A holistic order allocation strategy for profitability maximization: a simulation study

João Marinho Alves

Master's Dissertation

Supervisor: Prof. José Luís Moura Borges

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Integrated Master in Industrial Engineering and Management

2018-06-15

Abstract

Luxury e-commerce is a growing market with a very demanding customer base. With new players entering the market every year, competition is fierce for both incumbents and new entrants. This demands that e-tailers provide a flawless shopping experience. In such a competitive industry, serving customers through a marketplace further increases complexity since the fulfillment process is carried out by partners. Hence, when more than one partner can fulfill an order, the marketplace owner should be capable of selecting the best partner for the job.

In this thesis, with the aid of discrete event simulation, different strategies will be developed in order to achieve that goal. The featured approach is based on leveraging the data available from past orders and then testing different allocation strategies. To understand the impact each of these different strategies possess, research was conducted on what the most important and actionable fulfillment variables were. To account for the effects of some of them, a retention model was developed through a proportional hazard rates model. This was incorporated in a client lifetime value model.

This dissertation details several strategies, such as cost reduction, lead time reduction, among others, as well as their effects on both the company and their customers. This is then compared to the current methodology. The goal was to create a holistic algorithm that would be able to maximize the well-being of all the marketplace stakeholders. Such goal was achieved through the development of a profitability measure, capable of providing the best decision for the marketplace as a whole.

The simulating tool developed also brought more visibility and understanding to the allocation decision making, finding shortcomings in the current methodology, as well as paving the way for further development of profitability measure.

Resumo

O comércio eletrónico de luxo é um mercado em crescimento com uma base de clientes muito exigente. Com novas empresas a entrar no mercado todos os anos, a concorrência aumenta quer para os incumbentes, quer para as novas empresas. Esta concorrência implica que os comerciantes eletrónicos proporcionem a experiência de compra mais satisfatória possível.

Num setor tão competitivo, manter um marketplace acarreta grande complexidade, já que o processo de preparação e entrega da encomenda é executado por parceiros e não pela própria empresa. Deste modo, existindo mais de um parceiro capaz de satisfazer uma entrega, cabe à empresa selecionar o mais capacitado para esse fim.

Nesta tese, com o auxílio da simulação de eventos discretos, foram testadas diferentes estratégias com o intuito de selecionar os melhores parceiros. A abordagem utilizada baseia-se na alavancagem dos dados disponíveis de pedidos anteriores para testar diferentes cenários. Para entender o impacto das diferentes estratégias, foi realizada uma pesquisa com o intuito de descobrir quais seriam as variáveis mais importantes e passíveis de modificar no processo de preparação e entrega de uma encomenda. Para contabilizar os efeitos de algumas destas variáveis, modelos de retenção baseados em *proportional hazard rates* foram utilizados e os seus valores incorporados em modelos de *customer lifetime value*.

Esta dissertação detalha também estratégias como a redução de custos e do tempo de entrega, entre outras, e o seu impacto tanto na empresa como nos clientes. O objetivo consistiu em criar um algoritmo holístico que fosse capaz de maximizar o bem-estar de todas as partes envolvidas do mercado. Este objetivo foi alcançado através do desenvolvimento de uma medida de rentabilidade, capaz de proporcionar a melhor decisão para o mercado como um todo.

A ferramenta de simulação desenvolvida permitiu também fornecer uma maior visibilidade e compreensão para a tomada de decisões de alocação, encontrando falhas na metodologia atual e abrindo caminho para o desenvolvimento adicional de uma medida de rentabilidade.

Acknowledgements

I would like to thank Farfetch for giving me the opportunity to develop this project. To all the Farfetchers, especially the Operations Strategy team, I own a huge thank you for the way you integrated me in the company.

In particular, I owe my gratitude to my supervisor Carlos Gomes for all the support provided during this project as well as the guidance for my personal development, to João Ascensão for all the time invested in helping me give the first steps in Data Science, to Pedro Justo for the constant help and availability and to Augusto Gomes for the opportunities provided. It truly was an amazing experience.

I would also like to thank my dissertation supervisor, Prof.José Luis Borges, whose help was invaluable for the completion of this document. Thank you for the guidance, for the great feedback, the constant availability and for showing me the amazing field that Data Science is. To Rui Ribeiro a huge thank you for all the help in revising the project. I would like to thank all the professors at MIEGI for all the knowledge provided through the years. To professor Nuno Soares a special thank you is in order for all the patience and teaching provided in our discussions, I truly learned a lot from you.

During my enrolment in this course there was no shortage of great moments. To all the Friends done during this path, thank you so much for all those moments, patience and laughs. Without you, this path would be too dull.

To my family thank you for the constant support, to Joana Miguel for all the love and to the rest of my friends for always being there.

Contents

Acronyms and Symbols

KPI	Key Performance Indicator
VAT	Value added Tax
DES	Discrete Event Simulation
SQL	Structured Query Language
PHM	Proportional Hazard Model
WTP	Willingness to pay
CLV	Customer Lifetime Value
AFT	Accelerated Failure Time
pdf	probability density function
SoS	Speed of Sending
TiT	Time in Transit
LT	Lead Time
ROS	Rate of offline sales
NPS	Net Promoter Score
ROI	Return On Investment
FOB	Free on Board
CIF	Cost, Insurance and Freight

List of Figures

List of Tables

Chapter 1

Introduction

1.1 E-commerce luxury market

Despite the previous belief that e-commerce was confined to low and mid-priced products, (Linda Dauriz, Nathalie Remy and Sandri (2014)) in more recent times, luxury online retailers such as Farfetch, Net-a-Porter, Yoox, and Saks, have been desmistifying this concept. This is due to the successful creation and growth of an online market for luxury products.

Multi-brand e-tailers enable luxury brands to reach both consumers which are more time constrained, and therefore can not afford to shop around in single-brand sites, as well as consumers who are located in more rural areas with no easy access to urban boutiques. (Jennifer Schmidt, Karel Dörner, Achim Berg, Thomas Schumacher and Bockholdt (2015)).

This trend is confirmed by an increasing percentage of luxury online sales, set to reach 12% of all the sales by 2020 and 25% by 2025(Antonio Achille, Sophie Marchessou and Remy (2018)). As a consequence of this, the market will evolve towards an ever-growing online presence of brands and stores, which will stress competition between e-tailers.

Regarding growth drivers, two stand out: emerging nations and the increasing presence of millennials. First off, countries such as Russia, UAE and China have been contributing towards market growth with an average increase in luxury shopping of 73% in the last years. (Kate McCarthy, Ben Perkins, Nick Pope, Linda Portaluppi, Valeria Scaramuzzi (2017)) Furthermore, in a market where generation Y and Z were responsible for 85% of the growth in 2017, brands are starting to rethink their customer experience from product conception to sale (making it more appealing for the younger generations). (Claudia D'Arpizio, Federica Levato and de Montgolfier (2017)).

The adoption of online retailers by consumers will lead to higher complexity regarding delivery strategies in these companies. In such an environment, the efficiency of the delivery process will be of paramount importance. That is achieved through allocating each order to the right warehouse, making it reach the customer on the right timing, in order to maximize the profit per order.

This thesis arises from the needs of Farfetch, one of the biggest luxury e-tail players, for a more holistic strategy around their order allocation algorithm. The project is mainly focused on exploring different allocation strategies, such as maximizing profit per order, reducing lead time, among others, and come up with one that will enable the company to continue in the path to market leadership. This is done while increasing not only short-term profit, but also customer lifetime value, with each individual order allocation decision.

1.2 Company Description

Farfetch is a global leader in the online fashion industry, with the 900 boutiques and brands presented in its platform scattered around 190 Countries. The substantial number of partners allows Farfetch to position itself as the leading player in terms of product offering. This market positioning can be seen in figure 1.1 with the company occupying an unique positioning in terms of both product depth and breadth.

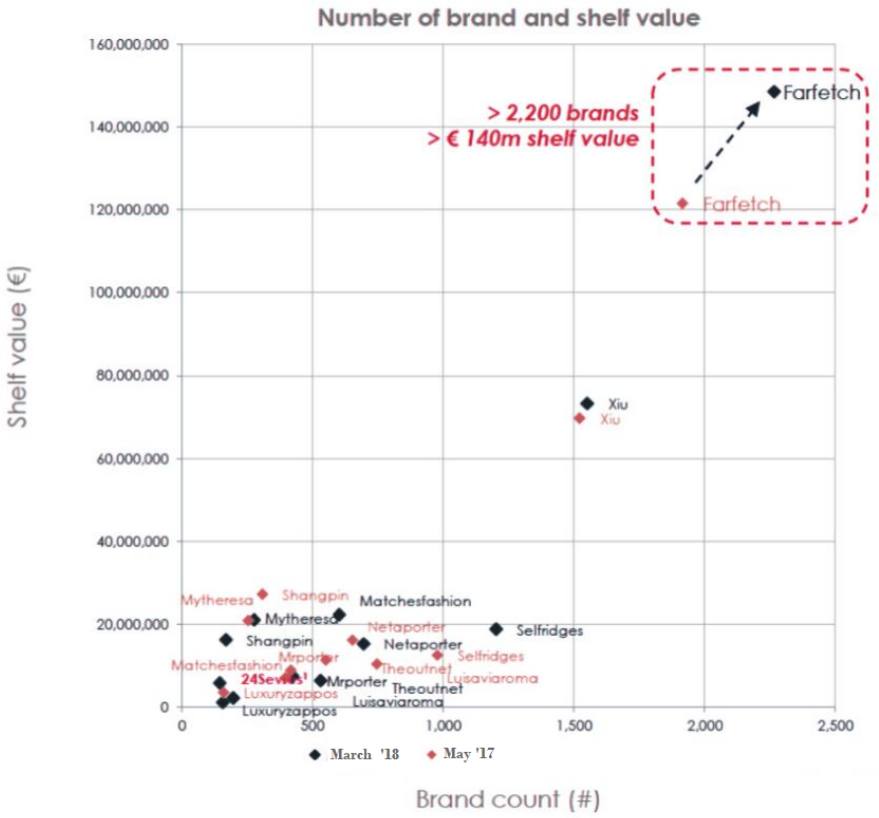

Figure 1.1: Farfetch Brand positioning strategy

Furthermore, this product availability creates a customer experience that very few companies can provide. In a single platform, customers can purchase from more than 2000 different brands from all over the world and have their products delivered to their houses, pick them up in the store or even have them brought directly to their boat.

The influx of so many customers from all over the world offer Farfetch economies of scale that few other players possess. This allows the company to further extend its offers and possess a larger leverage while negotiating with couriers and partners. This creates a virtuous cycle, again, improving the product offering, gaining more customers and starting the cycle over again.

The company positioning in terms of product offering is only possible because of their business model. Unlike its main competitors, which buy and stock their products in warehouses, this enterprise does not own any stock. It functions only as a marketplace that connects buyers and boutiques/brands.

However, the company is not your typical marketplace, such as Alibaba or Amazon. At Farfetch, the same product is never displayed twice, which sets it apart from other companies. Instead, when the same product is sold by more than one seller, the platform agglomerates them into a single offer. This translates into more of a seamless shopping experience for the user. Even though this is never hidden - the customers can see which store the product will be shipped from - it saves the hassle of looking at the same product more than once. Figure 1.2 provides an example of how this marketplace works. When a product - such as product A - is offered by two sellers - partner 1 and 2 - the company agglomerates them into a single offering. The price is not displayed since it depends on which seller is selected. This choice may vary from customer to customer. Nevertheless, should a product be offered by only one seller, like product B, it will appear directly in the website.

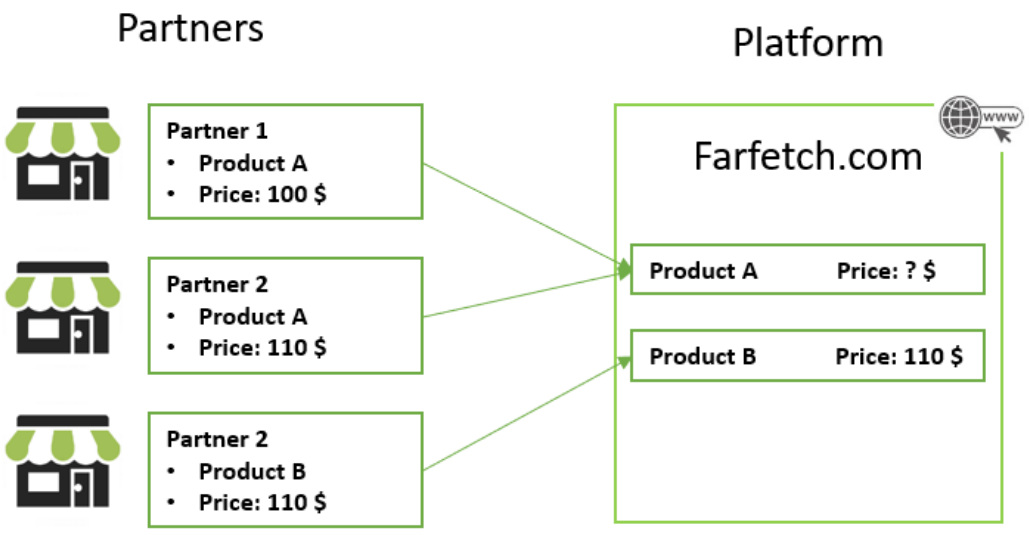

Figure 1.2: Farfetch business model - Several sellers can sell the same product but it will be agglomerated into a single offering, where the price to be displayed will be decided through an internal algorithm, based on the client location.

1.3 Motivation

The problem this thesis tackles originates from the need to select a single selling partner when more than one is providing the product. These products, the ones sold by different partners, are called duplicates.

As seen before in figure 1.2, for the duplicate products, the company needs to decide which seller will be selected for a given customer in order to display the product in the platform.

In most cases, deciding the seller also defines the price of the product showed. Those sellers might be brands, warehouses, or privately owned boutiques, which will be referenced as stock points. The fact that only a single product is shown means that there will be an internal competition between the sellers as to capture an order.

This competition occurs in two different moments, depicted in figure 1.3: firstly, when the client is navigating the products page and later when the client is finalizing the purchase. During the website navigation, a single price needs to be shown to the customers -even though the stock points may sell the same products at various prices - and this is decided by an algorithm called the Duplicates Algorithm. This algorithm will select a partner to provide that product for a given customer, but only temporarily.

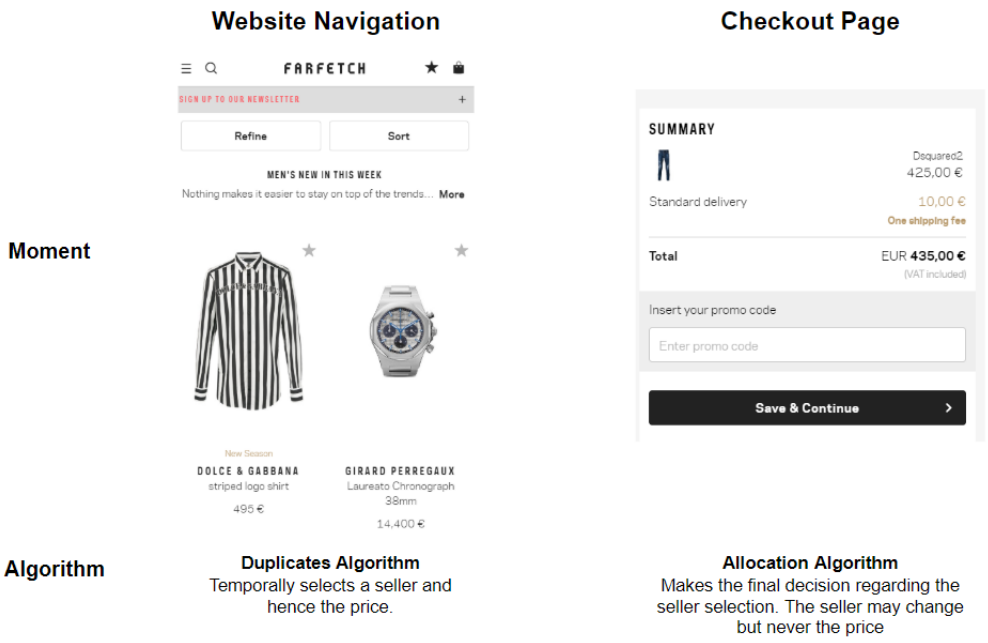

Figure 1.3: Farfetch decision algorithms

The term temporarily is employed here since this decision is not a permanent one. The partner selected by the duplicates algorithm is chosen solely for the purpose of providing the costumer with a price while he is browsing. This decision may be changed as the client reaches the checkout page. Whenever the basket of products is settled, and hence the client is already at checkout, a second algorithm is ran to decide which seller will provide which products - there may be multiple

products in the basket. As such, a product that is allocated to a seller through the duplicates algorithm may be allocated to another seller, depending on the decision of the second algorithm - the allocation one. However, the price charged to the customer never changes.

The need for two different algorithms is due to the dimension of the search space. When a client is browsing through the products, around 30% of them are duplicates. With around 100 000 duplicate products in the platform, selecting a single price for each and every product as the customer is navigating is computationally demanding. As such, a quicker algorithm is used in this phase, the duplicates Algorithm. Nevertheless, before the purchase is finished, this space is confined to products only in the client's basket, allowing for several calculations and new opportunities, such as consolidation - the possibility for a single partner to deliver all the products in the basket- and this is where the allocation algorithm comes in. A more detailed description of these algorithms is presented in the next section.

The scope of this thesis will be to improve the last aforementioned algorithm, that is, the allocation one. Different strategies will be tested, such as selecting stock points to reduced the lead time, increasing the profit per order, increasing stock levels, among others. In the end, a more holistic approach than the one currently used will be proposed. This approach will be based in providing more visibility towards the direct costs while also promoting higher priority to customer satisfaction which should, hopefully, positively affect the company's bottom line.

1.4 Methodology

A new allocation framework will be proposed. This framework is sustained by an extensive literature review regarding order fulfillment strategies, customer satisfaction, customer retention and customer lifetime value. The success of this project will be evaluated on several fronts. Firstly, an increase in the profit per order is expected. Secondly, customer satisfaction and its impact on retention will be measured and increased. Lastly, partners' health will be monitored and properly managed in order to increase their own performance and contribution towards the business.

For the purpose of testing new allocation strategies and their impact, a case study was elaborated around key products, available in more than one stock point, with the aid of discrete event simulation (DES). This tool, DES, is widely used in supply chain problems, especially at an operational and tactical level (Tako and Robinson (2012)).

Since the simulation would be done using real sales data, a data mining methodology was followed - Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al. (2000)), where the main steps can be seen in figure 1.4. This methodology functions as a guide to data mining endeavors. It formulates a series of steps that should be followed in a typical data mining project. In this project, the main phases on that methodology were followed since the needs were the same, except for the modeling techniques used (which were not delimited to data mining techniques).

Following the CRISP-DM methodology, a deep business understanding was obtained through a series of inductions and informal meetings with key stakeholders in the company.

Figure 1.4: Cross Industry Standard Process for Data Mining (CRISP-DM). Source: Chapman et al. [2000]

Next, data was obtained from several different sources scattered around the company databases. There were two reasons to resort to different databases: first of all, no single database had all the information needed. Secondly, to speed up the process, some queries could not be run on the standard company system - SQL server management studio - and had to be done using a cloud solution -Google Big Query[1]- with the running time of the queries on Google Big query reduced from several hours to minutes, as compared to the SQL server.

In short, the data needed to be gathered was from sales, stock levels, shipping costs, lead time, duties, marketing and stock points, to name the most relevant ones.

After selecting the different data sources an exploration process was started: the complexity showed by the data-sets forced the application of several pre-processing steps to homogenize the data and make it ready for the simulation. The bulk of the time spent around this project is related to these steps, since the intricacies of data collection and preparation required it, as is usual in data-science endeavors. This is known as the 80/20 data science dilemma, where 80% of the time is spent finding, cleaning, and reorganizing huge amounts of data. Finally, after the data is fully prepared, parts of it were directly fed into the simulation while others were used for modeling

[1] Google Big Query is a database cloud solution that functions like a SQL server as-a-service where the user pays per megabytes processed in each query. However, since it runs on Google servers, it eliminates the time-consuming work of provisioning infrastructure and reduces downtime with a serverless infrastructure. This handles all ongoing maintenance, including patches, and upgrades speed constraints. Lastly, it uses automatic scaling and high-performance streaming ingestion to load data, which results in much higher performance in terms of querying times

purposes to support the simulation (this will be explained later).

Regarding the simulation development methodology, after the data was ready, the simulation was built in a modular fashion, as to follow the principles of lean software development (Poppendieck and Poppendieck (2003)). That way, additional modules could be added to it in an iterative way. This leads to the development of new allocation strategies, as well as their implementation, and allows further complexity and variables to be added into the simulation. The process is then repeated.

1.5 Goals

The objective of this project was first and foremost to create a new allocation algorithm that is capable of maximizing profitability through the smarter selection of the right seller. In reaching this goal, the understanding on which of the allocation variables should be considered in the allocation decision, and how their impact can be modeled in terms of profitability, were of uppermost importance.

1.6 Thesis Outline

This thesis is organized into six chapters. Its outline is as follows:

Chapter 1 introduces the company and the subject of the thesis.

Chapter 2 aims to provide a theoretical background for the relevant subjects of this work. Firstly, research will be conducted in order fulfillment, major strategies and its impact on the final customer. Secondly, customer lifetime value literature will be reviewed along with customer retention literature. Lastly, a technique to model distributions, among other things, called B-splines will be the object of research and its inner workings explained.

Chapter 3 furthers the understanding of the current solution used by the company as well as explores what constitutes the problem variables to be addressed.

Chapter 4 describes the variables to be used to solve the problem. It further explains how the solution will work and how the work was divided.

Chapter 5 begins to show the quality of the simulator and the results obtained by the different solutions.

Chapter 6 is a summary and a reflection on the findings of this thesis.

Chapter 2

Literature Review

This section will be dedicated to an overview of the most relevant literature related to the subjects used in the implementation of the solution in this project. The topics considered most pertinent which are under investigation are: order fulfillment and its impact on customer satisfaction, customer lifetime value and modeling distributions.

The purpose of this part is not to serve as a compendium of the most relevant literature but instead to provide guidance to solve the problems faced during the project.

2.1 Order Fulfillment

As defined by Pyke, D.F., Johnson, M.E., Desmond (2001) "order fulfillment involves all of the activities from the point of a customer's purchase decision until the product is delivered to the customer and he or she is fully satisfied with its quality and functionality.". The scope of the research done on the topic order fulfillment was made under this definition.

More specifically, the scope of the research will be focused on the impact of delivery experience in the final customer.

Several studies have been conducted on the proverbial "last mile" of the retail supply chain – i.e., delivering products to the end-customer – and its impact on customer satisfaction is well known and documented (Shenton; Newton (2000); Pastore (1999)). In a study conducted by Hau L. Lee and Seungjin (2001) one of the main conclusions was that an e-tailer capacity to do order fulfillment and delivery on time could be determinant in its success.

Furthermore, in a empirical study conducted by Heim and Sinha (2001) a regression model was elaborated using both order fulfillment variables and procurement variables such as price, web site navigability,among others. In this study, it was concluded that fulfillment variables contributed the most to customer loyalty. More specifically variables such as product availability, easiness of returns and, again, timeliness of delivery where the biggest drivers of loyalty. Although the study was conducted in a B2C operations retailer, the principles are the same and last-mile fulfillment strategies should operate under this assumptions.

In essence, order fulfillment strategies can be optimized to positively contribute to customer satisfaction. Customer satisfaction has been proved not only to affect choice and purchase behaviour of the consumer (Homburg et al. (2005)) but it has also been shown that it highly correlates with higher levels of repurchase intent as well as customer retention (Lam et al. (2004); Mittal and Kamakura (2001)).

Then, we should attempt to further the reader's understanding of what customer satisfaction is and how it can, through order fulfillment strategies, maximize a firm profitability.

Oliver (1997) defined customer satisfaction as "the summary of psychological state resulting when the emotion surrounding disconfirmed expectation is coupled with a consumer's prior feelings about the consumer experience". In a similar manner, Kotler et al. (2009) also states "a person's feelings of pleasure or disappointment resulting from comparing a product's perceived performance (or outcome) in relation to his or her expectations". In essence, customer satisfaction originates from the comparison of the actual performance a company had in a interaction with a customer with a mental benchmark the customer had resulting from what it expects from that interaction. Based on this definition, if we were to give the customer exactly what he would expect, it would leave the customer in a neutral state, and customer satisfaction would only be positive if we where to exceed those same expectations.

A question then arises, what variables influence customer satisfaction the most. Liu et al. (2000) states than in a context of e-commerce, client satisfaction is mostly influenced by information quality, web site design, merchandise attributes, security and privacy, delivery, payment and customer service. However, for some geographies like China, delivery and customer services share a primordial role, having a higher impact than other variables in customer satisfaction. Furthermore, more authors, such as Collier and Bienstock (2006) ,Sheng and Liu (2010) ,Parasuraman et al. (2005) further highlight the importance of order timeliness, order accuracy and order condition, among others, in client satisfaction.

Other authors, such as Christopher (1993) further concentrate their efforts in proving that it's possible to enhance profitability through logistics alone by optimizing customer satisfaction, providing the following linkages:

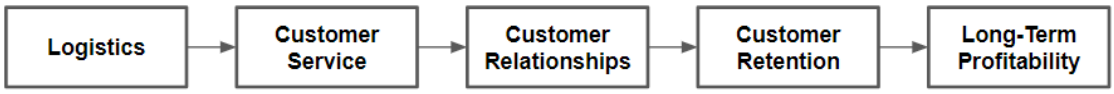

Figure 2.1: Logistics and Profitability - the linkages - Source: Christopher (1993)

Concluding that order fulfillment contributes to customer satisfaction, the question would be how can it be used to increase profitability. Lewin (2009) conducted some research on the topic and concluded that customer satisfaction could be used as a source of competitive edge for any company since it would lead to customer loyalty and repeated purchase. In a study conducted by Homburg et al. (2005) it was concluded that very strong evidence existed to forge a link between customer satisfaction and it's willingness to pay (WTP). Nevertheless, the relationship founded

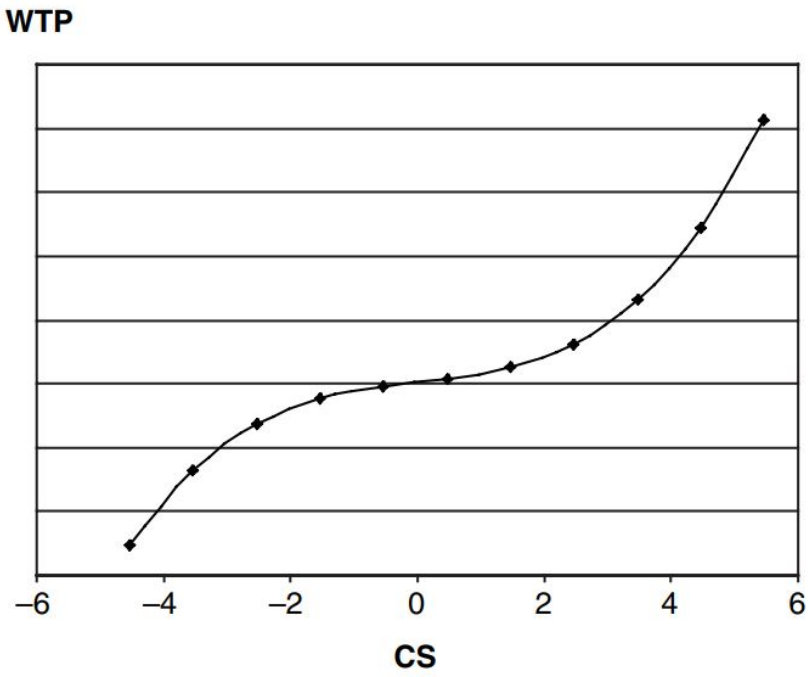

Figure 2.2: Empirical Relationship Between CS and WTP. Source: Homburg et al. (2005).

was not a linear one but a S profile instead, showed in figure 2.2. The implications are that the extremes of satisfaction/dissatisfaction contributed much more to customer willingness to pay.

Moreover, Fornell et al. (2010) proved that not only the customer WTP increased but that satisfied consumer really did spend more by being linking customer satisfaction to future spending growth.

In conclusion, order fulfillment strategies can be used to enhance customer satisfaction and the results are threefold: Increase in purchase rate/ customer retention, increasing customer spending and willingness to pay.

2.2 Customer Lifetime Value

Customer Lifetime Value (CLV) can be defined, in broad terms, as the present value of all the future profits generated from a customer over his relationship with a company. This definition has some variations according to some authors like Pfeifer et al. (2005) discussing that instead of profits other values like cash-flows generated could be used. However, using profit fitted better the needs of the project so the first definition will be used.

Although some variations exist (Berger and Nasr (1998)), a well established formula to calculate CLV, given by Gupta and Lehmann (2004) is:

$$CLV = \sum_{i=1}^{T} \frac{(p_t - c_t)\, r_t}{(1+i)^t} - AC \tag{2.1}$$

where

p_t = price paid by a consumer at time t,
c_t = direct cost of servicing the customer at time t,
i = discount rate or cost of capital for the firm,
r_t = probability of customer repeat buying or being “alive” at time t,
AC = acquisition cost, and
T = time horizon for estimating CLV.

Looking at equation 2.1, in order to calculate the CLV, it’s needed to know the margin (p_t- c_t), the acquisition cost, the retention rate for the time interval that will be used to calculate the present value of the future profits and the capital cost of the company.

This means that different models are needed to calculate the retention rate, the margin and the customer acquisition cost. The customer acquisition cost is normally used to evaluate marketing initiatives which is not the purpose of this project, meaning no review will be done on the topic.

In terms of retention rate some literature will be reviewed. Retention rate can be defined as "the chance that the account will remain with the vendor for the next purchase, provided that the customer has bought from that vendor on a previous purchase" (Jackson (1988)). Based on this definition, researchers have been constructing and relying in statistical models to infer the probability of retention.

Retention is so important, not only for the sake of having returning customers but because the profits generated by customers tend to increase as their tenure increase in the company. F. Reichheld and Jr. W. Earl (1990) attribute this phenomena to four main reasons:

1. Revenues from customers typically grow over time as they get more comfortable and trust the company more;
2. Customers become increasingly easier to serve since they already know the product mix well and don’t need so much help;
3. Satisfied customers work as referral who recommend the business’s to others
4. Lastly, in old industries, old customers pay effectively higher prices than new ones.

To infer the probability of retention there are two broad types of retention models, "lost for good" and "always a share" as defined by Jackson (1988). In the "lost for good" model, the base assumption is that either the customer is totally committed to the company or totally lost and committed to another company. This models usually rely on hazard models to predict customer defection. The second model, "always a share" the customer can also be committed to other companies and typically uses Markov or migration models (Gupta et al. (2006)).

Although some researchers argue that "lost for good" models understates CLV since they don’t allow defect customers to return, this caveat has been approached by other researchers as not a serious problem because they can always be treated as a renewable source (Drèze and Bonfrer (2009)) and defecting customers can be reacquired (Thomas et al. (2004)).

Furthermore, "always a share" retention models didn't provide the information needed for the project. These models estimate transition probabilities of a customer being on a certain state based on the recency of the last purchased. Although there are good arguments for the the use of purchase history like DuWors et al. (1990) that used customer purchase history to predict brand loyalty, those models are not causal models and wouldn't provide the causal relations needed for the purpose of the project.

For the "lost for good", typical hazard models or computer science models are used. The hazard models fall in two broad groups: Accelerated Failure Time (AFT) or Proportional Hazard (PH) models (Gupta et al. (2006)).

AFT models are formulated as follows Kalbfleisch and Prentice (1980):

$$ln(t_j) = \beta_j X_j + \sigma\mu_j \tag{2.2}$$

where t is the interpurchase time for customer j and X are the covariates. When $\sigma = 1$ and μ takes an extreme value distribution then the result is an exponential duration model with constant hazard rate. And so, different combinations of σ and μ lead to very different models.

PH models function in a different way. These models specify the hazard rate ($h_i(t;X_t)$) as a function of the baseline hazard rate ($h_i(t)$) and covariates (X),

$$h_i(t;X_t) = h_i(t) \times exp(\beta_i X_i) \tag{2.3}$$

The most interesting aspect of the use of PHM models is the capacity they have to capture the intrinsic temporal purchase pattern in the baseline hazard and using the covariates to control for the effect of other variables, such as marketing variables. In other words, the baseline hazard can be used to characterize the distribution of the interpurchase times after accounting for the effects of other variables (Seetharaman and Chintagunta (2003)).

Other approaches are also possible to estimate the time until the next purchase and, as a consequence, the retention rate. Namely, in recent years, a new trend has appear where instead of using parametric models - which are more favored by marketing literature for their easier understanding and being based on theory (e.g, utility theory) - computer science models are used. These models can be neural network models, classification and regression trees (CART), support vector machines(SVM),and others.

For the context of this work, no such models will be used, not because they lack predictive ability, but given the complexity of those, models that are easier to understand and interpret will be favored.

Namely, PHM for its ability to produce a baseline and describe, clearly, the impact of the co-factors in the baseline was the one selected.

Given the preference for PHM, a deeper dive will be done in the works of the model for a clear understanding of the results.

In machine maintenance literature, PHM was first proposed by Cox (1972) to study failure censored times in machines. In fact, hazard rate can be defined, in machine maintenance, as the

probability of a machine component failing given that it was working until the given moment. What would be beneficial for those cases would be that the hazard rate would be decreasing, that is, the component would be less likely to fail as the time passed. No such situation exists in reality.

Making an analogy with customer, the fails of the components are the purchases, and what is hoped is that the hazard rate increases, that is, the probability of failing - making a purchase - increases over time.

To use a PHM model, one of the first challenges is to decide which distribution will characterize the baseline hazard $h_i(t)$, and those can be parametric distributions such as weibull, exponential, Erlang, log-logistic, expo-power,among others. Since when using the model it may not be evident which distribution better fit the data-set, it may be needed to test different distributions or use flexible distributions like weibull, which is the most common use in the literature (Seetharaman and Chintagunta (2003)).

To estimate the parameters, the equation in 2.3, needs to be manipulated to the point where it's written as - see Appendix A for the complete demonstration:

$$f(t,X_t) = h(t) \times \exp(X_t \times \beta) \times \exp(-\int_0^t h(u)(X_u \times \beta)du) \tag{2.4}$$

Where $f(t,X_t)$ stands for the probability density function (pdf) corresponding to the customer i's hazard function at time t. Then, in order to estimate the parameters, suppose one had n interpurchase time observations of a given product, for a given customer, like $(t1,t2,t3,...,tn)$. For each interpuchase time, assume there is also a vector of covariates $(X1,X2,X3,...,Xn)$ where the covariates can be the existence of marketing activities at the moment t, the satisfaction of the client at time t, among others. The following likelihood function can be maximized to estimate the parameters of the PHM at the individual-level:

$$L = \prod_{j=1}^{n} f(t_j - t_{j-1}, X_j) \tag{2.5}$$

Having estimated the parameters we can easily estimate the survival function - see Appendix A - and, using the value β_i we can estimate the impact of each covariate. For example, if one covariate has a value of $exp(\beta) = 1.05$, the conclusion is that by increase that cofactor in one unit, we expect the hazard rate to increase, if all other factors remain equal, in 5% for a given time.

2.3 Approximating Distributions

A lot of work has been done in the field of Statistics regarding approximating sample data to know distributions as well as fitting unknown distributions. It can be useful, when using simulation techniques, to generate values based on the real distributions to ensure the variability presented in the real world is presented in the simulator.

During the development of this work, the need to model time distributions that did not follow a known distribution appeared. This happened because some data-sets are too rich to be modeled

using parametric models. When this happens, and the data cannot be fitted, other techniques are applied.

Those techniques can be either to interpolate, with several techniques available, the distribution function or use the empirical distribution, in order to generate values.

In this project, since the amount of data available would allow for it, it would be possible not to model the data but use the empirical distribution instead. The reason this solution was not used is that the data would have to be either available in memory which would waste resources, given the sample size - millions of observations - or would have to be loaded each time it needed to be accessed, which would make the tool using the technique too slow.

As such, interpolation was chosen to reduce memory usages. In the present study, B-splines where used to model this data instead of other models, such as Polynomial curve fitting, for their high overfitting capacity with little computational power required to compute the B-splines (R. and de Boor (1980); Diercx (1993)). Further advantages of B-Splines is that they are a straight forward extension of (generalized) linear regression models, conserve moments (means, variances) of the data and have polynomial curve fits as limits (Eilers and Marx (1996)).

To explain how B-splines work, inspiration was drawn from the work of Eilers and Marx (1996). First of all, B-splines are constructed from polynomial pieces, of any degree, joined at certain values of x, the knots.

A very simple explanation of how splines work can be done using figure 3.4. On the left side there's a B-spline of degree 1. It consists of 3 knots $x1, x2$ and $x3$ and two linear pieces connecting those dots: one from $x1$ to $x2$ and another from $x2$ to $x3$. On the right side of the figure there is another B-spline of degree 2. This, uses 4 knots $x1, x2, x3$ and $x4$ and three quadratic pieces, joining the knots. An important feature of the spline is that at the joining points not only the ordinates match but also their first derivatives are equal.

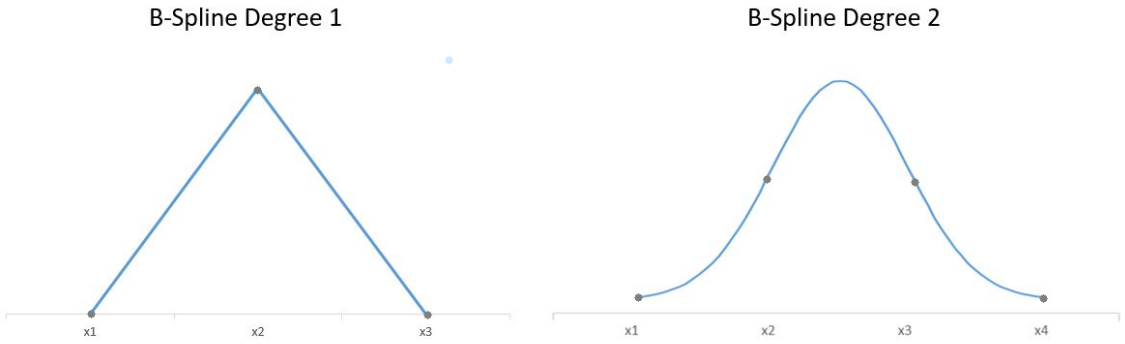

Figure 2.3: B-spline example. On the left, there is a B-spline of degree 1 with 3 knots and on the right, a B-spline of degree 2, with 4 knots.

Choosing the number of knots is no trivial task and much work as been done on the subject, such as Friedman and Silverman (1989) who tried to propose a optimization schemes for choosing the number and positions of the knots. The problem with choosing the knots is a trade off between underfitting and overfitting - choosing too many will lead too an overfitting and choosing too few

will lead to underfitting. An approach to prevent overfitting, and still highly used to date was first proposed by O'Sullivan (1988) using penalties on the second derivative as to restrict the flexibility of the fitted curve.

In general, a B-spline is a piecewise polynomial function of degree $n-1$ in a variable x. The B-spline gives values only for ranges between the first and the last knot, taking the value of 0 everywhere else in the domain.

Furthermore, the true usefulness of B-splines is achieved by the simple representation that any spline function of order n, on a given set of knots, can be represented as a simple linear combination of B-splines as follows:

$$S_{n,\mathbf{t}}(x) = \sum_i \alpha_i B_{i,n}(x). \tag{2.6}$$

In other words, B-splines serve as the basis function for the spline space, hence the name. This property results from the fact that all the pieces in the spline have the same continuity properties, within their boundaries, the knots.

Using the original R. and de Boor (1980) concept applied to splines, the curve fitting of the splines of any order to the data can be done using the a function composed of the Sum of B-splines and the error minimized using a least squares. Consider the regression of m data points (x_i, y_i) on a set of n B-splines $B_j(\cdot)$, then the minimization function would be:

$$S = \sum_{i=1}^{n} \left[y(x) - \sum_j^n \alpha_j B_j(x_i) \right]^2 \tag{2.7}$$

The coefficients α_i are the parameters to be determined. Although the knots may also be used as parameters when their number and positioning are to be calculated. For simplicity, only equidistant knots will be used, called uniform B-splines.

In a study done by O'Sullivan (1988) a penalty was added on the second derivative of the fitted curve to avoid overfitting, making the resulting spline partially constructed based on the fitting data and, partially constructed based on the penalties. The resulting objective function was :

$$S = \sum_{i=1}^{n} \left\{ y(x) - \sum_j^n \alpha_j B_j(x_i) \right\}^2 + \lambda \int_{x_{min}}^{x_{max}} \left\{ \sum_j^n \alpha_j B_j^{''}(x_i) \right\} dx \tag{2.8}$$

where λ works as a continuous control parameter for the smoothness of the fit, that is, it can be used to control the number of nodes and coefficients the splines need to fit the data. To calculate the first and the second derivative De Boor (1978) also provided a simple formula:

$$h \sum_j \alpha_j B_j^{'}(x) = \sum_j \alpha_j B_j(x) - \sum_j \alpha_{j+1} B_{j+1}(x) = -\sum_j \Delta\alpha_{j+1} B_j(x) \tag{2.9}$$

where h is the distance between the knots and $\Delta\alpha_{j+1} = \alpha_{j+1} - \alpha_j$. The second derivative can then be easily found by induction with:

$$h^2 \sum_j \alpha_j B_j''(x) = \sum_j \Delta^2 \alpha_j B_j(x) \tag{2.10}$$

Other authors propose different penalties (Eilers and Marx (1996)) but the a penalty applied to the second derivative became common place to avoid overfitting. The results of one of the applications Eilers and Marx (1996) did can be seen in figure 2.4, using B-splines with penalties to approximate helmet crash data.

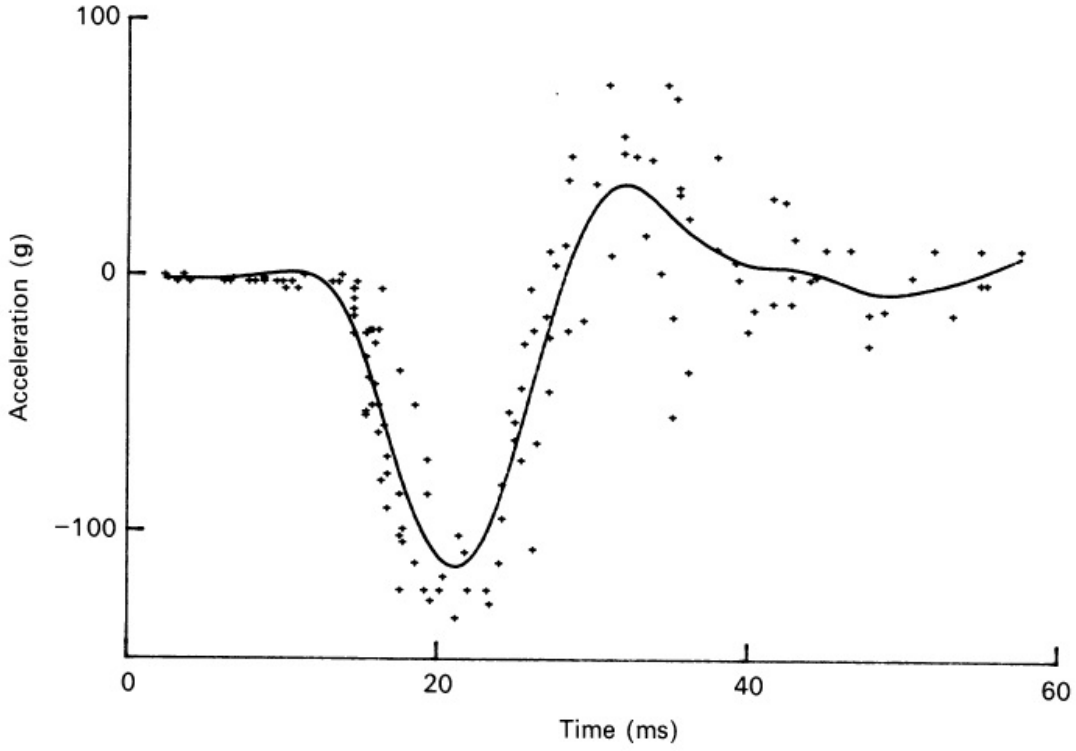

Figure 2.4: Motorcycle crash helmet impact data: optimal fit with B-splines of third degree, a second-order penalty and λ =0.5 Source : Eilers and Marx (1996).

In the same way, it could be used to fit any type of data of any shape, including cumulative distribution functions as it will be used in this study.

Chapter 3

Problem Statement

This thesis hopes to address the peculiarity of the marketplace Farfetch owns. The company's intent on displaying a sole copy of the product despite it being linked to various sellers requires a choosing algorithm. As highlighted in chapter 1, there are actually two algorithms in play, one when the customer is still navigating the website (duplicates algorithm) and another at checkout (allocation algorithm).

Even though this thesis focuses on developing a better alternative for the later, that is, the allocation algorithm, it still features an overview of the first. The reason being that, in it's current iteration, this algorithm will influence the decision taken by the allocation one. This influence can, nevertheless, be overridden by the allocation algorithm if needed be. Even though the price paid by the costumers remains stale, the decisions taken by the allocation algorithm can be independent from the duplicates one, should the circumstances demand it.

This section features an in-depth description of the problem. It begins with an explanation of the inner working the algorithms that were mentioned before. What follows is a concise explanation of the problem and the main stakeholders affected by it. Lastly, a review of the shortcomings of the current solution alongside a description of the solution proposed is done.

3.1 Algorithms

As mentioned before, the duplicates algorithm is the algorithm which selects the stock point price to be displayed to users scrolling through the various items. The allocation algorithm, however, selects the final stock point which will ship to the customer, which may, or may not, be different from the one chosen by the duplicates one.

3.1.1 Duplicates Algorithm

The Duplicates Algorithm works like a funnel, narrowing the number of possible stock points from level to level until only one stock point and price remains available.

These steps can be seen in figure 3.1 and work as follows:

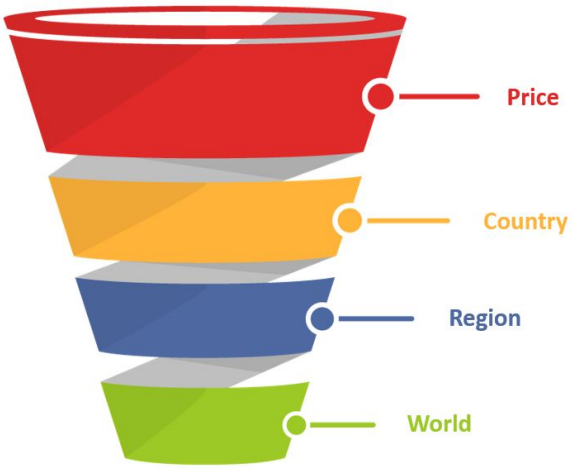

Figure 3.1: Steps in the duplicates algorithm.

1. Of all the stock points offering the same products, those with a price 5% higher than the cheapest are discharged. If only one is left, it is selected. Otherwise, we proceed to the second step.

2. In the "Country" step there are two distinct moments: the first one evaluates boutiques only, and, should there be boutiques from the same country, the one with the higher score is chosen[1]. Should there be no boutiques, the same logic is applied to brands. If no brand has a stock point in the same country they will not be selected, meaning only the stock points that belong to boutiques will go to the next step.

3. In the "Region" step the same logic that is applied to the "Country" one is applied but only to boutiques, since brands are dropped in the step before.

4. In the "World" step, the same logic that is applied in the two previous is used. Should no single boutique be selected, the boutique which has been a partner for longer is chosen.

3.1.2 Allocation Algorithm

The allocation algorithm, the focus of this thesis, works as a polynomial function which attributes a score to each stock point, assigning different weights to each variable in play. The function is defined, for n stock points with $Score_i$ (where i = 1,2, . . . , n), as:

[1] The score is calculated through averaging each boutique's average package sending speed over the last month, along with the number of orders they received which they could not fulfill and the average satisfaction of it's consumers . This is used as an empirical, yet unproven, way to optimize for customer satisfaction.

$$Score_i = HasStock \times \Big[click\&collect \times 10000 + BrandRanking \times 10000 + BasketCoverage \times 1000 + MerchantFromRequest \times 10 + MerchantRanking + \frac{StockDepth}{10} \Big] \quad (3.1)$$

Where,

HasStock takes the value of either 0 or 1 to check whether the boutique has available stock.

Click & Collecit takes the value of 1 when a certain stock point was selected as a pick up location, otherwise it is 0.

Brand Ranking, a binary variable, is used to flag if a stock point belongs to a given brand.

Basket Coverage, a binary variable, is given a value of one should the boutique be able fulfill all the items in the order.

Merchant from request, a binary variable, flags whether the stock point was the same as the one selected by the duplicates algorithm- this is the only connection between the routing algorithm and the duplicates one.

Merchant Ranking is applied when partners have more than one stock point that can fulfill the order and have therefore specified an allocation order for those stock points.

Stock Point Depth gives a higher score to stock points that have higher stock depth and attempts to minimize the number of no-stocks. No-stocks occur whenever the routing algorithm allocates to a stock point and, when the partner gets informed, the stock was already sold offline.

In the end, the stock point with the highest score is selected. The allocation algorithm is currently a complement to the duplicates one. It will respect the decision taken by duplicates algorithm, except for a very small number of cases where this would be deemed detrimental or due to business contracts.

3.2 The Problem

The reason why the allocation decision is not a trivial problem is due to three different stakeholders with different priorities. The problem can be considered a multi-objective maximization as it aims to maximize Farfetch's short-term profit, it's partners' potential as well as customer experience.

In order to illustrate this situation an example is provided. An item may be available to a client located in the United States in three different boutiques, all at the same price. One in Italy, one in the United States and one in Germany. The boutiques differ in the following aspects: the Italian gives a higher commission to Farfetch; the one in Germany has made a 1 million commitment in stocks to be sold in the portal and is seeing slow sales; and the one in the United States has a lead time which is, on average, 2 days smaller than the others.

This leads to three possible decisions depending on which perspective is maximized:

Client: Since the price will be the same, the faster the item arrives, the better. This sets the American boutique as the obvious choice - customers value smaller lead times (see section 2.1);

Partners: The partners make a large investment anticipating how much they expect to sell, which in turn means Farfetch will have both more stock depth and portfolio length. Should the ROI not be enough, the following year they will need to reduce their spending in stock. According to this perspective, allocating to the struggling boutique in Germany would be ideal;

Farfetch: as any company, Farfetch's motivation is maximizing profit. Short-term profit indicates that the boutique which bears the biggest cut should be chosen, in this case, the one located in Italy.

This is just a simple example of some criteria which could be used to make a decision, depending on the point of view each stakeholder holds. What follows is an extensive view of the variables and factors that may be important in the order fulfillment process, according to the three different perspectives.

3.2.1 Partners Perspective

Farfetch is a marketplace, meaning it relies on partners in order to do business. In other words, the company's product offering and, consequently, their sales volume, are highly correlated with the number of partners it has and how many products it offers. Keeping that in mind, it can be inferred that an increase in their product offering and/or their stock depth can be correlated higher sales.

The underlying hypothesis: higher partner satisfaction translates into sales.

Nevertheless, two problems arise from this hypothesis: how can partner's happiness be defined and how does one increase it without hurting business.

Regarding the first, there's an easy answer. A partner will be happy whenever ROI is positive and the bigger the increase in ROI, the happier they will be. This can be evaluated internally through a metric called F metric, that is expressed as:

$$Fmetric = \frac{AccumulatedStock}{Sales} \tag{3.2}$$

It is optimal that this metric reaches a value around 1, thereby guaranteeing that the boutiques will not be left with unsold stock.

Still, the second problem remains. Should an order be allocated to a boutique based solely on this metric, we may end up hurting Farfetch's bottom line, since it may be unreasonably costlier than cheaper alternatives. For this reason, we need to know exactly how valuable it is to balance the partners stock.

3.2.2 Client Perspective

Although the client is unaware of this, the decision from the allocation algorithm will influence his satisfaction and experience. Even tough price could be a determinant factor in this decision, as mentioned before, the algorithm's choice does not affect the price shown to the customer so it will

not be taken into consideration. Still, other factors can be considered which are known to have an impact on customer satisfaction.

Crossing factors overviews in the literature review with the ones that the allocation decision can affect, one can find a shortlist of possible variables to control.

First of all, the lead time for the order, will highly depend on the chosen boutique. After the client has made an order the store needs to follow a series of steps that conclude whenever it is ready to be shipped - this time is called speed of sending (SoS). After this, the selected courier picks the items from the store and delivers them to the customer - this is called the time in transit (TiT). The sum of these times gives us the lead time.

Some boutiques have, on average, a substantially higher speed of sending, which will directly increase the lead time. Furthermore, the boutique location itself also affects lead time, since boutiques are spread all over the world and some will be much farther from the client than others. The lead time needs to be controlled, since increasing it will be detrimental to customer experience, which will decrease the client satisfaction. That will, in turn, decrease retention rate, hurting future profits - as refereed in section 2.1.

The choice of boutique can further alter the experience the client has with Farfetch. The stock point will influence customer experience with their packaging quality, whether or not they ship the wrong item, whether they updated the stock levels in the system which may trigger no-stock events [2], among others. In oder to monitor this, Farfetch has a metric, Net Promoter Score (NPS), linked to each order. At the end of the experience, the client is asked to evaluate the service on a scale of 1 to 10. From the client's perspective, the orders should be allocated to stock points who can deliver the fastest and with the highest quality - higher NPS score.

3.2.3 Farfetch Perspective

As in any private company, the objective is profit maximization. This means that each order should aim to maximize profit. Profit depends directly on the following factors :

1. Shipping Cost: The shipping price paid by Farfetch to the courier, which, for the same product, varies depending on partner location and delivery type.

2. Commissions: The commission paid by boutiques to the company.

3. Recommended Retail Price (RRP): The price that the boutique charges Farfetch for the product. It varies from boutique to boutique.

And, with a smaller impact on:

1. Refunds: Some orders generate returns and refunds which may be contested by the boutiques,i.e, they will not be accepted and the company has to play judge in the case. When no

[2] No stock events happen because some boutiques don't give Farfetch automatic visibility of their offline sales, that is, sales in their brick and mortar stores. What happens then is that when an offline sale occurs, there is a delay until that missing unit is declared to Farfetch, so a sell may still be allocated to the store regardless. Should this happen, in case the store doesn't own any stock, it's called a no-stock event.

compromise can be arranged, a part of the cost may stay on the company's side. This results in some partners being regarded as less costly, refund wise, should they share a bigger part of those costs.

2. Overheads: Partners who suffer from delays shipping their orders cause more contacts from the customers, which increases the need for staff.

3. No-stock events: Since there isn't a full end to end integration with the stores, there is no real-time visibility of the offline sales from the boutiques which causes some orders to be allocated to stock points that already sold the product. Although once this situation is caught the order can be allocated to another boutique automatically, time is lost and customer experience deteriorated.

4. Rate of offline Sales (ROS) - Since each partner also features a brick and mortar business model, competition with offline clients, in an attempt to sell the boutique stock before offline clients are able to purchase it, will occur. However, should this be done indiscriminately and it may end up hurting the bottom line, since choosing boutiques using this variable only could end up having higher costs - due to possible increases in shipping costs, lead times, and others.

3.3 Problem Approach

The fact that each stakeholder has different priorities is a common problem in marketplaces. However, since the marketplace only works when all the parts benefit from the it, one could argue that a single perspective exists: the marketplace perspective. Instead of looking at the problem as a multi-objective maximization, one should look at it as a single objective problem - the need to maximize the marketplace profitability.

Since Farfetch owns the marketplace, maximizing the marketplace profitability is maximizing the company's profitability. In essence, all views contribute, directly or indirectly, to the company's profitability, with Farfetch prespectve featuring a short-term and narrower view of what profitability is and the other two being a broader and longer-term view of what profitability encompasses, as figure 3.2 shows.

However, in order to consider profitability in all its extension, a novel approach is needed regarding the long-term part. Instead of thinking about client satisfaction as the goal - such as aiming for reduced lead time - the satisfaction contribution towards client retention could be modeled and brought to the present. This would be directly comparable with short-term metrics.

Furthermore, the same could be done for all other variables, should their impact be modeled in terms of client retention and partners retention.

The proposed approach to the problem will do exactly that, meaning it will model and measure all the impact each variable has in profitability and provide a decision framework capable of guiding the allocation decision. This will increase the company's profitability.

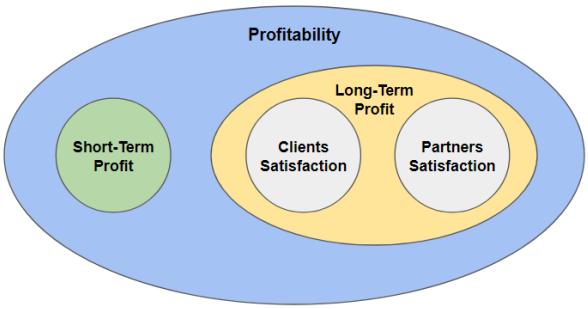

Figure 3.2: A Holistic view of what profitability encompasses.

3.3.1 Short-comings of current methodology

Before diving into the proposed solution, one should first understand why the current methodology used in the allocation algorithm is not ideal.

The algorithm is guided almost uniquely by the decision made by the duplicates algorithm and this decision is seldom changed besides the rare cases seen before. Otherwise, the short-comings that can be pinpointed to the duplicates algorithm will be shared by the routing one.

A quick overview of the short-comings would be as follows:

1. No bottom line protection: the algorithm has no way of calculating bottom line impact for each order, meaning no protection to maximize it is in place.

2. Customer satisfaction: Although the customer's country and region is in the algorithm, this doesn't directly correlate with a lower lead time. A customer that makes his order from the USA might have his order shipped by land from a boutique there, yet if the order was sent from Europe it would travel by plane, which, in some cases, would be faster.

3. Low mathematical accuracy in the score definition of the duplicates tiebreaker. The score is constructed using a weighted average, where weights are defined using an empirical intuition of what benefits customer experience. Those effects are not modeled in terms of customer satisfaction.

4. No modeling of the benefits. The algorithm was developed with the purpose of maximizing customer experience under the assumption that it would be the best alternative from a business standpoint. However, the formula developed is highly based on business intuition and not on models capable of measuring the actual impact.

3.3.2 Solution

To create a holistic profitability metric an understanding is required on exactly what variables contribute to profitability and how they can be modeled. In order to do this, a simulation of past sales was done to measure the impact when different allocation strategies were used.

In many cases, discrete event simulation is a natural approach in studying supply chains as their complexity obstructs analytic evaluation (Huang et al. (2003); Swaminathan et al. (1995)).

As such, the order fulfillment process will be modeled, step by step, into a simulation. The purpose is to evaluate what happens when different allocation strategies are deployed and help develop a strategy that maximizes profitability. To allow for the integration of some of the variables into the simulation, some modeling techniques will also be used.

Since, from the beginning, it was very hard to estimate how long it would take to model and run the simulator with all the variables that characterize the problem, the simulation was built in a modular fashion.

Even though the end goal is to develop a full profitability maximization strategy, by dividing the problem into smaller parts it allows for allocation strategies to be tested during development. That way, after a couple months, it was possible to deliver the first results and to deliver value to the organization while the simulator was still being further built, escalated and increased in complexity.

The iterative process that this work is based on will be reflected in the structure of the report, which will mimic the time line of the project. In order for the reader to fully understand how each step was made, the work documentation will be divided, into sections for every iteration that was performed. These amount to a total of three, so the results and methodology will be divided each into three sections as well.

The proposed solution involves a simulation which runs based on the general work-flow that can be seen in figure 3.3.

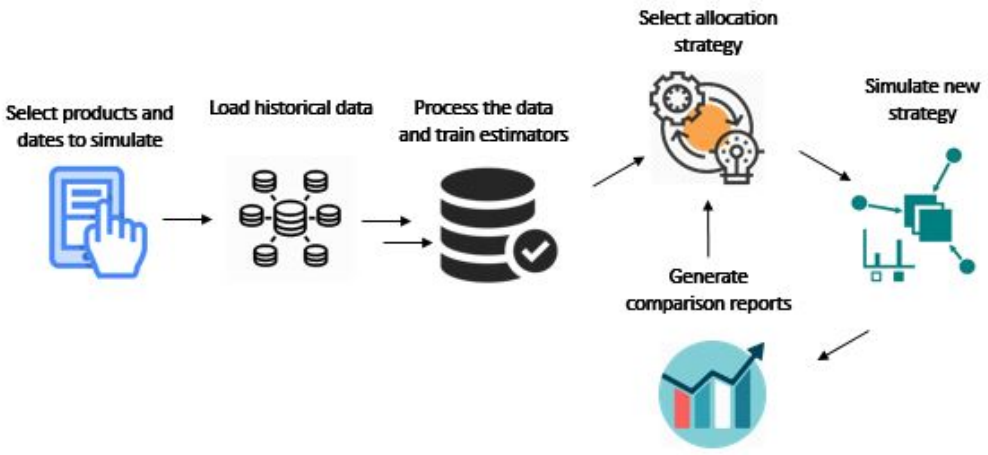

Figure 3.3: Simulation high-level overview.

As to ensure the simulator runs, two inputs are required: the products to be simulated and the time interval taken to look for orders to simulate. Since the purpose was to elaborate a case study,

the simulation always used the same products in the same period of time, for each strategy tried.

After the products and dates were selected, a script accesses the database and data begins to be loaded as needed - such as sales data, stock data, among others - , a process that is described in greater detail in the next chapter. When the data acquisition was finished, data cleaning processes were triggered and some variables where modeled, finishing when everything was ready to start the simulation process.

The simulation runs after a strategy is selected, such as minimizing the lead time, maximizing commissions, among others. As soon as an allocation strategy is selected the simulator starts allocating the orders following the strategy chosen. At the end of the simulation, reports are generated that allow for a comparison of the new solution with some benchmark.

Supporting this work-flow was the following hardware:

1. Processor Intel(R) Core(TM) i7-7820HQ CPU @ 2.90GHz, 2904 Mhz, 4 Core(s)
2. (RAM) 16,0 GB

Regarding the software and programming languages, the following were used:

1. SQL server 2014 management studio - the normal tool used in Farfetch to run SQL queries.
2. Google Big query - some queries accessed too many gigabytes of Farfetch information that were not allowed to be run as not to harm the other analysts work, in those cases, this cloud solution was used.
3. Python 3.6
4. SQL

Regarding the iterations that guided the development of this work, they all followed the same general process. As seen in figure 3.4, the first step would be to add variables to the simulator. These variables are modules that can be developed and connected to the simulator. After adding the new modules, an allocation strategy would be selected and ran. When the run ended, reports would be generated and evaluated. If the results would be satisfactory, an A/B test would be proposed, if not, other allocation strategies would be tested with the hope of finding better results. Finally, if the results found were satisfying, an A/B test would be proposed and a new iteration began.

In the next sub sections what will be done is explain which variables were used in each iteration. Each iteration further contributes to the building of a more holistic profitability metric since they add weight to the different perspectives.

3.3.3 1st Iteration: Direct Costs - Variables used

In the first iteration, only direct costs for Farfetch were considered, to see if a simple solution could find improvement opportunities. The variables - modules - added to the simulator in this iteration where:

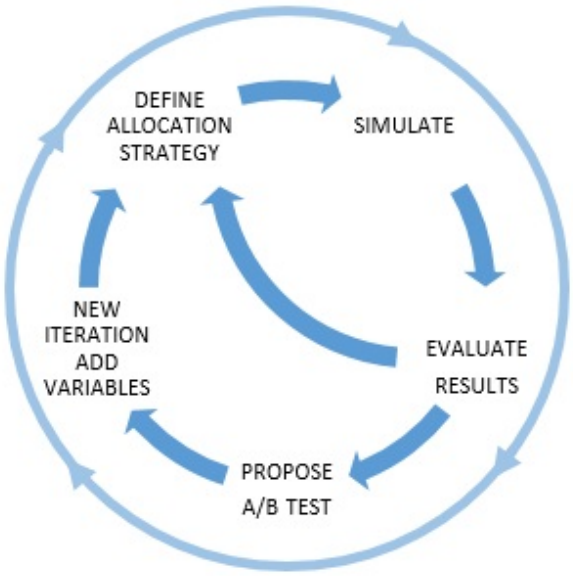

Figure 3.4: Iterative process of work development.

1. Price Paid –The amount, in GBP, that will be paid by the client to Farfetch deducted of any kind of credit card commission. This price includes the shipping costs, the duties,the VAT and the item price charged to the client. This costs charged to the client may be inflated[3] versions of the real ones paid by Farfetch.

2. Shipping Cost – The cost of shipping, in GBP, that Farfetch will have to pay to the courier - note that this is not necessarily the shipping cost that the client will pay. This price is calculated depending on the origin and destination, the weight of the package and the shipping type requested (Express, Standard, same-day and F90)

3. Local VAT – when this cost exist - only in purchases within the same country or within EU countries-, the cost, in GBP, that will be paid by the boutique/brand to the local government – again note that this is not necessarily the cost imputed to the customer but the actual cost.

4. Duties – when this cost exist- only in purchases between different countries, not including EU countries-, the cost, in GBP, that will be paid by Farfetch to the local government – again note that this is not the cost imputed to the customer but the actual cost. This cost is calculated based on country-specific tax laws and the base imputation is the RRP and may or may not include the shipping cost depending whether the importing country is a FOB or CIF country[4]

5. RRP–the item price that the boutique/brand charges on their stock, in GBP.

[3] When the client and the stock point use different currencies, the conversion rate applied is an inflated version of the real one.

[4] CIF and FOB mainly differ in who assumes responsibility for the goods during transit. In CIF agreements, insurance and other costs are assumed by the seller, with liability and costs associated with successful transit paid by the seller up until the goods are received by the buyer. FOB contracts relieve the seller of responsibility once the goods are shipped. In FOB countries, the shipping cost is not included. Source : Investopedia

6. Geo-Price or Fixed Price – Is the boutique/brand selling with a geo-price or not. The geo-price is a price dictated by the brand that creates the item, that reflects different demands and price sensibilities in different parts of the world. For example, a Gucci wallet may cost 1000 £ in Italy, 1200 £ in the USA and 1500£ in South Korea. When a price is dictated by a brand, the boutiques may agree to sell at that price or sell at a higher one.

7. Commission – the contracted commission that is paid to Farfetch.

8. Is Brand or Boutique – is the stock point a brand stock point or not. This is important since depending if it's a brand or a boutique the commission imputation base varies. If an item is not geo-priced, the money paid to the seller is calculated as follows:

$$Sellercut = RRP \times \left[1 - commission\right] \tag{3.3}$$

However, when an item is geo-priced, the formula differs if the seller is a brand, becoming :

$$SellerCut = \left[PricePaid - Duties - ShippingCost - VAT - \right] \times \left[1 - commission\right] \tag{3.4}$$

9. Farfetch cost – the cost that will be imputed to Farfetch with the purchase. This metric was created for this simulator and validated with key stakeholders. It's calculated as follows:

$$FarfetchCost = Duties + ShippingCost + VAT + SellerCut \tag{3.5}$$

3.3.4 2nd Iteration: Customer Experience - Variables used

In the second iteration, two lead time-related variables were added in order to start looking at the client perspective and see the costs related to it.

1. Speed of Sending – When an order is placed by the customer, the boutique/brand to which that order was allocated is immediately notified in order to start preparing the package for the courier to pick up. The time since they are alerted there is a new order until the time the courier picks the package for delivery is called the speed of sending and it's measured in days.

2. Time in Transit - The time, in days, since the order is picked up by the courier until it's delivered to the customer.

3.3.5 3rd Iteration: Partners Potential - Variables used

In the third iteration, the purpose was to try and optimize partner contribution to the business. As it was mention before, Farfetch is competing with brick and mortar clients for stock and that was the focus of this iteration. To evaluate this perspective, a new variable was created:

1. Rate of offline sales (ROS) - using the data from the stock levels, offline sales conducted by the boutiques could be estimated. The metric calculates the offline sales per week.

Chapter 4

Methodology

This chapter will be devoted to explaining the inner workings of the simulator. It will start with a deep dive into the simulation main steps hoping to inform the reader how it helped test new allocation strategies.

After explaining how the simulator works, there will be a section for each of the three iterations done in this work. Each section will clearly explain how the modules that estimate the variables - such as shipping costs, duties, among others - were developed.

4.1 Simulator

The main tool developed in this project was the simulator. The simulator occupies the spotlight of this work because it is what allowed to test new strategies and find new solutions. Given the allocation process complexity, only a holistic tool capable of monitoring all of the variables while trying new strategies would be able to reach any conclusion.

Furthermore, the simulator was built to take advantage of all the data available from past orders. What it does, in very simple terms, is take all the orders from a specific time period and simulate what would happen if different allocation strategies were employed. In the end, it allows for the direct comparison of what happened in reality with what would have happened if, for example, the allocation was done with the intent of reducing lead times. The way it works, in terms of information flow, is depicted in figure 4.1.

Before the simulation starts two inputs are demanded from the user: the products orders to simulate and the time interval. Once those are given, the program starts running. If it's the first time that input combination was given, a data preparation process has to be done. The process includes data acquisition, data pre-processing and cleaning. The two main processes worth mentioning are the collection of the original sales data and the reconstruction of the stock levels for the timeline selected. This guarantees that the allocation options available will be the same as in reality. After this process is done, a script is run to load the modules needed into the simulator.

In the next step, the simulation main loop starts and the orders from the timeline selected are retrieved and ran, one by one, through the simulator. When a new order enters the main loop, all

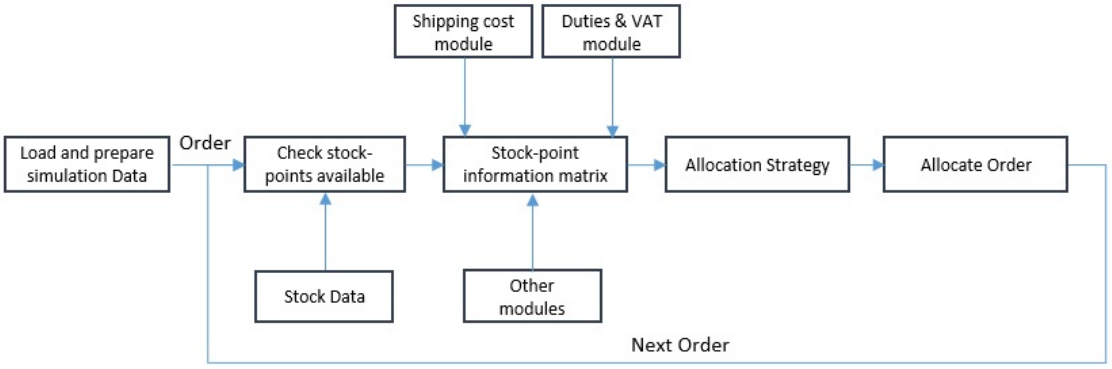

Figure 4.1: Simulator information flow overview.

the stock points that could fulfill the order are retrieved. Next, a matrix is created and filled with the expected values for that order, i.e the expected shipping cost, duties, VAT, RRP, and others -, for all the stock points available. Those estimations are done using the modules that were loaded. The way each module estimates the respective variable is explained in the next sections.

Once the decision matrix is completed, the simulator advances to the next step. Using the information in the matrix, a stock point to fulfill the order is selected based on the strategy that was defined. Again, this strategy can be the minimization of the Farfetch cost, the maximization of commissions or any other strategy deemed interesting. After the order is allocated, the main loop is repeated, until all the orders have been allocated.

As it can be seen in the figure 4.1, the stock point information matrix is generated using independent modules for each variable. This was crucial for the work developed since it allows for the work to be done in an iterative fashion. Further complexity could be added to the simulation simply by modeling new variables and creating, in the end, a new module to feed the matrix. Plus, since the results are in a matrix format, the implementation of different strategies is fairly easy, with simple tweaks in the code. To further the understanding of what the simulation matrix looks like, an example will be provided in chapter 5.

Finally, given that the purpose of this study was to find leads of potential areas of action, only a few products where selected, representing around 1 000 000£ in sales. The reason behind not using all the data available was due to performance problems. Nevertheless, since the simulator was done to test new allocation strategies, when there were new findings, they would be validated through A/B testing where the actual impact would be measured. This mitigated the need to scale the simulator.

4.2 1st iteration: Direct Costs

In the first iteration, only direct costs were subjected to analysis. The purpose was to start developing a profitability driven approach to the problem, using only short-term profit and start working from there.

As in all the iteration that will follow, the first step was to model the variables that would be used. After modeling, the simulator ran and allocated the orders to the same stock points which the orders were originally allocated. This allowed for the evaluation of the estimates the modules generated, comparing those against the original values.

In the first iteration the modules that were created allow the estimation of shipping costs, duties, VAT, geo-price, RRP, commission and brand status. The following subsections present how they were calculated.

4.2.1 Shipping Cost Module

There are, in short, 4 main types of delivery at Farfetch:

1. Express Delivery – Most of Europe and USA: delivery within 2-4 days. Rest of the world: delivery within 3-7 days. Represents 81,05% of the orders.

2. Standard Delivery - Only available within Europe: delivery within 5-7 days. Represents 18,85% of the orders.

3. Same day delivery - Only available in Berlin, London, Paris, Los Angeles, New York, Miami, Milan, Rome, Barcelona, Madrid, Dubai, and Hong Kong. Represents 0,09% of the orders.

4. F90 - delivers in 90 minutes but only available in Berlin, Dubai, Hong Kong, LA, London, Madrid, Miami, Milan, New York, Paris, São Paulo and Tokyo, in a very small radius. Represents 0,01% of the orders.

The main couriers used by Farfetch to fulfill the orders are UPS and DHL, with around 20% fulfilled by UPS and the rest 80% by DHL. However, in certain countries and shipping types, some other players are used. For example, F90 is mostly fulfilled by local couriers. Since the couriers only send the invoices and due to the nonexistence of an internal tool that estimates how much an order will cost, such a tool had to be developed. Nevertheless, time-wise, it would not be possible to integrate all the couriers and shipping types in that tool.

Given that Express delivery and Standard delivery represent around 99.9% of the orders, only those were deemed worth modeling. As such, orders fulfilled by F90 and same day services will not be processed on the simulator and the original allocation will be used instead.

To start developing the estimation tool, the first step involved talking with delivery analysts at Farfetch. The insights provided hinted that the main drivers of price, in both UPS and DHL, were distance, origin, destination, weight and the box dimensions. Using that knowledge, some regression models were tried but the results were not satisfactory as can be seen in Appendix B.

Discussing those results with the analysts, one suggested the use of two databases being constructed to validate invoices from the main couriers, DHL and UPS. The database for DHL required 4 inputs: Delivery Service, Box Weight, Origin, and Destination to estimate a price. The

UPS database required only 3: Delivery Service, dimensional weight, and zone. The zone is defined using the countries and the zip codes. The dimensional weight is calculated by dividing the cubic size of the package in centimeters by 5000 and increasing any fraction to the next half kilogram.

The zone estimation presented a challenge since the formula for its definition was not available. The easier solution would be to estimate all the shipping costs resorting to the DHL database since the couriers' charges tended to be similar. However, inside the United States, one of the top countries in order volume, the DHL service is not used and, therefore, not negotiated so a standard rate is applied. This rate, if used, would overestimate the cost in all orders when compared to the value charged by UPS. To solve the problem, a compromise was done. Make a module using the DHL tables to calculate all the shipping costs for orders that were not between states in the US. For the orders between US states, the UPS tables were used, as is shown in figure 4.2.

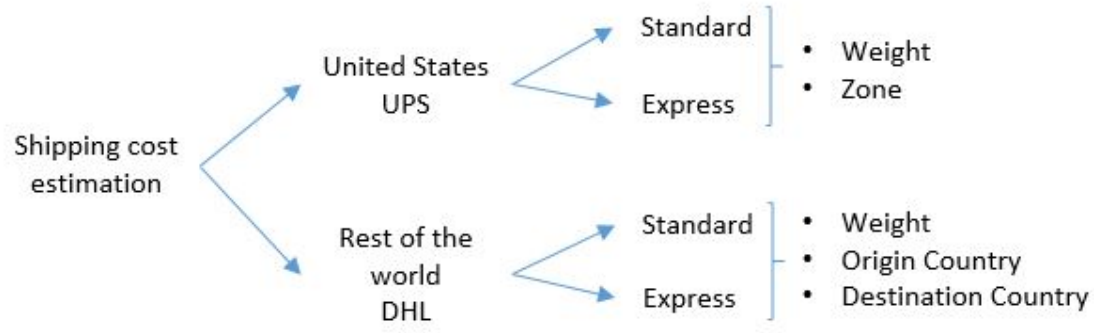

Figure 4.2: Shipping Costs module overview. It uses DHL estimation for all the orders except for those between USA states. For DHL, the shipping service - standard or express, weight, Origin Country and Destination country need to be specified. For UPS, shipping service, zone, and dimensional weight need to be specified.

To check if there were any caveats in using the DHL tables to estimate costs when it was delivered and charged by UPS, a statistic study was done. The results can be seen in appendix C, with no statistical evidence that there were significative differences between the two.

Regarding the UPS use in the US, the zones still had to be calculated using the zip codes. Fortunately, in the US, the zip codes follow a fairly consistent pattern where 6 digits are used and the more the zip codes sequences are similar, the closer are the zip codes. For example, the zip code 900001 is very close, location wise, to the 90002. With that information, and using historical data, the zones between postal codes could be identified by fetching the zones charged to those combinations in previous occasions.

However, a problem arose with the direct use of historical data. There were millions of possible zip code combinations and it would not be possible, performance wise, to keep millions of data entries in memory to estimate the zone.

In order to reduce the matrix dimension a quick analyze was run that can be seen in appendix D. The main conclusion was that using the combinations of the first 3 numbers in the postal codes, the average number of zones that existed was 1.57 per combination. To select a single zone for each combination, the zone that appears the most was used.

Before the simulation started, a script was run to calculate all the possible allocation combinations that could happen and 6 tables were created and kept in memory. Those tables allowed the simulation to fetch the shipping costs as needed. The tables included 2 tables for the DHL express and standard costs. They had all the possible country combinations as well as all the possible weights charged, as exemplified in table 4.1.

Table 4.1: Example of table used by the Shipping cost Module.

Origin	Destination	Weight(Kg)	Cost(£)
Portugal	London	2	10
Portugal	London	2.5	12
Portugal	United States	2	15
...	...	...	...

For the UPS standard and express costs, four more tables were used. Two of them with all the postal codes combinations of United States clients and stock points needed for that simulation and the respective zone - zones differ if it's express or standard, hence the need for two tables. Other two tables were used with all the possible zones, weights combinations present in the simulation and the respective cost.

4.2.2 Duties & VAT modules

Duties exist in international purchases, that is, purchases that are made between different countries. The only exception to that, in the luxury goods market at least, is when those countries are European Union countries. There, the transactions are done as if it was within the same country and only VAT exists. Otherwise, duties apply and the orders are VAT exempted.

Calculating duties is an intricate job due to the legislation differences between countries, still, a simplified approach can be done. The most relevant variables are: is the product geo-priced, is the country a Free on Board (FOB) our Cost, Insurance and Freight (CIF) country, the price paid and the tax bracket that is applied.

The geo-price defines the formula since when a purchase is geo-priced the duties are included in that price and calculated as:

$$C = WP - \frac{WP}{1 - DDP} \tag{4.1}$$

where

C are the duties value to be paid,

WP is the website price, that is, the price charged ,

DDP is the tax bracket for that product.

When it's not geo-priced, if it's a CIP country, the duties are calculated as :

$$C = \left[RRP(withoutVAT) + ShippingCost \right] \times DDP \tag{4.2}$$

,otherwise, in the FOB countries, the Shipping cost it's not included.

The problem with duties is to know whether the seller, in the client' market, sold at a geo-price or not and in which tax bracket that purchase falls into. In 15[1] markets, the most important, Farfetch keeps a daily snapshot of each product that is offered in those countries. The snapshot includes information regarding how much the sellers are selling them for, if it's fixed price, the DDP and VAT rate applied. For those, calculating duties and VAT is trivial. For all the other markets, it was not so much. In appendix E, there is a diagram of how the duties and VAT are calculated in those cases, resorting to the equations 4.1 and 4.2.

Again, before the simulation started, a script was run to import all the snapshots from the countries that were present for all the days and products analyzed. For the other countries, not included in the snapshot, a tool that accessed the database in real time was constructed and used.

4.2.3 Geo-price and RRP modules

These two variables were put together because the process to calculate both was exactly the same. As in the case of the duties, in the 15 countries that the company keeps a daily snapshot, it was only necessary to access the snapshot already loaded for the Duties & VAT. In the other cases, the countries not included in the snapshot, some approximations where done.

For those, the first step was to select the country zone, which is an internal classification for the countries done by Farfetch. Countries in the same zone tend to share similar prices and laws. After selecting the zone, information from countries that belong to that zone and were present in the snapshot was retrieved. Then an average RRP of the stock point for the countries in that zone was used - in the vast majority of the cases, the RRP of the stock point did not change. For the geo-price, a more conservative approach was made. If there was a country in the zone where the stock point had accepted geo-price, it was assumed that it had also used geo-price for the country being queried. Then, the average geo-price was retrieved from the countries in the snapshot belonging to the same zone and used. If no country in the same zone was subjected to geo-price, it was assumed there was no geo-price. This information was needed for duties calculation purposes. For this variables, the calculations were done in real time, fetching only the daily snapshot from memory.

[1] The countries included in the snapshot are : Australia, Brazil, China, France, Germany, Hong Kong, Italy, Japan, Republic of Korea, Macao, Portugal, Russian Federation, Switzerland, United Kingdom and United States.

4.2.4 Price Paid, Commission and Is brand Boutique modules

This last variables did not require any math to be calculated. In the case of the price paid, the routing algorithm is deployed at checkout where the price will not change, so the price paid is always the same.

Regarding the Commission and the brand boutique status, those are attributes from the stock points and are loaded at the beginning of the simulation from the database, and that information is fetched as needed.

4.2.5 Establishing a baseline

First and foremost, the quality of the estimations for the variables needed to be evaluated. In order to do this, the simulation was run using an as-is allocation strategy. This strategy allocated the orders to the same stock points used in reality. The estimations generated by the modules are compared to the original ones. In the figure 4.3 are depicted the results from the estimated variables using the as-is allocation strategy and the real values. What the as-is allocation does is allocate each order as it was allocated in reality.

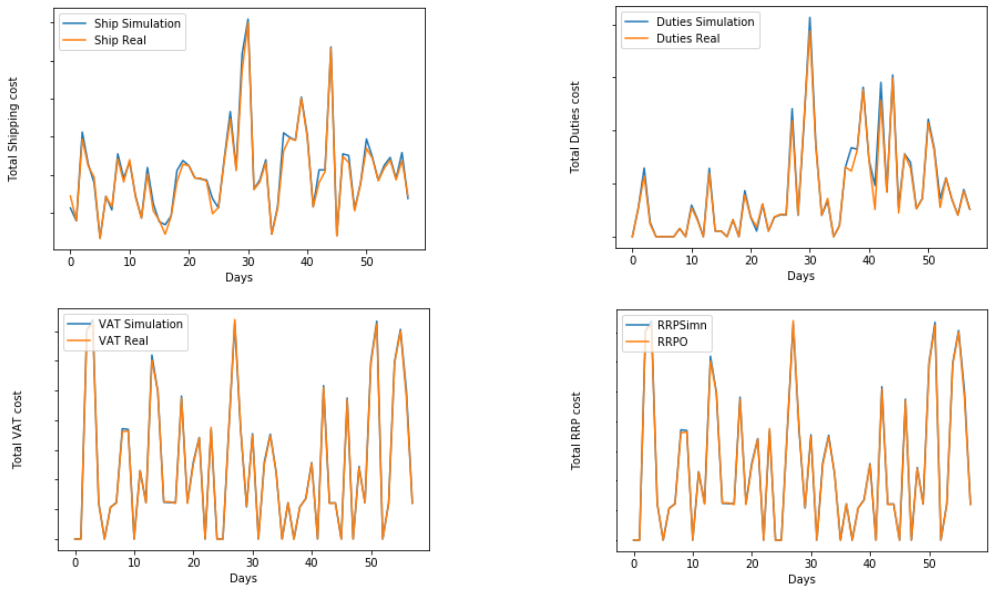

Figure 4.3: Comparison of the simulator values and the real ones to evaluate the quality of the estimations made by the modules.

As it can be seen, the estimations are not perfect since some approximations were used. Still, overall the modules are accurate enough to be used and test different allocation strategies.

In the next chapter, the results from testing different allocation strategies done in this iteration are showed.

4.3 2nd iteration: Customer Experience - Variables used

In the second iteration, lead time was added to the decision matrix, adding the first long-term profit variable thus making the first step towards the construction of a holistic profitability measure. At Farfetch, the two main lead time components are the speed of sending and the time in transit. The next subsections focus on explaining how lead times were calculated as well as how they were converted into profitability.

4.3.1 Time in transit

Two methodologies for creating a TiT module were thought. The first was to create a forecasting tool to predict how long an order would take to reach the final destination. Although this alternative would be academically interesting, the effort and time involved would be detrimental to the quality of the rest of the analyses. Since the main goal was to capture the variability that exists between the routes only, a much simpler tool would be able to deliver this results without requiring so much time. For that reason, the decision was to modulate the TiT distribution for each country and then use that distribution to generate TiT values for each order allocation.

Fitting to a know distribution was the logical approach and it was the first one tried.

To test if this approach was feasible, a sample of TiT values was extracted from the database. The samples consisted of one year worth of observations of the most commonly used route - Italy to the United States- using DHL express delivery service. The purpose was to fit the sample to a normal distribution. If that did not work, then a script would run to try and fit the data in up to 80 know distributions. After fitting, the approximated distribution that had the highest p-value in a Kolmogorov–Smirnov(KS) would be selected. This approach had an assumption that, at least one of the distributions would have a p-value higher than 5% since that was the significance level selected. As it will be shown, that assumption would not be correct.

The Kolmogorov–Smirnov(KS) is used to test whether two underlying one-dimensional probability distributions differ. The test hypothesizes are as follows:

H_0 : The two samples come from a common distribution.

H_a : The two samples do not come from a common distribution.

The histogram from the sample selected is shown in figure 4.4. In the histogram showed, outliers were removed. Samples with times in transit higher than 20 days - around 0.01% - were considered lost, following shipping annalist recommendations. Observations that had a negative time in transit - around 0.04% - were also removed.

Looking at the histogram two things pop out, the fact that the distribution is skewed to the left and the existence of depression between the main bars. In order to deal with the skewness of the distribution and try to normalize it, 3 transformations were tried: logarithmic, box and using a robust scaler. The results can be seen in appendix F.

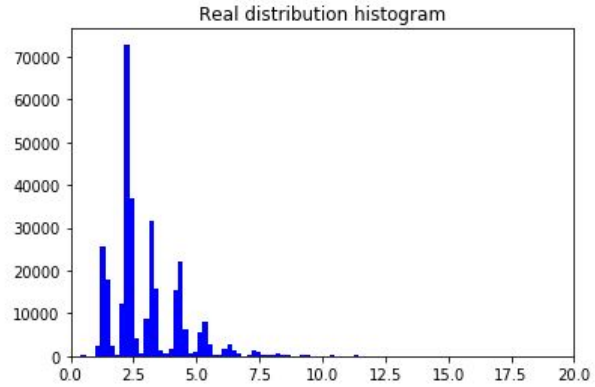

Figure 4.4: Time in transit histogram from route between USA and Italy for express shipping type.

Nothing was able to produce p-values greater than 0.05 by approximating to a normal distribution so the next approach was attempted, try to fit the distribution directly to one of the 80 know distributions the script would try.

Again, none was capable of a good fit. The problem was the existence of these depressions between the bars that no known distribution was capable of fitting. Knowing this, efforts were concentrated on understanding what was happening in the data that made the histograms look so peculiar.

A series of hypothesis where tested, and in each, a new level of granularity was reached. The data was divided by day of the week the order was sent, isolating the partner that was sending the package, the month in which it was being sent, among others, but no answer was found. Finally, the reason for the depressions was found and can be explained using the table 4.2.

Table 4.2: Cumulative frequency of picking hours in stores by the couriers.

Picking hour	Frequency	Cumulative Frequency
19	0,16	0,16
18	0,15	0,31
17	0,15	0,46
20	0,09	0,55
15	0,08	0,64
13	0,07	0,70
16	0,06	0,77
21	0,06	0,83
14	0,03	0,86
22	0,03	0,89

As it can be seen, most of the orders are picked up by the courier at the stock point during the afternoon. This results in either the location being close enough and the courier could deliver the package right away or the business hours window would end. The pattern repeats itself every day, either the courier is capable of delivering during the time window the business hours allow, or it can only be delivered the next day.

Seeing it was not a data problem but the reality of the situation, an approach capable of handling this peculiarity was needed. This is where the B-splines were used.

As seen in chapter 2, B-splines can be used to approximate functions resorting to a set of polynomial functions and a set of data points, the knots. The purpose was, with a limited number of coefficients, generate time in transit values that followed the same distribution of the original sample. In order to do that, the splines were used to model the inverse cumulative function generated from the sample. The reason to use the inverse and not the original one was solely due to, computationally-wise, being easier to generate values that way, as it will be explained later.

To calculate the splines coefficients, the package interpolate from scipy was used. As mention in chapter 2, the inputs, besides the data points, were the polynomial degree to use, the knots and the smoothing factor. The package author recommends the use of 3-degree polynomials for the splines, so those were used. Regarding the number of knots, those were chosen automatically and equidistant knots were used. The way the knots are chosen by the package is based on the smoothing factor chosen. A small smoothing factor allows for overfitting so a large number of knots is used and a high smoothing factor prevents overfitting and fewer knots are used.

Nevertheless, the purpose was to create a similar distribution, so the interest was to have as much overfitting as possible. However, overfitting required a lot of knots and coefficients, which required more memory, therefore, defeating the purpose of modeling. To solve this trade-off, a study was conducted by varying the smoothness factor and calculating the p-value using a KS test. The test used the real distribution and the one obtained from the use of the B-splines with the varying smoothing factor, which can be seen in figure 4.5.

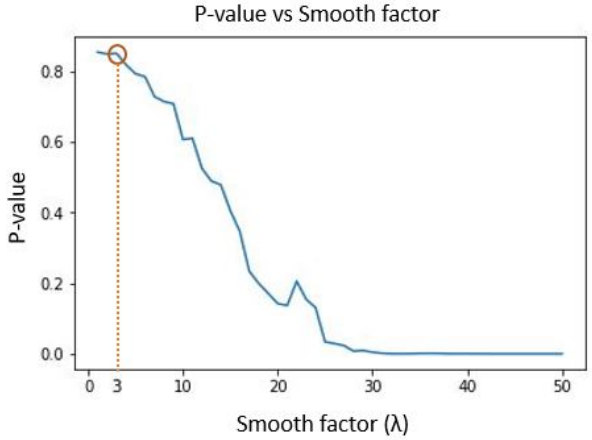

Figure 4.5: Comparison test with different smooth factors and respective p-values. The smooth factor that was considered as the best trade-off was $\lambda = 3$.

The smooth factor choose, $\lambda = 3$, allow for a reduction of dimensionality from 300000 data points to 150 knots and 149 coefficients.

The value generation, after having obtained the coefficients and knots for the B-splines is straightforward. Since the interpolated function was the inverse cumulative function, a number was randomly generated between 0 and 1, given to the function as an x value, and the y value

obtained would be the time in transit generated. In figure 4.6 this generation process is illustrated. The reason for using the inverse cumulative is that it's simpler to obtain the value of y where $y = f(x)$ than to obtain the x having the y value.

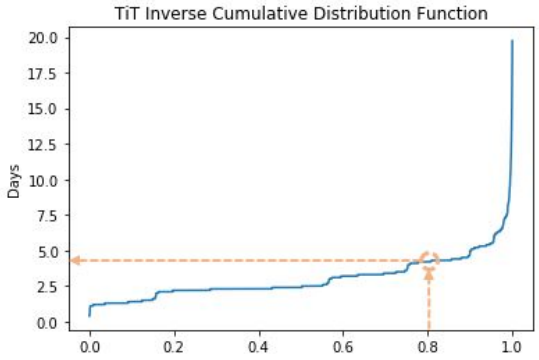

Figure 4.6: A random value, x, is generated, between 0 and 1. After generating the value, f(x) is obtained, giving the time in transit based on the distribution.

The quality of the results of using B-splines to approximate a distribution and generate values can be seen in figure 4.7. The results are from the test sample used with the TiT between Italy and US. The KS test was done with both distributions on figure 4.7. The results were a test statistic of 0.001695 and a p-value of 0.766, meaning that there was no statically significative evidence that the values came from two different distributions.

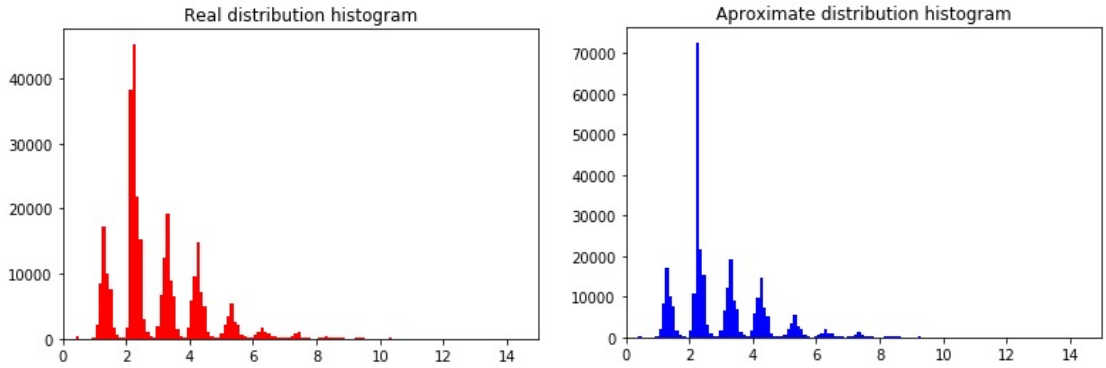

Figure 4.7: Comparison of the distribution histogram of the TiT between Italy and USA using the DHL express delivery service. Both samples have the same number of observations - around 300000. On the left there is the real distribution while on the right the one obtained using the B-splines to generate 300000 values.

Furthermore, before using the technique to create the estimations two preventive measures were put in the value generation. The first was to avoid that the values generated would be extreme values by chance. As such the simulation is run 1000 times and an average and confidence interval is defined for each value generated, to ensure that the TiT obtained would not be an extreme value. The second was done to prevent that the values generated would not be outside the interpolation

domain, where the B-splines do not behave well. Hence, the minimum and maximum values from each distribution were saved alongside the splines coefficients to ensure the values generated were located in that interval.

In the end, a module was constructed and had to be loaded before the simulation started. The module created four data-sets that where kept in memory during the simulation. Two of those were generated for the DHL express and standard delivery time. Each had all the possible country combinations for the orders being simulated and the subsequent B-splines interpolation parameters, fitted in the loading process.

For the shipments made using UPS, two more data-sets were generated for the corresponding shipping types, but instead of the country combinations, the zones were used instead.

4.3.2 Speed of Sending

To estimate the speed of sending, the same options available for the TiT existed. A predicting tool could have been developed or the empirical distribution could be modeled. Using the same line of thought as before, modeling was preferred since the tool was already developed for the TiT and no further adjustments would be needed.

Looking at the density plot from the speed of sending of a store - see figure 4.8- there are still depressions in the graph.

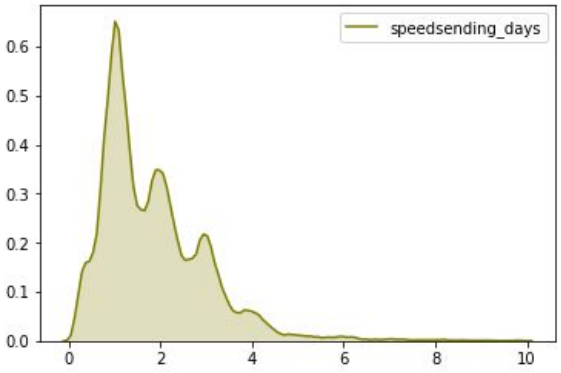

Figure 4.8: Density plot with the speed of sending of randomly selected store.

In this case, although the store shares the business hours problem, when the stock points are available to prepare the package, the problem is not so critical. The reason is that the orders come from all around the world, and with different times zones, they come in more evenly distributed times, as it can be seen in table 4.3. This helps in dampening the effects of the working schedule of the store.

Still, the depressions did not allow for the use of known distribution and B-splines where used again. The results demonstrating the quality of the approximation are shown, using a random store, in figure 4.9.

For that sample, a KS test was made with a test statistic of 0.003203 and a p-value of 0.774, meaning that there was no statically significative evidence that the values came from two different

Table 4.3: Cumulative frequency of time a order arrives to the store.

Order Hour	Frequency	Cumulative Frequency
15	0,06	0,06
14	0,06	0,12
16	0,06	0,17
17	0,05	0,23
13	0,05	0,28
12	0,05	0,33
11	0,05	0,38
18	0,05	0,43
10	0,05	0,47
19	0,05	0,52
9	0,04	0,56
20	0,04	0,60
21	0,04	0,65

distributions. The module for the stores' speed of sending was also loaded before the simulation started and the B-splines coefficients for each store in the simulation were kept in memory, as well as the minimum and maximum values.

4.3.3 Establishing a baseline

Using the same strategy as before, the simulation was run using an as-is strategy in order to check the modeling quality. In figure 4.10, a comparison between the average times and the average that resulted from running the simulator 1000 times with an as-is strategy are shown. Again, the differences were small so the approximation was deemed accurate and used.

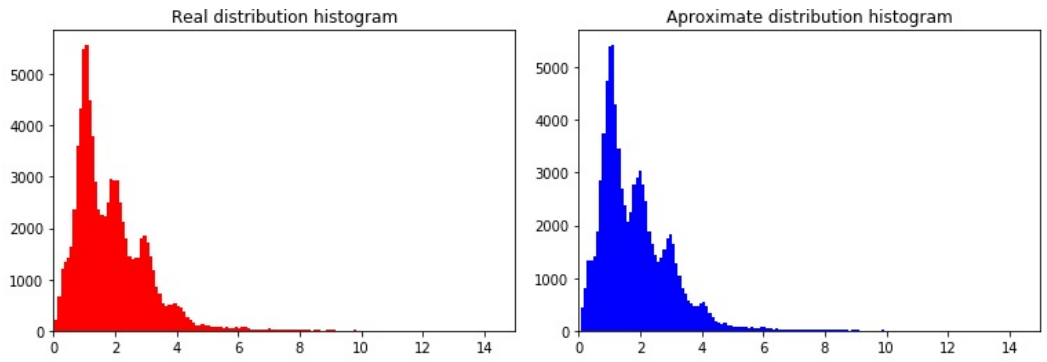

Figure 4.9: Comparison of the distribution histogram of the SoS of a random store. Both samples have the same number of observations - around 30000. On the left there is the real distribution while on the right the one obtained using the B-splines to generate 30000 values.

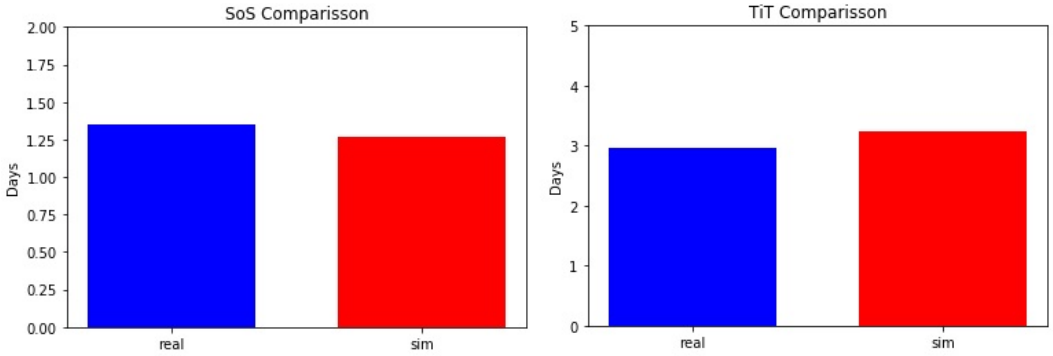

Figure 4.10: SoS and TiT comparison between the real and simulation values.

4.3.4 Converting to profitability

Creating a module capable of estimating the lead time(LT), calculated as:

$$LT = SoS + TiT \tag{4.3}$$

was not enough to use in the allocation decision. To explain why an example is elaborated around the data from table 4.4.

Table 4.4: Example of possible decison allocation where both costs and time need to be considered.

Stock Point	Farfetch Cost (£)	Lead Time (days)
1	200	2
2	205	1.9

From the example, there are two stock points that could fulfill the order. Stock point 1 has a lower cost and higher lead time and stock point 2 has the exact opposite, so the allocation decision is not straightforward. When making such a decision a multi-objective maximization problem is present. If lead time was being minimized, stock point 2 would be the obvious option while if costs were minimized, stock point 1 would be the chosen.

However, if the problem were to be formulated in such a fashion as -"How much is worth to the company to provide a better customer experience for the client?"- instead of having a multiple objective problem, there would be only costs and profits considered instead.

Although the question is easily formulated, it poses a challenge to provide an answer. The underlying hypothesis is that smaller lead times contribute positively to the customer experience resulting in higher retention. The hypothesis is exemplified in figure 4.11, with the time between purchases - signaled with the letter P - getting smaller for smaller lead times.

To measure the lead time contribution the difference, for a period of time, in profits of serving customers with different lead times would have to be known. Resorting to the knowledge acquired in chapter 2, what would be needed was the difference in the expected CLV when subjecting customers to different lead times.

As seen before, CLV is calculated as follows:

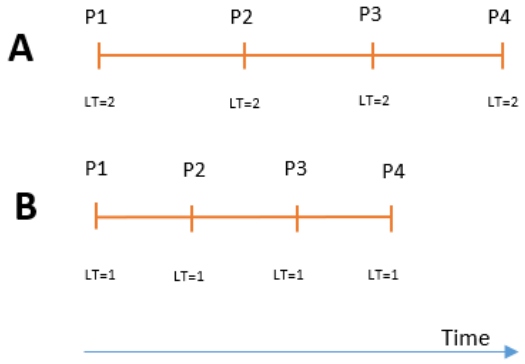

Figure 4.11: Comparison of purchase rates of two identical clients subjected to diferent lead times in each purchase. Client A has a lead time of 2 days in every purchase while B only has 1 day.

$$CLV = \sum_{i=1}^{T} \frac{(m_t)\, r_t}{(1+i)^t} - AC \tag{4.4}$$

where

m_t = is the margin of the purchases made in time t,
i = discount rate or cost of capital for the firm,
r_t = probability of customer repeat buying or being "alive" at time t,
AC = acquisition cost, and
T = time horizon for estimating CLV.

Using the example of figure 4.11, what needs to be calculate is :

$$\Delta CLV = CLV_B - CLV_A \tag{4.5}$$

Which can be written as , using equation 4.4 and 4.5 :

$$\Delta CLV = \sum_{i=1}^{T} \frac{(m_t)\, \Delta r_{B-A}}{(1+i)^t} \tag{4.6}$$

Looking at equation 4.6, if the increase in retention rate,Δr_{B-A}, resulting from lower lead times could be calculated, the difference in CLV, ΔCLV, would be how much the lower lead time is worth, at its present value. That way, it's no longer a multi-objective problem but a single objective, since the present value of the future impact is calculated, and the objective becomes maximizing the profitability of the order.

Thus, the question left to answer is how much does the decrease in lead time affects retention. To calculate this, it's only needed to calculate the coefficients for the proportional hazard rates review in chapter 2. To understand how the model works and how the coefficient was calculated the reader is urged to consult Appendix A an in-depth explanation is provided. Here, only the output of the model will be used, which gave the lead time a $exp(\beta) = 0.98$. This means that, if

all the other factors are kept the same, a reduction of one day in lead time, will increase retention by 2%.

Lastly, what's left is to define the time interval of the margins calculated, how many years to project the CLV and the capital cost of the company. The capital cost of the company used by the marketing personnel is 23%. Regarding the time interval to consider, at Farfetch the CLV is calculated using a yearly time interval and projecting for one year only. Although literature convened the use of perpetuity's, using the companies current approach seemed a better option since it would be more logical to follow what's already implemented.

And so, the formula to convert the lead times difference into a comparable metric with cost was the following:

$$\Delta CLV = \frac{m_t \times \Delta LT \times 0.02}{(1+0.23)} \tag{4.7}$$

where

ΔLT is the Lead Time difference between the lowest lead time in the decision matrix and the one being evaluated,

0.23 is the capital cost of the company

r_t = probability of customer repeat buying or being "alive" at time t,

0.02 is the impact one day in lead time has in retention

m_t is the average yearly profit the company makes with the average customer.

Having the ability to convert future impacts into short-term profits/costs, allowed for the creation of the profitability metric. The metric is defined in equation 4.8.

$$Profitability = PricePaid - FarfetchCost + \Delta CLV \tag{4.8}$$

Using the value of 145£ as the average yearly profit for the average customer, equation 4.7 takes the form:

$$\Delta CLV = \frac{145 \times \Delta LT \times 0.02}{(1+0.23)} = 2.36 \times \Delta LT \tag{4.9}$$

The results means, that if the company managed to reduce the lead time in a order by one day, that would result in an increase in profit of 2.36 £.

Lastly, table 4.5 provides an example of how a order could be allocated based on this criteria. In the example, although if looking solely at the short term profit stock point 1 would be selected, looking at profitability where the lead time value is represented, the best solution would be to allocate to stock point B.

Table 4.5: Profitability based allocation.

Stock Point	Price Paid (£)	Cost (£)	Lead Time	Profit (£)	$\Delta CLV\,(£)$	Profitability (£)
1	15	10	2	5	0	5
2	15	11	1.5	4	1.18	5.18

4.4 3rd iteration: Partners Potential

In the third iteration, the purpose was to look at strategies to increase partners potential. The first step was to see if different allocation strategies would result in higher stock levels in the system. To do this, the offline sales from the stock points were also needed since otherwise, it would not be possible to monitor the stock levels.

4.4.1 Offline Sales & Rate of Offline Sales

Since only stock information is provided by the stock points, the offline sales had to be estimated. The data available to do so were the online sales and the stock depth at the end of the day for each product, in each stock point.

First, all the sales of one day were added to the stock levels at the end of that day, giving what would be the stock levels, in each day, as if no sales in the portal occurred. Then, the differences between the stock depth from one day to the next were calculated. If the differences were positive, that is, there was more stock in the day before, that value was assumed to be the offline sales of that day. If the differences were negative, and the stock had increased versus the day before, it was assumed that new stock had been added to the system. This was an approximation of reality since both could have happened on the same day. However, since the stock is added periodically, it would not have a big impact.

In the end, the metric to be used was calculated, the ROS, estimating the rate of offline sales in each stock point. For the module, the value for each week, per product, per stock point, was kept in memory.

Chapter 5

Results

In this chapter results from running the simulator using different allocation strategies in each iteration will be showed and compared to the results from the original allocation.

The chapter is divided into sections, one devoted to each iteration done. As such, inside each section, the KPIs to be evaluated in that iteration will first be shown, then an example of how the decision matrix looks is provided. The purpose is to see, as new variables are added to the decision matrix, how new allocation strategies can be used and their impact in the KPIs. It should be noted that all the iterations resort to the same orders data-set, growing only in complexity the decision matrix.

5.1 1st Iteration: Direct Costs

In the first iteration, only direct costs are being analyzed, so the KPIs focus will be on direct financial profit per order resulting from different allocation strategies. The main KPI will be the profit per order, that is, the money that is left, on average, to Farfetch after all the expenses have been paid. The variables that contribute to this KPI will also be monitored to further the understanding of where the changes are coming from.

Running the simulator with the as-is allocation a baseline was established. The results for the KPIs are not showed since the company requested for the results to be showed in variations from the baseline. Nevertheless, the KPIs used are: average shipping costs (£), average RRP (£), average duties paid (£), average VAT paid (£), average Farfetch cost (£), average commissions captured (%) and average profit per order (£). These KPIs will serve as comparison for how the other strategies implemented are behaving against the original strategy used by the company.

Lastly, before diving into the strategies tested, the decision matrix for a given order is presented in table 5.1. This matrix is created for each individual order ran through the simulator. It has, for all the stock points that could be used to fulfill an order, an estimation of all the variables that may impact the decision. This estimations are generated using the modules, as explained in chapter 4. For example, the expected shipping cost, the expected amount paid in duties, among others. The

matrix serves as the base to test different allocation strategies since each order can be allocated knowing, beforehand, what will be the value of each variable.

Table 5.1: Decision matrix example

Stock-Point	Price Paid	Shipping Cost	RRP	VAT	Duties	GP	Commission %	Is Brand Boutique	Farfetch Cost
9499	171 £	14.1 £	219.8 £	0 £	102.1 £	FALSE	-	No	285.4 £
9148	171 £	4.7 £	138.3 £	27.6 £	0 £	TRUE	-	No	137.5 £
9904	171 £	8.1 £	138.8 £	30.5 £	0 £	TRUE	-	No	140.2 £
9843	171 £	8.1 £	140.7 £	30.9 £	0 £	TRUE	-	No	145.5 £
9597	171 £	8.1 £	139.7 £	30.7 £	0 £	TRUE	-	No	141.1 £
9640	171 £	14.1 £	234.8 £	0 £	108.6 £	FALSE	-	No	305.9 £

5.1.1 Maximizing Commissions

The first strategy that was tested was maximizing the commissions, meaning an order would be allocated to the stock point that had the higher negotiated commission. The rationale behind testing this strategy was that a similar one had been deployed in the company. The tested strategy was not a plain commission maximization since what was done replaced the score by the commission in the duplicates algorithm[1]. The results from the test previously done by the company were dubious. However, since it was an allocation strategy that would not require technological development and could be readily implemented, it was further tested.

Implementing the commission maximization was done using the matrix decision generated for a given order and allocating the order to the stock point with the highest negotiated commission. The results are presented in table 5.2, with the as-is comparison.

Table 5.2: Results from applying a maximizing commissions strategy.

KPIs	Avg. Shipping Costs	Avg. RRP	Avg. Duties	Avg. VAT	Avg. Cost	Avg. Commission	Avg. Profit per Order
Maximization	+1.8%	-0.5%	+4.8%	+2.7%	-0.4%	+2.1%	+1.1%

As seen in table 5.2, this new allocation would result in a higher profit per order, increasing the average profit per order in 1.1%. Although this is interesting, it provides little margin for error since the impact on customer satisfaction is not being controlled. For that reason alone, even though the result is interesting, it was not very actionable without further investigation.

5.1.2 Minimizing Costs

The next strategy tried made the most business sense, minimizing the costs in each order and, consequently, maximizing the profit per order, since the price paid does not change. The results from this strategy can be seen in table 5.3.

Comparing the results, the average profit per order could be increased by an astonishing 9.0%. This increased seemed too good to be true, therefore the decision matrix for each order was saved in order to be analyzed and the source of the gains inspected.

[1]The duplicates was using price, location and boutique score as some of the criteria to select a stock point. In that test, the score was replaced by the commission to try and understand if the commission captured would increase without hurting the customers.

Table 5.3: Results from applying a cost minimization strategy.

KPIs	Avg. Shipping Costs	Avg. RRP	Avg. Duties	Avg. VAT	Avg. Cost	Avg. Commission	Avg. Profit per Order
Minimization	+1.0%	-2.2%	-3.2%	+2.7%	-2.0%	+0.9%	+9.0%

In figure 5.1 the difference in profit per order for all orders between this strategy and the as-is is showed. The spikes seen indicate major gains could be achieved,where in some orders the gains could ascend to almost 220£.

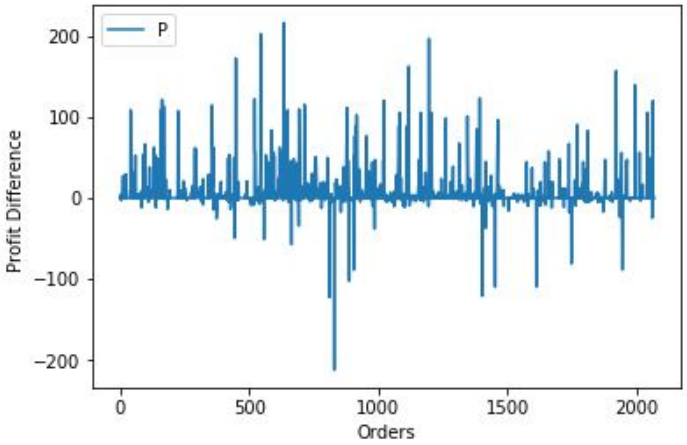

Figure 5.1: Differences in profit per order from minimizing costs vs the original allocation.

To understand were those gains were coming from, the saved decision matrices were used to identify some of the spikes and the results were validated by resorting to the analysis of the website snapshot for those days. The daily snapshots have, for some markets, the price at which the products are being sold by the different sellers. In table 5.4 there's the snapshot of one of the products analyzed on the day the simulator found a profit difference of around 60 £ between the strategies.

Table 5.4: Analysis of snapshot from the website where a spike occured.

Customer Region	Boutique Region	Stock Point	VAT	Geo Price	DDP rate	Website price (£)	RRP (£)
United States	United States	15288	0%	FALSE	0%	235	235
United States	United States	13592	0%	TRUE	0%	219	219
United States	Italy	15053	22%	TRUE	5%	219	139
United States	Italy	15142	22%	TRUE	5%	219	138
United States	Italy	15063	22%	TRUE	5%	219	138
United States	Italy	15314	22%	TRUE	5%	219	138
United States	Italy	15126	22%	TRUE	5%	219	138
United States	Italy	14656	22%	TRUE	5%	219	138
United States	Italy	15150	22%	TRUE	5%	219	138
United States	Sweden	15062	25%	TRUE	5%	219	134

What was found is quite interesting. The duplicate algorithm was not able to protect the bottom line of the company. That happened due to business changes,which ultimatly rendered obsolete the protection used. First, it will be explained how the duplicates algorithm was protecting the company's bottom line and then why it is failing.

The first step in the duplicates algorithm is to cast out all the stock points where the price is higher than 5% of the lowest price. Although this was meant to protect the client it also protected the company. The price was, in simplified terms, constructed using the formula in equation 5.1.

$$Price = RRP + ShippingSubsidy + LocalVAT + Duties \tag{5.1}$$

This meant that the main driver of the price was the RRP. The reason is that the shipping cost, for very high price purchases, it's not very significant. Also duties are dependent on the country and RRP itself. In conclusion, the price protection meant, partially, an RRP protection for the company.

However, as the company grew in size and started to be recognized as a luxury fashion player by the brands, they demanded that the pricing strategy followed the brands' strategy for each country. This meant the price was no longer constructed based on the formula above but set by outsiders. The reason this matters is that the price paid by the customer is the same when the item is geo-price, if a seller has a higher RRP, Farfetch is the one paying that difference.

The implications of this change were profound. As it can be consulted in table 5.4, if the duplicated algorithm would be applied to those stock points what would happen would be the following:

1. The algorithm would look at the price and remove only the first stock point (15288) that did not accept the geo-price and had a higher price. All the other stock points would go to the next step in the algorithm.

2. In the second step, the algorithm would look for the stock points that were in the same region as the client, in this case, the United States. Since only one stock point is from the US - stock point 13592 -, in this case, it would be selected.

Nevertheless, the problem is that since the price is not constructed from the RRP, there was no step in this algorithm that protected the bottom line. And, as a consequence, the resulting allocation was catastrophic. The stock point that got the order had an RRP of 218.6 £ while the most profit allocation had an RRP of 167.9£. Again, this difference is all imputed to Farfetch, given the price paid will be the same. This situation, as it can be seen from the comparison graph in figure 5.1, was not an outlier but a recurring problem.

The problem resulted from stock points also having to buy the brand products locally, meaning the price the items were sold to them by the brands was itself already subjected to the geo-price of the country.

The question left to answer was: how big could this problem be. To find an answer, an investigation was done into one of the main markets, the US market. On a given day, 54% of the orders in the United States are products that are duplicates. Around 46% of the duplicates are geo-priced. This means that, assuming a random distribution for the duplicates sold, 25% of the orders in the US are from geo-priced duplicates. From the duplicates that that are geo-priced, in 84% there is a price difference in RRP between stock points. The average difference between the stock point

selling at the highest RRP and the one selling at the lowest is 60.89£. In other words, making the same assumption, one would expect that 21% of the orders are geo-priced duplicates with an average RRP amplitude of 60.89 £.

Given the origin of the problem, in which stock points buy merchandise locally, and the fact that the duplicates algorithm prioritizes local stock points, the implications are profound.

Furthermore, analyzing the figure 5.1, more gains can be expected, although not so prominent as the ones related to differences in the RRP. These smaller gains were also inspected resorting to the decision matrices. They came from allocations that resulted in reduced shipping costs, fiscal gains when paying duties compensated not paying VAT and when even though the costs were slightly higher, the negotiated commission compensated that.

Seeing the presented algorithm had a design flaw, the results were immediately presented to the stakeholders who agreed there was a problem and an A/B test was requested. However, to construct the decision matrix the same way it was made in the simulation, there was a need to develop technology for that, so the implementation was not immediate, but the solution is in development.

Lastly, the savings expected from the algorithm implementation were estimated to be around 2% of gross profit.

5.2 2nd Iteration: Customer Experience

In the second iteration, lead time variables were added to the matrix. With that, several impacts could be tested. First, the impact on the lead time for the customer using the profit per order maximization strategy could be measured. Secondly, it could be evaluated the cost of implementing a lead time minimization strategy. Lastly and most important, it allowed for the testing of the first basic profitability strategy, which was the goal of the thesis.

5.2.1 Minimizing Costs

To start, the simulator was run using the as-is strategy so that a baseline for the lead time could be created. After, the cost minimization strategy was run again. The results can be seen in table 5.5.

Table 5.5: KPIs result from using a Farfetch cost minimization strategy.

KPIs	Avg. Shipping Costs	Avg. RRP	Avg. Duties	Avg. VAT	Avg. Cost	Avg. Commission	Avg. Profit per Order	Avg. LT
Minimization	+1.0%	-2.2%	-3.2%	+2.7%	-2.0%	+0.9%	+9.0%	-0.2%

While maximizing the profit per order, no significative differences are found in the lead times. Still, the question lingers, how much would cost the company, in profit per order, to minimize the lead time to the customer and provide a better experience.

5.2.2 Minimizing Lead Time

A lead time minimization strategy was tested and the results can be seen in table 5.6. The reduction in average lead time was around 0.36 days or 8.6h. However, this was achieved through a profit per order reduction of 6.78% when comparing to the as-is allocation. Comparing against the profit per order maximization, the difference was -14.4%. The question left to be answered was, would this reduction in profit be worth it.

Table 5.6: KPIs result from using a Lead Time minimization strategy.

KPIs	Avg. Shipping Costs	Avg. RRP	Avg. Duties	Avg. VAT	Avg. Cost	Avg. Commission	Avg. Profit per Order	Avg. LT
Lead Time Minimziation	+1.62%	+1.57%	+19.49%	-5.87%	2.58%	-0.24%	-6.78%	-7.79%

To provide the answer to the question, a new variable had to be added. The ΔCLV, which allows for the direct comparison of profit per order and lead time. This was possible using the profitability formula developed. A new strategy was then tried, a profitability per order maximization strategy, where profitability per order is defined in 5.2.

$$Profitability = PricePaid - FarfetchCost + \Delta CLV \tag{5.2}$$

Defining profitability allows for the incorporation of the impact of short-term decisions on long-term profit and paves the way for the integration of a holistic profitability measurement. Every variable that affects retention can be controlled through the predicted change in CLV while the variables that directly represent costs, can be controlled through the costs themselves.

The result of applying a profitability maximization strategy can be seen in table 5.7.

Table 5.7: KPIs result from using a Profitability maximization strategy.

KPIs	Avg. Shipping Costs	Avg. RRP	Avg. Duties	Avg. VAT	Avg. Cost	Avg. Commission	Avg. Profit per Order	Avg. LT
Profitability Maximization	+1.62%	-2.15%	-3.62%	+2.93%	-2.01%	+0.88%	+8.57%	-1.08%

The conclusions are fairly interesting. Employing a strategy completely focused on customer experience is not worth it since the value customers put in the reduced lead time does not compensate for the investment needed. This is confirmed because when profitability is considered, the average lead time is higher than when optimizing only for the lead time.

Additionally, the profitability maximization strategy indicates that maximizing only short-term profit neglects opportunities for a better customer experience. The reason is that sometimes it is worth to invest some short-term profit in bettering the customer experience which will pay off in future purchases.

5.3 3rd Iteration: Partners Potential

In the third iteration, the rate of offline sales was added to the matrix. The purpose was to see how much could potentially be gained if a strategy to prevent offline sales was used. Understanding how such a strategy could work is better done using a life-like analogy. Imagine there are 2 glasses with 25cl of water each and a third, empty, with a capacity of 50cl.

An individual has a small recipient to transport the water and is given the task to fill the empty cup with as much water as possible using the water from the other two cups. It would not matter from which cup he drew water first, in the end, the empty recipient would have 50cl of water and the other two would be empty. However, imagine that one of the cups had a small hole in the bottom that would start dripping water as soon as the exercise to fill the empty one began. In such a situation, if the purpose was to fill the empty cup with as much water as possible, the subject best strategy would be to first extract all the water from the dripping cup and then extract the water from the cup with no hole. That way, the cup would be filled with the most water. The reason would be that the amount of water dripping would be minimized.

Making the bridge to the real situation, the stock would be the water and the way to put the most stock in the system is to allocate the orders first to the stock points that are selling more offline, that is, dripping the most.

The first step was to define how the impact would be measured and it was decided that using the stock levels in the system was the best solution. If offline sales were to be prevented, then the result would be an increase in the available stock per day.

Nevertheless, a problem with the results should be kept in mind, the offline sales data is just an estimation and is not being predicted. This means that the offline sales that are used in the simulation are the ones that happen in reality. This is a problem since when allocation changes, the stock will be available in stock points when it was not available before. Hence, new offline sales could have happened during that time period that are not being accounted for. Still, since the stock will be available in stock points that sell less offline, the amount of product sold offline should be less than it would be if it was available in fast selling stock points.

To explore if such a strategy would be worth, the simulator was run using a ROS maximization strategy, that is, allocating the orders to the stock points that sold the most stock, in a given week, offline. Table 5.8 presents the comparison results of using such a strategy against the profitability maximization strategy. The decrease in profit per order was around 13%.

Table 5.8: KPIs result from using a RoS optimization strategy.

KPIs	Avg. Shipping Costs	Avg. RRP	Avg. Duties	Avg. VAT	Avg. Cost	Avg. Commission	Avg. Profit per Order	Avg. LT
ROS Maximization	-0.55%	+3.62%	+21.35%	-10.95%	+4.01%	-1.54%	-13.12%	+1.31%

However, to evaluate the merit of the strategy stock levels information is needed. This information is provided in figure 5.2 that shows the stock evolution when using each strategy.

Arriving at conclusions with this results is not easy. First and foremost, the stock gains we see in figure 5.2, would not translate directly into sales. As said before, the offline sales were not modeled, meaning some of that stock could be capture in offline sales.

Secondly, a way to measure the trade-off between the investment done in profit per order and the increase in possible sales is needed. To do this, a goal-seek approach was used. The goal was to establish the number of sales needed so that it would be worth to use such a strategy.

Comparing the profitability strategy against the ROS strategy, the profit reduction, for the time of the study, is around 10%. Knowing the average profit per order when using the ROS strategy,

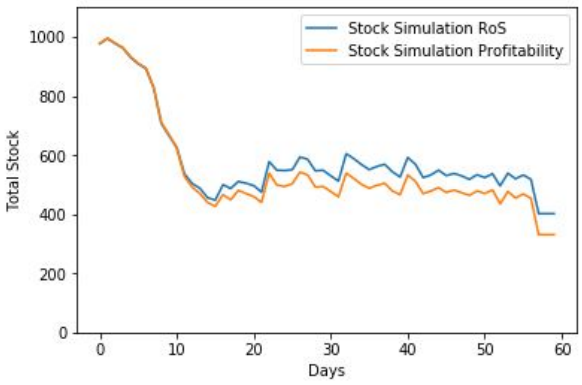

Figure 5.2: Stock Evolution using the profitability maximization strategy vs the ROS optimization strategy

around 340 more orders would have to exist for the strategy to be successful.

Comparing the stock graphic, for the time period of this study, the increase in stock available was of 71 units, well below the 340 units needed for the strategy to be worth it. This lead to the conclusion that implementing a ROS optimization strategy, for the products selected at least, would not be worth it.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This work addressed the problem of order allocation for an online marketplace. The problem exists when more than one seller is capable of fulfilling the order and the company needs to decide which one should do it. The project had the goal of developing an algorithm to improve this allocation decision, which was met with the implementation of such a system at e-tail merchant Farfetch.

During the development of this project, much was learned regarding which variables affect customer experience and the impact that experience has on repeated purchases. It allowed for the development of a holistic algorithm where both short and long-term profits could be put in the same equation, discarding the need for a multi-objective function.

To find a better solution, a simulator was developed where different allocation strategies could be tested before being implemented in the company. The development of this simulation was done in three different phases, or iterations, with each adding further complexity and testing new scenarios.

In all the iterations done, new knowledge was acquired. Starting with the parameterization of all variables related to direct costs which was a huge step forward in the decision evaluation. It allowed not only for the comparison of different alternatives, but it also gave visibility to the allocation decision that did not exist before. The change from an empirical to a mathematical decision would already had been a great improvement for the decision making. However, the gains of costs modeling did not end there. It was found that the allocation algorithm used was neglecting recent business changes in pricing which resulted in higher costs for the company. With the implementation of profit maximization algorithm to replace the current solution, which is happening at the moment, the forecasted gains are 2% in gross profit.

Furthermore, the results showed the need to save, for each order, the stock points that could have fulfilled it and why they were not selected. This would allow for the creation of an even more accurate simulator, since all the information needed would already be stored, and different strategies could be tried. Seeing that, at the moment of writing, such a storage solution is also being developed.

In the second iteration, it was discovered that the original line of thought, which focused on customer experience with little regard for the costs, was not the best solution. Although the empirical optimization of reducing lead time was not completely working - the lead times could be further reduced -, it was proven that this was not the best strategy for the company. The cost of minimizing lead time by a few hours, on average, reduced the profit per order in more than 14% when compared to the most profitable allocation. Using the knowledge provided by the PHM model, it is now known that a reduction of 1 day in lead time originates a profit of 2.36£, meaning that a strategy to minimize lead time at all costs is not worth it.

Moreover, a new concept was introduced to the allocation decision, the maximization of profitability. Considering profitability and all it encompasses provided a new way to look at the decision. It was no longer a multi-objective problem where different perspectives needed to be considered. It also ends an emotional discussion among stakeholders about what should be the focus of the company, minimizing lead times or maximizing profits. This approach paved the way for the integration of all variables related to customer experience with all the direct costs. Although during the development of this work, regarding customer experience, only lead time was evaluated, the methodology developed is adaptable to all the other variables mention and not studied. The only thing needed is to calculate how much a certain factor affects retention - which can be done using the already developed PHM methodology - and then assess how much that retention is worth - through the CLV.

Profitability, as the most holistic strategy employed, should be used in future iterations needing only the inclusion of all the variables that affect the allocation decision.

Additionally, the last iteration attempted to increase the contribution the partners have in the business through the increase of their stock available for the marketplace. Although the conclusions were not positive, since the costs would exceed the profits, this may not be true to all the products. The intuition that exists is that for products where the profits are higher and tend to be sold out early in the season, this kind of strategy could be interesting. Nevertheless, for such a strategy to be implemented, a forecasting tool capable of predicting future sales for each product would be needed, which was beyond the scope and time available for this thesis.

Lastly, the use of lean methodology contributed greatly to the success of this work. It allowed for iterations in the solution and to deliver value to the organization in a continuous way, as well as permitting the further development of the work by others. It would only be needed to model new variables.

6.2 Future Work

A great deal was accomplished in the development of this work but a lot of stones were left unturned as well. This section will provide a direction for future work that can be built using this project as its base.

First of all, some variables were not studied because of time constraints, such as no-stock events, returns, and overheads. Predictive models need to be created for those, as well as measurements of the impact of those events in retention. The integration of those in the simulation will be fairly simple, just by adding the models and keep implementing the profitability maximization strategy.

Secondly, the PHM model created is a model for the clients in general. However, different types of clients exist in the company, so different retention models should be created. This could be done using clustering techniques to create different profiles and then create retention models for each of the clusters.

Lastly, all the orders were considered to only have one product and studied independently. However, synergies may exist from shipping from the same stock point when an order has more than one product and those are available in the same stock point. This effects should be studied since gains could arise from them.

Bibliography

Antonio Achille, Sophie Marchessou and Nathalie Remy. Luxury in the age of digital Darwinism | McKinsey & Company, 2018. URL https://www.mckinsey.com/industries/retail/our-insights/luxury-in-the-age-of-digital-darwinism.

Paul D. Berger and Nada I. Nasr. Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*, 12(1):17–30, jan 1998. ISSN 10949968. doi: 10.1002/(SICI)1520-6653(199824)12:1<17::AID-DIR3>3.0.CO;2-K. URL https://www.sciencedirect.com/science/article/pii/S1094996898702506.

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-Dm 1.0. *CRISP-DM Consortium*, page 76, 2000. ISSN 0957-4174. doi: 10.1109/ICETET.2008.239.

Martin Christopher. Logistics and competitive strategy. *European Management Journal*, 11(2): 258–261, 1993. ISSN 02632373. doi: 10.1016/0263-2373(93)90049-N.

Marc-André Kamel Claudia D'Arpizio, Federica Levato and Joëlle de Montgolfier. Luxury Goods Worldwide Market Study, Fall–Winter 2017 - Bain & Company, 2017. URL http://www.bain.com/publications/articles/luxury-goods-worldwide-market-study-fall-winter-2017.aspx.

Joel E. Collier and Carol C. Bienstock. Measuring Service Quality in E-Retailing. *Journal of Service Research*, 8(3):260–275, feb 2006. ISSN 1094-6705. doi: 10.1177/1094670505278867. URL http://journals.sagepub.com/doi/10.1177/1094670505278867.

D R Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. URL http://links.jstor.org/sici?sici=0035-9246{%}281972{%}2934{%}3A2{%}3C187{%}3ARMAL{%}3E2.0.CO{%}3B2-6.

Paul Dierckx. Curve and Surface Fitting with Splines. *Curve and Surface Fitting with Splines*, page 308, 1993. ISSN 02724960. doi: 10.1093/imamat/1.2.164. URL https://dl.acm.org/citation.cfm?id=151103.

Xavier Drèze and André Bonfrer. Moving from customer lifetime value to customer equity. *Quantitative Marketing and Economics*, 7(3):289–320, sep 2009. ISSN 1570-7156. doi: 10.1007/s11129-009-9067-y. URL http://link.springer.com/10.1007/s11129-009-9067-y.

Richard E. DuWors, George H. Haines, and Jr. Event History Analysis Measures of Brand Loyalty. *Journal of Marketing Research*, 27(4):485, nov 1990. ISSN 00222437. doi: 10.2307/3172633. URL https://www.jstor.org/stable/3172633?origin=crossref.

Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with B -splines and penalties. *Statistical Science*, 11(2):89–121, may 1996. ISSN 0883-4237. doi: 10.1214/ss/1038425655. URL http://projecteuclid.org/euclid.ss/1038425655.

Frederick F. Reichheld and Sasser Jr. W. Earl. Zero Defections: Quality Comes to Services, 1990. URL https://hbr.org/1990/09/zero-defections-quality-comes-to-services.

Claes Fornell, Roland Rust, and Marnik Dekimpe. The Effect of Customer Satisfaction on Consumer Spending Growth. *Journal of marketing research*, 41(1):28–35, 2010. URL https://lirias.kuleuven.be/bitstream/123456789/205822/2/Theeffectofcustomer.pdf.

Jerome H. Friedman and Bernard W. Silverman. Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics*, 31(1):3–21, feb 1989. ISSN 0040-1706. doi: 10.1080/00401706.1989.10488470. URL http://www.tandfonline.com/doi/abs/10.1080/00401706.1989.10488470.

Sunil Gupta and Donald R Lehmann. Valuing Customers. *Journal of Marketing Research*, 41(1):7–18, 2004. URL http://www.jstor.org/stable/30162308http://www.jstor.org/stable/30162308.

Sunil Gupta, Dominique Hanssens, Bruce Hardie, Wiliam Kahn, V. Kumar, Nathaniel Lin, Nalini Ravishanker, and S. Sriram. Modeling Customer Lifetime Value. *Journal of Service Research*, 9(2):139–155, nov 2006. ISSN 1094-6705. doi: 10.1177/1094670506293810. URL http://journals.sagepub.com/doi/10.1177/1094670506293810.

Whang Hau L. Lee and Seungjin. Winning the Last Mile of E-Commerce, 2001. URL https://sloanreview.mit.edu/article/winning-the-last-mile-of-ecommerce/.

Gregory R. Heim and Kingshuk K. Sinha. Operational Drivers of Customer Loyalty in Electronic Retailing: An Empirical Analysis of Electronic Food Retailers. *Manufacturing & Service Operations Management*, 3(3):264–271, 2001. ISSN 1523-4614. doi: 10.1287/msom.3.3.264.9890. URL http://pubsonline.informs.org/doi/abs/10.1287/msom.3.3.264.9890.

Christian Homburg, Nicole Koschate, and Wayne D. Hoyer. Do Satisfied Customers Really Pay More? A Study of the Relationship Between Customer Satisfaction and Willingness to Pay. *Journal of Marketing*, 69(2):84–96, 2005. ISSN 0022-2429. doi: 10.1509/jmkg.69.2.84.60760. URL http://journals.ama.org/doi/abs/10.1509/jmkg.69.2.84.60760.

George Q. Huang, Jason S. K. Lau, and K. L. Mak. The impacts of sharing production information on supply chain dynamics: A review of the literature. *International Journal of Production Research*, 41(7):1483–1517, jan 2003. ISSN 0020-7543. doi: 10.1080/0020754031000069625. URL http://www.tandfonline.com/doi/abs/10.1080/0020754031000069625.

Barbara Bund Jackson. Winning & Keeping Industrial Customers: The Dynamics of Customer Relations. *Journal of Marketing*, 52(January):148–150, 1988. ISSN 00222429. doi: 10.2307/1251693. URL https://books.google.pt/books/about/Winning{_}and{_}keeping{_}industrial{_}customers.html?id=BvwTAQAAMAAJ{&}redir{_}esc=y.

Jennifer Schmidt, Karel Dörner, Achim Berg, Thomas Schumacher and Katrin Bockholdt. The opportunity in online luxury fashion | McKinsey & Company, 2015. URL https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-opportunity-in-online-luxury-fashion.

John D Kalbfleisch and R L Prentice. *The statistical analysis of failure time data*, volume 10. Wiley-Blackwell, mar 1980. ISBN 0-471-05519-0. doi: 10.2307/3315078. URL http://doi.wiley.com/10.2307/3315078.

Lisa Su Kate McCarthy, Ben Perkins, Nick Pope, Linda Portaluppi, Valeria Scaramuzzi. Global Powers of Luxury Goods 2017 The new luxury consumer 2 Global Powers of Luxury Goods 2017, 2017. URL https://www2.deloitte.com/content/dam/Deloitte/global/Documents/consumer-industrial-products/gx-cip-global-powers-luxury-2017.pdf.

Philip. Kotler, Kevin Lane Keller, Delphine Manceau, and Bernard Dubois. *Marketing Management*. Prentice Hall of India, 2009. ISBN 5 NIV 658.83 KOTL. doi: 10.1080/08911760903022556. URL https://books.google.pt/books/about/Marketing{_}management.html?id=eSUPAQAAMAAJ{&}redir{_}esc=y.

Shun Y. Lam, Venkatesh Shankar, M. Krishna Erramilli, and Bvsan Murthy. Customer Value, Satisfaction, Loyalty, and Switching Costs: An Illustration From a Business-to-Business Service Context. *Journal of the Academy of Marketing Science*, 32(3):293–311, jul 2004. ISSN 00920703. doi: 10.1177/0092070304263330. URL http://link.springer.com/10.1177/0092070304263330.

Jeffrey E. Lewin. Business customers' satisfaction: What happens when suppliers downsize? *Industrial Marketing Management*, 38(3):283–299, apr 2009. ISSN 0019-8501. doi: 10.1016/J.INDMARMAN.2007.11.005. URL https://www.sciencedirect.com/science/article/abs/pii/S0019850108000072.

Linda Dauriz, Nathalie Remy and Nicola Sandri. Luxury shopping in the digital age | McKinsey & Company, 2014. URL https://www.mckinsey.com/industries/retail/our-insights/luxury-shopping-in-the-digital-age{#}0.

Chang Liu, Kirk P Arnett, and Chuck Litecky. Design quality of websites for electronic commerce: Fortune 1000 webmasters' evaluations. *Electronic Markets*, 10(2):120–129, 2000. ISSN 1019-6781. doi: 10.1080/10196780050138173. URL https://pdfs.semanticscholar.org/a968/7de098fc1b816d8ffa5e7a19a3758af05d44.pdfhttp://www.tandfonline.com/doi/abs/10.1080/10196780050138173.

Vikas Mittal and Wagner A. Kamakura. Satisfaction, Repurchase Intent, and Repurchase Behavior: Investigating the Moderating Effect of Customer Characteristics. *Journal of Marketing Research*, 38(1):131–142, feb 2001. ISSN 0022-2437. doi: 10.1509/jmkr.38.1.131.18832. URL https://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=2344925http://journals.ama.org/doi/abs/10.1509/jmkr.38.1.131.18832.

C.J. Newton. e-Fulfillment does not require owning your own distribution center, 2000. URL http://www.webshippers.com/newart.htm.

R.L. Oliver. *Satisfaction: A Behavioral Perspective on the Consumer*, volume 14. McGraw Hill, 1997. ISBN 0765617706. doi: 10.1139/

h11-134. URL https://books.google.pt/books/about/Satisfaction.html?id=iCeQQgAACAAJ{&}redir{_}esc=yhttp://www.amazon.com/Satisfaction-Behavioral-Perspective-McGraw-Hill-marketing/dp/0071154124.

Finbarr O'Sullivan. Nonparametric Estimation of Relative Risk Using Splines and Cross-Validation. *SIAM Journal on Scientific and Statistical Computing*, 9(3):531–542, may 1988. ISSN 0196-5204. doi: 10.1137/0909035. URL http://epubs.siam.org/doi/10.1137/0909035.

A. Parasuraman, Valarie A. Zeithaml, and Arvind Malhotra. E-S-QUAL. *Journal of Service Research*, 7(3):213–233, feb 2005. ISSN 1094-6705. doi: 10.1177/1094670504271156. URL http://journals.sagepub.com/doi/10.1177/1094670504271156.

Michael Pastore. E-Commerce Faces Logistics Nightmare - ClickZ, 1999. URL https://www.clickz.com/e-commerce-faces-logistics-nightmare/78423/.

Phillip E. Pfeifer, Mark E. Haskins, and Robert M. Conroy. Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending, 2005. URL https://www.jstor.org/stable/40604472.

Mary (Mary B.) Poppendieck and Thomas David. Poppendieck. Lean software development: an agile toolkit [Book Review]. *Computer*, 36(8):89–89, 2003. ISSN 0018-9162. doi: 10.1109/MC.2003.1220585. URL http://ieeexplore.ieee.org/document/1220585/.

P. Pyke, D.F., Johnson, M.E., Desmond. E-fulfillment: it's harder than it looks. *Supply Chain Management Review 5*, 2001. URL http://digitalstrategies.tuck.dartmouth.edu/wp-content/uploads/2016/10/Press{_}SCMR{_}eFulfillment.pdf.

J. R. and Carl de Boor. A Practical Guide to Splines. *Mathematics of Computation*, 34(149):325, 1980. ISSN 00255718. doi: 10.2307/2006241. URL http://www.jstor.org/stable/2006241?origin=crossref.

P. B. Seetharaman and Pradeep K. Chintagunta. The proportional hazard model for purchase timing: A comparison of alternative specifications. *Journal of Business and Economic Statistics*, 21(3):368–382, 2003. ISSN 07350015. doi: 10.1198/073500103288619025.

Tianxiang Sheng and Chunlin Liu. An empirical study on the effect of e-service quality on online customer satisfaction and loyalty. *Nankai Business Review International*, 1(3):273–283, jul 2010. ISSN 2040-8749. doi: 10.1108/20408741011069205. URL http://www.emeraldinsight.com/doi/10.1108/20408741011069205.

J. Shenton. e-Business brings e-Fulfillment. URL http://www.globalmillenniamarketing.com/article{_}efulfillment{_}ecommerce{_}ebusiness.htm.

Jayashankar M Swaminathan, Stephen F Smith, and Norman M Sadeh. Modeling the Dynamics of Supply Chains. 1995. URL https://www.ri.cmu.edu/pub{_}files/pub2/swaminathan{_}j{_}m{_}1994{_}1/swaminathan{_}j{_}m{_}1994{_}1.pdf.

Antuela A. Tako and Stewart Robinson. The application of discrete event simulation and system dynamics in the logistics and supply chain context. *Decision Support Systems*, 52(4):802–815, 2012. ISSN 01679236. doi: 10.1016/j.dss.2011.11.015. URL http://dx.doi.org/10.1016/j.dss.2011.11.015.

Jacquelyn S. Thomas, Robert C. Blattberg, and Edward J. Fox. Recapturing Lost Customers. *Journal of Marketing Research*, 41(1):31–45, feb 2004. ISSN 0022-2437. doi: 10.1509/jmkr.41.1.31.25086. URL http://journals.ama.org/doi/abs/10.1509/jmkr.41.1.31.25086.

Appendix A

Proportional Hazard Model

The proportional hazard model (phm) is used to study purchase-timing behavior where a baseline hazard is established, that captures the clients intrinsic buying behavior and allows for the quantification of the effect of other buying variables in that purchase buying behavior.

The model is formulated by specify the hazard rate ($h_i(t;X_t)$) as a function of the baseline hazard rate($h_i(t)$) and covariates (X),

$$h_i(t;X_t) = h_i(t) * exp(\beta_i X_i) \tag{A.1}$$

The hazard function can also be written as, dropping the subscript, *i*, for notational ease:

$$h(t;X_t) = \frac{f(t;X_t)}{1-F(t;X_t)} \tag{A.2}$$

where $f(t;X_t)$ stands for the probability density function (pdf) corresponding to the client *i* hazard function at time t and $1-F(t;X_t)$ stands for the corresponding cumulative distribution function (cdf), as well as the survivor function, that is, the probability the client *i* "survives" a purchase until time t, represented by $S(t;X_t)$.

Substituting equation A.2 on equation A.1 and using the survivor function notation, one obtains the following equality:

$$f(t;X_t) = h(t) * exp(\beta X_t) * S(t;X_t) \tag{A.3}$$

One can further simplify equation A.3 to a more estimation friendly form as :

$$\frac{dF(t;X_t)}{1-F(t;X_t)} = h(t) * exp(\beta X_t) * dt \tag{A.4}$$

Which is a first-order differential equation and can be solved as:

$$\int_0^{F(t;X_t)} \frac{dF(u;X_u)}{1-F(u;X_u)} = \int_0^t h(u) * exp(\beta X_u) * du \tag{A.5}$$

where the lower limit of integration (i.e., 0) corresponds to the time of the client's previous purchase. Solving this yields:

$$F(t;X_t) = 1 - exp(-\int_0^t \beta X_u * du) \tag{A.6}$$

or, if one wants the survival function:

$$S(t;X_t) = exp(-\int_0^t \beta X_u * du) \tag{A.7}$$

Replacing from equation A.7 on equation A.3 one obtains the estimable version of the PHM:

$$f(t;X_t) = h(t) * exp(\beta X_t) * exp- \int_0^t \beta X_u * du \tag{A.8}$$

And using the following likelihood function one can estimate the PHM parameters, assuming a set of n interpurchase time observations, say, $(t_1, t_2, t_3, ..., t_n)$ with each of these interpurchase times is associated with a covariate vector, yielding a stacked set of covariate vectors given by $(X_1, X_2, X_3, ..., X_n)$:

$$L = \prod_{j=1}^{n} f(t_j - t_{j-1}, X_j) \tag{A.9}$$

A.1 PHM at Farfetch

After explaining how the PHM model can be trained, one can jump directly to the results obtained in this project when such a model was used. Some details will be omitted for the sake of simplicity.

It should be noted that the PHM is a failure model, where the failings, in this case, are the purchases being modeled. As such, a positive outcome is when a failure occurs, that is, the client doesn't survive in the survival model.

Once a model is trained, having the covariates values and using equation A.7 a continuous time surviving curve for the continuous retention model can be plotted, as can be seen in figure A.1. The vector X_t used was populated with the average lead time.

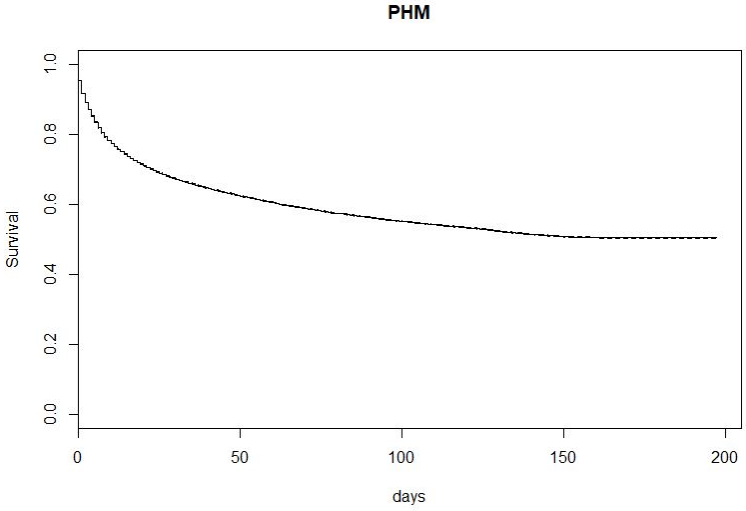

Figure A.1: Retention model for the average customer.

The usefulness of the PHM is related to its ability, by using the β values from the covariates, to estimate the increase in retention of a covariate if all the other covariates were kept the same. For the trained model, the covariate used was the lead time since that was the purpose the tool was constructed. The result for the model covariate can be seen in table A.1. The modeling quality

Table A.1: Values for the Lead time as a covariate in the model.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
leadtime	-0.018912	0.981265	0.002219	-8.525	<2e-16

was evaluated through a Likelihood ratio test that gave a p-value of 0%, affirming the significance of the model.

The results can be interpreted as follow:

1. Statistical significance. The column marked "z" gives the Wald statistic value. It corresponds to the ratio of each regression coefficient to its standard error (z = coef/se(coef)). The Wald statistic evaluates, whether the beta (β) coefficient of a given variable is statistically significantly different from 0. From the output above, we can conclude that the variable lead time has highly statistically significant coefficients.

2. The regression coefficients. The second feature to note in the model results is the sign of the regression coefficients (coef). A positive sign means that the hazard (risk of buying) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The beta coefficient for lead time = -0.018912 indicates that higher lead times have higher survival rates, and since in this case, surviving is not purchasing again lower retention.

3. Hazard ratios. The exponential coefficients (exp(coef) = exp(-0.018912) = 0.981265), also known as hazard ratios, give the effect size of covariates. For example, an increase in the lead time of 1 day reduces the hazard by a factor of 0.98, or 2%.

This effect can be seen in figure A.2, where different values of lead time are given to the covariate. As the lead time increases, a customer surviving chances, that is, the probability of not buying again, increase, by 2%. In other words, the retention rate decreases by 2% with the increase in lead time since $Retention = 1 - Survival$.

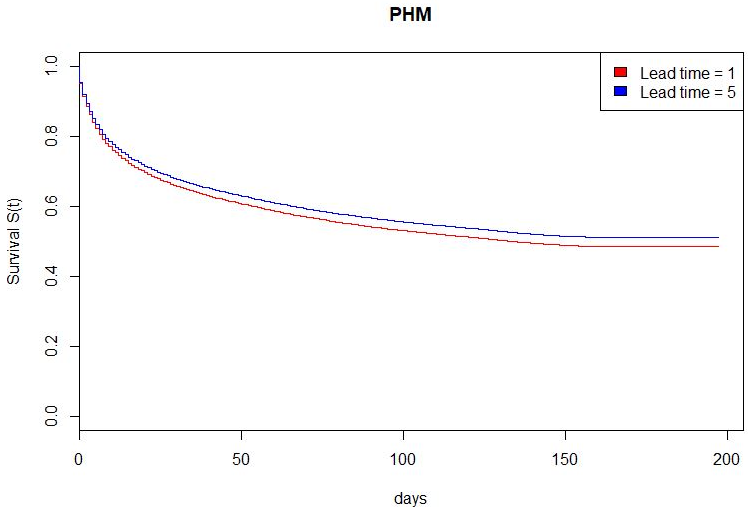

Figure A.2: Survival model lead time comparison.

Appendix B

Weights and Shipping Cost

The regression models tried involved three parameters available with each order, the box weight, the box used and the shipping cost. The objective was to see if a simple regression model would be able to explain the majority of the shipping costs in a order.

The first regression tried was using the box selected - Farfetch has 10 boxes available for the boutiques to choose from - to predict the shipping cost, for the DHL express shipping service. The results can be seen in figure B.1.

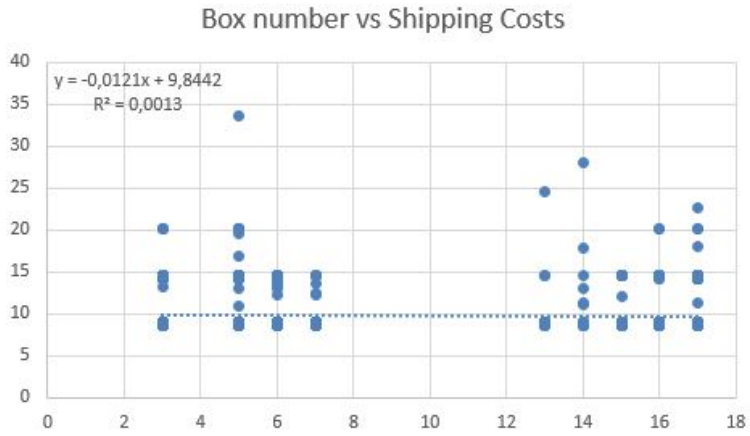

Figure B.1: Box selected vs Shipping Costs.

The results were poor, with the regression having a $r^2 = 0.0013$. Next a regression using the box weight measured by DHL to predict the shipping cost, for the DHL express shipping service was done. The results can be seen in figure B.2.

Again, the results were poor, with the regression having a $r^2 = 0.0304$.

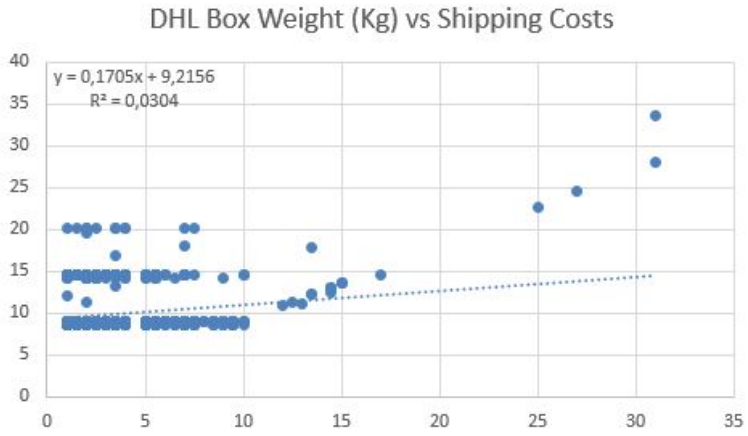

Figure B.2: Box weight vs Shipping Costs.

Appendix C

DHL Approximation

DHL shipping cost tool was used to estimate all the shipping costs except for purchases made between boutiques and clients both located in the USA. To ensure the estimation was good enough some orders that were shipped through UPS were selected randomly and the costs estimated with the DHL tool. Afterward, the results were compared using a Welch's t-test, which has the following hypothesis:

H_0 : Two samples have equal means.

H_a : Two samples don't have equal means.

The test had a $tstat = -0.9$ and $p - value = 0.36$ validating the use of the DHL tool.

Appendix D

Zones and zip codes study

The courier UPS uses a zoning system to define, together with the package volume, how much to charge to ship a package. However, estimating the zones, in the united states at least, is not an easy task since the courier didn't provide the company a tool to calculate the zones.

Mostly, the zones in the USA depend on the states to where the package will be shipped from and to. Nevertheless, for the same states, prices may vary depending on where the two points are located. This means there is a need to use postal codes to estimate the zone correctly.

The way the postal codes, composed of 6 numbers, are distributed in the USA is that, for each region, the first 2 to 3 numbers are the same and the last 3 or 4 numbers identify the exact location, like in figure D.1.

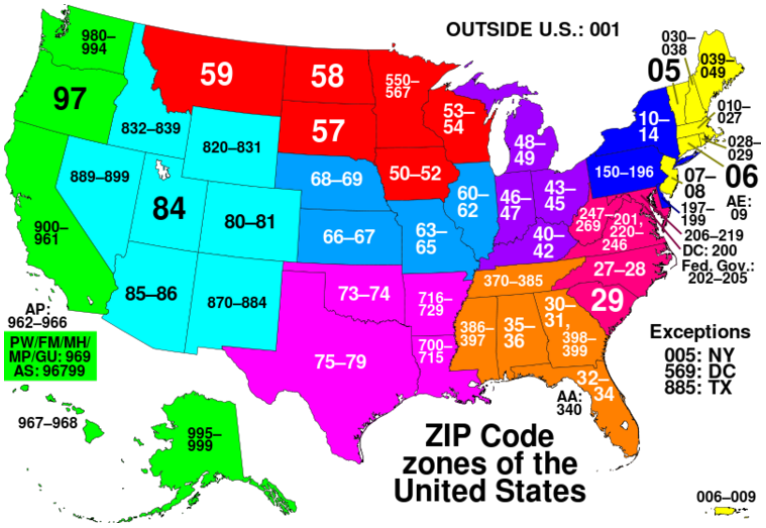

Figure D.1: Zip codes distribution in USA. The first 2 or three numbers can be used to identify the State. Source: http://feellikeyoubelong.com/blog/2018/3/1/zip-codes

What was required was a tool similar to table D.1, that had all the zip codes combinations and respective zone available. That way, when the simulator needed to calculate the zone, a simple search operation would be conducted.

However, such a table had a slight inconvenience. The number of combinations needed was in the millions range, which would require to much memory, so a little optimization was needed.

Knowing that close locations tended to share similar zones and that the closer the location more similar would be the zip code sequence a study was conducted. In it, only the first two zip

Table D.1: UPS Zones and Postal codes

Origin	Destination	Zone
560215	280125	001
543248	630251	002

numbers for each location were used and calculated how many zones existed per combination, resulting in a table similar to table D.2.

Table D.2: UPS Zones and Postal codes

Origin	Destination	Number of Zones
56	28	2
54	63	1

In figure D.2, a full overview of the number of zones that exist when zip codes are grouped based on the first two zip code numbers are shown.

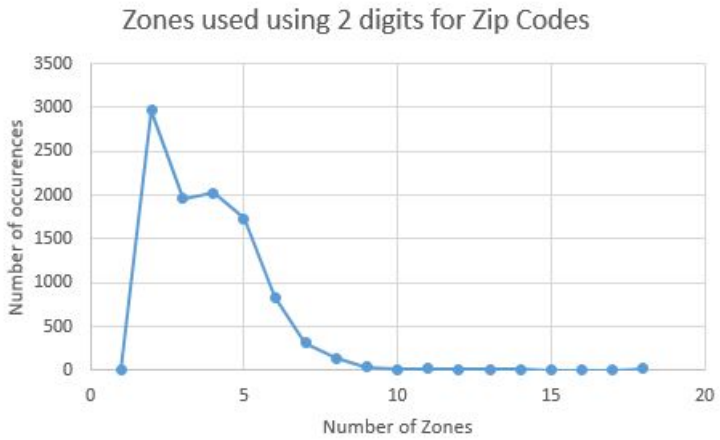

Figure D.2: Zones used when only the first two numbers of the zip codes are used. The average number of zones per combination is 2.8 zones.

Using only the first two numbers of the zip codes greatly reduces the number of combinations, from millions to around 10000. However, the trade-off doesn't compensate since per combination almost 3 zones exist, which gives room for improvement seeing the number of combinations to store in memory are few.

As such, next the combinations using the first three zip numbers, were tried, like in table D.2.

The results can be seen in figure D.3. The average number of zones per combination was reduced almost to a half, with 1.5 zones per combination and having around 80000 combinations. This result, given the trade-off, seemed satisfactory and so the first 3 numbers in the zip codes were used.

Table D.3: UPS Zones and Postal codes

Origin	Destination	Number of Zones
560	280	1
543	630	1

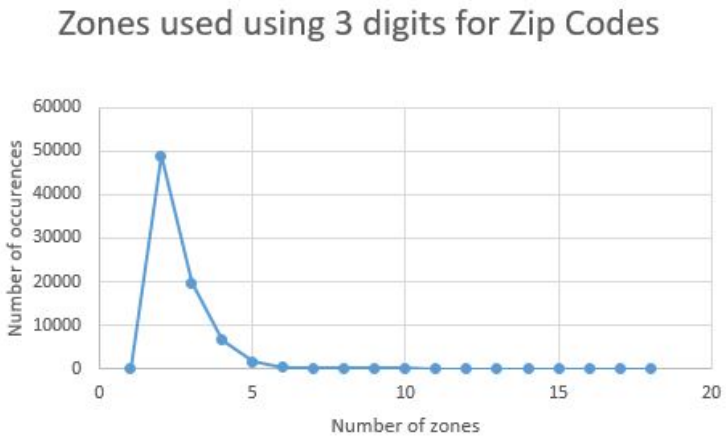

Figure D.3: Zones used when only the first three numbers of the zip codes are used. The average number of zones per combination is 1.5 zones.

Appendix E

Duties Calculation

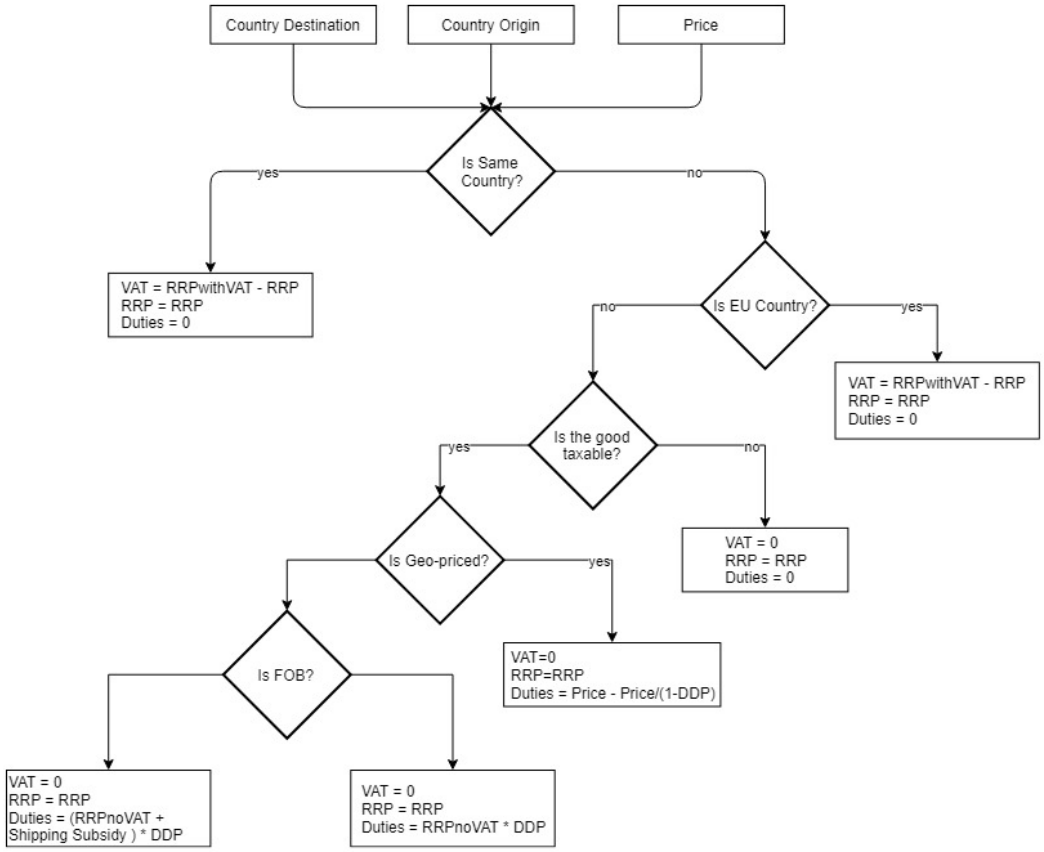

Figure E.1

Appendix F

Distributions Transformations and Fit

A series of transformations were tried to try to eliminate the depressions and skewness of the time in transit distribution, as it can be seen in figure F.1.

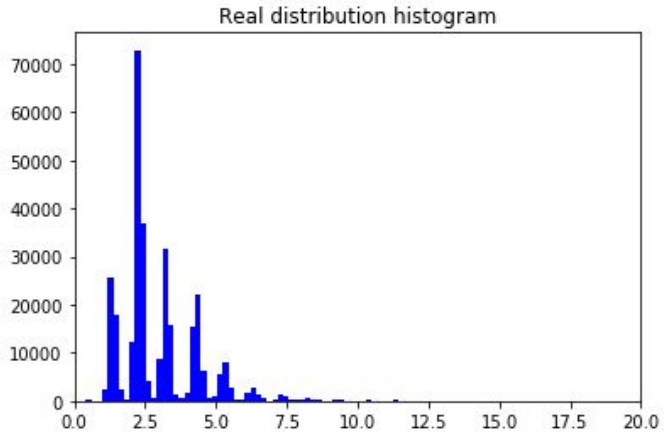

Figure F.1: Histogram with time in transit between Italy and the United States using DHL express service.

The Log transformation was the first to be tried. It is usually used when the data has a positively skewed distribution and there are a few larger values such as the one in this case. The transformation is done using the following conversion:

$$Y(s) = \ln(Z(s)), \tag{F.1}$$

The results from the transformation can be seen in figure F.2, with the approximation to a normal distribution being evaluated through a normal probability plot.

The deviations from the straight line in the normal probability plot are big enough to conclude immediately that a normal fit is not present and others transformations were tried.

The Box-Cox transformation was the transformation that was tried next. This transformation is done through the following conversion:

$$Y(s) = \frac{(Z(s)^{\lambda} - 1)}{\lambda}, \tag{F.2}$$

where $\lambda \neq 0$.

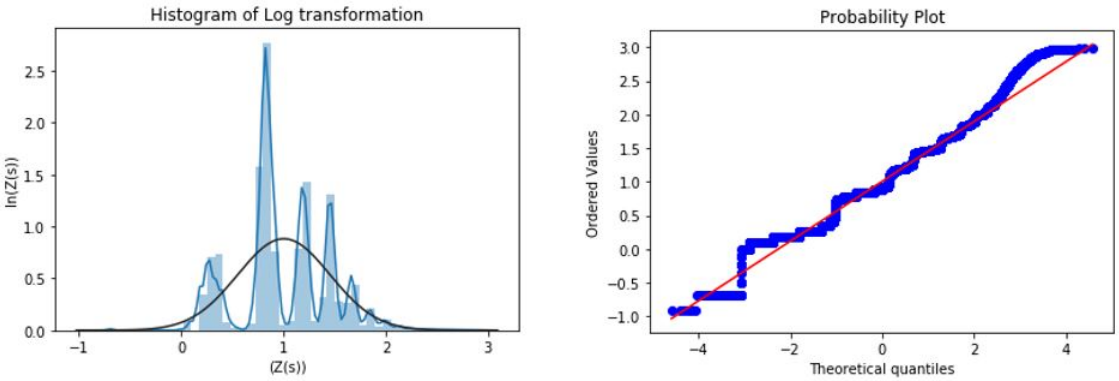

Figure F.2: Histogram with logarithmic transformation and respective normal probability plot.

This transformation was tried using the python package scipy with the value being chosen automatically that maximizes the log-likelihood function. The results can be seen in figure F.3.

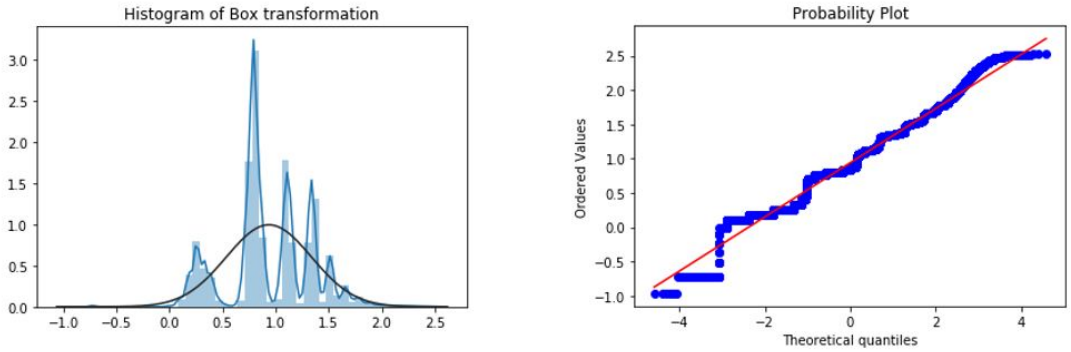

Figure F.3: Histogram with Box-Cox transformation and respective normal probability plot.

Again, the deviations from the straight line in the normal probability plot are big enough to conclude immediately that a normal fit is not present and one last transformation was tried.

The Robust Scaler was the last transformation that was tried. This transformation is a very similar method to the Min-Max scaler which makes it robust to outliers. It uses a interquartile range instead of the min-max, and follows the following formula:

$$Y(s) = \frac{x_i - Q_1(x))}{Q_3(x) - Q_1(x)}, \tag{F.3}$$

This transformation was tried using the python package scipy. The results can be seen in figure F.4.

Again, the deviations from the straight line in the normal probability plot are big enough to conclude immediately that a normal fit is not present and so other methods were tried.

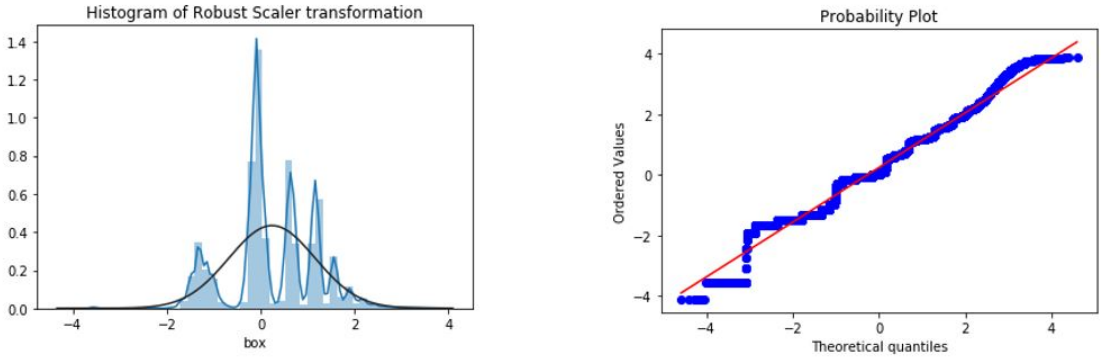

Figure F.4: Histogram with Robust Scaler transformation and respective normal probability plot.