling process was applied in the case environment. The results for second research question *How to utilize analytic modelling process in retail context?* are presented below in table 9.

Table 9. Results for RQ2

Research Question 2: How to utilize analytic modelling process in retail context?
The theoretical modelling process was utilized with two different aims and thus two models were also created. The first model focused on mapping the current position of the case company's customer network. Earlier, case company had segmented its customer into three segments and with this research more accurate view was needed. As the focus was on with the past and current customers, and how these have evolved over time, the focus of this process was on diagnostic modelling. The actual modelling was run by using computer software Gephi that is designed for network modelling. After the model was validated with test data set, customer networks were processed by feeding into the model data set that include all the transactions, item by item, during the years 2010-2014. The results of this process was a network graph, revealing 16 similarly behaving customer clusters. These clusters were unevenly spread over company's product portfolio and it was evident that some of the clusters were clearly generating more income than others. Second interesting remark from some of the clusters was that inside the clusters there were few customers that clearly generated the most income while most of the companies in these clusters were producing extremely small amounts compared to the dominant customer. When planning to drive business initiatives based on these clusters, case company must decide which of these clusters needs to be prioritized and what kind of actions to take with each of these clusters. The second model was designed to identify factors that drive the sales with predictive analytics. This was done by case framework set to find the key drivers

for predicting sales value from the existing data set. This modelling was executed by using IBM Watson advanced analytics program. As Watson is cognitive computing software able to understand natural language, the question “What influences the sales value?” was posed for the data. The prediction model identified in total of eight correlations on top predictors for sales value. Six out of these eight values had predictive score over 75 % and the remaining two had the predictive score below 20 %. The strongest correlation was found between variables invoice number and item, which drives sales value with predictive strength of 87.2% and the second strongest interaction was found between invoiced quantities and item together, that drive the sales value with predictive strength of 87.1%. Even these results are valid, since the company data includes only company’s internal transaction data, the overall picture remains quite blurry. Especially for predicting sales value, the data sets should include more customer related information as well as preferably information about the industry environment.

From the case company’s point of view to just starting to explore what they could do with their currently produced data and the opportunities analytics provides, the diagnostic customer modelling was perceived more usable at the moment. It gives solid knowledge about their current customer network as well as how this point has been reached. This provides valuable starting point for the case company to become more active on their customer relations management. However it was acknowledged, that when the case company starts collecting also external data about their industry environment, the predictive modelling would become extremely useful to become more agile in the highly competitive and volatile industry they operate in.

8.2 Managerial implications

The aim of all development actions in the companies are undertaken with the mindset to enable business growth. In this sub-chapter the managerial implications of this study are presented.

The network analysis clearly identifies similarly behaving customer cluster form the transaction data. Since these clusters are identified based on real buying behavior, this enables companies to more efficiently perform customer segmentation with similar customer profiles. With more detailed and accurate segmentation, the company's assets and resources can be targeted with more consciously by identifying cross- and up-selling option along with new business opportunities. To illustrate this with an example; in the customer cluster number four there are 12 items that are regularly purchased. Some of the customers in this clusters buy all of these and some of the customers buy only some of the products. With manual reviewing it can be confirmed if the customers are operating in similar industry and after this reviewing, offering at least these 12 products to all customers operating in similar industry since the model implies that these customers might have similar usage for the products. Of course there are variety of reasons why all the companies with similar profile are not buying these items from the case company, such as if they manufacture some of these items by themselves or supply from another company, but without further inquiries, clear selling opportunities are lost. In a wider picture network analysis enables more pro-active approach for deepening the customer relationship and retaining customers in long term.

Network analysis can also be applied to identify key products from the company's product portfolio. These are the items that are most vital for the company to operate, and one of the reasons why customers choose this company as their supplier. Key drivers for these items in the industry level may form on price, availability, namely whether the item is stored on the warehouse location or purchased from order, delivery terms, security of supply, added of course to the importance of good customer service abilities. From the warehouse management point of view, this might also mean new distinction between items which are constantly available from the warehouse and which should be by-order items. Accurate warehouse management might balance the usage of warehouse and even saving costs if inventory rotation speeds up.

Predictive analytics enables more accurate and cautionary warehouse management in terms of for example seasonal products. With the development of customer monitoring it is already possible to predict future customer behavior with machine learning programs, but at least so far these processes require more advanced machines, because the processes are heavy to process with normal everyday use computers. All models and processes described in this study can be built and run on a normal desktop or laptop. However these models are not 100% accurate and at least so far these should be used as supportive tools for decision-making.

But like Eric Siegel said:

> “Little prediction goes a long way” (Siegel, 2013, p. 11).

The more data is used in these systems, the more accurate these will become. These models do not have to be 100 % accurate, they just need to be better and more accurate than the currently used systems.

REFERENCES

Accenture, 2015. *Accenture Technology Vision 2015: Digital Business Era: Stretch Your Boundaries.* [Online] Available at: http://techtrends.accenture.com/us-en/downloads/Accenture_Technology_Vision_2015.pdf [Accessed 15 August 2015].

Alstete, J. W. & Cannarozzi, E. G. M., 2014. Big data in managerial decision-making: Concerns and concepts to reduce risk. *International Journal of Business Continuity and Risk Management,* 5(1), pp. 57-71.

AT & T, 2015. *AT&T Company information.* [Online] Available at: http://www.att.com/gen/investor-relations?pid=5711 [Accessed 8 September 2015].

Barrenechea, M., 2013. *Forbes Tech: Big data: Big hype?.* [Online] Available at: http://www.forbes.com/sites/ciocentral/2013/02/04/big-data-big-hype/ [Accessed 28 August 2015].

Barton , D. & Court, D., 2012. Making advanced analytics work for you. *Harvard Business Review,* 90(10), pp. 78-83.

Berry, M. J. A. & Linoff, G. S., 2000. *Mastering data mining.* 1st ed. New York: Wiley.

Burawoy, M., 1998. The Extended Case Method. *Sociological Theory,* 1998(1), p. 4.

Chen, Y., Li, J. & Wang, J. Z., 2004. *Machine Learning and Statistical Modeling Approaches to Image Retrieval.* 1st ed. Boston: Kluwer Academic Publisher.

Coalition Against Insurance Fraud, 2014. *By the numbers: fraud statistics.* [Online] Available at: http://www.insurancefraud.org/statistics.htm#.ViJOlPmqqko [Accessed 3 August 2015].

Computerworld, 2013. *Watson and the future of cognitive computing.* [Online] Available at: http://www.computerworld.com.au/article/522302/watson_future_cognitive_computing/ [Accessed 15 September 2015].

Croon Fors, A., 2010. The Beauty of the Beast: The matter of meaning in digitalization. *AI & Society,* 25(1), pp. 27-33.

Danneels, E., 2010. Trying to Become a Different type of company: Dynamic capability at Smith Corona. *Strategic Management Journal,* 2010(32), pp. 1-31.

Davenport, D., 2013. Analytics 3.0. *Harvard Business Review,* 91(12), pp. 64-72.

Davenport, T. H., 2006. Competing on Analytics. *Harvard Business Review,* 84(1), pp. 98-107.

De Jonge, E. & Van Der Loo, M., 2013. *An Introduction to data cleaning with R.* Hague, Statisctics Netherlands.

Dell, 2015. *Data Mining Techniques.* [Online] Available at: http://documents.software.dell.com/Statistics/Textbook/Data-Mining-Techniques [Accessed 28 August 2015].

Deloitte University Press, 2014. *Tech Trends 2014: Cognitive analytics.* [Online] Available at: http://dupress.com/articles/2014-tech-trends-cognitive-analytics/ [Accessed 10 September 2015].

Demirkan, H. & Delen, D., 2013. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision support systems,* 55(2), pp. 412-421.

Edelstein, H. A., 1999. *Introduction to data mining and knowledge discovery.* 3rd ed. Potomac: Two Crows Corp.

Eisenhardt, K., 1989. Building theories from case study research. *The academy of management review,* 14(4), pp. 532-550.

eMetrics Summit, 2013. *Leveraging Customer Data to Drive Business Strategy.* [Online] Available at: http://www.emetrics.org/predictiveanalytics/ [Accessed 3 August 2015].

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthutusamy, R., 1996. *Advances in knowledge dicovery & data mining.* 1st ed. Campbridge: MIT Press.

Forbes tech, 2011. *How IBM's cognitive computing works.* [Online] Available at: http://www.forbes.com/sites/alexknapp/2011/08/26/how-ibms-cognitive-computer-works/2/ [Accessed 4 9 2015].

Ghauri, P. & Gronhaug, K., 2010. *Research Methods in Business Studies.* 4th edition ed. Harlow: Pearson Education Limited.

Gibbons, P., 2015. Voice of the customer: the risky business of predictive analytics. *Customer Relationship Management Magazine,* Issue February, p. 7.

Given, L. M., 2008. *The Sage Encyclopedia of Qualitative Research Methods.* [Online] Available at: https://srmo.sagepub.com/view/sage-encyc-qualitative-research-methods/n186.xml [Accessed 4 October 2015].

Guba, E. & Lincoln, Y., 1994. *Competing paragidm in qualitative research. Handbook of qualitative research.* 2nd ed. Thousands Oaks: Sage Publication.

Gummesson, E., 1991. *Qualitative methods in management research.* London: Sage Publications.

Hand, D. J., 1999. Statistics and data mining: Intersecting Disciplines. *SIGKDD Explorations,* 1(1), pp. 16-19.

Hardesty, L., 2015. *System that replaces human intuition with algorithms outperforms human teams.* [Online] Available at: http://phys.org/news/2015-10-human-intuition-algorithms-outperforms-teams.html [Accessed 16 October 2015].

Harrison, L. & Callan, T., 2013. *Key research concepts in politics and international relations.* 1st ed ed. Thousand Oaks: SAGE Publications Ltd.

Hastie, T., Tibshirani, R. & Friedman, J. H., 2001. *The elements of statistical learning: Data mining, inference, and prediction..* 1st ed. New York: Springer.

Hirata, D. & Matsumura, T., 2011. Price leadership in a homogenous product market. *Journal of economics,* 104(3), pp. 199-217.

Hirsjärvi, S., Remes, P. & Sajavaara, P., 2009. *Tutki ja kirjoita.* First edition. Helsinki: Tammi.

Hsieh, N.-C., 2004. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications,* Issue 27, pp. 623-633.

Huffington Post Tech, 2014. *Amazon Just Patented Shipping Items Before They're Even Ordered.* [Online]

Available at: http://www.huffingtonpost.com/2014/01/18/amazon-anticipatory-shipping-items-before-ordered_n_4623499.html [Accessed 10 8 2015].

IBM , 2015. *IBM Watson analytics.* [Online] Available at: http://www.ibm.com/analytics/watson-analytics/ [Accessed 4 October 2015].

IBM Research, 2015. *Cognitive computing.* [Online] Available at: http://www.research.ibm.com/cognitive-computing/#fbid=Xh8B_6NTVcI [Accessed 8 9 2015].

IBM, 2015. *INFANT Centre at University College Cork to use IBM Big Data & Analytics for real time monitoring of babies in neonatal intensive care.* [Online] Available at: http://www-03.ibm.com/press/uk/en/pressrelease/45921.wss [Accessed 22 9 2015].

IBM, 2015. *What is Watson.* [Online] Available at: http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html [Accessed 8 September 2015].

IEEE Life Sciences, 2013. *Big data in neo-natal intensive case.* [Online] Available at: http://lifesciences.ieee.org/articles/347-big-data-in-neonatal-intensive-care [Accessed 10 9 2015].

Joe F, H. J., 2007. Knowledge creation in marketing: the role of predictive analytics. *European Business Review,* 19(4), pp. 303-315.

Kamber, M. & Han, J., 2000. *Data Mining: Concepts and Techniques.* 1st ed. New York: Morgan-Kaufman.

Kantardzic, M., 2011. *Data Mining: Concepts, Models, Methods and Algorithms.* 2nd edition ed. Hoboken: John Wiley & Sons Inc..

Klatt, T., Schlaefke, M. & Moeller, K., 2011. Integrating business analytics into strategic planning for better performance. *Journal of Business Strategy,* 32(6), pp. 30-39.

Knapp, A., 2011. *Forbes Tech; How IBM's cognitive computing works.* [Online] Available at: http://www.forbes.com/sites/gordonkelly/2015/10/16/microsoft-accident-forces-windows-10-onto-windows-7-windows-8/ [Accessed 26 September 2015].

Kotler, P., Rachham, N. & Krishnaswamy, S., 2006. Ending the war between sales and marketing. *Harvard Business Review,* Volume 84, pp. 68-78.

Krill, P., 2012. *Analytics redraw CRM lines*, Path: Infoworld Enterprise applications.

Laerd Dissertation, 2012. *Reliability in research.* [Online] Available at: http://dissertation.laerd.com/reliability-in-research.php#first [Accessed 4 October 2015].

Laerd statistics, 2013. *Ona-Way Anova.* [Online] Available at: https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php [Accessed 10 October 2015].

Lancer, R., Coats, P., Shanker, C. & Fant, L., 1995. A neural network for classifying the financial health of a firm. *European Journal of operational Research,* 85(1), pp. 53-65.

Lemieux, V. L., Gormly, B. & Rowledge, L., 2014. Meeting Big data challenges with visual analytics: the role of records management. *Records Management Journal*, 24(2), pp. 122-141.

Linoff, G. S. & Berry, M. J. A., 2000. *Masterinh Data Mining.* 2nd ed. New York: Wiley.

Lin, Y. T., 2002. *Attribute feature completion: The theory of attributes from data mining prospect.* San Jose, IEEE.

Liyakasa, K., 2013. Predictive analytics: The futurists' formula. *Customer Relationship Management Magazine,* Issue May, pp. 28-31.

Luciano, S., Couso, I., Otero, J. & Palacios, A., 2010. Assessing the evolution of learning capabilities and dirorders with a graphical exploratory analysis of surveys containing missing and conflicting answers. *Neural Network World,* 20(7), pp. 825-838.

Lynch, J., Mason, R. J., Beresford, A. K. & Found, P. A., 2012. An examination of the role for business orientation in an uncertain business environment. *International Journal of Production Economics,* 137(1), pp. 145-156.

Madison, M. C. et al., 2012. Knowledge encapsulation framewework for technosocial predictive modeling. *Security Informatics,* 1(10), pp. 1-18.

Malhotra, N. K., 2009. *Marketing research; An applied orientation.* 6th edition ed. New Jersey: Prentice hall.

McAfee, A. & Brynjolfsson, E., 2012. Big Data: The Management Revolution. *Harvard Business Review,* 90(10), pp. 61-68.

McKinsey & Company, 2012. *Perspectives on digital business,* New York: McKinsey center for business technology.

McKinsey Global Institute, 2011. *Big data: The next frontier for innovation, competition and productivity,* Washington: McKinsey Global Insitute.

McKnight, N., 2008. Looking beyond the score: Predictive analytics and Subrogation. *Claims Magazine,* Issue November, pp. 42-44.

Miles, M. & Huberman, A., 2014. *Qualitative data analysis: A methods sourcebook.* 3nd edition ed. Thousand Oaks: Sage Publications.

Modha, D. S. et al., 2011. Cognitive computing. *Communications of the ACM,* 54(8), pp. 62-71.

Montgomery, H. & Svenson, O., 1992. Process and structure in human decision making. *Journal of marketing research,* 29(February), pp. 151-153.

Provost, F. & Fawcett, T., 2013. *Data Science for Business.* 1st ed. Sebastopol, CA: O'Reilly Media Inc.

Ranjit, B., 2009. Advanced analytics: Opportunities and challenges. *Industrial Management and Data Systems,* 109(2), pp. 155-172.

Roberts, P., 2006. Reliability and validityin research. *Nursing Standard,* 20(44), pp. 41-45.

Saunders, M., Lewis, P. & Thornhill, A., 2009. *Research methods for business students.* 5th edition ed. Essex: Pearson Education Limited.

Siegel, E., 2013. *Predictive analytics: The power to predict who will click, buy, lie, or die.* 1st ed. Hoboken, New Jersey: John Wikey & Sons, Inc.

Smith, S. M. & Albaum, G. S., 2012. *Basic Marketing Research: Volume1. Handbook for research professionals.* 1st edition ed. Provo: Qualtricks Labs.

The Guardian, 2014. *Insurance fraud worth £3.5m uncovered every day.* [Online] Available at: http://www.theguardian.com/money/2014/may/30/insurance-fraud-industry-figures [Accessed 18 9 2015].

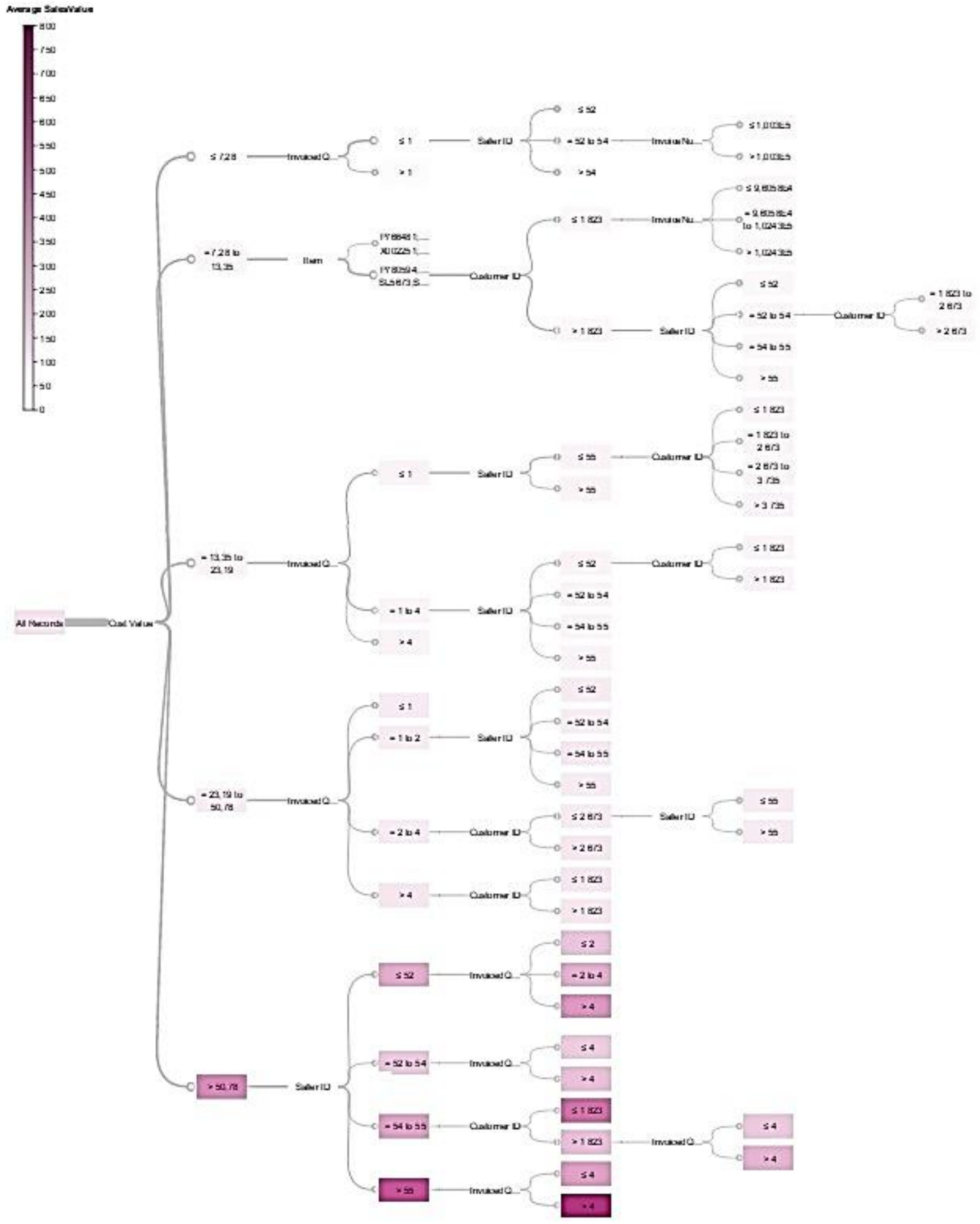
The UK Cards Association, 2014. *Annual Report 2014,* London: The UK Cards Association.

Wang, M. & Chen, W., 2015. A data-driven netwrok analysis approach to predicting customer chpice sets for choice modeling in engineering design. *Journal of Mechanical Design,* 137(July), pp. 071410-1 - 071410-11.

Varho, E., 2015. *YLE: Hengenvaarallinen verenmyrkytys uhkaa pikkukeskosia – Watson-tekoälystä etsitään turvaa.* [Online] Available at: http://yle.fi/uutiset/hengenvaarallinen_verenmyrkytys_uhkaa_pikkukeskosia__watson-tekoalysta_etsitaan_turvaa/8316195 [Accessed 12 October 2015].

Yin, R., 2002. *Case study research.* 3rd edition ed. Thousand Oaks: Sage publishing.

APPENDIX A: Overview of the decision tree to predict sales value


Average SalesValue
800
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
All Records
Cost Value
≤ 7,28
Invoiced Q...
≤ 1
Seller ID
≤ 52
> 52 to 54
Invoice Nu...
≤ 1,003E5
> 1,003E5
> 54
> 1
> 7,28 to 13,35
Item
FY8848 1,...
X00225 1,...
FY8094...
SL5873,S...
Customer ID
≤ 1 823
Invoice Nu...
≤ 9,6058E4
> 9,6058E4 to 1,0243E5
> 1,0243E5
> 1 823
Seller ID
≤ 52
> 52 to 54
Customer ID
> 1 823 to 2 673
> 2 673
> 54 to 55
> 55
> 13,35 to 23,19
Invoiced Q...
≤ 1
Seller ID
≤ 55
Customer ID
≤ 1 823
> 1 823 to 2 673
> 2 673 to 3 735
> 3 735
> 55
> 1 to 4
Seller ID
≤ 52
Customer ID
≤ 1 823
> 1 823
> 52 to 54
> 54 to 55
> 55
> 4
> 23,19 to 50,78
Invoiced Q...
≤ 1
> 1 to 2
Seller ID
≤ 52
> 52 to 54
> 54 to 55
> 55
> 2 to 4
Customer ID
≤ 2 673
Seller ID
≤ 55
> 55
> 2 673
> 4
Customer ID
≤ 1 823
> 1 823
> 50,78
Seller ID
≤ 52
Invoiced Q...
≤ 2
> 2 to 4
> 4
> 52 to 54
Invoiced Q...
≤ 4
> 4
> 54 to 55
Customer ID
≤ 1 823
> 1 823
Invoiced Q...
≤ 4
> 4
> 55
Invoiced Q...
≤ 4
> 4