A data driven inventory decision model for the early life cycle of product offerings in an e-tail environment

van Bussel, J.M.

Award date:
2017

Link to publication

A data driven inventory decision model for the early life cycle of product offerings in an e-tail environment.

A Master Thesis Project at ComCom

By Jur van Bussel
Student ID 0878478

In partial fulfilment of the requirements for the degree of

Master of Science

in Operations Management and Logistics

Master thesis (1CM96)
Eindhoven University of Technology, Augusts 2017

Supervisors TU/e
Dr. A.E. (Alp) Akçay
Dr. N.R. (Nevin) Mutlu
Dr. W.L. (Willem) van Jaarsveld

Supervisors ComCom
Ir. S. (Sander) van den Broek
Ir. W. (Wiebe) Konter

Abstract

This study includes the development of a data driven inventory decision model for the early life cycle of product offering in an e-tail environment. The development is the result of the initiated problem about the inventory decisions during the early life. The overall model includes a data driven model that decides whether a product should be purchased, how much to purchase for the first time and the optimisation of the (R,S)-policy during the early life cycle of the e-tailer. The developed model maximises the profit, is fully data driven and has limited human interaction during the early offering phase. The results indicate that implementing the model will increase the current profit and decrease the stock keeping units in the assortment.

Management summary

This report is a result of a master thesis project that is conducted at ComCom in Eindhoven. The master thesis presents a research on data driven inventory decision models for the early life cycle of product offerings in an e-tail environment. As a case study Company D is chosen from the customer network of the supervising company ComCom.

Problem Statement

ComCom provides supply management software to companies with webshops, the software advises the electronic retailer (e-tailer) the amount of products to purchase. ComCom has indicated that there is no sufficient method available for the purchase decision of products that are available for the first time in a webshop and that there is no acceptable method for controlling the inventory during the early offering phase of this product. The early offering phase is the moment the new stock keeping unit (SKU) arrives at the warehouse for the first time until the moment the SKU reaches five sales data points. Five sales data points happen when the specific SKU is sold five different times (one sales data point can have more than one product of a SKU). After the fifth sales data point is reached, a SKU is called mature.

The general idea of ComCom is to minimize the human intervention and maximise the automation during the purchase process. ComCom attempts to 'replace the logistic expert in the company' with their model. In contrast with the general idea of ComCom, the decisions if a product should be purchased and how much to purchase are currently made by the e-tailer based on his expert judgement. The current expert judgement methods are not in line with the general idea of ComCom. ComCom also indicated that method for controlling the inventory during the early offering phase of this product is not satisfactory. ComCom uses the (R, S) inventory policy this is a periodic review policy, the inventory level is observed at time intervals of the review period (R). When the inventory position (IP) is lower than the order-up-to-level (S) on moment (R), (S) – (IP) products are ordered. The review period (R) depends on the supplier. The products of ComCom are classified in A,B or C classes based on the contribution to the overall profit. The A-classified products are fast movers and the C-classified products are slow movers. ComCom automatically places products which are in the early offering phase in the C-class. The products in the C-class have a (R,1)-inventory policy, the policy advices to order one product when the inventory on hand drops to zero. As long as a product is classified in the C-class, the product does not react to changes and trend. The e-tailers believe that the new product offering is a fast mover and need to react fast to changes and trend. Reacting fast to a sudden change is not possible for new product offerings in the current algorithm of ComCom.

The case study company called 'Company D' is a company with one webshop and one central warehouse. Company D has around 7900 stock keeping units (SKU) and a high diversity of products. Currently 37% of all the SKU of company D are never sold before and only 38% of all the SKU are profitable. This leads to the following research question:

Develop a data driven model that decides whether a product should be purchased, how much to purchase the first time and how to control the inventory during the early offering phase of the e-tailer.

New model development

When the e-tailer finds a potential new product, several inventory decisions have to be made. Figure 1 visualises the inventory decisions for product introductions.

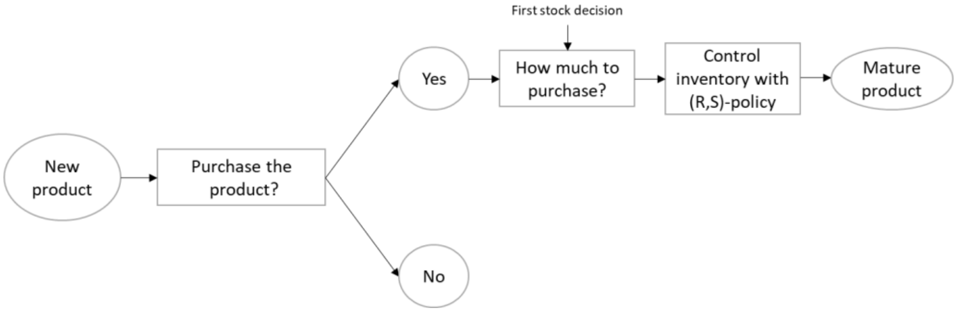


Figure 1: Inventory decisions during the early offering phase.

The first decision is to decide if the product is worth to purchase. Next, when a product is marked as "worth to purchase", the decision needs to be made how much products to purchase, this decision is called the first stock decision. At last when the product arrives at the warehouse the inventory of the SKU needs to be controlled with the (R,S) inventory policy until the product reaches maturity. The decisions are translated into a three step model, the steps are visualised in Figure 2.

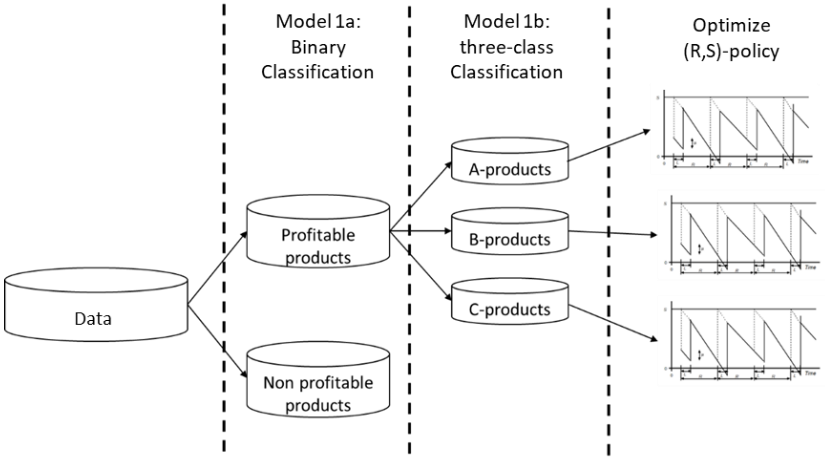


Figure 2: Three step model

Case study

For the binary classification model (1a) five different classification methods are tested. The model tested are the classification and regression tree (CART), random forest, radial based function (RBF) support vector machine, logistic regression and multi-layer perceptron (MLP) neural network. The second model (1b) the three class classification model classifies a SKU in an A, B or C-class. The models tested are classification and regression tree (CART), random forest, radial based function (RBF) support vector machine and multi-layer perceptron (MLP) neural network. The models (1a & 1b) are tested based on the interpretability, time to train, accuracy, profit and consistency. Before three class classification model (1b) is tested during the early offering phase, the optimal order-up-to-levels (S) need to be calculated. The order-up-to-levels are used for testing the models. The review period (R) is an already known variable. The optimal order-up-to-levels are used as input for testing the different three-class classification models.

Results
The models (1a & 1b) are simulated 50 times with each time a randomly chosen training and test set of a fixed size. For the binary classification, the classification tree outperformed the other models based on average highest profit, interpretability and consistency. The MLP neural network outperformed the other models based on average accuracy and was the most times the model with the highest profit over the 50 simulations. The model with the highest average accuracy (MLP neural network) does not outperform the classification tree based on average profit for binary classification. The optimal order-up-to-levels for each class are calculated based on the highest profit. For controlling the inventory the optimal order-up-to-level is eight for class A, and five for B and C. The maximal profit when using the optimal order-up-to-levels is on average €11.64 per stock keeping unit during the early offering phase. For the three class classification model, the random forest outperformed the other models based on accuracy, average profit and consistency.

Conclusion
Since profit is the leading performance indicator the classification tree model performs the best on the data set for binary classification. When using the classification tree instead of the current policy, the number of SKUs will decrease with 55% and the profit will increase with 26.2%. The random forest clearly outperforms the other models in the three class classification. The optimal order-up-to-levels for each class based on the total profit are eight for class A, five for class B and five for class C. The profit during the early offering phase when using the random forest model and the optimal order-up-to-levels as input is on average €7.62. In comparison to the profit using expert judgement this is a bit lower, the current profit is €7.71 per SKU during the early life cycle. However, the profit of the expert judgement policy is tested on SKU which do not have lost sales. Lost sales have a significant influence on the profit, what makes it hard to compare. In addition to this, the model is able to have a comparable performance as the e-tailer. For the e-tailer it is a time consuming process to decide how much products to purchase of each SKU, when using the random forest model the e-tailer has more time on hand for other jobs.

The automated decision tool
The model is an automated decision model for the early offering phase. First the classification tree gives an automated purchase decision. Second the random forest algorithm classifies the SKU in a ABC-class. Lastly the (R,S)-policy with the optimal order-up-to-levels for each class, controls the inventory control policy during the early offering phase. The models together are a data driven model that advices the users what to do during the early offering phase.

Recommendations:

- It is recommended to use the decision support model to assist in the decision making process. Using the tool will lead to increase in profit, decrease of assortment and controls the inventory in the early offering phase.
- ComCom uses the ABC-classification method to classify the products, however the classification is not always correctly since it does not matter how the product performs if it has less than five sales data points it is placed in the C-class. We would recommend to use an extra class for products that have not yet reached five sales data points.

Preface

The preface is the final written chapter of my master thesis. This report describes the results of my master theis, which is conducted at ComCom in Eindhoven. It concludes my master in Operations Management and Logistics and brings my student time at the Eindhoven University of Technology to an end.

First of all I want to thank my company supervisors Sander van den Broek and Wiebe Konter. Thanks to you both for let me conduct my master Thesis at your company. I learned much about new product introduction in an e-commerce environment. It has been a good time discussing how we could overcome problem situations by looking at it from different perspectives and attempting to find the best practical solution.

Furthermore, I particularly want to thank my first supervisor, Akçay. I really appreciate his amount of feedback to help me out when I got stuck. Especially his quick responses and clear communication ensured a pleasant cooperation. Moreover, I would also like to thank my second supervisor, Nevin Mutlu. The feedback she gave me during our meetings were very helpful. I would also like to thank my third supervisor Willem van Jaarsveld for taking the time to review my report and showing interest in my work.

My student life would not be as great without my fellow study mates and friends. I would like to thank them all for supporting me during the years as a student.

Last but not least, I would really like to thank my parents and my girlfriend who supported me along the way and teaching me that the sky is the limit. Without them it would not be possible to finish my masters.

List of contents

List of Figures

List of Tables

List of abbreviations and definitions

ANN	Artificial Neural Network
BFGS	Broyden–Fletcher–Goldfarb–Shanno algorithm
CART	Classification and Regression Tree
CHAID	Chi-square Automatic Interaction Detection
CP	Complexity Parameter
CRISP-DM	Cross Industry Standard Process for Data Mining
DT	Decision Tree
EOP	Early Offering Phase
FSD	First Stock Decision
HL	Hidden Layers
LR	Linear Regression
MDA	Mean Decrease Accuracy
MLP	Multi-Layer Perceptron
NN	Neural Network
NPO	New Product Offering
OOB	Out Of Bag
PPN	Probabilistic Neural Network
RBF	Radial Based Function
RF	Random Forest
ROI	Return On Investment
RSP	Randomly selected predictors
SADP	Sales Data Points
SDP	Sales Data Point
SKU	Stock Keeping Unit
SMO	Sequential Minimal Optimisations
SMOTE	Synthetic Minority Over-sampling Technique
STDP	Stock Data Point
SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate
USDP	Unique Stock Data Point
WMS	Warehouse Management System

1. Introduction

The worldwide sales of the online retailers (e-tailers) are expected to grow by at least 10% each year from 2016-2020 (eMarketer, 2016). In other words, the value of e-commerce sales in 2016 will be doubled in 2020. E-tailers try to keep up with the market and attract new customers by increasing their assortment. The assortment can be increased with new product offerings. In this master thesis, the definition of a new product offering is:

The introduction of a product that is available in the webshop of the e-tailer for the first time.

Considering that, the product is not new to the world but is already sold by other (electronic) retailers. The introduction of new product is important for e-tailers, Thomas (1998) concludes that the introduction of new products to the market can result in a long-term financial return on investment (ROI) as well as a competitive advantage. However, the most important reason is to attract new customers. Customers are continuously seeking for novelty, especially in the online segment (Agrawal & Smith, 2015). For the traditional retailer (a physical shop owner), the introduction of new products can be expensive, due to costs as shelf space requirements, production of shelf signs, production of price tags and extremely high handling costs (Rao & Mclaughlin, 1989). In contrast, for the e-tailer the costs of inventory marketing and handling are significantly lower. On the other hand, the e-tailer has other challenges to overcome. For example, an online customer can easily switch to any other e-tailer who sells the same product, because of low switching costs (L. Zhou, Dai, & Zhang, 2007). In the (electronic) retail market, a product introduction is a complex decision. The main reason for the complexity is that there is no historical data available about the product. There might be a large amount of data available from earlier products, how to use this data to make decision for new product offerings is another practical challenge.

This master thesis is an extensive research to find a solution for the inventory decisions for new product introductions of products that are already in the market but for the first time sold by the specific e-tailer. When the e-tailer finds a potential new product, several inventory decisions have to be made. Figure 3 visualises the inventory decisions for product introductions.

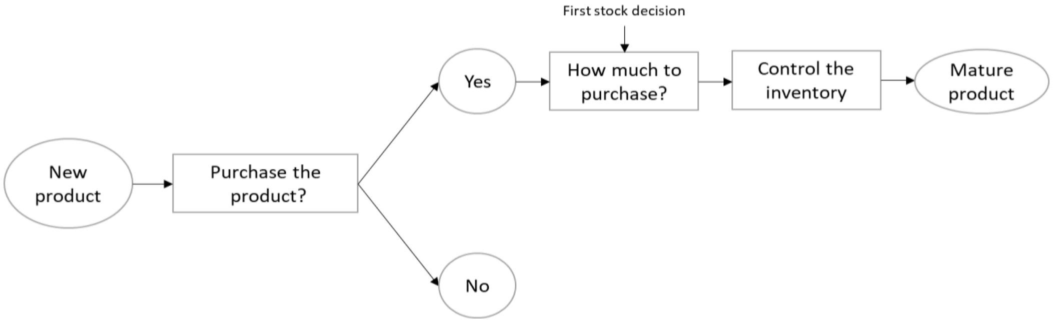


Figure 3: Inventory decisions of new product offerings

The first decision is to decide if the product is worth to purchase. Next, when a product is marked as "worth to purchase", the decision needs to be made how much to purchase of this specific product. The first decision is called the first stock decision. As last when the product arrives the inventory of that product needs to be controlled until the product reaches maturity. A stock keeping unit (SKU) reaches maturity when it has five sales data points. Five sales data points happen when the SKU is ordered five different times (a SKU order can have several of the same products). The moment the product arrives at the warehouse for the first time until the SKU reaches maturity is named the early offering phase.

In summary this research searches for a model that decides whether a product should be purchased, how much to purchase and how to control the inventory during the early offering phase of an e-tailer.

1.1 Introduction to ComCom and case study company

ComCom was founded in 2015 by two recently graduated students from the Technical University Eindhoven and has offices located in Amsterdam and Eindhoven. The start-up provides supply management software, which advises how many products the e-tailer must purchase. Currently, all the customers are online retailers, also known as e-tailers. ComCom sells his product not in “one time buy” packages, but in monthly subscriptions. The main reason for the monthly subscription is that the algorithm is continuously improved and reviewed. Figure 4 visualises the process of the product ComCom offers; (1) The e-tailers have access to an online application, the online application has an underlying model that is integrated with the e-commerce software. (2) The model, analyses the assortment, creates a forecast for each product and advises how many products the e-tailer must purchase. A dashboard visualises the model for the customer. (3) The decision tool has some Unique Selling Points (USP); minimise the lost sales, lower the inventory costs and automation of the supply.

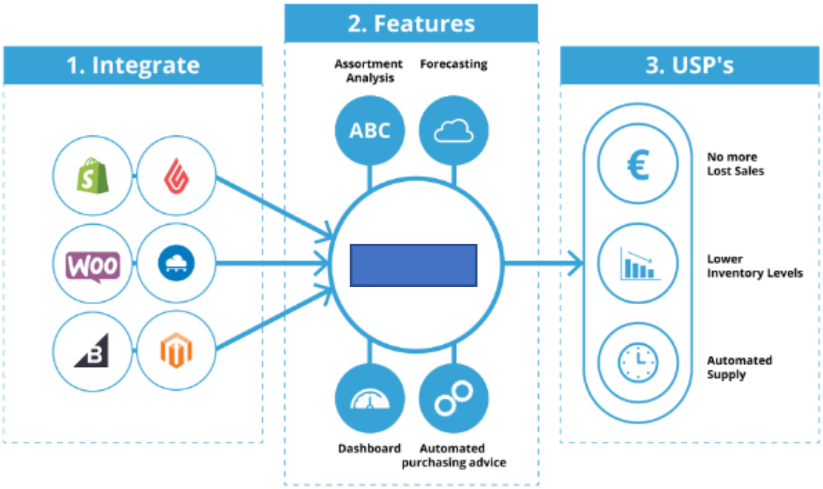


Figure 4: Process of the product of ComCom

Process

The algorithm of ComCom works in R-studio, R-studio is an open-source integrated development environment for R. R is a programming language for statistical computing and graphics. Figure 5 describes the structure of the model of ComCom. The model consists of 3 layers, a tactical, operational and cache layer.

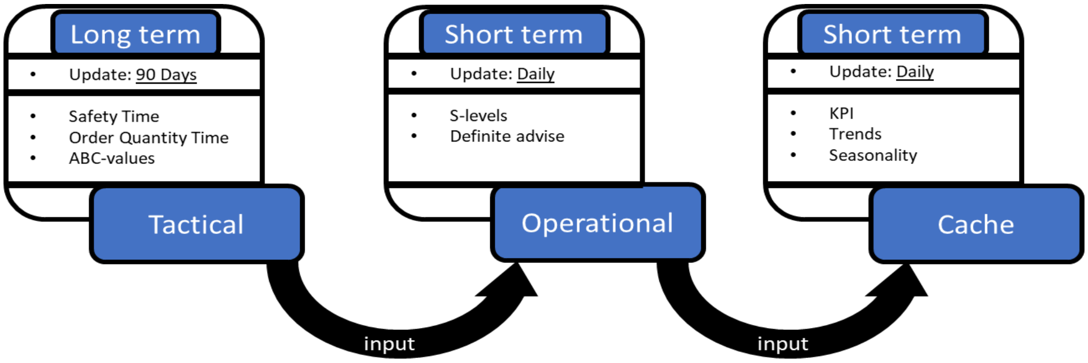


Figure 5: Model structure algorithm ComCom

The tactical level is updated once every 90 days and calculates parameters such as safety time and order quantity time, ABC-values and the agenda of the specific e-tailer. The operational level part, which is updated every day at 00:00, takes the tactical inputs for calculating different parameters, for example, the S-levels (order up to levels). Lastly, the Cache level, which also updates every day, calculates parameters such as Key Performance Indicators (KPI), revenue trends and seasonal factors. Some of the parameters need further explanation:

- **ABC-classification**: ComCom classifies a product in an (A), (B) or (C) class. In the tactical level, the algorithm of ComCom decides based on the price, purchase margin and number of sales in which classification category the product belongs. Appendix 1 explains the ABC-classification in detail.
- **S-levels**: the S-levels are part of the (R,S)-policy. The (R,S)-policy is a periodic-review, Order-Up-To-Level system and is also known as a replenishment cycle system. In short, at every replenishment period (R) the policy checks how much to order to raise the inventory position to the order up to level (S). Silver, Pyke, & Peterson (1998) give a detailed description of the inventory policy.
- **Safety time**: a way to represent safely stock as number of days demand. The safety stock is an additional quantity of an item held in stock in order to reduce risk that the item will be out of stock

Company D

The case study focusses on one company, this company is named company D. Company D is an e-tailer, has no physical store and has one central warehouse located in the Netherlands. Company D has around 7900 stock keeping units (SKUs), and a high diversity of products. Some examples of product classes are, kitchen equipment, garden tools, electronics, camping accessories, car equipment and sport equipment. The company offers a "before 20:00 ordered, next day at home" policy and offers free delivery when the order value is above €50. Company D focusses on the Dutch market, and occasionally sell products in Germany.

1.2 Motivation of study and problem statement

ComCom has indicated that there is no sufficient method available for the purchase decision of products that are available for the first time in a webshop and that there is no acceptable method for controlling the inventory during the early offering phase of this product. The general idea of ComCom is to make data driven decisions, and minimize expert judgement. Currently the decisions if a product should be purchased and how much to purchase are both made by the e-tailer based on his expert judgement. ComCom places all new product offerings in the C-class of the ABC-classification. The C-class products are marked as slow movers and have a lower service level than products of other categories, which result in a lower stock level. The inventory of the C-Class is controlled with a (R,1)-policy, the (R,1)-policy advices to order one product when the inventory on hand drops to zero. Since the classifications happen at the tactical level, a new product offering can only be re-classified from a C-class to A- or B-class, when the tactical level is updated (once every 90 days). As long as a product is classified in the C-class, the product reacts slow to changes and trend. The e-tailers believe that the new product offering is a fast mover, and need to react fast to changes and trend. Reacting fast to a sudden change is not possible for new product offerings in the current algorithm of ComCom.

The literature states new product introductions fail regularly (Eilander, 1997). The data given in paragraph 5.4 shows that 37% of all the stock keeping units of company D, are never sold before. The products that are never sold indicate that a significant amount of products are currently not profitable. This results in the following main assignement:

Develop a data-driven decision model that decides if a product should be purchased, indicates how much to purchase and the model should control the inventory during the early offering phase.

The new products are available for the first time in the webshop (the product is already available in the market), so the products are never sold before by the e-tailer.

1.3 Research Outline

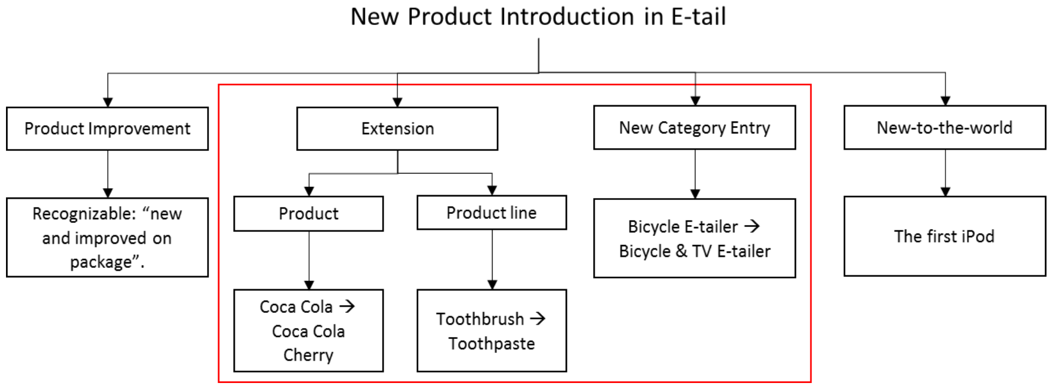
The first chapter provides a description of the company with a problem statement that is initiated by the company. The second chapter explains the new product offering process in detail. The literature based model choices is given in chapter three, the next chapter explains the research design. In the fifth chapter the data is prepared and analysed, the model building is done in the following chapter. Next in chapter seven the models are tested and evaluated. The conclusion, limitations and recommendations are given in chapter eight, and finally the implementation plan is described.

2. The new product offering process

Before the literature is reviewed the new product introduction proces is explained. It is not always clear what "kind" of new product introduction the literature means; there is not a clear definition of the new product introduction. In the first paragraph, the new product introduction is renamed and explained. In the second paragraph, the introduction phase is explained. Next, some customers of ComCom are interviewed about the new product introductions. In the last paragraph, the current product introduction process in numbers is explained.

2.1 Definition of product introduction process

There are enough papers available which discuss the introduction of new products. Kahn (2006) gives an overview of the five known different product introductions. Figure 6 shows the five different kinds of product introductions with examples.



New Product Introduction in E-tail
Product Improvement
Recognizable: "new and improved on package".
Extension
Product
Coca Cola → Coca Cola Cherry
Product line
Toothbrush → Toothpaste
New Category Entry
Bicycle E-tailer → Bicycle & TV E-tailer
New-to-the-world
The first iPod

Figure 6: New product introduction categories

The customers of ComCom are small online retailers; retailers in general usually sell products that are already sold somewhere in the world. The supplier/manufacturers who sell the products to the e-tailer, also sell their products to other (electronic) retailers. In other words, the products are widely available in the market, because of the availability the new-to-the-world products are out of scope. Furthermore, the product improvements are also out of scope. Firstly, the product improvement is not a very common occurrence in the retail. Secondly, the product is easier to control because of the availability of historical data of the version before the product. As showed in Figure 6 with the red square, the research focusses on the three product introductions: product extension, product line extension and new category entry. Since especially the literature sees a new product introduction as a new-to-the-world product, in this report the product introduction process of an e-tailer is defined as a **new product offering (NPO)** by the e-tailer. In other words, the definition of a new product offering is:

A product that is available for the first time in the webshop, but is already sold by other (electronic) retailers.

2.2 New product offering phase

There is a certain moment in time that a product is not a new product offering any more, the product will become a mature product. The moment the product reaches maturity, the early offering phase is over. To distinct the new product offering phase from the mature product offering stage, the early offer phase (EOP) of a product that is available for the first time must be defined. For ComCom both the moment in time (T) and the number of sales data points (SDP), distinguish the early offering phase from the mature phase. A detailed explanation of sales data points in given in Appendix 2. When a customer orders a certain amount of a product, an extra sales data point becomes available.

For ComCom, a product is in its early offering phase when there are not enough sales data points available to forecast the product. Currently, ComCom uses the exponential smoothing state space mode (ESSSM) to forecast the products (R. J. Hyndman, Koehler, Snyder, & Grose, 2002)(R. J. Hyndman, Akram, & Archibald, 2008).

Early offering phase for ComCom

Hyndman and Kostenko (2007) state that you need more observations than parameters, to use the exponential smoothing state space model. The ESSSM uses four exponential smoothing parameters (α, β, γ and ϕ), this means that the model needs five sales data points to make a forecast model. Since ComCom uses the ESSSM model, the early offering phase of ComCom is from the moment the product is first in stock till five sales data points are reached. The early offering phase of ComCom is visualised in Figure 7.

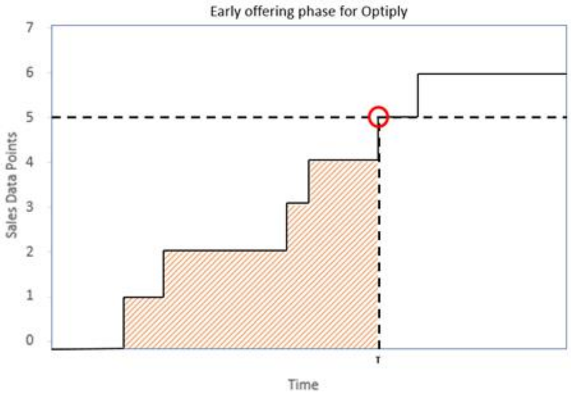


Figure 7: Early offering phase for ComCom

2.3 The current new product offering process of the e-tailer

Each e-tailer uses different methods during his product introduction process. Three different e-tailers are interviewed on how products are introduced in their company, Table 1 summarises the different approaches. The process has several steps: find the product, decide how much products to purchase and set the price. At the moment the product arrives at the warehouse, the company places the product immediately online for sale.

Table 1: Product introduction process for different e-tailers

Company	**X**	**Y**	**Z**
Find the product	(Scrape) Supplier Website Offered by supplier Expert Judgement Offered by supplier	Offered by supplier Facebook (trend) Expert Judgement Searching Internet	Via Supplier Searching internet Expert Judgement Based good selling categories Visit fairs Spy tools
Number of products	Lead Time Batch sizes Bulk discount Delivery Costs Costs per product Supplier information Expert Judgement.	Lead Time Delivery Costs Costs per product Expert Judgement Supplier information	Lead Time Expert Judgement Buy only a few products when LT is short (trial and error), Margin per product
Set the price	Using a Price Tracker, Product does not has to be the cheapest.	Around the internet price.	Compare the prices online Try to set price in between. Number of competitors decides price (little competition = high margin)

The e-tailers point out that the model of ComCom reacts slowly to trend changes for product introductions in webshops. All the e-tailers indicate that the experience and judgement of the purchaser are conclusive in the decision process.

2.4 The new product offering process in numbers

ComCom provides data of company D, which has in total 7900 stock keeping units. All the SKUs are new product offerings at some point. However, not all the data is reliable enough. In paragraph 5.2 the steps that are performed to clean the data are explained. After the data cleaning, there are in total 1448 new product offerings which are in stock for at least 65 days. The data gives an indication of the importance of the new product offering process. In total there are 1448 new product offerings analysed, the results of the analysis are given in Table 2. There are a few noteworthy results of the analysis:

- 38% of the products are never sold before (547/1448).
- 72% of the products are still in there early offering phase (1049/ 1448)
- On average the mature products reach maturity (reach five sales data points) in 158 days.

Table 2: New product offering in numbers of company D

company	***Total New Product Offerings (NPO)***	***Total NPO With Sales Data Points (SDP)***	***Total NPO Without Sales Data Points***	***Total NPO, Which are Still in EOP***	***Total NPO Reached: Maturity***	***Average Early Offering Phase Time (Days)***
D	1448	901	547	1049	399	158

3. Literature based model choices

This literature review discusses the assortment selection, the classification problem and the identified gaps in the literature. Since a general literature study is already performed on the new product introduction, the literature is specified to the current master thesis subject. The model decisions are based on the literature and the available data.

3.1 Assortment Selection

One part of assortment selection is the decision to offer new products to the market. Honhon, Jonnalagedda and Pan (2012) discuss the retailers product decision process which is largely driven on the assumption how customers make choices in brick and mortar (physical) stores. The brick and mortar businesses which are widely discussed in the literature are the fashion and supermarket business. In Appendix 3 the e-tail market is compared with the supermarket and fashion market. The comparison of the different markets, shows that the supermarket and fashion markets are in some categories comparable, however in the most categories significant different. In summary Appendix 3 describes that the e-tailers sell products only online, enter new markets more often, have online data available (e.g. reviews, clicks), limited assortment restrictions, have only one central warehouse, handling costs are much lower, no shelf restrictions and easily lose customers because of the low switching costs and purchase in smaller volumes. The conclusion of the comparison is that the assortment selection for the e-tailer is significant different than the brick and mortar businesses.

3.2 Classification

It is necessary to link the new products to the historical data of other old products since no historical data about the new product is available. Deciding whether to purchase (binary classification) a product and the classification of the product (multiple classification), are both classification problems. Classification learns patterns from past data in order to place new instances into their respective groups or classes (Thomassey & Happiette, 2007). A classification learns the function between the independent variable and their output variable through a supervised learning process, both of the variables are presented to the classification algorithm. There are many classification methods available, but globally they can be categorised into statistical and machine learning methods. The statistical methods are practical and easy to interpret (De Andrés, Landajo, & Lorca, 2005), but can have difficulties when applying in real life because the researcher needs to appoint structures to different models (e.g. linearity for regression models). Many studies compared the traditional methods with the machine learning methods and indicate that machine learning outperforms the traditional methods regularly but not always (Sexton & Dorsey, 2000). When the relationships in the data are complex and/or non-linear, statistical methods become unreliable. With complex and non-linear data, the literature recommends machine learning techniques (Altman, Marco, & Varetto, 1994).

Statistical based classification

For comparison reasons one statistical method is modelled. Logistic regression and linear discriminant analysis are two of the most wildly used methods for classification (Pohar, Blas, & Turk, 2004). Peng, Lee, & Ingersoll (2002) overview the logistic regression studies in 10 years and conclude that logistic regression can be a powerful analytical technique when the outcome is binary. Pohar, Blas & Turk (2004) compare linear discriminant analysis and linear regression in a simulation study, the authors conclude that logistic regression is more flexible and more robust when the data is noisy. When the data is noisy, the data has a large amount of additional meaningless information and the data is called corrupt.

Logisstic regression model is applied in this master thesis because, the data is noisy (paragraph 5.2.1), the method is widely used and the method is statistical based.

Machine learning based classification

Since the data is real world data, high complexity and non-linearity in the data is expected. There are different machine learning classification methods that can handle the real world data. Commonly used machine learning classification techniques are decision trees (DT), neural networks (NN) and support vector machines (SVM) (Zopounidis & Doumpos, 2002). Numerous studies compare the differences of the techniques and conclude that it is highly dependable on the context of the data which technique performs better. The ability to generalise and handle numeric data, neural networks and support vector machines are preferred over decision trees (Z. H. H. Zhou, Wu, & Tang, 2002). On the other hand, decision trees outperform the neural networks and support vector machines regarding interpretability, sensitivity to reduction in sample size and non-numerical data (Z. H. Zhou & Jiang, 2004). The paper of Byvatov, Fechner, Sadowski, & Schneider (2003) concludes that SVM outperforms the NN in prediction accuracy for a binary classification problem. Random forest (RF) is another method for classification introduced by Breiman (2001), and has comparable performance with support vector machine (Díaz-Uriarte & Alvarez de Andrés, 2006). Random forest shows excellent performance when most predictive variables are noisy, maintain accuracy when lots of data is missing and can balance errors in unbalance data sets (Caruana, Karampatziakis, & Yessenalina, 2008) (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012)

The available data when offering a new product is all numeric data (paragraph 5.4), both the neural network and support vector machine show good results in handling numeric data. Although the literature states the support vector machine outperforms the neural network in accuracy for binary classification, both the neural network and support vector machine are applied for classification in this master thesis. The neural network is included, because not only the prediction accuracy, but also the profit, interpretability and model speed are important performance metrics. Interpretability is important for ComCom, when a model is easy to interpret ComCom can easier convince the e-tailers the added value of their algorithm. Because interpretability is important the decision tree is added to assess the classification. Finally, the random forest model is used for classification, because random forest shows excellent performance even when the data is noise. The next part explains logistic regression, decision tree, support vector machine, neural networks and random forest in detail.

Decision Trees

There are numerous of well-known decision trees algorithm. Classification and regression trees (CART), C4.5 and chi-squared automatic interaction detector (CHAID) are all well-known decisions tree algorithms (Turban, Sharda, Delen, & King, 2010). Loh (2011) compared the different decision tree methods; the comparison is given in Table 3.

Table 3: Comparison of different decision tree algorithms

-	CART	C.4.5	CHAID
Splitting Criteria	Gini index	Information Gain	Chi-Square test
Branches/split	2	≥2	≥2
Attribute type	Handles both categorical & numeric value	Handles both categorical & numeric value	Handles both categorical & numeric value
Missing values	Handles missing values	Handles missing values	Handles missing values
Pruning Strategy	Cost-complexity pruning	Error based pruning	No pruning
Outlier detection	Can handle outliers	Susceptible to outliers	Susceptible to outliers
Not recommended	Small data set, binary tree is not always suitable	Very large data set, tree size increases with dataset size.	Difficult to set up, Tree size sensitive to settings, classification performance.
Information availability in R	Yes	Yes	Low information

The information availability in R is important for choosing the decision tree method, since CHAID has low information this method is not taking in to account any further. As can be seen in Table 3, the CART and C4.5 differ in splits, pruning strategy and outlier detection. Since the data set is not to small and the availability of handling outlier the CART algorithm is chosen in this master thesis for the decision tree.

A decision tree starts with all the data in the root (top) node and checks all the variables for the best split, the split is based on minimizing the impurity of a node. The CART algorithm keeps splitting until all data points are classified into mutually exclusive classes. The impurity measure is based on the Gini impurity index. The decision tree grows until no predictors can be used or the impurity of each group at a final class cannot be reduced further. A detailed explanation of the CART decision tree can be found in the paper of Breiman, et al. (1984).

Random Forest

Random forest is based on the algorithm with decisions trees (CART), the algorithm builds multiple classification trees with different samples and different initial variables. The multiple decision trees are merged into one general random forest. Random forest uses bootstrap aggregating, this technique repeatedly selects a random sample with replacement of the training set and fits the tree to these samples. The random forest algorithm is introduced by Breiman (2001), the paper gives a detailed explanation of the random forest method.

Artificial Neural Networks

Artificial neural network (ANN) often called a neural network, is a pattern recognition methodology. Several ANN are used for classification; multi-layer perceptron (MLP), radial basis functions (RBF), and probabilistic neural networks (PNN) are the most common and discussed in detail (Turban et al., 2010). Bishop (1995) stated in his research that the MLP with only two hidden layers or less is sufficient to almost any real-life problem. The RBF function trains faster than MLP, however, RBF is a more sensitive to multi dimensionally. Both the MLP and RBF are feed-forward neural networks; the PNN is completely different from the feed-forward networks as can be seen in Table 4.

Table 4: Comparison of neural network algorithms

	MLP	RBF	PNN
Layers	3≥	3	4
Hidden Layers	1≥	1	1
Trained	Back-propagation	k-mean and RLS	Spread optimization
Output Layer	Non-linear	Linear	Competitive layer
Activation function	Sigmoid (nonlinear)	Gaussian (nonlinear)	Bayesian
Time	Medium	Medium	Fast
Information availability in R	Yes	Medium	Yes

(Gorunescu, Gorunescu, El-Darzi, & Gorunescu, 2008)

The literature stated that the best performing neural network algorithm is highly dependable on the features of the data set. However, with real world cases, the MLP is more accurate than other architectures (Vali, Ramesht, & Mokarram, 2013). Since the literature states that a MLP neural network sufficient to almost any real-life problem, the MLP is applied in this master thesis.

The basis of a neural network are artificial neurons or nodes, each connected circle in Figure 8 is a node. A node (Figure 9) typical has many inputs, the inputs are all individual weighted (w_i). All the inputs are multiplied with their specific weights, the sum of this is passed into the activation function. The activation function converts the input to a more useful output. All the nodes in the model make the neural network. A neural network has a input layer, output layer and optional hidden layer(s). Figure 8 is a multilayer feed-forward network, also called a multilayer perceptron (MLP). Each layer of nodes receives inputs from previous layers. The neural network is optimized with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Fletcher, 1987).

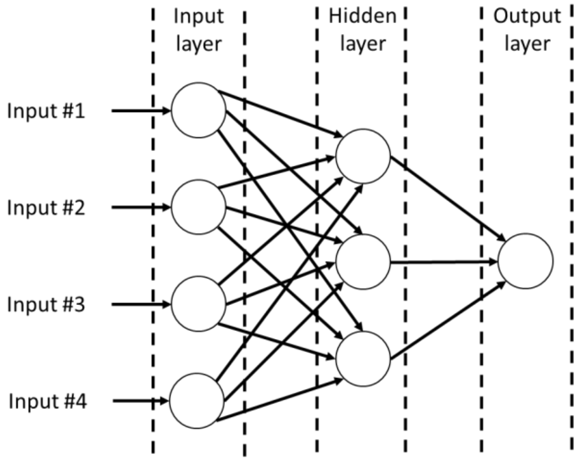


Figure 8: MLP neural network

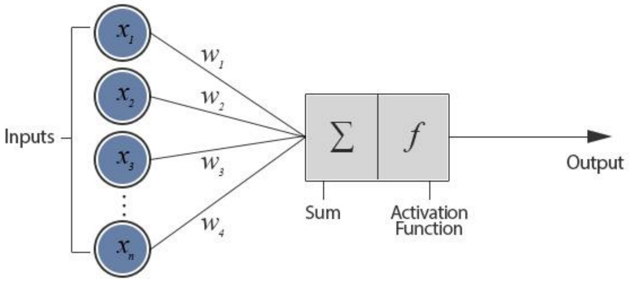


Figure 9: Node of a neural network

Support Vector Machine

Support vector machine uses a kernel trick; this is a method for using a linear classifier algorithm to solve a nonlinear problem by mapping the original nonlinear observation into a higher dimensional space. An important decision in creating the support vector model is selecting the kernel. The literature tells that the radial based function (RBF) is a good first option, because can handle non-linearity and is less complex than other kernels (Turban, Sharda, & Delen, 2011)**.** The classification with support vector machines (RBF) is based on the value of the linear combination of input variables. Figure 10 is an example of a simple two dimension problem. The examples shows the construction of maximum-margin to optimal separate different classes from each other based on the training set. A hyperplane aims to maximize the distance between support vectors. The example of Figure 10 is a simple two dimension problem; most real-world problems have data points in more than two dimensions. The parameters of the maximum-margin hyperplanes are calculated with the Platt's sequential minimal optimisations (SMO), which is described in the paper of Platt (1998).

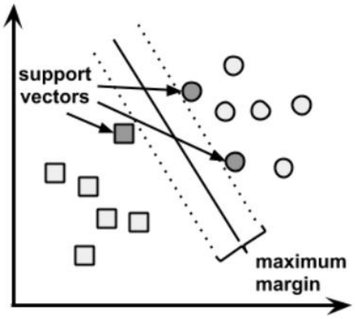


Figure 10: Support vector machine

Logistic Regression

The logistic function is the key algorithm of logistic regression. The logistic function is a sigmoid function; it is an S-shaped curve that can take a real value and map it to a value between 0 and 1. A logistic regression function uses input values (X) in combination with coefficients (β) to predict the output value (y). The coefficients are estimated for the training data with the minimization algorithm maximum-likelihood estimation (MLE) (Wilks, 1938). If the estimated probability is greater than a certain threshold, the outcome is assigned to the success group (1); otherwise it is classified into the failure group (0). Hosmer and Lemeshow (2000) give a detailed description of logistic regression.

3.3 Identified gaps in the literature

- The new product offering models are tested in some similar environments (fashion) that have some similar characteristic to e-tail. However, no literature is found on the new product offering in a e-tail environment.
- In the literature the classification models are tested based on the performance measure accuracy. The models are not measured based on operational profit or interpretability as implication of the model performance.
- There are some models in the literature that use (external) online data as demand forecasting method. However, none of the models use the online data for the decision whether to purchase a product or not. In addition to this, none of the models use the online data to make the inventory decision of how much to purchase.

4. Research design

This chapter describes the research design that is the foundation of the Master Thesis Project Research. In paragraph 4.1 the research assignment with the research objective is defined. The next paragraphs describe the methodology of the master thesis. Finally the deliverables and scope are given.

4.1 Research Assignment

Based on the motivation of the study and the identified research gaps, the research objective is defined as followed:

> **Develop a data driven model that decides whether a product should be purchased, how much to purchase the first time and how to control the inventory during the early offering phase of the e-tailer.**

Research questions:

R1: What kind of online data of competitors can be used?

R2: What is the best method to prepare the available data for analysis?

R3: Which classification models can be used to classify the data?

R4: What are the optimal S-levels for controlling the inventory during the early offering phase?

R5: What is the best model for deciding whether a product should be or not based on profit, interpretability, speed and predictive accuracy?

R6: What is the best model for deciding the amount of the first stock decision based on, interpretability, speed and predictive accuracy?

4.2 Methodology

In order to provide an answer to the research objective, the CRISP-DM methodology is used to guide through the master thesis process. CRISP/DM is a well-known data mining process methodology and stands for, Cross-Industry Stand Process for Data Mining (CRISP-DM). Figure 11 illustrates the CRISP-DM process, which is a sequence of six steps that starts with a good business understanding and ends with the deployment of the solution. Since data mining is driven by knowledge and experience of the analyst, the process can be very iterative.

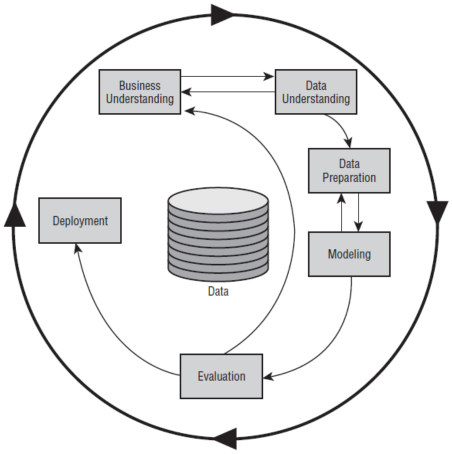


Figure 11: The Six-Step CRISP-DM Data Mining Process

The methodology is subdivided into six steps;

1. **Business understanding**: the key element is to know what the study is about. The research proposal and literature study give the insights in the business.
2. **Data understanding:** the main activity of a data mining process is to identify the relevant data from many available databases. The available data needs to be analysed to understand the data better.
3. **Data preparation:** the purpose of data preparation is to take the identified data and prepare them for analysis by data mining methods. In Figure 12, the four main steps of converting raw, real-world data into minable data sets are showed. The master thesis will follow the steps to convert the data into well-formed data. The data collection is important in the master thesis process since the data is limited external data sources must be scraped from the web first.

4. **Model building**: various modelling techniques are selected and applied to the dataset to address the specific business need. There should be a well-defined experimentation and assessment strategy to identify the "best" method for a given purpose. The data mining task can be a prediction (classification or regression), an association, or a clustering type.
5. **Testing and evaluation:** the developed models are assessed and evaluated for their accuracy and generality. Visualization techniques can be very useful to interpreted the patterns.
6. **Deployment:** it is important for the customer to understand what actions need to be carried out to make use of the created models. So, an implementation plan will be created to guide ComCom through the process. (Turban et al., 2010)

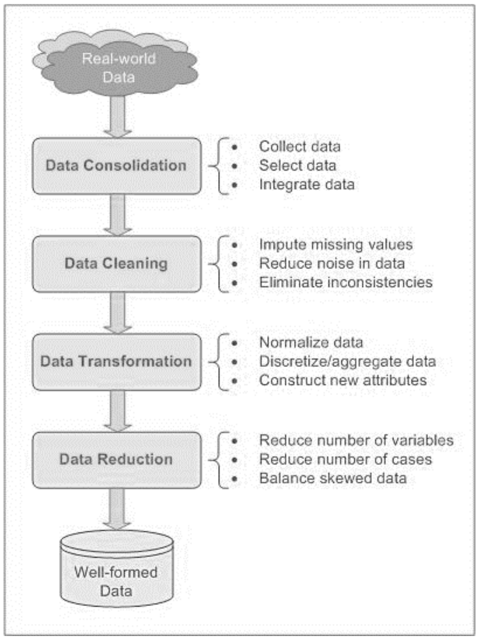


Figure 12: Data preprocessing tasks (Turban et al., 2010)

4.3 Deliverables and Scope

Deliverables:

ComCom demands some deliverables for the master thesis project:

- An inventory decision model that regulates the first purchase decision and controls the inventory during the early offering phase for new product offerings without human interference and is programmed in R studio.
- Implementation plan for the inventory decision model.

Scope:

The master thesis project needs a clear scope:

- Within the e-tail industry the focus is on; new category entries, product extensions and product line extensions.
- This master thesis focuses on the "early offering phase" of a product, so before the product is "mature" for ComCom.
- Sample and pre-sales data are not in the scope of the research. There is no special pre-sales or sample sales period; the research focusses on direct entering the market.
- Backordering is not within the scope of this project.
- The price decision is not part of the scope, the e-tailer specifies the price.
- ComCom works with an (R,S)-policy, comparing different policies are outside the scope of the project.
- ComCom indicates that new products are ordered on top of an already outstanding order. The fixed ordering and stocking costs are spread over a full order.

5. Data preparation and data analysis

For the case study company D is chosen. First the data preparation is performed. In Figure 12 the data preparation steps are given to transform the real-world data to well-formed data, these steps are used in the first subparagraphs. In Figure 13 the preparation steps are given, the 7900 SKUs are reduced to 1448 SKUs that well-formed are ready to be analysed. In the last paragraph the data is analysed.

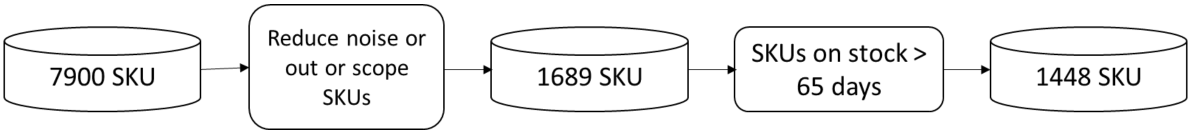


Figure 13: Data preparation SKU

5.1 Data consolidation

5.1.1 Select Data

ComCom provides the data; Table 5 shows the important variables.

Table 5: Variables

Sales Data	Stock Data	Supplier Data
Product ID	Product ID	Supplier ID
Sales ID	Stock ID	Review period (supplier)
Sales date + time	Stock date + time	Review period (user)
Amount of products sold	Inventory on hand	Fixed ordering costs
	Inventory position	Lead time

Product Data	Order Data	Company Data
Product ID	Order ID	Required service levels
Sales price	Supplier ID	Required margins for ABC
Purchase price	Amount ordered	Inventory costs margin
Supplier ID	Order Date	
EAN-code	Delivery Date	
ABC-status		

Some of the variables need some further explanation:

- *Inventory on hand:* the amount of stock available for sale at a particular time.
- *Inventory position:* the inventory on hand and the outstanding orders
- *Review period of supplier (set by ComCom):* the time between the successive evaluation of inventory to determine whether to reorder, which is determined by the algorithm of ComCom.
- *Review period of supplier (set by user):* the time between the successive evaluation of inventory to determine whether to reorder, which is determined by the user. The user is able to overrule the review period that is set by ComCom.
- *Fixed ordering costs*: all the costs for each order when ordering by a supplier, the handling costs of stocking the products are included in the fixed costs.
- *ABC-status*: the classification of a product (see Appendix 1).
- *Required service levels*: t the service levels set by the company.
- *Required sales margins for ABC*: the sales margins of A, B and C (see Appendix 1).
- *Inventory costs margin:* the percentage of inventory costs for a year and is set to 20%.

Company D has data of 7900 stock keeping units available. However, the data contains noise and needs to be cleaned before it is well-formed data. The reason for the noise in the data is that the majority of the data is entered manually by the e-tailers. The e-tailers do not always see the importance of having clean data. For example, a EAN code is 12 or 13 characters long and only contains numbers, the data shows that only 78.8% of the EAN-codes meet the criteria.

5.1.2 Collect data

ComCom does not have online data available from competitors, for this reason a tool is built to collect online data. A tool to gather online data of bol.com is built. Bol.com is the biggest player in the e-commerce market with a revenue of 730 million(Twinkle100, 2006). The tool makes a connection with a Application Programming Interface (API) of bol.com. When the API connection is established the tools request with a EAN code the data that is available. The tool is able to search if a product is available on bol.com based on their EAN-code. When a product is available on bol.com, 35 variables are extracted from the website. The 35 variables are explained in detail in Appendix 4.

5.2 Data cleaning

The sales data is available since the founding of the e-tailer, company D. However, the stock data is only available from the moment a warehouse management system (WMS) is introduced. During the WMS introduction period, all the current products need to be scanned into the system. When a product is scanned into the system during this introduction period, the data does not tell when the product was in stock the first time. In other words, it is unknown when the webshop introduces a product when the product is scanned during the WMS introduction period.

A new product offering for company D

Table 6 shows the number of stock data points (STDP), sales data points (SADP) and the unique stock data points (USDP) of one company. The STDP, SADP and USDP are explained in detail in Appendix 2. For a company, the data from the first moment the WMS is introduced (“13-1-2016”) till 3 months later is analysed (“13-4-2016”).

Table 6: WMS introduction

DATE	SALES DATA POINTS	STOCK DATA POINTS	UNIQUE STOCK POINTS	DATE	SALES DATA POINTS	STOCK DATA POINTS	UNIQUE STOCK POINTS
13-1-2016	0	3036	3036	**5-2-2016**	209	1756	1219
14-1-2016	144	11199	5355	**6-2-2016**	135	1	1
15-1-2016	150	581	512	**7-2-2016**	187	0	0
16-1-2016	127	0	0	**8-2-2016**	248	1152	210
17-1-2016	114	0	0	**9-2-2016**	217	235	0
18-1-2016	148	157	0	**10-2-2016**	263	234	0
19-1-2016	127	118	0	**11-2-2016**	217	218	0
20-1-2016	176	41	0	**12-2-2016**	113	1122	312
21-1-2016	138	111	0	**13-2-2016**	95	0	0
22-1-2016	169	120	0	**14-2-2016**	156	0	0
23-1-2016	91	0	0	**15-2-2016**	232	542	27
24-1-2016	102	0	0	**16-2-2016**	195	280	0
25-1-2016	197	178	0	**17-2-2016**	233	178	0
26-1-2016	103	124	0	**18-2-2016**	197	212	0
27-1-2016	165	165	0	**19-2-2016**	150	195	0
28-1-2016	165	206	0	**20-2-2016**	146	33	1
29-1-2016	139	120	0	**21-2-2016**	200	1	1
30-1-2016	120	0	0	**22-2-2016**	268	308	1
31-1-2016	133	0	0	**23-2-2016**	341	289	2
1-2-2016	140	290	0	**24-2-2016**	306	260	1
2-2-2016	129	128	3	**25-2-2016**	230	289	2
3-2-2016	132	9099	255	**26-2-2016**	201	223	0
4-2-2016	116	110	0				

As can be seen in Table 6 in the first row at “13-1-2016” there are 0 sales but 3036 changes in stock data points and 3036 changes in unique stock data points. The e-tailer does not introduce at one day 3036 SKU, especially not in the beginning. This day is marked as a WMS introduction day, and can be seen as a day that SKU are scanned into the WMS. As explained in Appendix 2 a stock change happens when a product is sent to the consumer or an order arrives at the warehouse. This means that the sum of the number SADP and the USDP should be the STDP. However, in the weekends (in the past) there was no personnel to process the order so the order was processed after the weekend. Processing the order after the weekend means that the stock changes are also after the weekend, this explains the differences in the STDP. As can be seen in Table 6, after “15-02-2017” the WMS introduction period is over, because the SADP and USDP balance out with the STDP and there are not a large amount of product introductions on one day. In summary from the “15-02-2017” the warehouse introduction period is over and *from this moment on each SKU that is a unique stock data point is a new product offering*. The SKU that are for the first time in stock before the WMS introduction date are filtered from the data.

5.2.1 Reduce noise and out of scope data

The available data has noise in the data or the data is out of scope. Noise in the data means that the data contains meaningless data, which cannot be interpreted correctly by the models. During this steps the SKU which contain noises and the data which is out of scope are filtered:

- The stock on hand cannot be negative, often there are user mistakes involved when the stock on hand is negative. The user mistakes make the data noisy. Because of the noisy data all the SKUs which have a negative stock on hand are not included for the analysis.
- The stock position needs to be 0 or more since back orders are not within the scope of the project, the SKU that have at least once a backorder are deleted from the data.
- As explained in the paragraph above there is a certain point in time for each company that the introduction period of the warehouse software is finished. The SKU that are introduced during the WMS introduction period are excluded from the data.
- Filter out products that have an average inventory on hand of 0, the products are in the system but never sold or have negative stock on hand.
- In the data, the date when a product is sold for the first time is occasionally earlier than the first time a product is in stock. When this occurs, the data is not included in further analysis.
- There are around 7900 different SKU available, each of the SKU has a field with an EAN code. Since frequently the EAN code is not entered correctly we do not know if bol.com sells the SKU. Due to the incorrect EAN codes, the EAN codes need to meet certain standards: numbers between 0-9 and have 12 or 13 digits.
- ComCom replaces outliers in the 99th percent interval, all observations above 99 percent interval are considered outliers and replaced with the mean value of all other observation (that are not outliers).

In Table 7 the number of products that are removed with the steps are given. After the cleaning there are 1689 products left. The number of products in the table can have overlapping parts, for example a product that has stock mistakes can also have a negative on hand.

Table 7: Data cleaning removed products

Step	**Number of products**
Remove stock that has a negative on hand	927
Remove products with back orders	848
Remove products that are in stock for the first time before the WMS introduction date	4121
Stock mistakes	90
Remove products which have a wrong EAN code	1675

Finally the last step in the data reduction is to remove the SKU which are not long enough in stock to make a correct analysis. The SKUs that are in stock for a short time and not sold are non-profitable. Despite these SKUs could be profitable in the near future. As a result of this, only products that are in stock for at least 65 days are analysed. The 65 days are based on the average time before a SKU is sold for the first time. The products that are in stock for less than 65 days are filtered from the data.

5.2.2 Impute missing values

ComCom chooses to impute missing values:

- When the sales price and/or purchase price is equal to zero or not defined, the price is replaced with the mean of the relevant company.
- When the lead time, review period, fixed costs or purchase price are not defined are replaced with the mean of the relevant variable of the specific company.

5.3 Data transformation

5.3.1 Construct new attributes

Some new and more informative variables need to be extracted from the existing variables. The variables are defined in Table 8.

Table 8: Parameters and definition

Parameter	Definition
$\boldsymbol{i}$	Unique product
$\boldsymbol{d}$	Day index
$\boldsymbol{\delta_i}$	The supplier which supplies product i
$\boldsymbol{e_i}$	The time window between the first stock decision and today.
$\boldsymbol{de_i}$	The last day of the time window e_i of product i
$\boldsymbol{P_{e_i}}$	The total profit for product i during e_i
$\boldsymbol{FC_{\delta_i}}$	The fixed order and stocking costs
$\boldsymbol{SP_i}$	The sales price of product i
$\boldsymbol{PSC}$	Packaging and sending costs (this is a fixed price).
$\boldsymbol{M_i}$	The margin of product i
$\boldsymbol{PP_i}$	The purchase price of product i
$\boldsymbol{QO_{\delta_i}}$	Total amount of products ordered from supplier δ
$\boldsymbol{QD_{\delta_i}}$	Total amount of deliveries from supplier δ
$\boldsymbol{in}$	Interest percentage for inventory costs
$\boldsymbol{I_i^{oh}(d)}$	Inventory on hand of product i at day dth day after introduction
$\boldsymbol{QS_i(d)}$	Total amount of products i sold at the dth day after introduction
$\boldsymbol{QD_i(d)}$	Total amount of products i delivered at the dth day after introduction
$\boldsymbol{IC_i}$	Inventory costs of product i
$\boldsymbol{t_i}$	The early offering window for product i in days
$\boldsymbol{dt_i}$	The last day of the time window t_i of product i
$\boldsymbol{LS_i}$	Lost sales of product i at the dth day
$\boldsymbol{R_i}$	Review period of product i
$\boldsymbol{S_i}$	Order up to level of product i
$\boldsymbol{IP_i}$	Inventory position of product i
$\boldsymbol{LC_i}$	Lost sales costs of product i
$\boldsymbol{FSD_i}$	The first stocking decision of product i
$\boldsymbol{ABC_i}$	The current classification of product i
$\boldsymbol{LT_i}$	The lead time of product i

Profit of the binary classification problem

We want to create a profit function, one that calculates profit between the first stock time and today this period (e_i) is visualised in Figure 14.

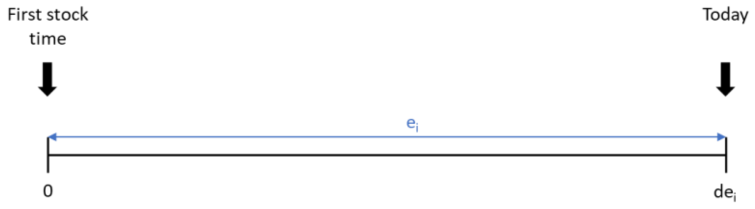


Figure 14: Time line first stock time till today

Important to mention is that the historical values of some variables are not know, when this happens the value today is used. For example only the current sales price (SP_i) is available, not the sales price at a certain moment in the past. In the profit calculation the costs of lost sales are ignored.

Calculate the current profit:
First the ordering and stocking costs per supplier per product need to be calculated. The ordering and stocking costs are fixed for a supplier. In other words, the order and stock costs are the independent of the order size. The ordering and stocking costs are divided by the average amount of products in an order.

$$OSC_{\delta_i} = \frac{FC_{\delta_i}}{\left(\frac{QO_z}{QD_z}\right)}$$

For the ordering and storing costs the assumption is made; <u>the ordering process in the past is similar to the future ordering process.</u> Next, the margin today is calculated with the sales prices minus the purchase price minus the packing and sending costs.

$$M_i = SP_i - PP_i - PSC$$

Subsequently, the total inventory costs during the first stock time and today is calculated (e_i). The average inventory on hand is multiplied with the number of days the products is in stock, the inventory interest percentage and the purchase price.

$$IC_i(e_i) = \frac{\sum_{d=1}^{de_i} I_i^{oh}(d)}{de_i} \cdot \frac{de_i}{365} \cdot in \cdot PP_i$$

$$IC_i(e_i) = \frac{\sum_{d=1}^{de_i} I_i^{oh}(d)}{365} \cdot in \cdot PP_i$$

Finally the profit period between the first stock time and now is calculated.

$$P_{e_i} = \sum_{d=1}^{de_i} QS_i(d) \cdot M_i - \sum_{d=1}^{de_i} QD_i(d) \cdot OSC_{\delta_i} - IC_i(e_i)$$

$$Total\ profit = quantity\ (i)\ sold \cdot margin(i) - quantity\ (i)\ delivered \cdot ordering\ \&\ stocking\ costs\ (i) - inventory\ costs\ (i)$$

Multiple classification problem
As explained in paragraph 6.3 the optimal order-up-to-levels during the early offering phase need to be calculated with the SKUs that have real demand and have reached the maturity phase. The SKUs with real demand are never out of stock during the early offering phase. Again the profit function is used, the only difference is that <u>lost sales are not ignored</u> because real demand is used. Only with the real demand the real optimal order-up-to-levels can be calculated. Figure 15 visualises the period from the first stock time until the fifth sales data point (t_i).

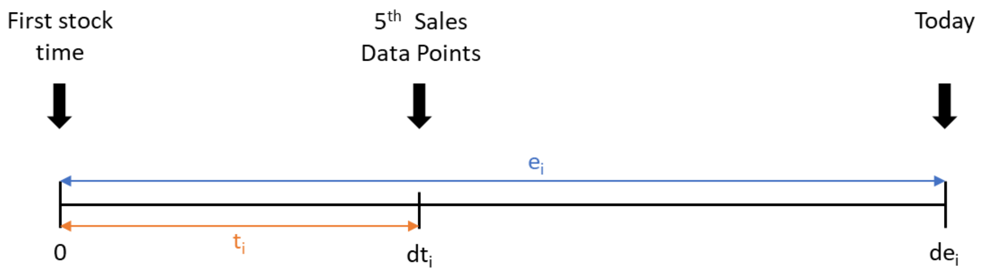


Figure 15: Time line first stock time and 5th SDP

The inventory costs of product i during the early offering phase:

$$IC_i(t_i) = in \cdot \frac{\sum_{d=1}^{dt_i} I_i^{oh}(d)}{365} \cdot PP_i$$

The lost sales costs of product i during the early offering phase:

$$LS_i(t_i) = PP_i \cdot \sum_{d=1}^{dt_i} LS_i(d)$$

The profit of product i during the early offering phase:

$$P_{t_i} = \sum_{d=1}^{dt_i} QS_i(d) \cdot M_i - \sum_{d=1}^{dt_i} QD_i(d) \cdot OSC_{\delta_i} - IC_i(t_i) - LS_i(t_i)$$

$$\begin{aligned} Total\ profit \ &= quantity\ (i)\ sold\ \cdot margin(i) - quantity\ (i)\ delivered \\ &\cdot ordering\ \&\ stocking\ costs\ (i) - inventory\ costs\ (i) - lost\ sales\ costs\ (i) \end{aligned}$$

5.3.2 Normalize data

Sola and Sevilla (1997) found out that normalisation of the input variables is crucial to obtain good results as well as fasten the calculations for neural networks. The same reasons account for support vector machines. In this master thesis the normalisation method that is described in the paper of Becker, Chambers and Wilks (1998) is used. The normalisation method transforms the data, such that every feature in the transformed dataset has a mean of 0 and a variance of 1. When classifying neural networks or support vector machines, the data is normalised.

5.3.3 Discretize or aggregate the data

Some of the variables, need to be converted for classification. Important to know if a product is profitable after at least 65 days or not (Profit Binary).

$$PB_i(e_i) = \begin{cases} 1 & P_i(e_i) > 0 \\ 0 & P_i(e_i) \leq 0 \end{cases}$$

Another important variable needs to added is if the product is currently available on bol.com (Bol Binary):

$$BB_i = \begin{cases} 1 & bol\ data\ available \\ 0 & no\ bol\ data\ available \end{cases}$$

The margin percentage is also a created variable :

$$MP_i(d_i) = \frac{M_i(de_i)}{PP_i(de_i)}$$

5.4 Data analysis

In this paragraph the data is analysed, Table 9 shows some of the characteristics of the cleaned and reduced data of company D. There are 1448 SKU left ready to be analysed.

Table 9: Data characteristics

Characteristic	Value	% total	Characteristic	Value
Total number of products	1448	100%	**Database date**	21-06-2017
			Number of suppliers	19
Number of profitable products	545	38%		
Number of unprofitable products	903	62%	**Mean sales price per product (pp)**	€22.00
			Mean purchase price pp	€9.28
Amount of A products	116	8%	**Mean ordering and handling costs pp**	€0.67
Amount of B products	411	28%	**Mean lead time pp**	32 days
Amount of C products	921	64%	**Mean review period pp**	119 days
Amount of A products with profit	103	89%	**Total costs**	€55370
Amount of B products with profit	291	71%	**Total revenue**	€85749
Amount of C products with profit	151	16%	**Total profit**	€30379
Products with sales	901	62%	**Mean first stock decision**	6.77
Product without sales	547	38%	**Mean first stock decision A products**	6.78
			Mean first stock decision B products	6.23
Products with bol.com data	594	41%	**Mean first stock decision C products**	7.02
Products without bol.com data	854	59%		

The data shows the necessity of a good purchasing process and to control the inventory in the early offering phase:

- 38% of all the SKUs that are in stock for at least 65 days are currently never sold.
- Only 38% of all the SKUs are profitable and 62% of the SKUs is not profitable.
- 62% of the products that are in stock for at least 65 days are currently not profitable
- Only 16% of all the C-classified products are currently profitable.
- The average number of (C-classified products), is higher than the average first stock decision of fast movers (A-classified).
- Appendix 5 shows that 65% of the products that are profitable are not sold on bol.com.

As can be seen in Table 9 only 16% of the C-products is profitable, this interpretation is wrong. Paragraph 1.2 mentions that all the products that do not have reached five sales data points are classified in the C-class. In other words, products that have at least five sales data points can be classified in a class correctly. Because of this reason also the products that have at least five sales data points are analysed, the products with five sales data points are visualised in Table 10. From the products that have reached five sales data points 64% of the products are profitable. And again 65% of the products that are profitable are not sold on bol.com (Appendix 5).

Table 10: Data characteristics of SKU with 5 SDP

Characteristic	**Value**	**% of total**	**Characteristic**	Value
Total number of products	399	100%	**Database date**	21-06-2017
			Number of suppliers	19
Number of profitable products	255	64%		
Number of unprofitable products	144	36%	**Mean sales price per product (pp)**	€15.53
			Mean purchase price pp	€6.77
Amount of A products	107	27%	**Mean ordering and handling costs pp**	€0.47
Amount of B products	254	64%	**Mean lead time pp**	5 days
Amount of C products	38	9%	**Mean review period pp**	19 days
			Average time to reach 5 SDP	159 days
Amount of A products with profit	95	89%		
Amount of B products with profit	156	61%	**Total costs**	€43803
Amount of C products with profit	4	11%	**Total revenue**	€71330
			Total profit	€27527
Products with bol.com data	125	31%		
Products without bol.com data	274	69%	**Mean first stock decision**	7.26
			Mean first stock decision A products	7.05
Products without stock outs during EOP	222	56%	**Mean first stock decision B products**	7.50
Products with stock outs during EOP	177	44%	**Mean first stock decision C products**	6.29

The e-commerce companies have often a large range of different products in their assortment because the products have limited storage and shelf restriction. Company D has 7900 different SKUs, for the case study in total 1448 SKUs are used. The large assortment indicates a long tail, which means that the SKU in the long tail have non-voluminous and often intermittent demand. Products with intermittent demand experience several periods of zero demand, and when demand occurs it is small and highly variable in size.

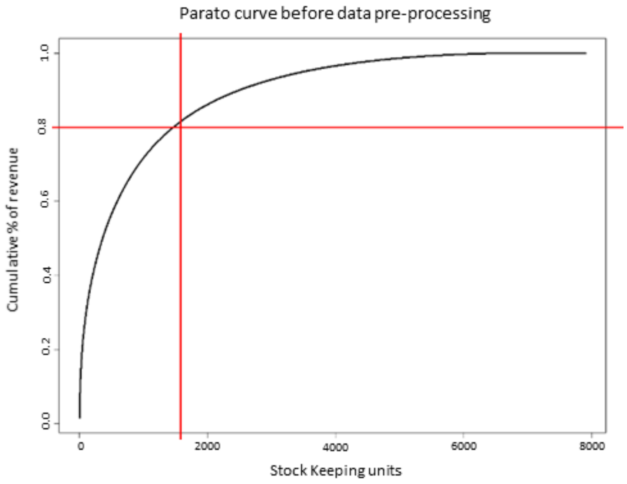


Figure 16: Pareto before data preprocessing (7900 SKU)

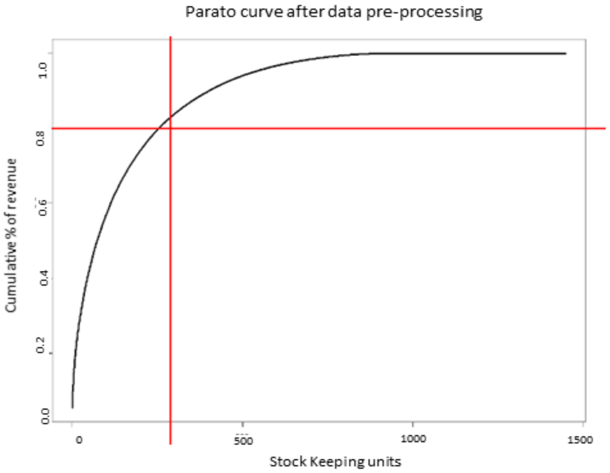


Figure 17: Pareto after data preprocessing (1448 SKU)

Long tail can be tested with the Pareto principle; 80% of the long tail assortment makes up for 20% of the total revenue of product sales (Brynjolfsson, Hu, & Simester, 2011). Figure 16 and Figure 17 show that both the data sets (before and after pre-processing) meet the Pareto criteria. In addition to this, the data set after pre-processing gives a good reflection of the reality.

Variables

The variables in Table 11 are available and relevant at the beginning of the new product offering process.

Table 11: Variables at beginning of new product offering process

Variable names	Variable	Value type
Bol.com binary	BB_i	Binary
Margin percentage	MP_i	Percentage
Margin	M_i	Number
Lead time	LT_i	Positive integer
Review period	R_i	Positive integer
Unique supplier	δ_i	Categorical value
Purchase price	PP_i	Positive number
Sales price	SP_i	Positive number
Mean handling and ordering costs	$FC_{i,z}$	Positive number

Simple split

The simple split partitions the data into two mutually exclusive subsets, namely the training and test set. It is common to assign two-thirds of the data to the training set and the remainder, one-third to the test set. The training set is used to build the model, the model is assessed with the test set (Turban et al., 2010). Figure 18 visualises the process.

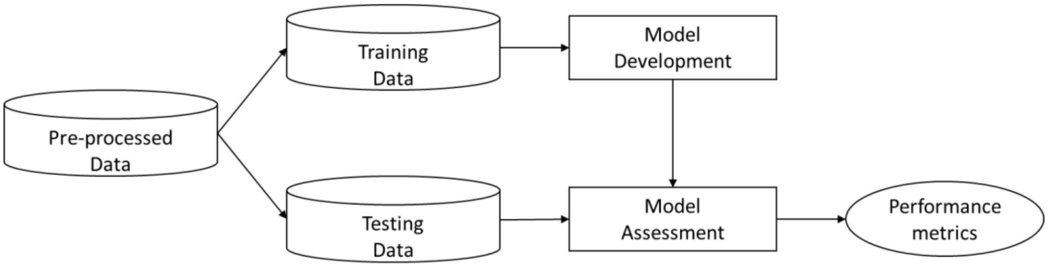


Figure 18: Simple split

Feature selection

The feature selection process selects a subset of features; the optimal features are selected based on a certain criterion. Usually, the goal of feature selection is to improve the prediction accuracy of the model and while reducing the computational costs. Additionally, feature selection can improve the data understanding and can help with visualisation of the data. The literature describes three different methods of feature selection; filter, wrapper and embedded methods (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2015)(Liu & Yu, 2005). The Boruta Algorithm is used in this master thesis because it takes into account multi-variable relationships, follows an all variable selection method, can handle interactions between variables and can deal with fluctuation nature of random (Kursa & Rudnicki, 2010). In Appendix 6 the nine steps the Boruta Algorithm uses are given.

Unbalanced data set

An unbalanced data set happens when the classes are not represented equally. Classifiers tend to provide severely imbalance degree of accuracy when the minority class is under 10% of the total data set (He & Garcia, 2009) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). As can be seen in Table 10 there are 8% A products, 28% B products and 64% C products, the data set is unbalanced. The synthetic minority over-sampling technique (SMOTE) is used to balance the data. This method is explained in detail in Appendix 7.

Performance metrics

Measuring the performance of the model on different levels gives the possibility to compare the models. The following performance metrics are used:

- Predictive accuracy: the ability of the model to predict the class label of new data correctly. The accuracy estimation is done with the confusion matrix, Appendix 8 explains the confusion matrix in detail (Platt, 1997).
- Interpretability rank: the level of understanding and insight provided by the model based on the literature and the expert judgement of the writer. The models are ranked from easy (1) to interpreted until hard to interpreted (5).
- Profit: for each simulation the profit in euro of the models can be calculated, the total profit is important in deciding which model is superior.
- Speed: the computational costs of training the model in seconds.
- Consistency: based on the standard error with confidence interval. With the simulation the standard error and the 95% confidence interval are calculated.

The profit is the leading performance metric. The e-tailer always tries to maximize the profit.

6. Model Building

The general model building exists out of three building steps, the binary classification model, the three class classification model and the optimize inventory model. Figure 19 visualises the model building process. The model (1a) decides whether to purchase a product or not, only products that are profitable in the future should be purchased. The second model (1b) decides if a product will be an A, B or C-product in the future. The last (R,S)-model is an already existing model and can only the-order-up-to-level can be optimized. With the optimal order-up-to-levels (S) for each class the performance of model 1b can be measured.

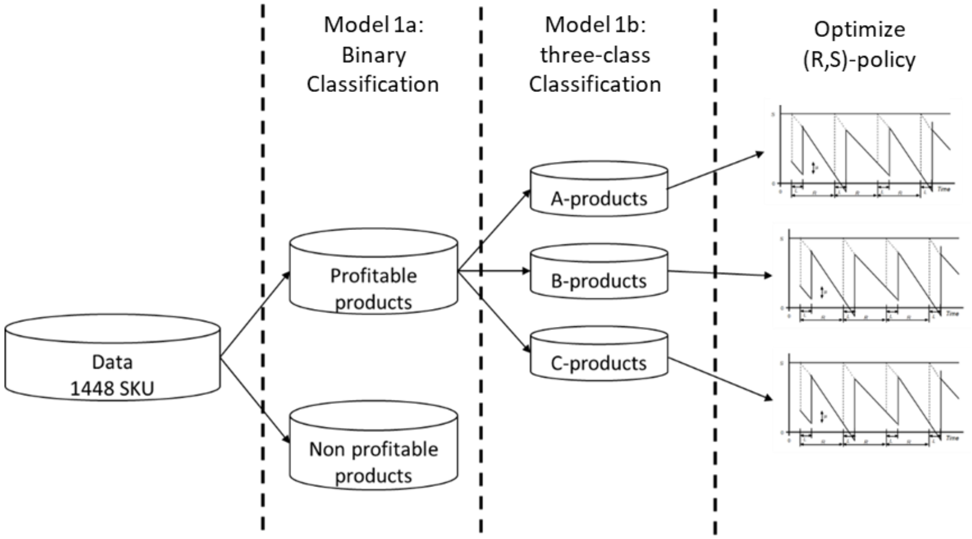


Figure 19: Three step model

Overfitting and underfitting

Several machine learning methods are used for the classification, one of the problems with machine learning is overfitting. Overfitting occurs when the model performs better on a training set than another simpler model but does worse on the test set. When a model performs better on a training set, the noise or random fluctuations in the training data is picked up and learned by the model. The machine learning techniques include parameters to limit and constrain the details of the models to avoid overfitting. Underfitting happens when a model cannot model the training data and is not able to generalise the new data. Underfitting is detected when there are low-performance metrics. The overfitting is avoided with parameter tuning, that is explained later (Caruana et al., 2008).

Bias and variance

The bias error, variance error and irreducible error together are the prediction error in the machine learning algorithm. The irreducible error is not possible to reduce and is the consequence of the chosen framing of the problem. The bias is the error from wrong assumptions in the learning algorithm. Models with a high bias learn fast and are easy to interpret. However, the more complex the model, the lower the predictive performance. Statistical machine learning methods, like logistic regression, have a high bias. Models with a low bias, are slow to learn and hard to understand but can deal better with complex models. Support vector machines and neural networks are examples of models with a low bias. The sensitivity to small changes in the training set is the variance error. When the machine learning model is strongly influenced by the specifics of the training data, the model has high variance. Models with a high variance are models with a low bias and vice versa (Yu-Wei, 2015).

The ultimate goal is to have a low bias and low variance, the bias-variance trade off. The parameters in the models can be changed to reach a low bias and low variance. For example, a support vector machine has a low bias and a high variance. With increasing the cost tuning parameter the bias is increases, but the variance decreases. The trade-off between bias and variance is an important trade off that needs to keep in mind during the model building.

Tuning

Figure 20 shows that dealing with the trade-off of bias and variance is in line with dealing with over- and underfitting. To prevent a model from overfitting the models are tuned; the tuning removes the details the model learned. To limit the overfitting k-fold cross validation is used, this is a resampling technique (Yu-Wei, 2015). Resampling decreases the change of overfitting and decreases the high variance.

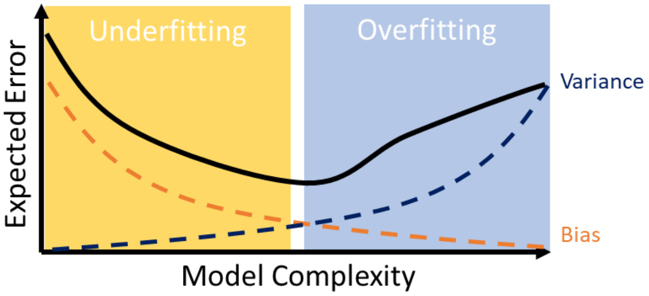


Figure 20: Trade-off bias, variance, underfitting and overfitting

6.1 Binary Classification problem

In this part the binary classification models are build, Figure 21 visualises the binary classification model.

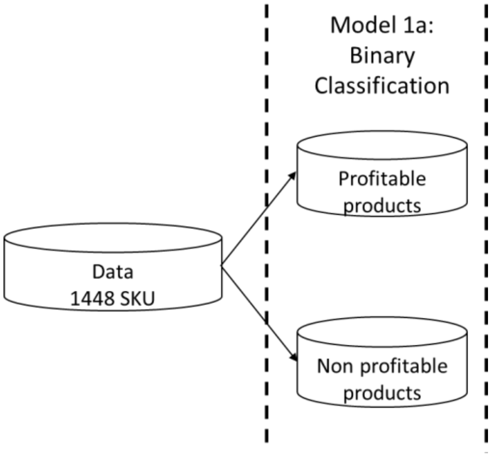


Figure 21: Binary Classification Model

For the binary classification five different models are build, CART tree, random forest, logistic regression, RBF support vector machine and MLP neural network. Paragraph 3.2 explains the decisions. All the models start with nine input variables that are available at the start of the new product offering process: bol binary, margin, margin percentage, lead time, review period, supplier, purchase price, sales price and the ordering and stocking costs per product. The goal variable is the profit binary, all the variables can be found in Appendix 9. The data is divided into a training and test set, as mentioned in Figure 18. All the models have tuning parameters and are given in Table 12.

Table 12: Binary classification models

-	DT	RF	SVM	LR	NN
K-fold	**10**	**10**	**10**	**10**	**10**
Tune parameters	• **Complexity parameter**	• **Randomly selected predictors (RSP)**	• **Gamma** • **Costs**	• **None**	• **Hidden layers (HL)**
Core Algorithms	• **Gini index**	• **Bootstrap Aggregating**	• **SMO**	• **Logistic function**	• **BFGS**
Performance measures	• **Profit** • **Speed** • **Interpretability** • **Predictive accuracy** • **Consistency**				

The outcome of the model decides if the SKU should be purchased or not. Based on the outcome of the model that predicts if a product in the test data is profitable or not, the profit is calculated. First, the new variable set N is defined. Variable set N is the set of products i where the outcome of the model is predicted profitable (1). Next the total profit of the products that should be purchased is defined:

$$\sum_{i \in N} P_{e_i}$$

The profit is one of the performance metrics, the other performance metrics are speed, interpretability and predicative accuracy and are given in Table 12. The profit, predictive accuracy and consistency are calculated based on the outcome of the test set, the speed of the model building is based on the trainings set and the interpretability is based on the literature and expert judgement. A simulation over 50 instances is used to generate the performance metrics.

6.2 Three class classification

The second model is a three class classification model, because the next classification is classified in A, B and C products. Figure 22 visualises the three class classification. Not all the models described in paragraph 6.1 are suitable for three class classifications. All the models besides logistic regression are suitable for three class classifications. The CART tree, random forest, RBF support vector machine and MLP neural networks models are used for modelling the data. The same input variables are used as in the binary classification model, only the goal variable is the ABC status (ABC_i)**.** Since the data is imbalanced, the synthetic minority over-sampling technique (SMOTE) is used, which is explained in Appendix 7. SKUs are only classified in the right class when a SKU is not in the early life cycle. Only SKUs (399 SKUs) that have reached the maturity phase are used for building the three class classification model. Table 13 describes the tuning parameters and the performance measures.

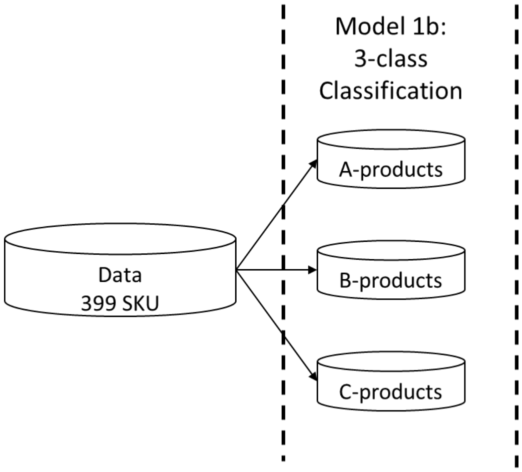


Figure 22 : Three class classification model

Table 13: Three class classification models

-	DT	RF	SVM	NN
K-fold	**10**	**10**	**10**	**10**
Tune parameters	• **Complexity parameter**	• **Randomly selected predictors**	• **Gamma** • **Costs**	• **Hidden layers**
Core Algorithms	• **Gini index**	• **Bootstrap Aggregating**	• **SMO**	• **BFGS**
Performance measures	• **Profit** • **Speed** • **Interpretability** • **Predictive accuracy** • **Consistency**			

Likewise the binary classification models the performance is tested, with different measures given in Table 13. The profit can only be calculated with an order-up-to level; the (R,S) inventory model is used to control the inventory during the early offering phase. For the testing and evaluation of the model based on the profit the optimal S-levels need to be calculated. The three class classification model predicts if a SKU will be classified as an A, B or C product, each classification has an optimal order-up-to level during the early offering phase. When the optimal value is known the order-up-to level can be used as input to calculate the profit during the early offering phase. A simulation over 50 instances is used to generate the performance metrics. In the next paragraph the (R,S)-inventory policy and the method for calculating the optimal order-up-to-level is explained.

6.3 (R, S)-model

The (R, S) inventory policy is a periodic review policy, the inventory level is observed at time intervals of the review period (R). When the observed demand (y) is lower than then the order-up-to-level (S) at moment (R), the policy orders (S) – (y) products to bring the inventory position back to the order-up-to-level. After the lead time (L), the replenishment order is delivered. Figure 23 shows the inventory position with the dotted lines and the inventory on hand with the solid lines.

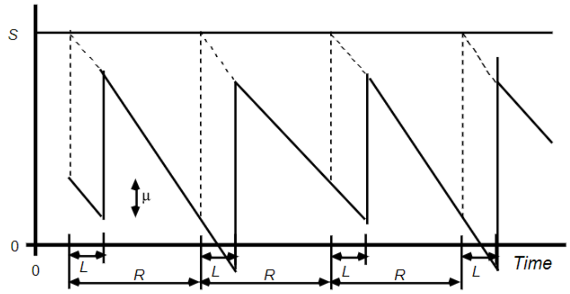


Figure 23: (R,S)-model

For each A, B and C classified SKU an optimal order-up-to level (S) during the early offering phase based on the profit needs to be calculated. The order-up-to-level is the same as first stock decision during the early offering phase. The demand of the 399 SKUs is not always known, because when a SKU is out of stock there could be demand that cannot be fulfilled. For the simulation of finding the optimal order-up-to-levels for each class the true demand is necessary. From the 399 SKUs, there are 222 SKUs that are never out of stock during the early offering phase (Table 10). For the calculation only SKU which are currently not in the early offering phase (EOP) anymore are used, these SKUs have at least five sales data points. In addition to this, only products with real demand (never out of stock during the early offering phase) are used. The SKUs that are not out of stock during the EOP have a more reliable simulation because it mirrors the reality. Figure 24 shows the steps to calculate the optimal order-up-to-levels (S).

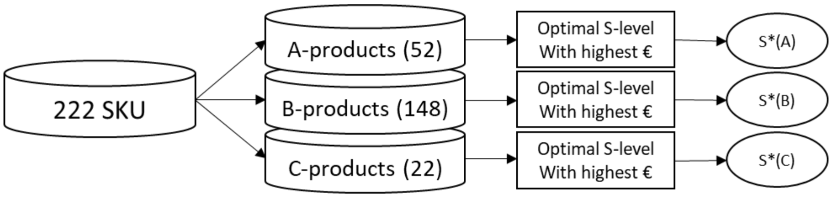


Figure 24: Optimal order level calculation

In total 222 SKUs are used for the calculation; the optimal level is based on the profit function. The profit function can be found in paragraph 5.3.1. The calculation is compared with the current inventory policy (expert judgement and R,1).

6.4 Four class classification model

Another option is the four-class classification model (model 2), the model is visualised in Figure 25. However, because of the following reason we choose to not build this model. When using the four class classification model, only SKUs that are correctly classified in a A,B or C class can be used. As mentioned earlier only products that have five sales data points can be correctly classified, in total there are 399 SKU that have reached five sales data points.

The binary classification model that decides whether a SKU is profitable or not can also include the products which are still in the early offering phase. There are in total 1049 SKU which are still in the early offering phase, 759 of these products are not profitable. In other words 72% of the products that are in the early offering phase (and in stock for at least 65 days) are not profitable. In addition to this 52% of the 1049 SKU that are in the early offering phase are never sold. In summary, there are lots of products still in the early offering phase that have a negative impact on the profit of company D. The four class classification model is not able to recognize when a SKU fails in the early offering phase, because the model is trained based on products that have reached at least five sales data points. In other words, the importance to make a good purchase decision is essential in this research. That is the reason thi smodel is not discussed.

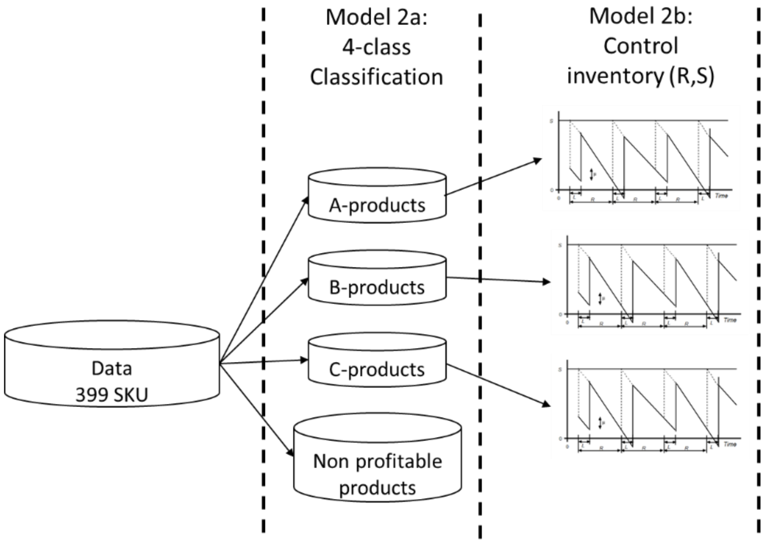


Figure 25: The four class classification model

7. Testing and evaluation

The model testing and evaluation are executed in this paragraph. Figure 26 visualises the steps that are performed to testing and evaluation both of the models.

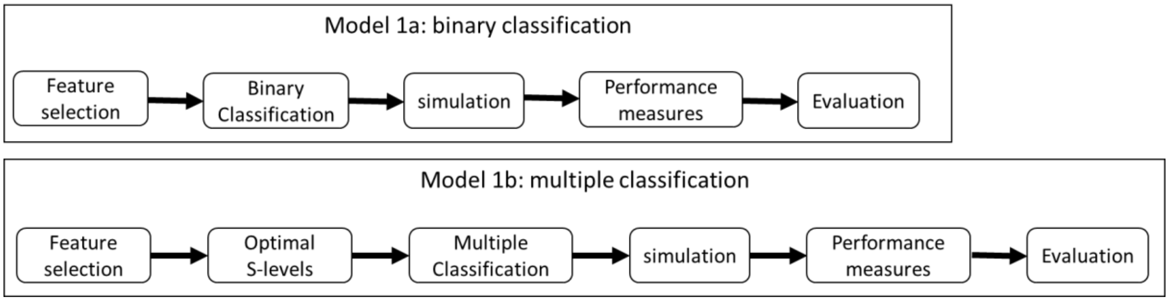


Figure 26: Steps for testing and evaluating model 1a and 1b.

In the first paragraph of chapter seven the feature selection is performed. Next the binary classification model is simulated and the performance measures are calculated. In the third paragraph the optimal s-levels are calculated, next the multiple classification model is simulated and the calculation of the performance measures is done. As last both the models are evaluated.

7.1 Feature selection

For the feature selection, the Boruta Algorithm is used. The feature selection is used to improve the prediction accuracy of the models and thereby reducing the computational costs. Both the binary classification and the three class classification model started with nine input variables (see Table 14).

Table 14: Input variables for feature selection

Variable names	Variable	Value type
Bol.com binary	BB_i	Binary
Margin percentage	MP_i	Percentage
Margin	M_i	Number
Lead time	LT_i	Positive integer
Review period	R_i	Positive integer
Unique supplier	δ_i	Categorical value
Purchase price	PP_i	Positive number
Sales price	SP_i	Positive number
Mean ordering and stocking costs	OSC_{δ_i}	Positive number

The algorithm uses the mean decrease accuracy (MDA) to evaluate the importance of each variable, with the MDA the Z-score is computed. In Appendix 6 the Boruta feature selection algorithm is explained and the numerical representation of Figure 27 is given. For the binary classification the Bol binary (BB_i) is rejected. The reason the Bol binary is rejected can be explained with Appendix 5. A large difference between the profitable products that are sold on bol.com (35%) and unprofitable products that are sold on bol.com (45%) was expected. Since the difference is small it is logical the bol.com binary is rejected. Therefore, when a SKU is sold on bol.com it has no significance influence on the profitability. For the three class classification the same feature selection method is used. The feature selection is visualised in Figure 28. The lead time(LT_i), review period (R_i) and bol binary (BB_i) are rejected. Appendix 6 gives the numerical representation of the figure.

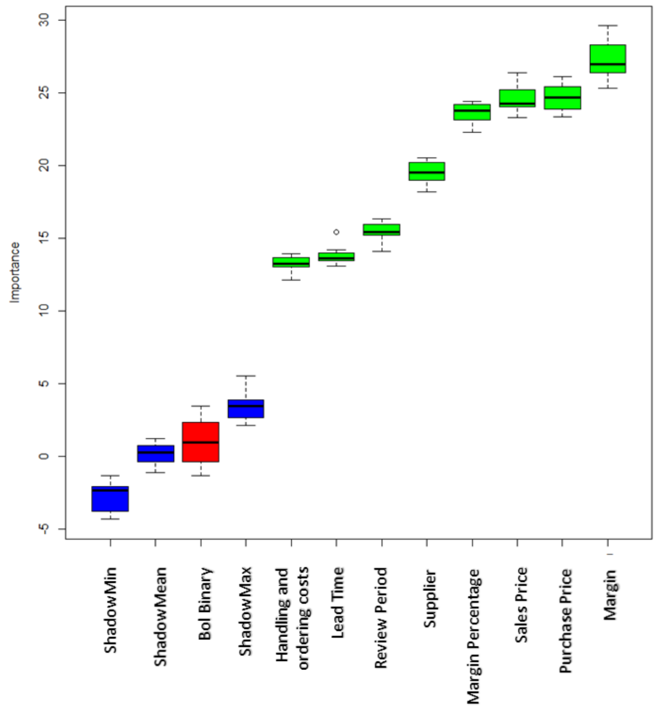


Figure 27: Boruta feature selection binary classification model

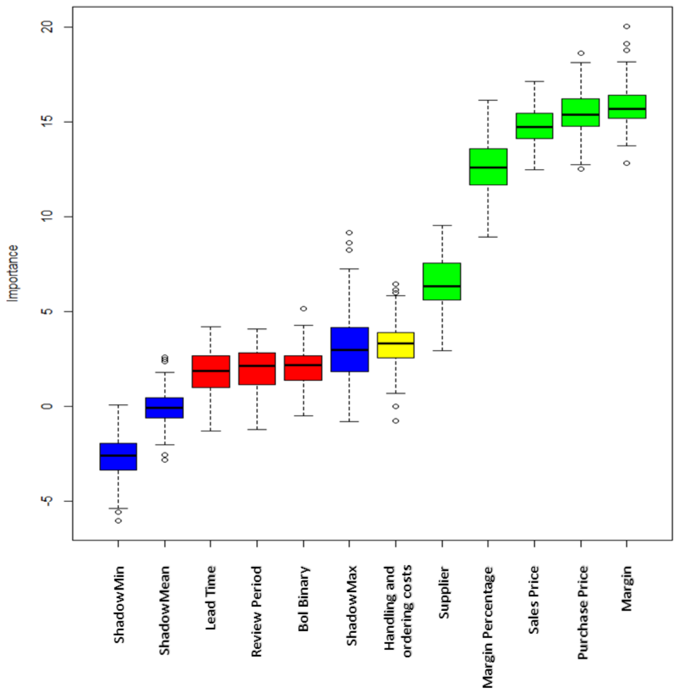


Figure 28: Boruta feature selection thee class classification model

7.2 The binary classification models

The results of the models in paragraph 7.2.1 are the results of one specific training and test set, the results in paragraph 7.2.2 are the results over 50 simulations over different test and training sets. Before building the model the balance of the data is tested. Table 15 presents the balance of the total, training and test set. A stratified random split of the data is used; this means that the train and test set have about the same partition of the profitable and no profitable SKU as the full set. As can be seen in the table, the data is not strongly imbalanced for the 1448 SKUs. Since there is no unbalance data set, SMOTE is not needed.

Table 15: Class balance

	all	train	test
No profit	0.624	0.623	0.624
Profitable	0.376	0.377	0.376

7.2.1 Model building with training data

Decision Tree (DT)

The CART decision tree under the specific training set is visualised in Figure 29. The decision tree model is built with the training set. The binary decision tree classification model starts with eight input variables given in Table 16.

Table 16: Input variables with examples

Variable names	Variable	SKU 999
Margin percentage	MP_i	200%
Margin	M_i	€5
Lead time	LT_i	10 days
Review period	R_i	20 days
Unique supplier	δ_i	110011
Purchase price	PP_i	€5
Sales price	SP_i	€15
Mean ordering and stocking costs	OSC_{δ_i}	€0.50

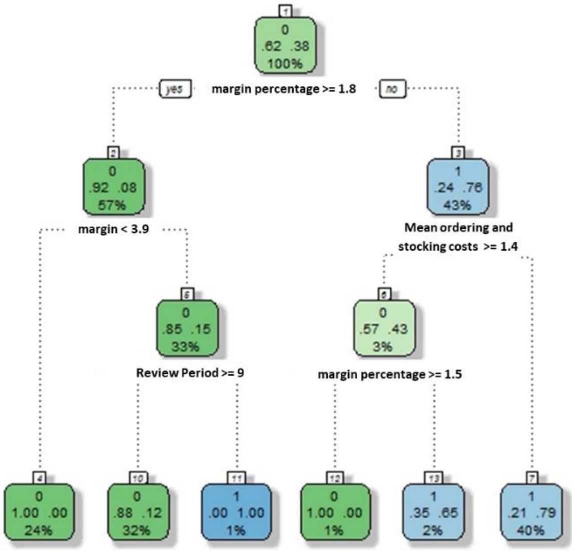


Figure 29: Classification tree

With a small example of SKU 999, Figure 29 is explained; the top of the tree asks if the margin percentage is greater or equal to 180%. The margin percentage is 200%, the decisions tree goes to the left with yes. Next the margin (5 days) is not smaller than 3.9 days so the tree goes right. As last the review period (20 days) is greater than 9 days. The products will end in the 0-class, this means no profit. Based on the training set 32% of all the SKUs will end in this "box". From all the products classified in this box, 88% is classified right (no profit) and 12% is classified wrong (profitable). The model is tuned with the complexity parameter (CP), the complexity parameter for the decision tree is given in Table 17. Also the training time and interpretability can be found in this table.

Random forest

Visualisation of the random forest makes no sense since there are several decision trees tied together. The model is tuned with the randomly selected predictor (RSP) parameter and can be found in Table 17. The model has in total 500 trees, and the out of bag error rate (OOB) of 14.39%. The speed and interpretability are also given in Table 17.

Support vector machine

The support vector machine with a radial basis function kernel is tuned with the tuning parameters costs and gamma. After tuning based on the training set the gamma and costs parameters are calculated, the parameters can be found in Table 17. The trained model has in total 350 support vectors, the time to train the 350 support vectors is showed in Table 17.

Logistic Regression

The logistic regression function only select input variables that are significant, the variables the model selected are the margin percentage (MP_i), purchase price (PP_i) and the mean ordering and stocking costs (OSC_{δ_i}). This results in the following coefficients:

Intercept	$\boldsymbol{MP_i}$	$\boldsymbol{PP_i}$	$\boldsymbol{OSC_{\delta_i}}$
7.549	-4.148	-0.029	-0.880

Also the time to train the logistic regression model is timed this is given in Table 17.

Neural Network

The neural network that is trained by the training data when it is tuned is a 7-5-1 network; the hidden layer has a size of 5 nodes. The tuning parameters, the number of hidden layers (HL), the time to train and the interpretability rank can be found in Table 17.

Tuning, speed and interpretability

The tuning, speed and interpretability parameters are given in Table 17. Based on the literature and visualisation the decision tree is the most easy to interpreted. Next the most interpretable model is the logistic regression it is clear and when the coefficients are available the calculations are easy. Random forest is harder to interpreted but is easier to explain because it are several decisions trees. As last both the SVM and NN are hard to explain because of the difficult calculations and visualisation.

Table 17: Important parameters of each model

	DT	**RF**	**SVM**	**LR**	**NN**
Tune parameters	CP = 0.0124	RSP = 2	Gamma = 0.934 Costs = 1	-	HL = 1
Time to Train	6.30 sec	30.28 sec	9.22 sec	2.18 sec	19.73 sec
Interpretability Rank	1	3	4	2	4

7.2.2 Testing the models

The model is tested, with the test data. Three performance metrics are calculated with the test data, accuracy metrics, new profit versus current profit and consistency. Figure 30 gives a representation of the steps we use to build and assess the binary classification model; in total 1448 new product offerings are used to build and test the model.

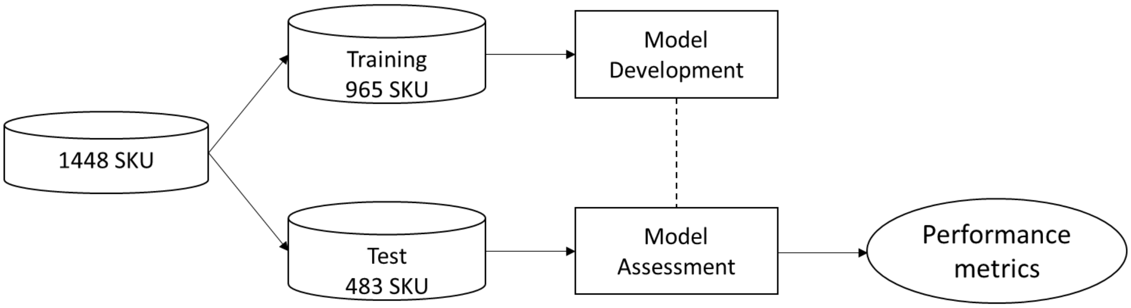


Figure 30: Binary classification model building

One of the performance metrics are accuracy metrics (accuracy, true positive rate and true negative rate), the accuracy metrics are calculated with the confusion matrix. The confusion matrix of the decision tree is visualized in Table 18. For each model a confusion matrix is build and with the confusion matrix the accuracy metrics are calculated.

Table 18: Confusion matrix

		True Class	
		Profit	No profit
Predicted class	Profit	249	52
	No profit	17	164

The second performance metric is the difference in profit; the current profit of the test data is compared with the profit of the model. All the metrics that are based on the test sets are given in Table 19.

Table 19: Performance metrics of one random sample

-	DT	RF	SVM	LR	NN
Accuracy	0.857	0.830	0.861	0.861	0.832
True Positive Rate	0.936	0.864	0.921	0.912	0.914
True negative Rate	0.759	0.773	0.779	0.788	0.731
Current Profit	€11638.76	€11638.76	€11638.76	€11638.76	€11638.76
New Profit	€14198.54	€12259.54	€14191.42	€14069.89	€14029.83
Profit	€2559.78	€620.79	€2552.66	€2431.13	€2391.07

Next the simulation is done over 50 different splits in the data, Table 20 shows the simulation results. The simulation gives as result based on the accuracy, true positive and times highest profit the neural network is the best model. The random forest outperforms the rest based on his true negative rate. The logistic regression model is the model needs the least time to build the models and is good to interprete. The decision tree has the highest average profit, the lowest minimum profit, the highest maximum profit, the lowest standard error and is also good to interpret.

Table 20: Average performance metrics of 50 simulations

-	DT	RF	SVM	LR	NN
Times highest profit	21	0	3	4	22
Average Profit diff.	€2606.95	€816.42	€1513.23	€2075.39	€2158.80
Average accuracy	0.852	0.854	0.864	0.851	0.865
Average TPR	0.912	0.886	0.914	0.904	0.915
Average TNR	0.771	0.803	0.794	0.777	0.794
Average current profit	€10374.375	€10374.375	€10374.375	€10374.375	€10374.375
Average new profit	€12981.324	€11190.794	€11887.609	€12449.762	€12533.171
Minimum profit	€-201.80	€-4394.73	€-1515.57	€-528.22	€-1474.09
Maximum profit	4276.32	3200.21	4219.24	3824.88	4234.52
Average speed	7.20 sec	35.21	8.50	3.19	21.83
Interpretability Rank	1	3	4	2	5
Predicted profitable	216	231	223	214	222
Standard deviation	818.91	1346.57	1311.02	963.56	1370.48
Standard error (mean)	115.81	190.43	185.41	136.27	193.81
lower 95% limit	2432.86	637.51	1173.29	1904.80	1845.68
upper 95% limit	2886.84	1384.02	1900.08	2438.97	2605.43

The true negative rate is the same as how much of the profitable products are classified as profitable. Another notifiable metric is the decrease in assortment. In total 486 SKUs (test set) are used for each simulation and only products that are predicted profitable are advised to purchase. From the 486 SKUs, on average 216 SKUs are predicted profitable for the decision tree model. That is a decrease in stock keeping units of 55%. In the evaluation in paragraph 7.5 the decision tree and neural network model are compared to decide which model outperforms the others.

7.3 The R, S model

For each classification, a new product has an optimal first stock decision which is the same as the order-up-to-level (S). In this paragraph the optimal order-up-to-level (S) for each A, B and C class during the early offering phase is calculated. Real demand (products that are never out of stock during their early offering phase) is used for calculations. The simulation looks for an optimal S-level which has the highest profit. First, the optimal order-up-to-level is calculated when there is no distinction between the classification classes (A, B and C). Figure 31 presents that when simulate all the 222 SKUs the S-level with the highest profit (€2527.55) is 5.

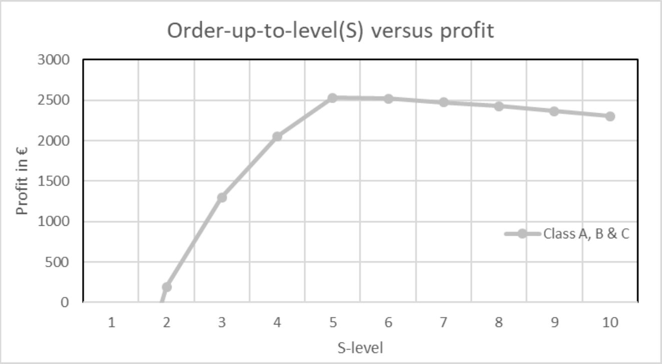


Figure 31: Profit simulation for S-level

Table 21: Comparisons of simulations A, B and C

	Optimal Profit	S-level
ABC	€2527.55	5
A	€1327.53	8
B	€1440.04	5
C	€-183.96	5
Total	€2583.71	

Next, each class is split and the costs per class are calculated. Appendix 10 shows the visualization of each simulation when the classes are simulated separately. The results of both simulations are compared in Table 21. So in other words, if all the 222 SKUs are classified in the right class the optimal profit would be €2583.71, that is on average €11.64 profit per SKU during the early offering phase. Next the real demand with the current policy (expert judgement) is simulated. Since only products that have real demand are used, there are no lost sales costs when using the current policy. The results are given in Table 22.

Table 22: Current profit real demand

	Current Profit
A	€1237.13
B	€956.51
C	€-481.14
Total	€1712.50

Again the C products are non-profitable during the early offering phase. The average profit when using the expert judgement method of the products during the early offering phase is €7.71. Next, the optimal S*(A), S*(B) and S*(C) are used for the profit calculation of the three class classification model. The order-up-to-levels for each class are 8 for S*(A), 5 for S*(B) and 5 for S*(C).

7.4 The three class classification model

To calculate the profit during the early offering phase of the thee class classification model, the early offering phase is simulated with the optimal order-up-to-levels that are calculated in paragraph 7.3. Again only products that have five sales data points (399 SKU) are used for training the model. For the profit calculations, only test products with real demand are used. These steps are shown in Figure 32.

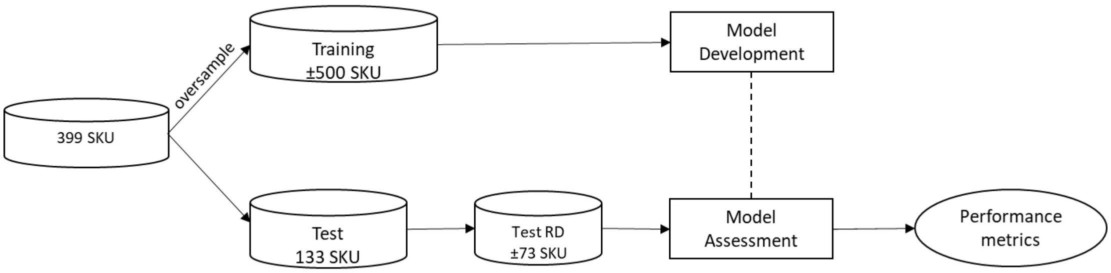


Figure 32: Thee class classification model building

Since the data is unbalanced see Table 23, SMOTE is used to oversample the training data.

Table 23: A,B and C balance

Class	Number of SKU	percentage
A	107	26.8%
B	254	63.7%
C	38	9.5%

The result of the simulation is given in Table 24. The decision tree has the shortest training time. The random forest has the highest accuracy, the highest average profit, the highest lower and upper limit and outperforms the rest 40 times based on the profit. The random forest and neural network are compared in the next paragraph to decide which model is superior to the others.

Table 24: Performance three class classification models

	DT	RF	SVM	NN
Times highest profit	1	40	0	9
Average accuracy	0.40	0.51	0.45	0.43
Average TPR	0.47	0.47	0.47	0.46
Average TNR	0.75	0.73	0.74	0.74
Average speed (sec)	1.81	9.05	3.21	9.92
Average Profit diff.	€531.25	€558.25	€537.81	€541.26
Average products	73.24	73.24	73.24	73.24
Average profit pp	€7.25	€7.62	€7.34	€7.39
Standard deviation	240.43	242.56	243.46	243.48
Standard error	34.00	34.30	34.43	34.43
lower 95% limit	€464.61	€491.02	€470.32	€473.77
upper 95% limit	€597.90	€625.49	€605.29	€608.74

7.5 Evaluation

Binary classification

The results give some important insights in the new product offering process of Company D. First of all for the binary classification; the decision tree outperformed the other models based on average profit and interpretability. Neural network outperformed the others based on times the highest profit. accuracy and true positive rating. The simulation result of the profit of the decision tree and the neural network are given in Figure 33. Table 20 shows that the decision tree has a more stable prediction than the neural network. The 95% confidence interval of the decision tree lies between the €2433-€2887, for the neural network the interval is €1846-€2605.

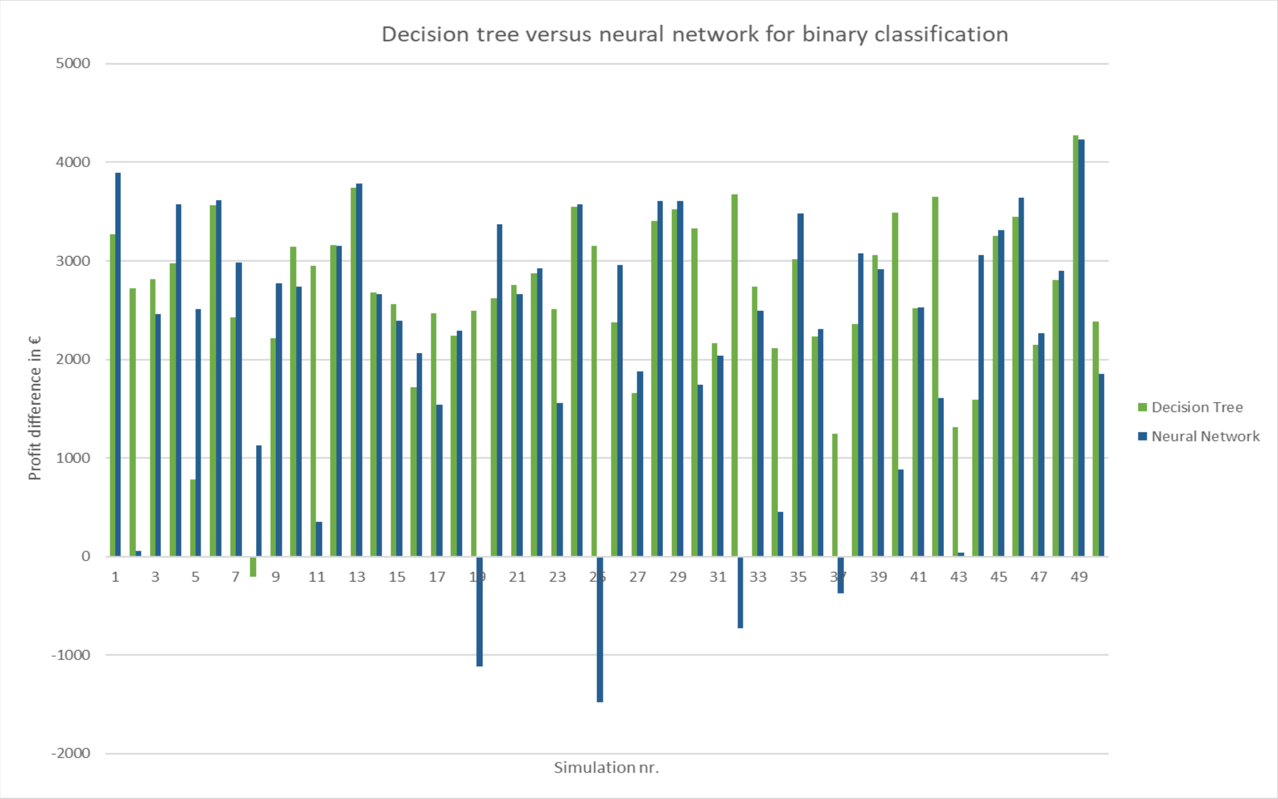


Figure 33: Comparison Decision tree versus Neural Network for binary classification

In addition to this, the neural network has four negative simulations, where the decision tree is only negative once. In other words, the decision tree outperforms the neural network based on profit, interpretability and speed. When using the decision tree for the purchase or not decision, the profit would be 26.2% higher instead of the current method (expert judgement). In addition, in the old situation 38% of the SKUs in the assortment were profitable, when using the new model 77% of the SKUs in the assortment would be profitable.

Optimize order-up-to-levels

The order-up-to-levels for each class with the lowest inventory and lost sales costs are given in Table 25.

Table 25: Optimal S

S*(A)	8
S*(B)	5
S*(C)	5

Three class classification
Finally the three class classification model are evaluated. The random forest has the highest accuracy, the highest average profit, and outperforms the rest 40 times (80% of the cases) based on the profit. The neural network model outperforms the rest 9 times (18%). The random forest model and the neural network model are compared based on profit during the early offering phase in Figure 34. When looking at Figure 34, the random forest and neural network have very comparable distribution. However, the profit of the random forest are in most cases higher than the neural network model.

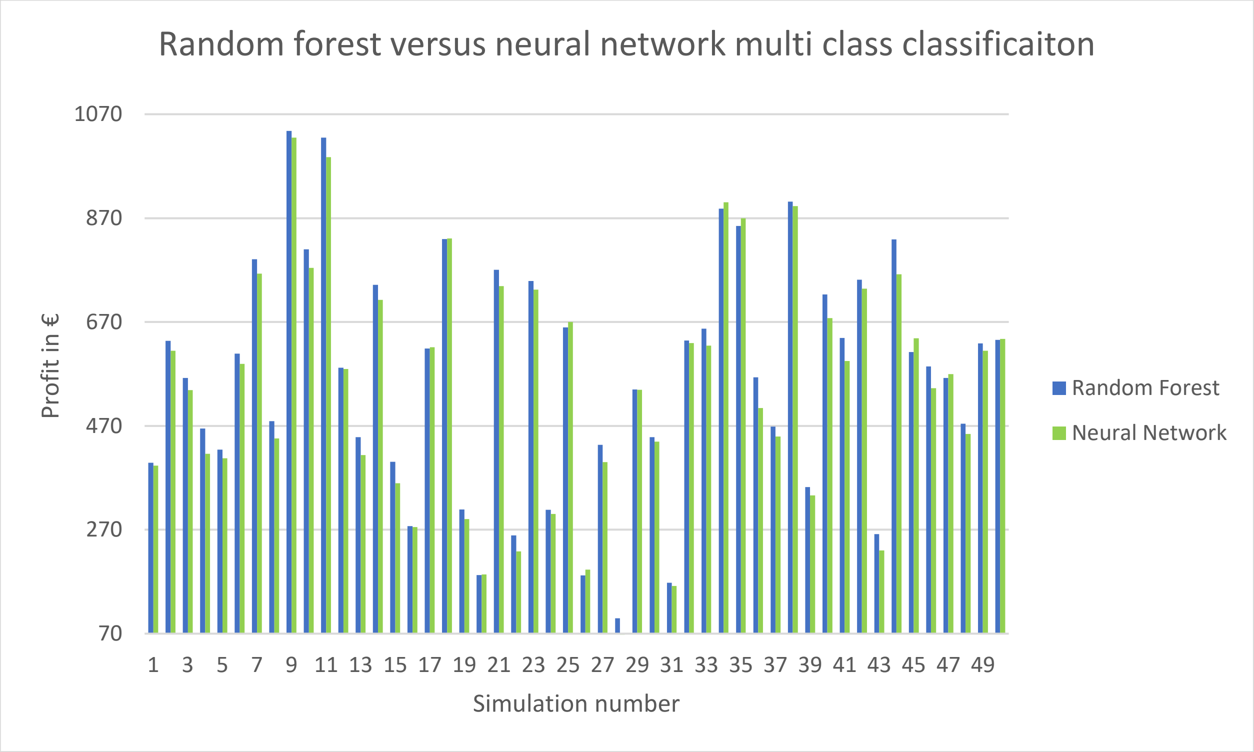


Figure 34: Comparison random forest and neural network

The random forest model results in average profit of €7.62 per SKU during the early offering phase, when using the neural network model the average costs are €7.37 per SKU. The random forest outperforms the neural network based on accuracy, interpretability, speed and profit.

8. Conclusion, limitations and recommendations

This chapter describes the conclusions, recommendations and limitations of the research. The research was performed to accomplish the following main assignment:

> *Develop a model that decides whether a product should be purchased, how much to purchase the first time and how to control the inventory during the early offering phase of the e-tailer*.

8.1 Conclusion

The initial aim of the research is to develop an automated decision model for the first purchase decision and to control the inventory during early offering phase of an e-tailer. The automated decision model for the early offering phase exist out of three different steps; the binary classification model, the three class classification model and the order-up-to-level (R,S) inventory policy.

Binary classification model

For deciding if a product should be purchased or not, several binary classification techniques are tested. The decision tree (CART), logistic regression, support vector machine (RBF), random forest and neural network (MLP) are tested on the data of company D. Despite the CART tree model has not the highest average accuracy, the CART tree outperforms the rest based on the average profit, has the highest interpretability and the is the consistence. The current purchase policy is based on the expert judgement of the e-tailer. The CART model is compared with the current policy. In summary of the comparison when using the CART tree instead of the expert judgement, the profit would be 26% higher, the assortment decreases with 55% and the percentage of stock keeping units that are profitable will increase with 39% to 78%. In conclusion the decision tree is an excellent option for the purchase decision in an e-tail environment. In addition the literature often test models based on prediction accuracy, for this data set the model with the highest accuracy is not the model with the highest profit.

Optimal order-up-to levels

The optimal order-up-to-levels for each class is calculated based on the highest profit. For controlling the inventory the optimal order-up-to-level is eight for class A, and five for B and C. The optimal profit when using the (R,S)-policy is on average €11.64 per stock keeping unit during the early offering phase. With the optimal order-up-to-levels the three class classification model is tested based on profit.

Three class classification model

For the decision to predict to which ABC-class the SKUs would be assigned in the future, four different classification techniques are tested. The tested techniques are CAR, random forest, RBF support vector machine and MLP neural network. The random forest outperformed the other models on accuracy, profit and has the highest upper and lower limit in the 95% confidence interval. The profit during the early offering phase when using the random forest model for classification is on average €7.62. In comparison to the profit of the real demand using expert judgement this is a bit lower, the profit of the real demand is €7.71. However, the profit of the expert judgement policy is based on products which do not have lost sales. Lost sales have a significant influence on the profit, what makes it hard to compare. In addition to this, the model is able to have a comparable performance as the e-tailer. For the e-tailer it is a time consuming process to decide how much products to purchase of each SKU, when using the random forest model the e-tailer has more time on hand for other jobs.

Online data

For the binary and three class classification model the online data did not have a significant effect on the outcomes. The feature selection algorithm rejected the variable bol.com binary, this binary tells whether a product is sold on Bol.com (1) or not (0). When the SKU is available on bol.com, it has no significant influence on the binary classification or three class classification model.

ComCom was looking for an automated decision model for the early offering phase, the CART tree gives an automated purchase decision, the random forest algorithm classifies the products in a ABC-class and the (R,S)-policy with the optimal order-up-to-levels control the inventory during the early offering phase. The models together is a data driven model that advices the users what to do during the early offering phase.

8.2 Limitations and future recommendations

There are several future recommendations and limitations

- It is recommended to use the decision support model to assist in the decision making process. Using the tool will lead to increase in profit, decrease of assortment and controls the inventory in the early offering phase.
- The advertisement costs are unknown of each new product introduction. Interesting would be to include the advertisement costs of a product into the model.
- Recommended is to use in the future google analytics during the early offering phase.
  - E.g. use the clicks and customer information to control the inventory
- A future recommendation is to look at the (S,s)-policy for new products to react fast to changes
- Analyse the possibility when new data becomes available during the early offering phase, products could be reclassified.
- Recommended is to include other external variables as weather and vacations.
- The data is very noisy, most of the mistakes are human mistakes. We would recommend to point out the importance of good data to the customers of ComCom.
- A limitations is that the fixed ordering costs, are the total costs of ordering and stocking of the whole order. The costs do not depend on the order size.
- There is no historical price information available, this means that only products can be classified based on the current information.
- There is no historical ABC classification available, in the past a product can be A product at a certain time, this can help with deleting the product from the assortment when life time is over.
- ComCom uses the ABC-classification method to classify the products, however the classification is not always correctly since it does not matter how the product performs if it has less than five sales data points it is placed in the C-class. We would recommend to use an extra class for products that have not yet reached five sales data points.

9. Implementation plan

This is the implementation plan for ComCom, the implementation plan exist out of several steps that need to be taken into account when building the model. All the steps are performed with R-studio.

1. Gather all the data available of a specific company
2. Gather the online data from the tool that is built for this master thesis.
3. Perform the data pre-processing steps described in paragraph 5.2
   - Set the WMS-introduction data
   - Filter the noise and out of scope data
   - Filter the products which are in stock for a smaller time period than the average time the first sale happen.
4. Transform the data with the steps in paragraph 5.3.
   - Construct new attributes
   - Normalize the data for the support vector machine and neural network (the caret package does this automatically).
   - Discretize the data
5. Train the CART algorithm with all the available data (use the caret package)
6. The trained model is the definite model which decides whether to purchase a product or not.
7. Next select the same data but only of products that have at least five sales data points.
8. Train the Random Forest algorithm with the available data (use the caret package)
   - Use the simulation model that is built for this master thesis to decide the optimal order-up-to-levels for each class.
9. Use the (R,S) policy for controlling the inventory of SKU during the EOP
   - Make a new classification level for products that have are new product, the new class is called N.

Every time the decision tree advises to introduce a new product offering is classified as N(C). Next the new product offering is classified as a A,B or C product with the random forest algorithm. If the product is classified as A, the product is name in the classification is N(A). The simulation model in step 8 gives the optimal order-up-to-level for each class. For example if the optimal order level for class A is 10, the order-up-to-level is set to 10. Since the order-up-to-level is also the first stock decision, 10 products are ordered. When the product arrives the product is controlled with a (R,10)-policy during the early offering phase. At the moment the product reaches maturity, the current algorithm of ComCom takes over.

Appendices

Appendix 1

ABC classification

ComCom uses the ABC-classification, each company sets a personal classification margin for each of the (A), (B) and (C) classes.

Company	**margin A**	**margin B**	**margin C**
X	0.70	0.25	0.05

The cumulative product sales margin is used to classify the products.

$$product\ sales\ margin = (sell\ price -\ purchase\ price) * total\ sales\ in\ last\ 365\ days$$

The table underneath shows an example of the classification of the products; the products are ranked based on the product margin. If a product has no sales data, then they are classified as (C).

Product	**Product margin**	**Cumulative product sales margin**	**Classification**
C	30	30	A
A	20	50	A
B	20	70	A
E	15	85	B
D	10	95	B
F	5	100	C

Appendix 2

Sales Data Point, Stock Data Point and Unique Stock Data point explanation

With the following an example in Table 26 the important terms sales data points, stock data point and unique stock data point are explained. As can be seen in Table 26 there are in total two orders in the sales data table, order A1 and order A2. A order is ordered by a customer of the e-tailer it is a business to consumer (B2C) order. A sales data point (SADP) is connected to a stock keeping unit, for example product (11111) has in total three sales data points at “01-01-2017 00:00:00”, “01-01-2017 12:12:12” and at “02-01-2017 12:34:56”. It does not matter how much number of products that are ordered in one order, it is always one sales data point.

Table 26: Sales Data

Time stamp	Number of products	Product ID/SKU	Order ID	Sales Data Point (SADP)
01-01-2017 00:00:00	3	P11111	A1	1
01-01-2017 00:00:00	2	P22222	A1	1
01-01-2017 00:00:00	1	P33333	A1	1
01-01-2017 12:12:12	10	P11111	A2	1
02-01-2017 12:34:56	1	P11111	A3	1
02-01-2017 12:34:56	99	P33333	A3	1

Next we explain the stock data points, a stock data point (STDP) happens when an products that is ordered by the e-tailer (a business to business order) arrives or when a products is send to a consumer. When the order ID starts with a B, the order of the e-tailer arrives at the warehouse. For example at “01-12-2016 00:00:00” an order (B1) arrives at the warehouse (100 products of P11111, 20 products of P22222 and 100 products of P33333). Every SKU that arrives is a stock data point, no matter how much products there are arrive. As mentioned earlier when a product is send to a consumer it is also a stock data point. As can be seen in Table 27, at “01-01-2017 12:12:12” 10 products are sold. In Table 27 at “01-01-2017 12:20:00” the products on hand changed from 97 to 87, also this change is one stock data point.

Table 27: Stock data

Time stamp	Number of products on hand	Product ID/SKU	Order ID	Stock Data Point (STDP)	Unique Stock data Point (USDP)
01-12-2016 00:00:00	100	P11111	B1	1	1
01-12-2016 00:00:00	20	P22222	B1	1	1
01-12-2016 00:00:00	100	P33333	B1	1	1
01-01-2017 04:01:00	97	P11111	A1	1	0
01-01-2017 04:02:00	18	P22222	A1	1	0
01-01-2017 04:03:00	99	P33333	A1	1	0
01-01-2017 12:20:00	87	P11111	A2	1	0
02-01-2017 12:50:00	86	P11111	A3	1	0
02-01-2017 12:50:08	0	P33333	A3	1	0
10-01-2017 12:00:00	50	P33333	B2	1	0
15-01-2017 12:00:00	10	P44444	B3	1	1
15-01-2017 12:00:00	15	P55555	B3	1	1
15-01-2017 12:00:00	32	P22222	B3	1	0

The unique stock data points (USDP) or first stock observation, is the first time a SKU arrives at the warehouse. Table 28 shows that there are in total 5 SKU (P11111-P55555), each SKU has a time stamp for the first time the SKU arrived at the warehouse. The unique stock data points are showed in Table 28.

One SKU always has one unique stock point and one unique stock date. The unique stock data point time can also be referred as the first stock date. And the amount of products ordered of the unique stock date point is named the first stock decision.

Table 28: Unique Stock Data

Time stamp	Number of products ordered	Product ID/SKU	Order ID	Unique Stock data point (USDP)
01-12-2016 00:00:00	100	P11111	B1	1
01-12-2016 00:00:00	20	P22222	B1	1
01-12-2016 00:00:00	100	P33333	B1	1
15-01-2017 12:00:00	10	P44444	B3	1
15-01-2017 12:00:00	15	P55555	B3	1

Appendix 3

Comparison of different markets

	Fashion Retail	**e-tail**	**Supermarket**
Products	Appeal items "one shot" items	All kinds of products	Food oriented, medical supplements, cleaning products.
Handling	High handling costs	Low handling costs	Very high handling costs
Lead Time	Long (months)	Short (days-weeks)	Very short (Days)
Selling Places	Online & in shops	Online	Shop (majority)
Classification of new product	More comparable products groups (t-shirts, brand etc.)	The classification can be hard because of the different kinds of products.	Classification often in food product groups (perishable vs non)
Types of Introductions	Product Extension, product line extension, Product replacements	Product extension, Product line extension & new category entry	Product extension, Products line extension
Data Availability	Point of Scanner data	Clickstream data, Reviews	Point of scanner data
Lost Sales	Very hard to measure because of the number of shops	Medium to measure, because only 1 warehouse and extra data is available (click stream).	Hard to measure, because re-stocking often in evening
Life Cycle	Fashion products often sold for one season.	Different kinds of life cycle (washing machine, till medical supplements)	Different kinds of life cycle (perishable vs. medical supplements)
New market	New product is still in the same market.	It is easier and more common that a retailer tries to expands his business with exploring new markets.	New products are in the majority in same market
Product type	Non-durable	Non-durable & durables	Most non-durable
Number of items (buy-in)	Large (1000+)	Small (1-1000)	Medium (100-1000)
Limitation	Limit shelf space, product not always visible, depends on weather data, do not how much exactly on stock or is not available	Low switching costs, easy to find concurrent, good product visibility and easy to find, cannot try product	Limit shelf space, product not always visible, do not how much exactly on stock or not available
Replenished	Often not replenished	Replenished	Replenished
Advantages	Can try product in shop, can ask "experts" their opinion	Reviews available, target advertisement, findability on google, direct stock availability, product deliver to home.	Can see and feel product in shop.
Location	In city centers	One warehouse	Near city center

(Choi, Hui, & Yu, 2014) (Wang, Head, & Archer, 2002)

Appendix 4

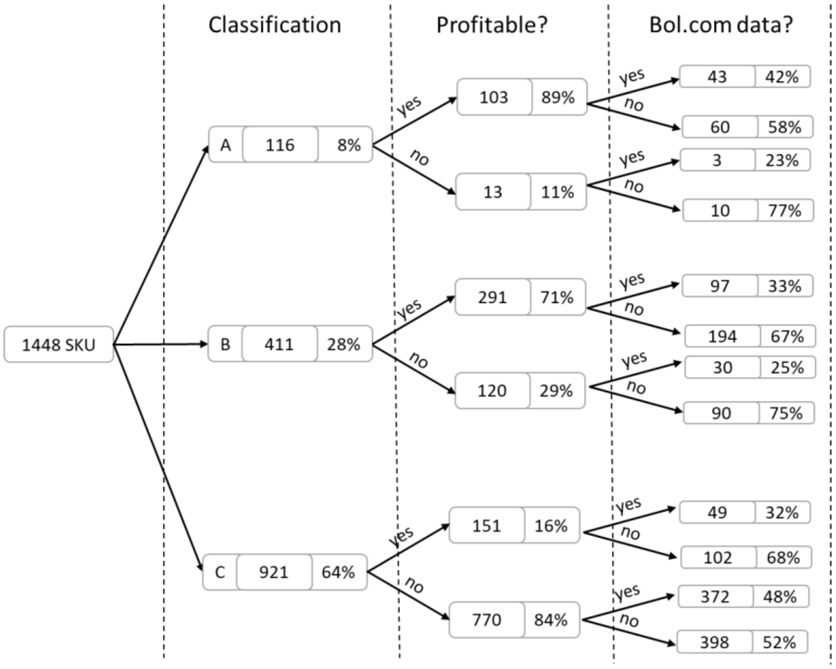
Bol.com variables

Bol.com data (external)

EAN	Electronic Article Number
GPC	Global Product Classification
Product Name	Name of Product
SpecsTAG	Author, Artist, Brand, Manufacturer or Publisher.
Rating	Rating of the product out of 50
URL	The link to product to bol.com
ImageURL	The link to the image of the product
SoldByBol	Product sold by Bol (yes=1, no=0)
SoldByNonProf	Product sold by Non-Professional (yes=1, no=0)
SoldByProf	Product sold by Professional (yes=1, no=0)
Condition	The condition of the product
StandardPrice	The standard price, not in sale
ListPrice	The current price
AvailabilityCode	The availability code of bol.com (see file)
SellerID	Bol.com seller ID
SellerType	“Grootzakelijke” or “Kleinzakelijke” verkoper
TopSeller	Indicates whether a seller is a top seller
RatingMethod	“3 months”
SellerRating	Average rating of seller out of 10
ProductInfoRating	Average product info rating out of 10
DeliveryTimeRating	Average delivery time rating out of 10
ShippingRating	Average shipping rating out of 10
ServiceRating	Average service rating out of 10
RecentPositiveRating	Amount of positive ratings in RatingMethod
RecentNeutralRating	Amount of neutral ratings in RatingMethod
RecentNegativeRating	Amount of negative ratings in RatingMethod
RecentTotalRating	Amount of total ratings in RatingMethod
AllPositiveRating	Amount of positive ratings
AllNeutralRating	Amount of neutral ratings
AllNegativeRating	Amount of negative ratings
AllTotalRating	Amount of total ratings
BestOffer	If product is the best offer; True or False
ReleaseDate	The release the product
CategoryID(1-6)	Product hierarchy ID
CategoryName(1-6)	Product hierarchy name big(1)->small(2)

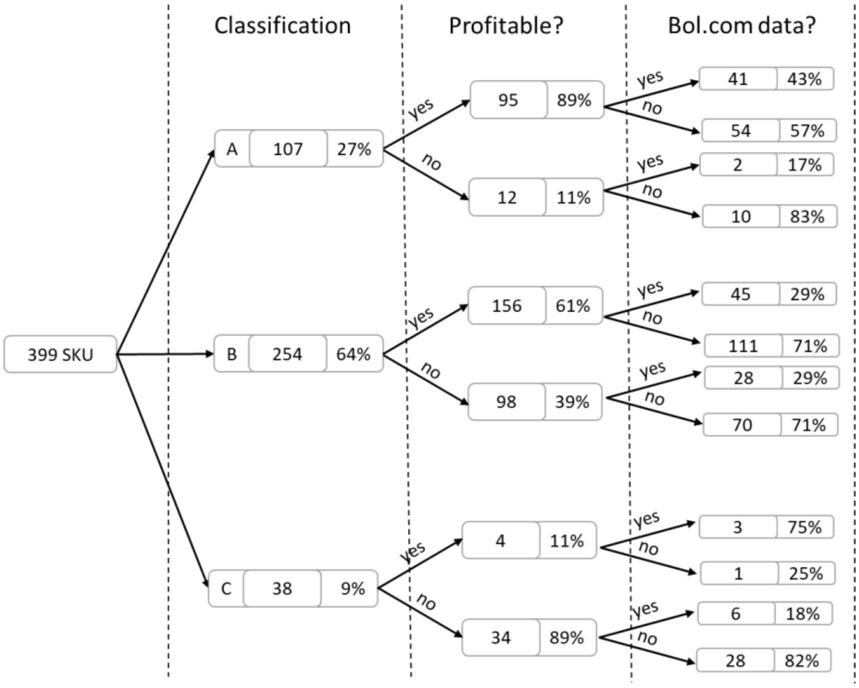
Appendix 5

Bol.com data analysis.



$$\% \, profitable \; products \; sold \; on \; bol = \frac{Profitable \; products \; sold \; on \; bol.com}{Profitable \; produts} = \frac{43+97+49}{103+291+151} = 35\%$$

$$\% \, unprofitable \; products \; sold \; on \; bol = \frac{unprofitable \; products \; sold \; on \; bol.com}{unprofitable \; produts} = \frac{3+30+372}{13+120+770} = 45\%$$



Appendix 6

Feature selection

The Boruta algorithm exist out of the following steps

i. Extend the information system by adding copies of all variables (the information system is always extended by at least 5 shadow attributes, even if the number of attributes in the original set is lower than 5)

ii. Shuffle the added attributes to remove their correlations with the response.

iii. Run a random forest classifier on the extended information system and gather the Z scores computed.

iv. Find the maximum Z score among shadow attributes (MZSA), and then assign a hit to every attribute that scored better than MZSA.

v. Find the maximum Z score among shadow attributes (MZSA), and then assign a hit to every attribute that scored better than MZSA.

vi. Deem the attributes which have importance significantly lower than MZSA as `unimportant' and permanently remove them from the information system.

vii. Deem the attributes which have importance significantly higher than MZSA as `important'.

viii. Remove all shadow attributes.

ix. Repeat the procedure until the importance is assigned for all the attributes, or the algorithm has reached the previously set limit of the random forest runs.

(Kursa & Rudnicki, 2010)

	meanImp	medianImp	minImp	maxImp	normHits	decision
Bol Binary	0.97	0.93	-1.36	3.43	0.07	Rejected
Margin Percentage	23.63	23.77	22.29	24.39	1	Confirmed
Margin	27.12	26.97	25.30	29.59	1	Confirmed
Lead Time	13.76	13.61	13.08	15.40	1	Confirmed
Review Period	15.48	15.44	14.10	16.34	1	Confirmed
Supplier	19.49	19.50	18.18	20.54	1	Confirmed
Purchase Price	24.71	24.68	23.35	26.13	1	Confirmed
Sales Price	24.57	24.24	23.30	26.38	1	Confirmed
Average Handling and Stocking costs	13.26	13.26	12.10	13.93	1	Confirmed

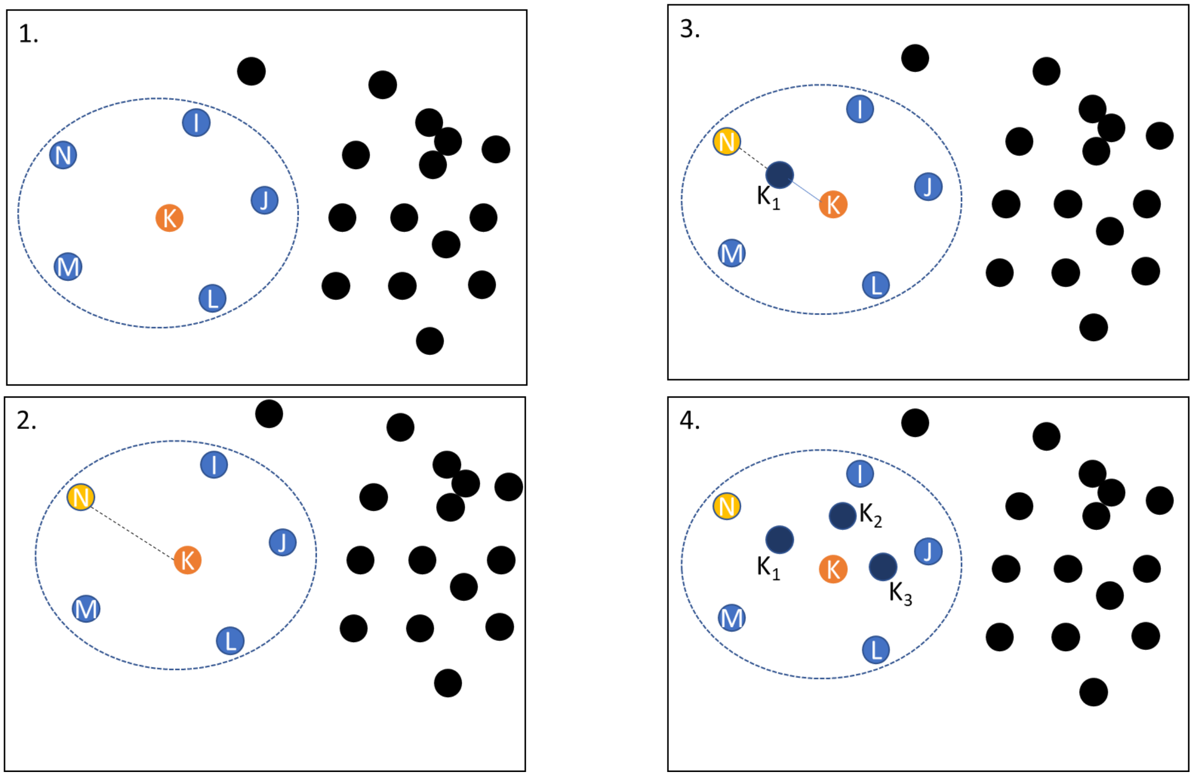
	meanImp	medianImp	minImp	maxImp	normHits	decision
Bol Binary	2.07	2.17	-0.47	5.17	0.18	Rejected
Margin Percentage	12.59	12.61	8.94	16.13	1.00	Confirmed
Margin	15.79	15.67	12.82	20.02	1.00	Confirmed
Lead Time	1.82	1.88	-1.27	4.21	0.19	Rejected
Review Period	1.98	2.15	-1.22	4.11	0.19	Rejected
Supplier	6.43	6.34	2.94	9.56	0.93	Confirmed
Purchase Price	15.50	15.40	12.51	18.64	1.00	Confirmed
Sales Price	14.77	14.72	12.48	17.15	1.00	Confirmed
Average Handling and Stocking costs	3.22	3.32	-0.74	6.44	0.53	Tentative

Appendix 7

SMOTE

Smote uses the following steps:

1. For each minority example k compute the nearest minority class examples (I,J,L,M,N)
2. Randomly choose an example out of the 5 closest points
3. Synthetically generate event K_1, such that K_1 lies between k and n
4. The data set after applying SMOTE 3 times.

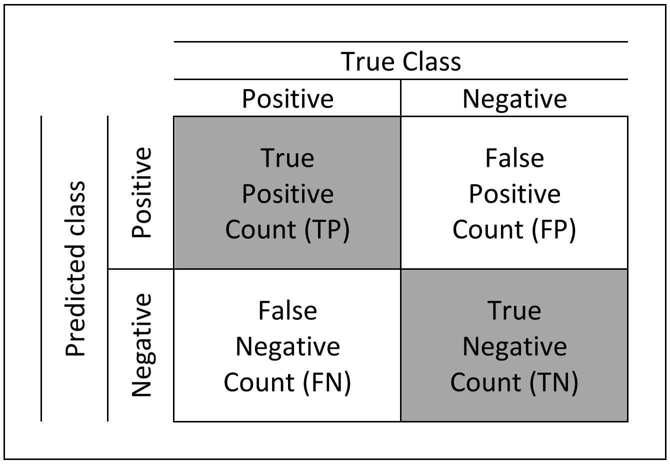


(Chawla et al., 2002)

Appendix 8

Confusion matrix

The figure shows an example of a confusion matrix of a two-class classification problem. The true positive (TP) and true negative (TN) represent the correct decisions, and the false positive (FP) and false negative (FN) represent the errors.



		True Class	
		Positive	Negative
Predicted class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

When there are more than two classes, the confusion matrix gets bigger, and accuracy metrics becomes limited per class accuracy rates and overall classifier accuracy. Each of the accuracy metrics calculations are given the table

True positive rate (TP)	$\frac{TP}{TP + FN}$
True negative rate (TN)	$\frac{TN}{TN + FP}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$

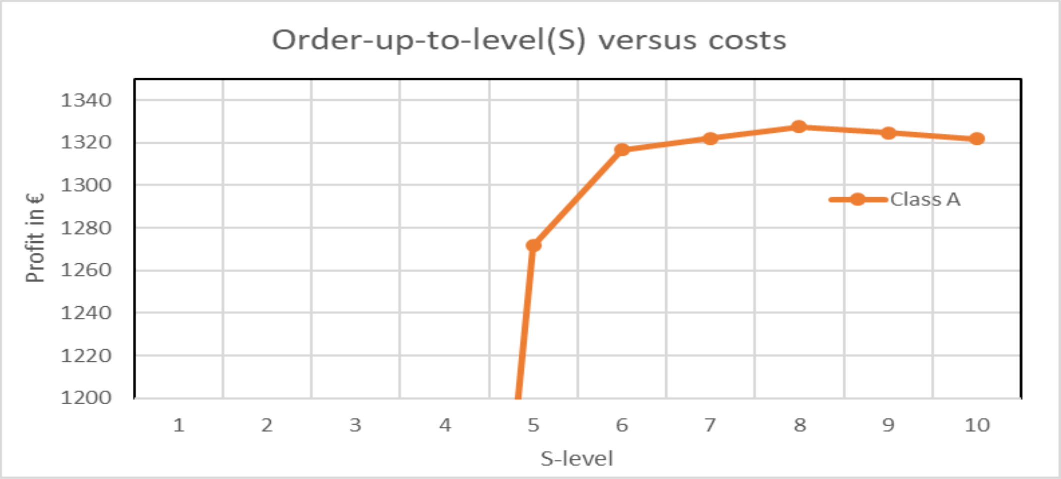
Appendix 9

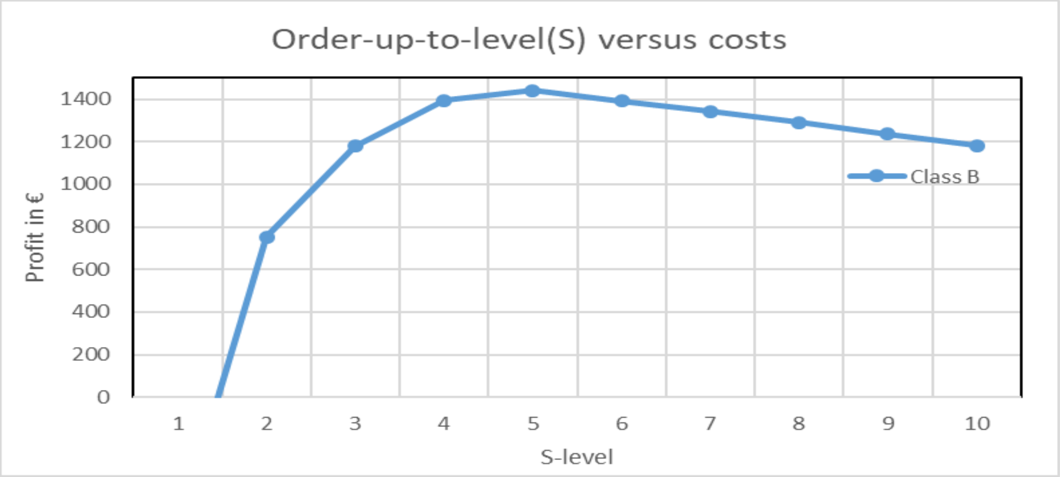
Variables list

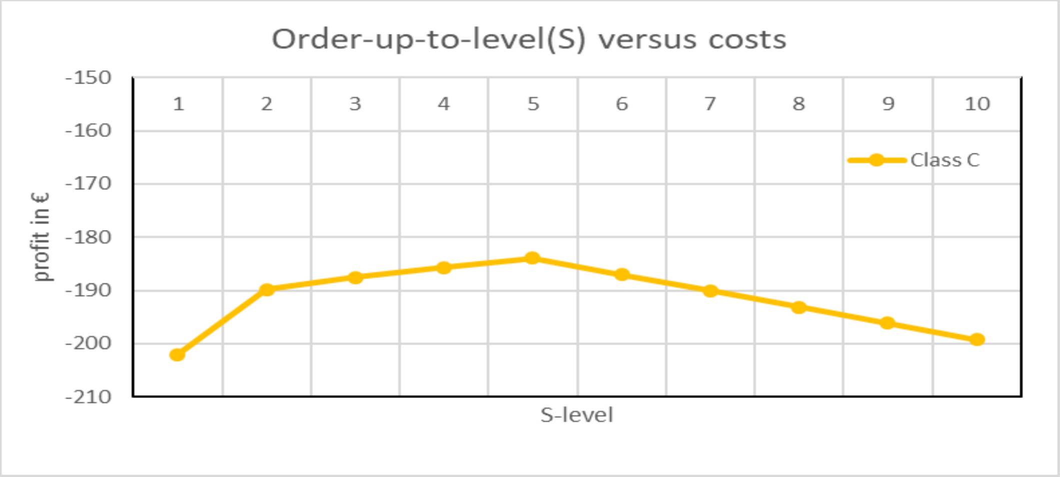
Parameter	Definition
$\boldsymbol{i}$	Unique product
$\boldsymbol{d}$	Day index
$\boldsymbol{\delta_i}$	The supplier which supplies product i
$\boldsymbol{e_i}$	The time window between the first stock decision and today.
$\boldsymbol{de_i}$	The last day of the time window e_i of product i
$\boldsymbol{P_{e_i}}$	The total profit for product i during e_i
$\boldsymbol{P_{t_i}}$	The total profit for product i during t_i
$\boldsymbol{FC_{\delta_i}}$	The fixed order and stocking costs
$\boldsymbol{SP_i}$	The sales price of product i
$\boldsymbol{PSC}$	Packaging and sending costs (this is a fixed price).
$\boldsymbol{M_i}$	The margin of product i
$\boldsymbol{PP_i}$	The purchase price of product i
$\boldsymbol{QO_{\delta_i}}$	Total amount of products ordered from supplier δ
$\boldsymbol{QD_{\delta_i}}$	Total amount of deliveries from supplier δ
$\boldsymbol{in}$	Interest percentage for inventory costs
$\boldsymbol{I_i^{oh}(d)}$	Inventory on hand of product i at day d^{th} day after introduction
$\boldsymbol{QS_i(d)}$	Total amount of products i sold at the d^{th} day after introduction
$\boldsymbol{QD_i(d)}$	Total amount of products i delivered at the d^{th} day after introduction
$\boldsymbol{IC_i}$	Inventory costs of product i
$\boldsymbol{t_i}$	The early offering window for product i in days
$\boldsymbol{dt_i}$	The last day of the time window t_i of product i
$\boldsymbol{LS_i}$	Lost sales of product i at the d^{th} day
$\boldsymbol{R_i}$	Review period of product i
$\boldsymbol{S_i}$	Order up to level of product i
$\boldsymbol{IP_i}$	Inventory position of product i
$\boldsymbol{LC_i}$	Lost sales costs of product i
$\boldsymbol{FSD_i}$	The first stocking decision of product i
$\boldsymbol{ABC_i}$	The current classification of product i
$\boldsymbol{LT_i}$	The lead time of product i
$\boldsymbol{OSC_{\delta_i}}$	Ordering and stocking cost per product i from supplier δ
$\boldsymbol{PB_i}$	The profit binary of product i
$\boldsymbol{BB_i}$	The bol.com binary of product I
$\boldsymbol{MP_i}$	The margin percentage of product i

Appendix 10

A,B and C order up to level simulation







References

Agrawal, N., & Smith, S. A. (2015). *Retail Supply Chain Management* (2nd ed., Vol. 122). Springer. https://doi.org/10.1007/978-0-387-78902-6

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*, *18*(3), 505–529. https://doi.org/10.1016/0378-4266(94)90007-8

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1998). *New S Language*. Wadsworth & Brooks/Cole.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). *Feature Selection for High-Dimensional Data*. *Springer*. https://doi.org/10.1007/978-3-319-21858-8

Breiman, L. (2001). Random Forests, 1–33.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. New York: CRC Prees.

Brynjolfsson, E., Hu, Y. (Jeffrey), & Simester, D. (2011). Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales. *Management Science*, *57*(8), 1373–1386. https://doi.org/10.1287/mnsc.1110.1371

Byvatov, E., Fechner, U., Sadowski, J., & Schneider, G. (2003). Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *Journal of Chemical Information and Modeling*, *43*(6), 1882–1889. https://doi.org/10.1021/ci0341161

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, 96–103. https://doi.org/10.1145/1390156.1390169

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(January), 321–357. https://doi.org/10.1613/jair.953

Choi, T., Hui, C., & Yu, Y. (2014). *Intelligent Fashion Forecasting Systems: Models and Applications*. https://doi.org/10.1007/978-3-642-39869-8

De Andrés, J., Landajo, M., & Lorca, P. (2005). Forecasting business profitability by using classification techniques: A comparative analysis based on a Spanish case. *European Journal of Operational Research*, *167*(2), 518–542. https://doi.org/10.1016/j.ejor.2004.02.018

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*, 3. https://doi.org/10.1186/1471-2105-7-3

Eilander, G. (1997). Nieuwe producten falen vooral wegens niet waargemaakte belofte. *VMT*, *23*(35–7).

eMarketer. (2016). Worldwide Retail Ecommerce Sales Will Reach $1.915 Trillion This Year. Retrieved from https://www.emarketer.com/Article/Worldwide-Retail-Ecommerce-Sales-Will-Reach-1915-Trillion-This-Year/1014369

Fletcher, R. (1987). *Practical methods of optimizations*. New York: John Wiley & Sons.

Gorunescu, F., Gorunescu, M., El-Darzi, E., & Gorunescu, S. (2008). A statistical evaluation of neural computing approaches to predict recurrent events in breast cancer. *2008 4th International IEEE Conference Intelligent Systems, IS 2008*, *3*(July), 1138–1143. https://doi.org/10.1109/IS.2008.4670506

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Honhon, D., Jonnalagedda, S., & Pan, X. a. (2012). Optimal Algorithms for Assortment Selection Under Ranking-Based Consumer Choice Models. *Manufacturing & Service Operations Management*, *14*(2), 279–289. https://doi.org/10.1287/msom.1110.0365

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley.

Hyndman, R. J., Akram, M., & Archibald, B. C. (2008). The admissible parameter space for exponential smoothing models. *Annals of the Institute of Statistical Mathematics*, *60*(2), 407–426. https://doi.org/10.1007/s10463-006-0109-x

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439–454. https://doi.org/10.1016/S0169-2070(01)00110-8

Hyndman, R., & Kostenko, A. V. (2007). Minimum Sample Size Requirements For Seasonal Forecasting Models. *Foresight: The International Journal of Applied Forecasting*, (6), 12–15. https://doi.org/10.1.1.218.6474

Kahn, K. B. (2006). *New Product Forecasting: An Applied Approach*. (M. E. Sharpe, Ed.). New York.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal Of Statistical Software*, *36*(11), 1–13. https://doi.org/Vol. 36, Issue 11, Sep 2010

Liu, H., & Yu, L. (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. *TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *17*(4), 491–502. https://doi.org/10.1109/TKDE.2005.66

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research*, *96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Platt, J. C. (1997). Selecting and interpreting measure of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89.

Platt, J. C. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Advances in Kernel Methods*, 185–208. https://doi.org/10.1.1.43.4376

Pohar, M., Blas, M., & Turk, S. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. *Metodološki Zvezki*, *1*(1), 143–161. https://doi.org/10.1198/004017005000000661

Rao, V. R., & Mclaughlin, E. W. (1989). Modeling the Decision to Add New Products by Channel Intermediaries. *Journal of Marketing*, *53*(January), 80–88. https://doi.org/10.2307/1251526

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*(1), 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002

Sexton, R. S., & Dorsey, R. E. (2000). Reliable classification using neural networks. *Decision Support Systems*, *30*(1), 11–22. https://doi.org/10.1016/s0167-9236(00)00086-5 T4 - A genetic algorithm and backpropagation comparison M4 - Citavi

Silver, E. A., Pyke, D. F., & Peterson, R. (1998). *Inventory Management and Production Planning and Scheduling* (Third). Danvers.

Sola, J., & Sevilla, J. (1997). Importance of Input Data Normalization for the Application of Neural Networks to Complex Industrial Problems. *Nuclear Science, IEEE Transactions on*, *44*(3), 1464–1468.

Thomas, R. J. (1998). *New product development: Managing and forecasting for strategic succes*. Wiley.

Thomassey, S., & Happiette, M. (2007). A neural clustering and classification system for sales forecasting of new apparel items. *Applied Soft Computing Journal*, *7*(4), 1177–1187. https://doi.org/10.1016/j.asoc.2006.01.005

Turban, E., Sharda, R., & Delen, D. (2011). *Decision Support and Business Intelligence Systems*. Pearson.

Turban, E., Sharda, R., Delen, D., & King, D. (2010). *Business Intelligence*. Pearson Education (US).

Twinkle100. (2006). Twinkle100. Retrieved from http://twinkle100.nl/

Vali, A. A., Ramesht, M. H., & Mokarram, M. (2013). The Comparison of RBF and MLP Neural Networks Performance for the Estimation of Land Suitability, *2*(3), 74–78.

Wang, F., Head, M., & Archer, N. (2002). E-tailing: An analysis of web impacts on the retail market. *Journal of Business Strategies*, *19*(1), 73–93.

Wilks, S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62.

Yu-Wei, C. (2015). *Machine Learning with R Cookbook*. Birmingham: Packt publishing. https://doi.org/10.1017/CBO9781107415324.004

Zhou, L., Dai, L., & Zhang, D. (2007). Online Shopping Acceptance Model - a Critical Survey of Consumer Factors in Online Shopping. *Journal of Electronic Commerce Reserach*, *8*(1), 41–62. https://doi.org/10.1086/209376,

Zhou, Z.-H. H., Wu, J., & Tang, W. (2002). Ensembling Neural Networks: Many Could Be Better Than All. *Artificial Intelligence*, *137*(1–2), 239–263. https://doi.org/10.1016/S0004-3702(02)00190-X

Zhou, Z. H., & Jiang, Y. (2004). NeC4.5: Neural ensemble based C4.5. *IEEE Transactions on Knowledge and Data Engineering*, *16*(6), 770–773. https://doi.org/10.1109/TKDE.2004.11

Zopounidis, C., & Doumpos, M. (2002). Multicriteria classification and sorting methods: A literature review. *European Journal Of Operational Research*, *138*, 229–246.