LIBERA UNIVERSITÀ INTERNAZIONALE DEGLI STUDI SOCIALI

Department of Business and Management

Chair of Marketing Management

Implementing Machine Learning in Customer Experience: An empirical study on Kickstarter

SUPERVISOR
Professor Marco Francesco Mazzù

ASSISTANT SUPERVISOR
Professor Luigi Laura

CANDIDATE
Giulia Canciani Graziani
684371

Academic Year: 2017/2018

Abstract *(Italian)*

The following Master thesis has the main aim to analyse whether and how the features of Kickstarter projects influence the Customer Experience and individuals' preferences.

First, the Thesis provides a Literature Review that illustrates the development of Marketing theories on Customer Experience, previous research of Machine Learning applications in Marketing and Customer Experience in Kickstarter. Then, an introduction on Machine Learning provides the basic knowledge to understand what Machine Learning is and the value that its implementation may bring to companies and customers to improve the quality of *touchpoints.*

Relying on this introduction, the study examines how four features of Kickstarter projects – (1) *Main Category,* (2) G*oal*, (3) *Country* and (4) *Currency* – influence the Customer Experience, analysing how combinations of these variables affect projects' success rate.

First, each dependent variable is analysed separately and then used to build step-by-step a predictive model, made with three of the four independent variables – (1) *Main Category*, (2) G*oal*, and (3) *Country* – and based on a multivariate logistic regression. Then, the model is trained on 70% of the dataset, and then tested on the remaining 30%.

The Predictive Model has an accuracy of 61%, meaning that it is capable of predicting the outcome of a campaign 6 times out of 10.

The findings of the research prove that customers are more attracted towards projects that offer artistic contents, with realistic goals, and that they are influenced by the creators' country of origin.

Abstract *(Chinese)*

本文旨在分析Kickstarter项目的特征是否以及如何影响客户体验和个人偏好。

首先，文献综述阐述了有关客户体验的营销理论的发展以及Kickstarter项目中机器学习在营销和客户体验中应用的以往研究。接下来是对机器学习的概述，包括机器学习的基本知识，以及其可能为企业和客户带来的价值。

依据概述内容，本研究考察了Kickstarter项目的四个特征 - （1）主要类别，（2）目标，（3）国家和（4）货币 - 如何影响客户体验并分析这些变量的组合对项目成功率的影响

本文对每个因变量进行单独分析，然后将建立的四个模型在331.675个项目上进行测试。其中拟合度最好的模型是包含三个分类变量 （主要类别、国家和目标 ）的多变量逻辑回归模型。 该模型使用了70%的数据来建立，并在剩余的30%上进行测试。

预测模型的准确率为61%，这意味着每10次预测中有6次的结果是准确的。

研究结果表明，客户更倾向于具有艺术性且目标实际的项目，并且他们的偏好受创作者国籍的影响。

Table of Contents

Table of Figures

Table of Tables

Introduction

The following Master Thesis has the main aim to investigate whether specific features of Kickstarter projects have an influence on Customer Experience and individuals' preferences.

The first chapter provides a literature review on the development of Customer Experience theories, on the implementation of Machine Learning in Marketing and Customer Experience, and on previous studies and researches made on Customer Experience on Kickstarter.

In the 1920s, Marketing experts began to recognize for the first time the importance of the buyer's point of view (Strong, 1925). This current of thoughts leads to a new challenging thesis into the marketing theory: the belief that what people really desire are not products but satisfying experiences (Abbott, 1955) and that the reason why people buy product is not only for what they do, but also for what they mean (Levy 1959).

In the 1960s, experts began to question about the customer decision processes and experience when buying products, developing models to represent the stages from the need recognition to the final purchasing decision. In the 1970s, some studies began to focus on customer satisfaction (Wilkström, 1983) and on customer loyalty (Day, 1969; Bass, 1974). In the 1970s, the rising awareness of the importance of services and intangible assets lead to the SERVQUAL Model (Parasuraman et al., 1988) and to the Hedonic Consumption Theory (Hirschmann et al., 1982). In the 1990s, the importance for businesses of strengthening the relationship with customers emerged with the development of the Customer Relationship Management. The 2000s brought forth a stronger focus on value extraction from the customer relationship (Lemon et al., 2016). The 2010s are characterized by the switch to an Omni-Channel Customer Management, defined as *the synergetic management of the numerous available channels and customer touchpoints, in such a way that the customer experience across channels and the performance over channels is optimized* (Verhoef et al. 2015). This Omni-channel environment, together with the digitalization, gave companies the opportunity of collecting huge amount of data and information. The use and analysis of these is experimenting an evolution in the last decades. Traditional statistical models make several stringent assumptions on the types of data and their distribution, that may limit the potentiality of the analysis (Cui et al., 2006) and – when applied to real data – the models incur in some issues because the key assumptions of the research methods are often violated (Bhattacharyya 1999). Therefore, Machine Learning evolved with the aim of eliminating the dispendious and time-consuming processes needed to develop knowledge-based system (Bose et al., 2001).

In Customer Experience, Machine Learning is used to identify purchasing intentions for different products and services (Crone et al., 2006), or to analyse customers' responses to direct marketing in order to build a model capable of predicting individuals' purchases (Cui, 2006).

Another application of Machine Learning permits to analyse texts to provide a holistic perspective of textual and nontextual information (Mikroyannidis et al., 2006), for instance to analyse customer experience feedbacks (Ordenes et al., 2014)

Machine Learning is also used to daily life tasks, such as spam detection (Crawford, 2015), and chatbots (Serban et al., 2018).

In Kickstarter, previous researches already tried to understand which projects are preferred by customers, and why. In fact, some studies analysed the personality and features of backers (Greenberg et al., 2013), other studies those of the founders (Zvilichovsky et al., 2013).

Furthermore, psychological studies tried to explain what drove people towards funding on online platforms (Gerber et al., 2012) and whether they preferred to invest in project that were almost funded (Wash, 2013). Yet, since now, no analysis was based solely on those data that are available to customers only through the webpage of a project. In fact, the webpage is the first and often the only *touchpoint* between a project and a customer.

The second chapter provides an overview on Machine Learning. Machine Learning consists in building models using training data or past experience. The belief is that, behind the large amount of data, there are simple patterns that can be discovered and used by machine learning to build up a model that may be *predictive* - if it is used to make prediction on future outcomes, or *descriptive* – if it is used to gain better understanding on the data, or both.

The ability of algorithms to label data *(supervised learning)*, find pattern in data *(unsupervised learning)* and learn from own experience and results *(reinforcement learning)* increased the possibility to transform raw data into useful information and, among others, to analyse the relationship between touchpoints – services, product features, descriptions, calls, ads – and customer experience. The gained information helps companies to deliver appealing, customized and real-time products that are able to improve the quality of the customer experience across multiple channels and numerous touchpoints.

The third chapter develops the four hypotheses of the Master Thesis that help to answer the research question. Each hypothesis is related with one of the projects' features that have the following characteristics:

- Being visible on the webpage of the project
- Do not change but remain constant during the whole campaign

The hypotheses are developed accordingly to the idea that people are influenced in their investing decision by the outline of a project. They might perceive some characteristics as symbols for more quality, they might prefer to invest in a project rather than in another because its features are more appealing. Therefore, the hypotheses are the following:

- *H1: Individuals prefer to become "backers" of campaigns that are mostly about creative projects*

- *H2: individuals are influenced in their investing decision by the value of the goal of a project*

- *H3: individuals' investment decisions are influenced by geographic boundaries*

- *H4: Individuals prefer to invest in their own currency and therefore projects that ask for diffused currencies are more successful*

The newness of this research is to use those variables to find the best combination to build a predictive model through the implementation of Machine Learning, in order to predict the success of a project taking as main assumption the positive or negative impact that the features of the project itself had on the customer.

The fourth chapter uses descriptive statistics in order to provide a detailed insight of the database on which the research is based on. First, the 331.675 projects are divided into two set, the *training set* and the *test set.* The training set is used to draw the descriptive statistics of the projects, focussing on the four independent variables:

- Main Category
- Goal
- Country
- Currency

Main Category, Country, and *Currency,* are *Categorical Variables,* while *Goal* is a *discrete variable.*

The descriptive statistics provide deep information about each specific variable.

For the categorical variables, the frequency and the success rate are computed for each value. For the goal, some value ranges are used in order to show differences in frequency and success rate depending on the value of the funding goal.

The fifth chapter illustrates the methodology and processes implemented to build the predictive Model. The chapter starts explaining the preparation and cleaning process on the dataset. Then, it provides a deep analysis of four logistic regressions build alone on each independent variable and with the success / fail state as dependent variable. Then, the best Predictive Model was built by adding one variable at a time and testing the consistency of the model.

The sixth – and final – chapter of the thesis illustrated the final Predictive Model, draws the main findings and uses the results of the research to comment the four initial hypotheses.

1. Literature Review

The first chapter of the Master Thesis has the aim to provide the necessary knowledge basis through a deep insight of the Existing Literature.

The chapter is divided in three parts, that gradually move from a more general framework to the narrower outline about Kickstarter.

The first part explains the evolution of the Marketing theories related with the Customer Experience, from the 1920s until nowadays. What emerges is how the customer became increasingly meaningful for Marketing strategies.

The second part takes a more specific look on previous studies and Research made in the Customer Experience field through the implementation of Machine Learning techniques.

The third and last part of the chapter introduces studies and discoveries regarding Customer Experience on Kickstarter from different perspectives. The sum of these three parts, together, provides the fundament for the development of the hypothesis that help responding to the research question: whether Kickstarter's projects features influence Customer Experience and individuals' preferences.

1.1 Customer Experience theories

Customer experience has broadly been defined by scholars and marketers as a *multidimensional construct that involves cognitive, emotional, behavioural, sensorial, and social components* (Schmitt, 1999).

Recently, with the increasing number of contact points between a company and its customers, the attention towards customers became visible with the monitoring of the experiences originated by those contact points (Gentile et al., 2007). Thus, it gains more and more importance to consider aspects that are part of the emotional and irrational side of the customer behaviour (Holbrook et al., 1982).

The classical economic theory sees the customer as a logical thinker whose decision process is based on rational problem solving; instead, the new Marketing theories advocate that the perceived value for customers is strongly linked to intangible and irrational elements.

The importance of customer experience lies in the fact that it encompasses every aspect of a company's offering (Meyer et al., 2007): anyhow, the path that lead to this current of thought has been long, and multiple theories have been developed in the last decades.

In the last century, companies and marketers have been forced to re-think and adapt their marketing strategies to the changes of the market. In “*The Psychology of Selling and Advertising”* (1925), E.K. Strong illustrates how the selling procedure was experiencing some main changes because of the growing recognition of the importance of the *buyer’s point of view***.** Further on, he attributes to Elias St. Elmo Lewis, an American advertising advocate, the 1898 slogan *“attract attention, maintain interest, create desire”* and the later on added *“get action”.* This formula constituted the basis for the development of the *AIDA Model* (Rawal, 2013)**.**

The model considers the sequence of the four phases that occur when a consumer engages with advertising:

- *Attention*
  The company needs to grab the consumer’s attention from the very beginning of the advertising

- *Interest*
  The attention has to be converted into interest, creating an emotional connection with the consumer

- *Desire*
  The customer has to desire what you are offering. This means, he should develop a strong motivation for buying the company’s product even if there is no need

- *Action*
  The desire has to lead to action: the customer should eventually buy the product

Nevertheless, the model became soon obsolete. In fact, for E. K. Strong, many salesmen were not able to recognize that *honest service* to a buyer in terms of his needs means that the buyer’s interests must be first, not the seller’s. Moreover, even less salesmen ignore that the easiest way to achieve this goal is simply for the seller to develop his proposition in order to “fit” with the buyer’s point of view.

This current of thoughts leads to a new challenging thesis into the marketing theory: the belief that what people really desire are not products but satisfying experiences (Abbott, 1955). Following this path, four years Levy (1959) affirmed that the reason why people buy product is not only for what they do, but also for what they mean.

This has been the seed for subsequent theories, that saw the customer experience becoming the turning point of marketing strategies:

1. *Customer buying behaviour process models*: understanding customer experience and customer decision making as a process (1960s–1970s)

2. *Customer satisfaction and loyalty*: assessing and evaluating customer perceptions and attitudes about an experience (1970s)

3. *Service quality*: identifying the specific context and elements of the customer experience and mapping the customer journey (1980s)

4. *Relationship marketing*: broadening the scope of customer responses considered in the customer experience (1990s)

5. *Customer relationship management (CRM)*: linkage models to identify how specific elements of the customer experience influence each other and business outcomes (2000s)

6. *Customer centricity and customer focus*: focusing on the interdisciplinary and organizational challenges associated with successfully designing and managing customer experience (2000s–2010s)

7. *Customer engagement*: recognizing the customer's role in the experience (2010s)

1.1.1 Customer buying Behaviour Process Models (1960s – 1970s)

In the 1960s, marketers began to question about the customer decision processes and experience when buying products, developing models to represent the stages from the need recognition to the final purchasing decision.

One of the most relevant theories of these years (Howard et al., 1969) tries to explain the brand choices and the customer buying process. The elements of the buying process are identified as:

- Set of motives, specific to a product class and reflecting the underlying needs

- Alternative courses of action, depending on the brands taken into consideration for the selection. Two important aspects of alternative brands, are that they may not be in the same product class as defined by the industry, and that each consumer will choose only from a limited number of brands he knows, the so called *evoked set*, that is strongly personal

- Decision mediators through which the motives are matched with the alternatives, serving to structure both the buyer's motives and the various brands based on their potential to satisfy the ordered motives

The theory and its graphical representation (see *Figure 1: A theory of buying behaviour*) highlight how a high number of variables and even more interconnections and relations between them influence and are responsible for the customer's final choice.

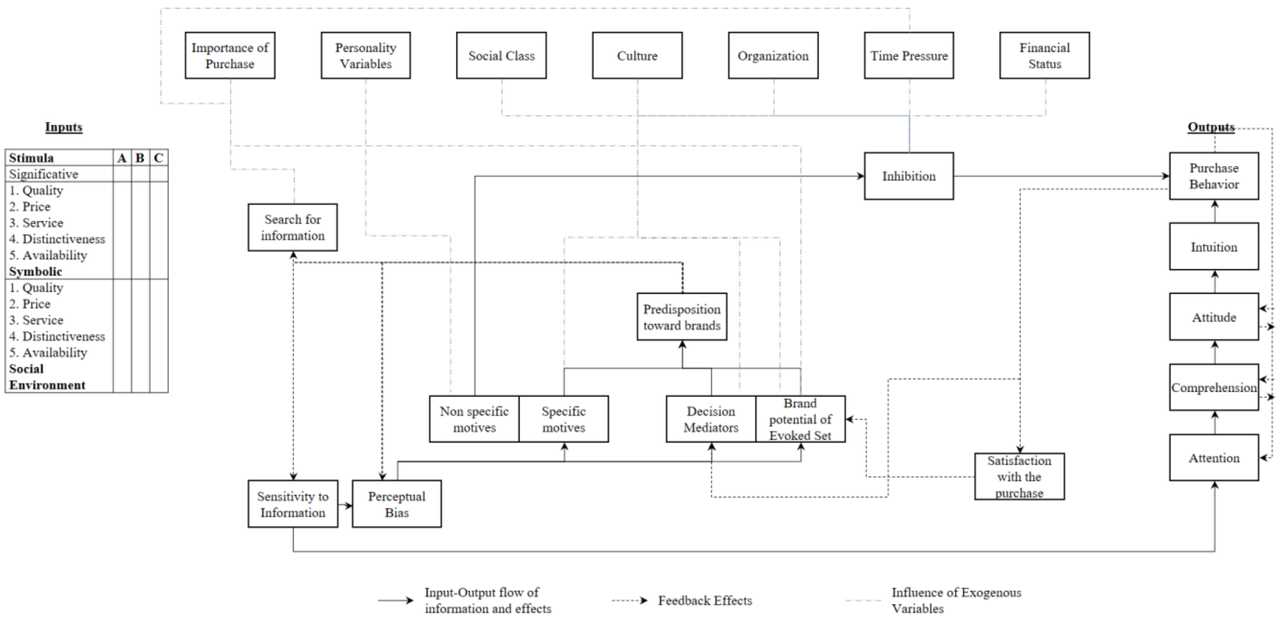

Figure 1: A theory of buying behaviour
Source: Howard et al. (1969)

Instead, in business-to-business Marketing, a theory defines the organizational buying process as a decision-making process carried out by individuals, in interaction with other people, in the context of a formal organization (Webster et al., 1972). Four classes of variables are assumed to influence the buying process: individual, social, organizational and environmental. A further distinction can be made between variables that are directly related to the buying process (*task variables*), and those that extend beyond the buying process (*nontask variables*).

Those models show how the attention towards the buying process was rising: in fact, it began to be identified as a complex sum of processes. Therefore, analysing this process has been an excellent starting point for developing effective marketing and selling strategies.

Those first theories, focussing on the customer decision making process, provided the foundation for a holistic thinking about the customer experience (Lemon et al., 2016).

1.1.2 Customer Satisfaction and Loyalty (1970s)

In the 1970s, some studies began to focus on *customer satisfaction.*

For instance, a study of Wikström (1983) based on customers' interviews held in 1978 both in US and Sweden, illustrates how interviewers believe that customer satisfaction has a critical importance, and how the right strategies depend on the specific market, because of the influence of cultural variables and national characteristics.

Customer Satisfaction has primarily been conceptualized as the difference between the actual delivered performance and the previous customer expectations. This disconfirmation (positive or negative) has been empirically shown to create customer satisfaction (Lemon et al., 2016).

Customer loyalty was approached from a behavioral perspective: repeated purchases were seen as a manifestation of customer loyalty. Nevertheless, this point of view lacked a conceptual basis (Day, 1969) and did not consider the stochastic component in every buying decision. Bass (1974) suggest that repeated purchases are not alone a sufficient indicator for customer loyalty, yet it might be useful to determine the major influences through which the stochastic decisions and preferences of the customer can be influenced.

1.1.3 Service Quality (1980s)

The Service Quality theories started to develop as a discipline separate from Marketing in the 1980s. Marketers began to realize how an increasing percentage of the monthly wage of consumers started to be allocated in *intangible assets*, and that services should be considered differently as goods (Rathmell, 1966).

The increased attention on this new discipline is reflected into the development of *SERVQUAL* (Parasuraman et al., 1988), a 22-item instrument for assessing customer perceptions of service quality in service and retailing organizations. As a matter of fact, no model existed yet to estimate the consumers' perception of quality in services and its implementation was intended to serve the purpose to track the service quality trends. Furthermore, it might be useful to determine the *relative* influence of the five dimensions (tangibles, reliability, responsiveness, assurance, empathy) on the customers' quality perception.

Furthermore, the *Hedonic consumption theory* by Elizabeth Hirschman and Morris Holbrook (1982) analyses those sides of consumer behaviour that are linked to the *multisensory, fantasy* and *emotive aspects* their experience with products. The paper expresses some concerns about the visible limitations of the traditional approach to marketing research, because of its incapability to recognize the complex, imaginative and emotion-laden nature of the customers' behaviour.

1.1.4 Relationship Marketing (1990s)

Although the first mention of "Relationship Marketing" can be tracked back to a 1983 paper, where Berry (1983) defined it as *attracting, maintaining and – in multi-service organizations – enhancing customer relationships*, it is only in the 1990s that the importance for businesses of strengthening the relationship with customers emerged. In the business-to-business marketing the concept of *trust* (Morgan et al., 1994) gains consistent relevance, as it is discovered to be important – together with commitment – for achieving cooperation. Furthermore, in the business-to-consumer marketing, a study (Reichheld et al., 1990) demonstrated that it is possible to increase profits from 25 to 85 percent by lowering the customer defection rate by only 5 percent. In fact, loyal customers generate more revenues for a longer period of time and the costs associated with maintaining existing customers are lower than those necessary to acquire new customers

1.1.5 Customer relationship Management (2000s)

The 2000s brought forth a stronger focus on value extraction from the customer relationship (Lemon et al., 2016). The further step that Customer Relationship Management (CRM) took compared to customer relationship, is that CRM is more focussed on the optimization of customer profitability and Customer Lifetime Value. For example, Reinartzt and Kumar (2000) highlight how stating that long-life customers are more profitable is an oversimplification, and that it is important how to make revenues from customers, independently from their loyalty to the company.

In 2005, Payne and Frow (2005) developed a single, process-based framework with the main aim to provide practical insights to help companies achieve greater success with CRM strategy development and implementation. An important aspect of this model is that it focuses on the *most profitable customers*, rather than simply on loyal customers.

Moreover, in these years, Rust, Lemon and Zeithaml (2004) developed the first broad framework for evaluating return on marketing, in order to make marketing financially accountable and to trade off competing strategic marketing investments on the basis of financial returns. The financial return is related to the firm’s customer equity, intended as the change in its current and future customers’ lifetime value.

1.1.6 Customer Centricity and Customer Focus (2000s – 2010s)

In Marketing, the new millennium began with a shift towards customer-centric marketing, that is developed to fulfil the needs and wants of each individual customer. The new technologies and the availability of systems capable of storing the huge amount of data produced daily by consumers and of software able to analyse and interpret those data made the focus on individual users easier.

The main consequences of this shift of marketing strategies towards a customer centric vision are: (a) development of a demand-driven supply management, (b) diffusion of unprofitable customer outsourcing to another company, (c) increase in interaction between firms and consumers for the “cocreation marketing”, and (d) reduction of transaction costs. (Sheth t al., 2000).

Some managerial tools have been developed to make the shift to customer centricity easier. For example, a buyer (or customer) persona is *a semi-fictional representation of your ideal customer based on market research and real data about your existing customers* (Kusinitz, 2014). Furthermore, another tool is the *jobs-to-be-done perspective*, that examines those circumstances in the customers' lives that influence or determine the purchasing of a product (Christensen, 2014).

1.1.7 Customer Engagement (2010s)

Currently, the main focus of Marketing theories is on customer and brand engagement.

Customer engagement is defined using five propositions: (*FP1*) it reflects a psychological stat that occurs by virtue of interactive customer experiences with a focal agent/object within specific service relationships, (*FP2*) it occurs within the cocreation process with the firm, (*FP3*) it plays a central role within a nomological network of service relationships, (*FP4*) it comprehends relevant cognitive, behavioural and emotional dimensions, and (*FP5*) it develops different levels of intensity and/or complexity, depending on individual and situational conditions (Brodie et al., 2011).

Furthermore, the broad development of digital and social media platform has strengthened the focus on the way people are affected by the engaged customers, directly or indirectly. In fact, customers have an increased number of interaction channels, and some researchers believe that there is a strong correlation between C2C interactions and firm profitability (Sterne, 2017).

The theories mentioned above – developed in the last decades – can be divided into three big research areas: (1) customer experience and the customer journey, (2) customer experience measurement, and (3) customer experience management. Customer experience has clearly gained a central position in Marketing strategies: it depends on the interactions between customers and firms through multiple channels and media, making it more complex for businesses to ensure to their customers a high quality end-to-end experience.

Researchers have been interested in the individualization and analysis of those interactions, also known as *touchpoints*. One commonly diffused believe (Jim Sterne, 2017) is that through the implementation of the right strategy during a specific touch point, it might be possible to show the right message in front of the right person at the right time in the right context on the right device and figuring out whether any of the work that was done had an impact on the buying decision.

The communication channels have changed considerably in the last decades. In a 2007 study, Verhoef, Neslin, and Vroomen (2007) considered three types of channels: offline channels (e.g..: stores), online channels (e.g.: Web store), and traditional direct marketing channels (e.g.: catalogues). Yet, in some specific retail markets the online channel has taken a dominant role (Christensen et al., 2003), while in other markets this shift has been less remarkable (Verhoef et al., 2015). Nevertheless, it became an important focus of marketing strategies whether to add those new channels in the existing Marketing Mix. The main problem linked with this decision was the fear of many companies of a cannibalization effect given by the adoption of the internet channel (Deleersnyder et al., 2002). From the opposite view, digital players started to wonder whether they should start to expand their presence also on the offline market (Avery et al., 2012). The above described phenomenon highlights the nature *of multi-channel marketing*, when different channels are deployed together in the same marketing strategy. In fact, marketing channels are not mutually exclusive: for instance, even though online migrations present cost savings opportunities, it might affect negatively customer satisfaction and brand health. More specifically, migrating customers to online channels may create resistance and customer dissatisfaction, as customers may feel forced to use new channels (Reinders et al., 2008).

From this perspective, *Multi-channel Customer Management* can be defined as the *design, deployment, coordination and evaluation of channels to enhance customer value through effective customer acquisition, retention, and development*. (Verhoef et al., 2007). This rigid distinction between channel typologies is justified by the changes that were happening in the market: the online channel was growing as something in opposition to traditional channels, such as stores and catalogues. Thus, it results clearer why they were often also managed separately, with only limited integration.

Anyway, in the last years, we are assisting to a further digitalisation in marketing, that is leading to a shift from multi-channel to omni-channel marketing (Rigby, 2011). The traditional distinction between retail stores, where customers can touch and feel merchandise, and online stores, where shoppers can choose between many products at convenient prices – is becoming less relevant. Online and offline are merging in a unique *omni-channel* retailing experience (Brynjolfsson et al., 2013). The new digital technologies and the diffusion of the mobile have been key drivers for a disruptive change in the retail environment. Compared to the previous multi-channel phase, the *omni-channel* phase shows significant differences not just for the increased number of channels but rather for the disappearing of the natural borders between them. Furthermore, the distinction between two-way communication (*interactive*) and one-way communication is not so clear.

Consumers have the possibility to use channels in an interchangeable and complementary way. For instance, shoppers may look on internet for further information about something that they just noticed in the store *(showrooming)*, or, on the opposite, they may go to the store to buy offline something that they sought and found online *(webrooming)*. In this omni-cannel environment, *touchpoints* may be one-way or two-way between companies and customers, of a more or less intensive nature. Touchpoints are moments of contact / communication between an organisation or brand and an individual consumer or stakeholder (Jenkinson, 2007).

Hence, *Omni-channel Customer Management* may be defined as *the synergetic management of the numerous available channels and customer touchpoints, in such a way that the customer experience across channels and the performance over channels is optimized* (Verhoef et al. 2015).

The table below (

Table 1: Multi-channel versus Omni-channel Management*)* provides with a more detailed comparison of the evolution from the Multi-channel to the Omni-channel Customer Management.

	Multi-channel management	*Omni-channel Management*
Channel focus	Interactive channels only	Interactive and mass-communication channels
Channel scope	Retail channels: store, online website, and direct marketing (catalog)	Retail channels: store, online website, and direct marketing, mobile channels (i.e., smart phones, tablets, apps), social media Customer Touchpoints (incl. mass communication channels: TV, Radio, Print, C2C, etc.).
Separation of channels	Separate channels with no overlap	Integrated channels providing seamless retail experiences.
Brand versus channel customer relationship focus	Customer – Retail channel focus	Customer – Retail channel – Brand focus
Channel Management	Per channel	Cross-channel

Objectives	Channel objectives (i.e., sales per channel; experience per channel)	Cross-channel objectives (i.e., overall retail customer experience, total sales over channels)

Table 1: Multi-channel versus Omni-channel Management
Source: P.C. Verhoef et al. (2015)

Nevertheless, from a customer point of view, there is not such attention on a distinction between channels. In fact, what they truly care about is finding an answer to their current need in a way that is convenient, easy, intuitive and valuable both for their use of time and money.

Furthermore, omni-channel customers are better informed, make use of technology and demand more from those retailers they do business with (Cook, 2014). The strong connection and integration between different channels is well represented by a quantitative survey of Sands et al. (2010), through which it was made evident how the online channel helps to improve the engagement with the retailer in a way that leads to a higher in-store spending.

In this highly challenging environment, where interactions between firms and customers happen through a huge number of channels, it is crucially important for a firm to understand customers' needs and the touchpoints through which it may influence the customers' experience. Therefore, the development of a *Customer Journey Map* might help to fulfil this purpose.

A customer journey map (*CJM*) is a visual depiction of the sequence of events through which customers may interact with a service organization during an entire purchase process (Rosenbaum, 2016). An example is provided in the figure below (*Figure 2: Customer Journey Map*)

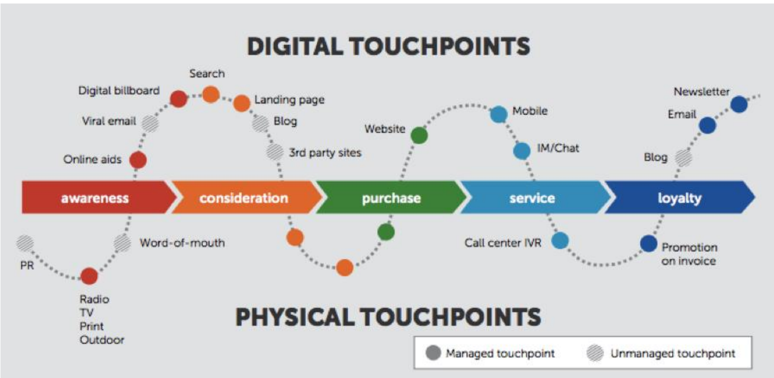

Figure 2: Customer Journey Map
Source: www.theclearing.com

It lists all the possible touchpoints a customer may encounter during his experience with the customer: this helps to improve the quality of the underperforming touchpoints in order to increase the overall end-to-end performance.

The customer journey mapping was first mentioned in the 1989 book "Service Wisdom" by Chip Bell and Ron Zemke as a *cycle of service mapping.* In a recent interview (Nasoi, 2017), Bell said that:

> *"The goal of customer journey mapping is to create and retain a deep understanding of the customer's experiences while he or she is traversing the path taken between having a need and getting that need met. Its intent is to "get inside the customer's head" to "see," and therefore, understand the customer's experiences. Armed with that perspective, organizations are better able to craft or recraft processes and encounters to become more customer-centric. It is essentially an evergreen effort since the needs and expectations of customers are constantly changing."*

Typically, the touchpoints are set on the *horizontal axis* following the timeline, and are part of one of three periods: pre-service, service, post-service.

Instead, on the *vertical axis,* should figure relevant strategic initiatives associated with every touchpoint. Other researches (Lingqvist, Plotkin, & Stanley, 2015) suggest that the vertical axis should be dedicated to the feelings, emotions and beliefs of the customer, transforming the CJM into an *empathy mapping exercise* (Tschimmel, 2012).

No customer journey map will ever be equal to another one: its outline depends on the company, on the products and services offered and on the potential actions that are considered.

Regardless of how many studies on the development of a customer journey have been made, three limits have not yet been overcome: (a) few examples of its practical application, (b) wrong basis assumption that each customer goes through every touchpoint and that all touchpoints have the same weight, and (c) managers' lack of understanding how to implement and use the vertical axis.

1.1.8 Overcoming touchpoints: Overall Experience

In recent years, some concerns about the effectiveness of customer journey maps are starting to arise. Ewan Duncan et al. (2010), for example, believe that to improve customers' experience it is necessary to move the focus from touchpoints to the overall experience. In fact, it may happen that the customer experience is not excellent, though individual touchpoints perform good.

In fact, it is still common for managers to underestimate the importance of an overall commitment of the company instead of focusing only on individual aspects, and of the internal cultural changes instead of adapting superficially.

In financial terms, implementing successful projects for optimizing the customer experience leads generally to a 5 to 10% revenue growth and to a 15 to 25% cost reduction in just two or three years. Furthermore, in comparison, companies can exceed their competitors' gross margins by 26%, make their employees happier and simplify their end-to-end operations.

Nevertheless, many companies do not provide their customers' needs as much relevance as they should. A study of Edelman Insights (2013) found that 90% of consumers believe that marketers should implement their *brand share* more. Brand share is intended as the value exchange between firms and their customers, and was assessed comparing customers' needs and brand behaviours.

Brandhshare was measured through six different variables – shared dialog, shared experience, shared goals, shared values, shared product, and shared history – and highlighted the existence of a link between effective brand sharing and business value. The research shows a consistent gap between the level of importance to customers of some actions that a company can take to share its brand (e.g. Asking people about their needs, communicating openly and transparently about how products are sourced and made…) and the performance.

What can companies do in order to improve their overall performance? Those companies that were able to create successful projects have been found to share six main pillars:

- *Define a clear customer-experience aspiration and common purpose*
  A company should provide its employees with a *state of intent* in order to make clear which are the customers' true needs and how to satisfy them. The state of intent should be coherent with the company's shared vision and aspiration. Furthermore, it should be translated into a set of standards to guide the employees across the whole end-to-end experience of the customer.

- *Develop a deep understanding of what matters to customers*
  Companies should understand which are the key factors that matter the most to customers and have the highest economic value.

Nowadays, the development of new technologies and the diffusion of social platforms such as Twitter, Facebook, TripAdvisor and Instagram provided customers with new ways to externalize feelings, comments and feedbacks about their experience with products and services. Furthermore, the development of *big data*, the creation of big data sets and the improvement of analytic methods for their analysis (e.g. *machine learning*) is helping organizations to understand their customers' needs and behaviour.

- *Use behavioural psychology to manage the customers' expectations*
  Companies should use behavioural psychology in order to shape the way in which they construct the consumer's end-to-end experience. In fact, behavioural psychology may be a useful tool to understand how to design each interaction with the customer in a way that creates a positive impact on his perception of the service.

- *Reinvent customer journeys using digital technologies*
  The digital era has revolutionized the customers' perception on products and services. Furthermore, the high competitive market allows customers to choose basing their decision on the current offers, even if this means buying every time from a different player. Digital technologies can guarantee a high competitive advantage to those companies that are able to offer an interactive, simple and functional digital experience to customers during their customer journeys. In fact, this may increase the willingness of the customers to interact with the company, improving the possibility to transform the interaction in a purchase.

- *Use customer journeys to empower the front line*
  Companies should choose and prepare carefully those employees that work on the front line: that means, those employees that are directly engaged in activities together with the customers. For example, it is very important to select employees with a good attitude, addressing their behaviour, not imposing strict rules. In this way, they will have the freedom to operate in order to maximize the customers' experience.

- *To improve constantly, establish metrics and a governance system*
  The knowledge about customers, their preferences and experiences should be analysed through metrics and performance indicators. In this way, it is possible to keep track of actions and their effects on the customer experience. The implementation of this kind

of system needs a strong leadership and governance, to be included downward in the whole organization.

In conclusion, the last decades have been theatre of a continuous development of the customer experience importance and understanding. Nowadays, the focus on the customer and the ability to deliver a compelling and satisfying experience to the customer are key drivers for a firm's economic success.

The number of recorded and registered data is increasing. Yet, those raw data are useless without a correct interpretation. Anyway, the traditional computing processes are often uncapable of dealing with large amount of data and of providing a solution to questions where the rules are not clear (e.g.: what differentiates a legitimate from a spam email? When is customer's review positive or negative?).

Thereafter, the deployment of *machine learning* algorithms in Marketing has been fundamental to achieve new results and to deliver a more compelling customer's experience, from a better customer segmentation to a more appealing advertisement.

1.2 Customer Experience and Machine Learning

Nowadays, the engagement between a company and its customers happens among a high number of touchpoints. These touchpoints differ for moment, channel, length, intensiveness, nature. The digitalisation lead to an increasing number of online channels with different facets, that influence the customer experience, the buying decision process and the overall brand image in front of the customer's eyes. Hence, it is becoming more and more important to shape those *touchpoints* in order to create a positive feeling that makes individuals want to buy, to become loyal customers and to spread positive comments because of the outstanding experience they had. One way of shaping customers' experience according to individuals' needs and preferences, is through the implementation of the latest technologies, systems and devices that the technological revolution made available in the last years.

Until recently, traditional statistical models have dominated the analysis of customer preferences and responses to direct Marketing (Berger et al., 1992). Yet, statistical models make several stringent assumptions on the types of data and their distribution that may limit the potentiality of the analysis (Cui et al., 2006). Furthermore, when applied to real data, the model incurs in some issues because the key assumption of the research methods are often violated (Bhattacharyya 1999).

Also the increasing amount and variety of data collected from customers makes it almost impossible to draw manual solutions (Bitran et al., 1996). In fact, the Omni-Channel Environment multiplies the touchpoints between customers and companies (Bitner et al., 2000).

Nevertheless, customers itself generate a large amount of data that, if properly analysed, may be a valuable source of insights for companies (Belkahla et al., 2011).

The question is, if the implementation of *Machine Learning* might truly help companies to overcome the above-mentioned limits and to exploit the potential of the huge amount of data and information collected on a daily basis about customers habits, choices, actions and preferences.

Machine Learning evolved with the aim of eliminating the dispendious and time consuming processes needed to develop knowledge-based system (Bose et al., 2001). In fact, an accurate analysis of this information with traditional tools requires a significant time contribution (Janasik et al., 2009).

In Customer Experience, Machine Learning is used to identify purchasing intentions for different products and services (Crone et al., 2006), or to analyse customers' responses to direct marketing in order to build a model capable of predicting individuals' purchases (Cui, 2006).

Another application of Machine Learning permits to analyse texts to provide a holistic perspective of textual and nontextual information (Mikroyannidis et al., 2006), for instance to analyse customer experience feedbacks (Ordenes et al., 2014)

Machine Learning is also used to daily life tasks, such as spam detection (Crawford, 2015), and chatbots (Serban et al., 2018).

What becomes clear from the above-mentioned studies, is that ML techniques are applied across different industries and for multiple purposes.

1.3 Customer Experience on Kickstarter

Crowdfunding platforms, such as Kickstarter, offer project founders the possibility to ask for funding for their idea from all internet users active on these open online services (Mollick, 2014).

Kickstarter was launched in 2009 with the mission of *helping to bring creative projects to life.*

Kickstarter helps artists, musicians, filmmakers, designers, and other creators to find the resources and support they need to make their ideas a reality. In fact, starting a campaign, creators can ask the *crowd* for funding.

Before launching the campaign, the creators of the project set the goal they want to achieve: if this amount is achieved, then the project receives the backers' funding. If the campaign misses the goal, the pledged amount is not charged on the bank account of the backer. This "all or nothing" procedure makes it fundamental for project founders to use the few elements and variables they can control for the presentation to the *crowd* to deliver as much value as possible, to achieve the financing goal. Many studies and researches have already tried to analyse and understand the relation between the variables of a campaign and the success of a project.

Depending on the funded amount, the backers may receive a "reward". Small amounts (generally below $10), are without any type of reward. Higher amounts generally permit to receive something that goes from a customized "thank you", to one or more prototypes of the backed product, that can come in different models.

Typically, rewards can be divided into four categories (Qiu, 2013):

- Recognition
- Items produced by the project
- Specially customized items
- Creative experiences (e.g.: studio visit to a gaming company)

Kickstarter projects can vary one another from many perspective: they can differ for the country the founders come from, the goal, the mission, the kind of product, the reward, the network and background of the founders, and many others. Thus, many studies have been developed on those different features.

From a general perspective, some studies compare profit and non-profit crowdfunding initiatives (Lambert et al., 2010).

Statistic variables of project and founders have also been widely analysed. Elements such as founders' friends on Facebook and followers on Twitter were insert in the model (Greenberg et at., 2013), while other more comprehensive studies focussed on project-specific and founder-specific variables (Koch et al., 2015).

Some studies focus on the importance of the reward scheme, the rewards that a backer offers depending on the pledged amount (Xiao et al., 2014), for achieving higher funding.

Another study introduced the perspective of seeing investors as customersand the need for innovativeness (Chan et al., 2014).

The question that drove other researches was whether information about the founder influences the funding success on future projects (Zvilichovsky et al., 2013). Kickstarter itself recommends backers to check the founders of a project, to know their history and eventually their previous successfully launched projects, that might be seen as a sign of their capability of transforming an idea into something real.

Because of the importance of visibility on the platform, previous literature also focused on which variables and elements lead to increasing probability of being picked by the Kickstarter staff as a “project we Love” (Gutsche et al., 2018). Furthermore, natural language methods permitted to identify sentences that stimulate investment decisions (Mitra et al., 2015). A marketing and customer experience perspective is used to analyse the distinction between normal donations and donations that are made to an almost funded (Wash, 2013), while a psychologic approach is used to study the motivations that push people to participate on computer mediated crowdfunding platforms (Gerber et al., 2012): in fact, a reason for participation may be found in the intrinsic gratification received by supporting a project (Ordanini et al., 2011).

From a managerial perspective, a study (Belleflamme et al., 2013) says that entrepreneurs should build the “right” community of crowdfunders accordingly to their operating and financing needs. In fact, entrepreneurs should provide to their backers to with community benefits for their participation in the campaign. This non-monetary reward is the key driver for achieving the necessary price discrimination that makes it possible for crowdfunding to become a cheaper and more convenient for of financing than banks or large equity investors.

An interesting research on music artists that launch a campaign on Kickstarter (Wang, 2016) affirms that Kickstarter does often just leverage on pre-Kickstarter social capital and relationships. Instead, it is the opposite for scientific research: in fact, a crowdfunding platform such as Kickstarter might represent a great opportunity for diffusing knowledge and improving science education. This means, that the true potential of crowdfunding – at least in the scientific field – lies in the possibility to reach a broader audience, helping to fight the misunderstanding of science among the general public (Wheat et al., 2013).

Furthermore, through a semantic text analytics approach, a team of researchers started to build a model to analyse which features of the description were more likely to help projects to obtain funding (Yuan et al., 2016).

From a data mining and Machine Learning perspective, only a few analyses have already been made.
A research (Mitra et al., 2014) uses Natural Language Processing techniques in order to analyse the textual content of Kickstarter projects, while another study leverages the updates of projects to determine their success rate (Xu et al., 2014).
Etter et al. (2013) build prediction models relating the probability if success of a project with the number and quality of Tweets publicly available.

On Kickstarter, each project provides the same kinds of information on its presentation webpage. The aim of this research is to find some hidden relationship between those information, that constitute the only *touchpoint* between the project, and Customer Experience and individuals' preferences. In fact, because of the nature of crowdfunding platforms themselves, whoever is not part of the FFF category (family, friends and fools) knows about a project only those information that have been made publicly available on the internet site. Thus, believing that project features may influence Customer Experience and individuals' preferences, the following study has the aim of building a Predictive Model with the implementation of Machine Learning algorithms.

The variables taken into account are those visible on the webpage and that do not change with time, but stay constant, that are: Category, Country Currency and Goal.

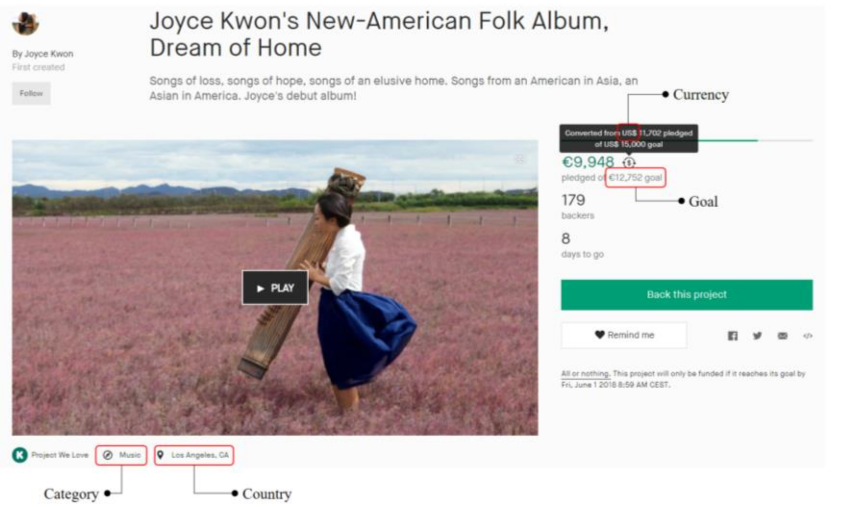

Figure 3: Sample Project on Kickstarter

To manipulate the data, R – a language and environment for statistical computing and graphics – was used.

In order to help those that are passionate about Machine Learning to understand how to apply the basics of R programming, the algorithms that were used in the project are always reported, step after step.

2. Introduction to Machine Learning

This chapter has the aim to provide an introduction on Machine Learning and on its main features, types and applications.

First, it is necessary to understand what Machine Learning is, and what is the meaning behind *learning.* The chapter provides an overview of those tasks where Machine Learning provides a higher efficiency compared to normal computer programs, and an insight of the typologies of Machine Learning: Supervised Learning – when the computer learns from labelled data, Unsupervised Learning – when the computer learns from unlabelled data, and Reinforcement Learning – when the computer learns to maximize a reward signal. The chapter provides also an insight of Deep Learning, a subset of Machine Learning, that deserves to be mentioned for its increasing implementations.

2.1 Features of Machine Learning algorithms

The new digital devices are recording a huge amount of reliable data, while computer technologies make it possible to store it and to access it from physically distant locations over a computer network (Alpaydin, 2010). Experts believe that there are some common patterns that explain the data, and that they do not have a completely random nature. *Consumer behaviour* offers a good example: usually, people buy less ice cream in winter than in summer and when they buy a beer they could buy also chips.

Developing *a good and useful approximation* may help in achieving some interesting discoveries, such as patterns or regularities.

This is exactly what *Machine Learning* is all about: the automated detection of meaningful patterns in data (Shalev-Shwartz, 2014). In fact, machine learning is making computers build a model and its parameters using the training data or past experience. The belief is that, behind the large amount of data, there are simple patterns that can be discovered and used by machine learning to build up a model**.**

The model may be *predictive* - if it is used to make prediction on future outcomes, or *descriptive* – if it is used to gain better understanding on the data, or both.

Eventually, Machine Learning may be preferred over normal programs in those situations where there is:

- *Complexity*
  - *Tasks performed by human/animals.* Tasks that are easy and automatic for humans and animals, but that do not have a clear explanation on how they happen (e.g. image recognition, driving...)
  - *Tasks beyond human capabilities.* Tasks that are related to the analysis of big and complex data.

- *Need for adaptivity*

  Machine *Learning* permits to overcome the rigidity of a standard computer program, offering tools that are able to adapt to a changing environment.

Machines are powerful. For instance, in a paper published in 2014 by Telenor Group and MIT (Sundsøy, 2014), a machine learning program was used to segment Asian customers of a MNO (Mobile Network Operator) for text-marketing purposes. In this field, the most common practice to target potential customers is to follow the "gut-feeling" of marketing experts. Through the implementation of a machine learning program, it was possible to achieve a 13 times better conversion rate.

The awareness towards *Machine Learning* is rising. *Figure 4: Machine Learning on Google Trends* provides an overview of this phenomenon. On Google Trends, data are normalized: the numbers represent the frequency (number of searches / population) of searches in the selected period, where 100 indicates highest frequency ever registered.

Yet, it is still difficult for people to distinguish between Machine Learning and Artificial Intelligence, and to notice how much both are present in their lives. Machine Learning is a subset of Artificial Intelligence: Artificial Intelligence is a machine able of doing something *smart*, that imitates or interacts with humans (Triggs, 2018); instead, machine learning is a training mechanism by which computers are able to learn and adapt themselves to various situations.

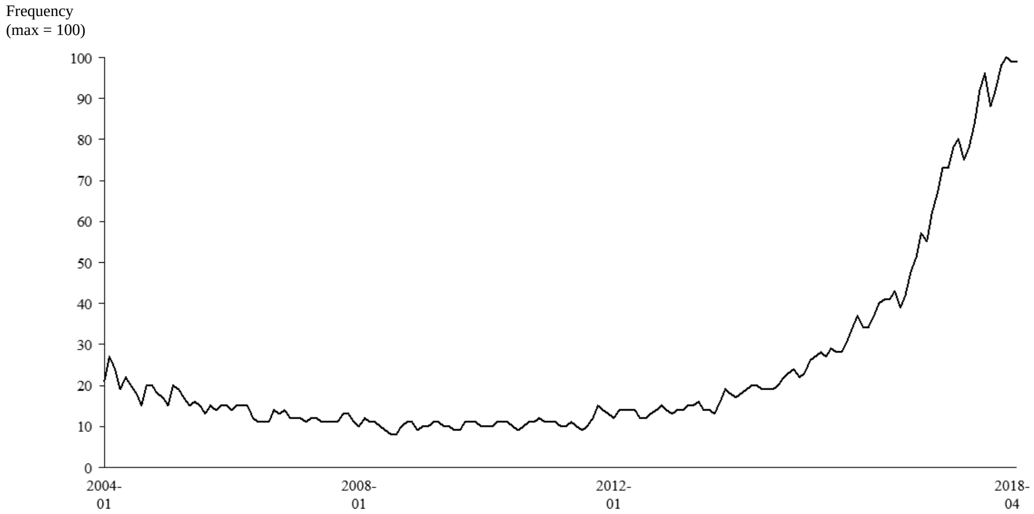

Figure 4: Machine Learning on Google Trends
Source: Google Trends

To implement a learning system, its design shall be structured properly. Making the right choices about the design is fundamental in order to assure the quality of the model and its ability to predict future outputs.

The first relevant choice is about the type of training experience (Mitchell, 1997).

The training experience has the following attributes:

- *Direct or indirect feedback*

  The training experience can provide a direct feedback about the choices made by the learning system. In this case, the system knows immediately whether it made the choice correctly or not. Alternatively, the feedback may be indirect if the system must *infer* indirectly from the final outcome the rightness of specific choices. This second scenario presents an additional problem, the *credit assignment:* determining the degree to which each choice is determinant for the quality of the final outcome.

  This task has a tricky nature: for example, good initial choices can lead to a negative performance anyway, if they are followed by bad choices.

  Therefore, direct feedbacks usually provide better information than indirect feedbacks do

- *Degree of control of the sequence of the training examples*

It is the degree to which the programmer knows which choices to make and whether they considered right or wrong.

- *Similarly of training set with the test set*
  It refers to the degree to which the distribution of the training set is similar to the one of the test set: higher the degree of similarity, higher the reliability of the results

Then, it is necessary to establish which kind of knowledge will be used by the program. The aim is to create a very expressive representation as close as possible to the ideal target function. However, the more expressive the function, the more data will be required in order to make the program capable of choosing among the alternative hypothesis.

2.2 Types of Machine Learning

There are three types if Machine Learning: Supervised Machine Learning, Unsupervised Machine Learning and Reinforced Machine Learning.

A key point that deserves a short focus is the importance of *high-quality data*. This is the reason why data preparation is a fundamental stage of data analysis. A lot of raw data is available in various data sources and on the Web, and companies and firms are interested in "cleaning" data to shape it into a high-quality data set that can be analysed and used to generate profits (Zhang et al., 2003).

In fact, *real world data* may be (Zhang et al., 2013):

- *Incomplete*, when there are attribute values or attributes of interest missing when it contains only aggregate data

- *Noisy*, when there are errors or outliers

- *Inconsistent*, when there are discrepancies in codes or names

Preparing those data to generate a quality dataset comprehends actions as selecting only relevant data, removing anomalies filling the missing value. At the end of the whole process, the quality of the data will provide a higher quality of the pattern.

2.2.1 Supervised Learning

Supervised Learning is based on the idea of learning from experience.
Figure 5: Supervised Machine Learning provides a representation of the steps that occur in the implementation of Supervised Machine Learning algorithms.

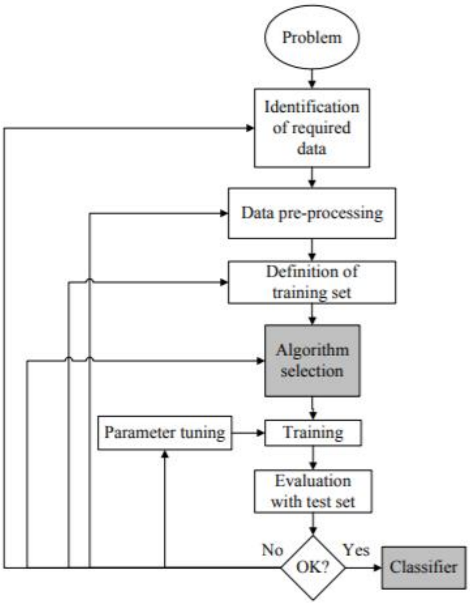

Figure 5: Supervised Machine Learning
Source: S. B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques

Usually, the computer is provided with two sets of data, a *training set* and a *test set.*

In the training set, the data are already labelled. In this way, the program can learn from the training set and try to label as correctly as possible the data in the test set. The aim of the program is to understand how to classify properly the unlabelled data through the example it has been provided with.

An example could be a set of images of cats and dogs, all properly labelled. After running the training set, the program should have enough knowledge to be able – in the training set – to distinguish cats and dogs.

The structure of the two sets is as follows:

- *Training set*

It consists of *n* ordered pairs (x_1; y_1), (x_2; y_2), … (x_n; y_n), where x_i is the *measure* or a *set of measures* of a single example data point, and y_i is the *label* for that specific data point (learned-Miller, 2014). For example, x_i could be a group of measurements for a patient in a hospital (e.g.: height, weight, temperature…) and the corresponding y_i could be the classification of the patient as "healthy" or "not healty"

- *Test set*
  It consists of another set of data with *m* measurements without labels (x_{n+1}, x_{n+2}, x_{n+m}). The goal is to label correctly the data basing the decision on the measurements and on the model built upon the training set. For example, the model should be able to tell – knowing the measurements for a patient in a hospital, if the patient is "healthy" or "not healthy"

Supervised learning algorithms can be deployed for *Classification* and *Regression* problems. In classification problems, the aim is to predict whether the output will be in a certain way or not (the labels of the data are categorical). Instead, in regression problems, the aim is to predict the value of the output (the labels of the data are numerical).

2.2.2 Unsupervised Learning

While in supervised learning the labels are given, in the unsupervised learning the only available information are the data. The aim is to find some patterns in the data, analysing the input space and searching for some structures that happen to be present more often than others. One of the most common ways to find patterns is using *clustering*.

Clustering

Clustering is one method to find cluster or groups of data with similar characteristics. The aim is to create groups such that the objects of one group are similar (or related) to one another and different from (or unrelated to) the objects in other groups. The goal is to minimize the *intra-cluster distance* (the distance between the elements in the same cluster) and to maximize the *inter-cluster distance* (the distance between elements in different clusters). A division in groups based on those characteristic is called *customer segmentation.* Starting from this segmentation, the company might decide to develop specific products or strategies differentiated for a specific or for each segment.

The customer segmentation may also allow to identify *outliers* – customers that differ strongly among the others, that could indicate the existence of a niche market with possibility of exploitation for the company (Alpaydin, 2010).

Other important applications of clustering are the following:

- Image processing, clustering images based on their visual content
- Web mining, clustering users based on their research or webpages based on their content

Nevertheless, the concept of cluster remains ambiguous: given a set of data, it is not always simple to evaluate the proper number of cluster and how to collocate each object.

Clustering can be *partitional* or *hierarchical,* as represented in *Figure 6: Graphical representation of hierarchical and partitional clustering*.

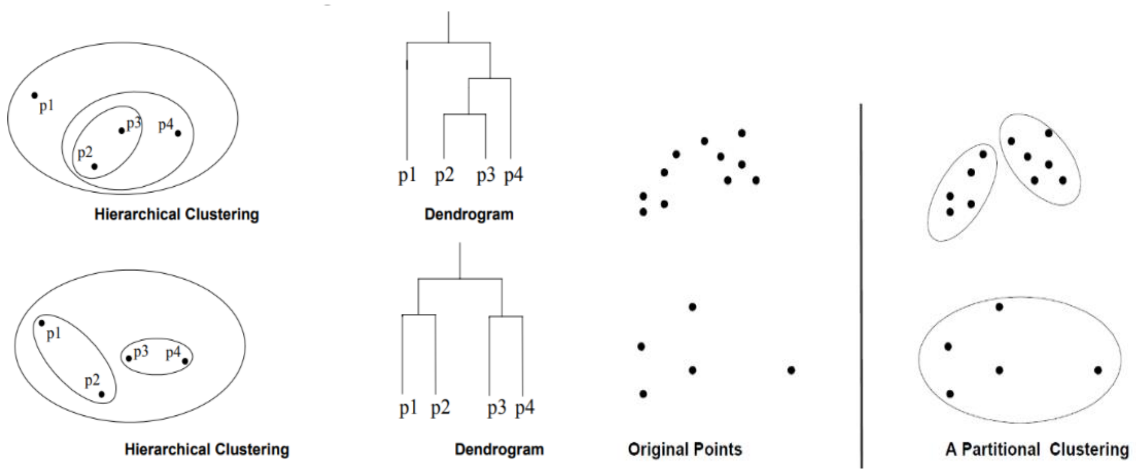

Figure 6: Graphical representation of hierarchical and partitional clustering
Source: Arthur et al. (2007)

- *Partitional Clustering*

  When Each object belongs in exactly one cluster

- *Hierarchical Clustering*

  A set of nested clusters organized in one tree.

  Hierarchical Clustering can be:

– *Agglomerative*

Starts with the points as individual clusters, and at each step the closest clusters are merged until just one cluster remains

– *Divisive*

Starts with a cluster, and at each step the cluster is split until there is just one point for each cluster or a K number of clusters.

At the beginning, researchers expressed some concerns regarding the use of clustering for marketing research. In fact, they have discussed problems with determining the right measure of similarity and the correct number of clusters (Green et al., 1967), or believe that clustering may be effective only when heterogeneous and strongly different groups can be identified (Wells, 1975). Nevertheless, this scepticism can be explained by the confused approach of Marketing Researchers towards clustering methods (Punj et al., 1983).

2.2.3 Reinforcement Learning

Reinforcement Learning is about analysing the environment, in order to understand which actions to take in order to maximize a reward signal.

It is at the same time the problem, the class of solution methods that work well on the class of problems, and the field that studies these problems and their possible solutions (Sutton et al., 2017).

It is possible to identify six main components in a reinforcement learning system:

- *Agent*

  The entity that has to analyse the environment and to understand which actions are those that lead to the best reward

- *Environment*

  The context in which the agent is acting

- *Policy*

The definition of how the agent behaves at a specific time, correspondingly to the perception of a given state of the environment. Policies may be stochastic

- *Reward signal*

  The single number that is sent back to the agent by the environment after each step, defining whether the action has positive or negative effects. The reward is what influences the policy: a low reward may lead to a change in the policy, in order that the next time another action – with a potentially higher reward – will be selected. Reward signals may be stochastic function of the environment and of the taken actions

- *Value function*

  The indicator of what is good in the long period. It can be considered as the total value of rewards given by the current state and the states that are probably following. This means that a state with a low reward can still be chosen by the agent if, together with the states that are likely to follow and their higher rewards, it may still provide the best value function. Despite the extremely high importance of value, it is difficult to determine them. In fact, while the reward is given directly by the environment, values must be estimated over and over again from the increasing sequence of actions made by the agent.

- *Model of the environment (optional)*

  The tool that allows to presume how the environment and the agent will behave, trying to predict future states, rewards and values.

  Methods that count on models are *model-based,* while those supported by the trial-and-error logic are *model-free* methods

Reinforcement Learning differs both from *supervised* and *unsupervised learning.*

Supervised learning is about learning from a labelled set of data. Nevertheless, in interactive problems, it is often difficult or even impossible to obtain labelled examples that are both correct and representative of all the situations the agent might have to deal with. Therefore, learning from the past experience is something that may happen to be useful in some contexts, as those in which interaction is required.

Instead, unsupervised learning is about finding patterns in unlabelled data, but it does not give any indication about how to maximize a rewarding signal.

Reinforcement Learning solves *close-loop problems*, that means, problems where the learning system's actions influence its later inputs.

Furthermore, the agent does not know which actions are the best, but it has to discover it learning from its experience. Given the correlation between different actions, the first action may influence not only the immediate reward but also the sequent action and, because of this, the following reward.

Summarizing, the three main and distinguishing features of Reinforcement Learning are: (a) being closed-up, (b) not having precise instructions on the next move and (c) and what consequences on what period of time each action will have.

Figure 7: Reinforcement Learning provides a representation of the Reinforcement Learning mechanism.

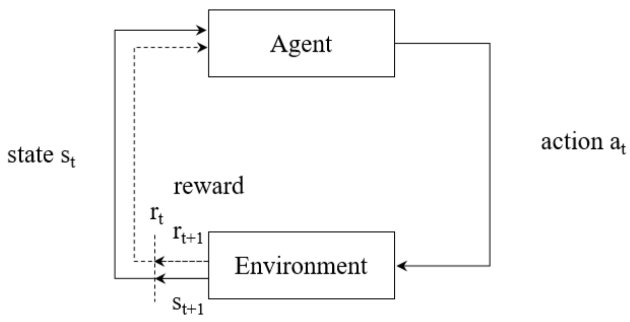

Figure 7: Reinforcement Learning

In state s_t, the agent chooses action a_t accordingly to a policy. Assuming that the model follows Markov's property (*the future is independent of the past given the present*, that means, that the future states of the process only depend on the present state, not on the sequence of previous events), the state s_{t+1} and the reward r_{t+1} depend exclusively on s_t and a_t.. Reinforcement learning methods determine how to change the policy learning from experience to maximize the cumulative reward (Pednault et al., 2002).

A unique challenge that only reinforcement learning has to face is the trade-off between exploitation and exploration. The agent has to exploit the best solutions in order to maximize the reward, but in the meanwhile it has also to explore further alternatives that can lead to a higher reward and that can be exploited further on.

Deep learning may be used for maximizing marketing campaign results. Usually, cost-sensitive learning methods learn policies that attempt to minimize the cost of a single decision.

However, in many applications, sequences of decisions must be made over time, considering the interaction between decisions (Pednault et al., 2002).

Current marketing strategies tend to maximize profits considering each campaign separately. This is a *greedy approach* - selecting the best possible solution at each step, without taking into consideration the overall performance – that might lead to undesired effects, as for example overmailing with offers and newsletters. A better approach would be to maximize profit over a series of campaigns.

2.2.4 Deep Learning

Even though Deep Learning is not properly a *type* of Machine Learning but more a *subset* of it, it deserves to be mentioned.

There are some tasks that appear natural and immediate for the human brain: recognizing a face, understanding whether a person is angry, knowing the difference between the picture of a cat and of a dog. If people were asked to explain how they could were able to solve those tasks, they would not know which process they followed: it just happened.

Deep Learning is a subset of Machine Learning that helps computers to deal with these kinds of tasks, easier to perform than to describe. The computer understands the world in terms of a hierarchy of concepts, allowing the computer to learn complicated concepts by building them out of simpler ones (Goodfellow et al., 2016).

Deep Learning works through the implementation of Artificial Neural Networks, structures that have been inspired by the human brain. An artificial neural network (ANN) is made by many simple, connected processors (neurons), each producing a sequence of real-valued activation. Input neurons get activated through sensors perceiving the environment, other neurons get activated through weighted connections from previously active neurons (Schmidhuber, 2015).

Deep learning has two properties:

- Multiple layers of nonlinear processing units
- Supervised or unsupervised learning of feature presentation on each layer

In the Neural Network structure, the *input* and the *output layers* are divided by *n hidden layers.* The layers are connected through *synapses,* each of whom has a *weight* for *neural activation.*

The Artificial Neural Networks can be:

- *Feedforward,* when the information moves only in one direction, forward, from the input to the output nodes
- *Feedback (or Recurrent),* when the information does not move only in one direction, but the neurons can use their internal state (memory) to process sequences of inputs

The layers can be:

- *Fully-connected*, when all input neurons are connected to all output neurons
- *Sparsely connected,* when all input neurons are not connected to all output neurons

Self-driving cars are an example of Deep Learning application.

In an experiment (Bojarskiet al.,2016), a team of researches taught a Convolutional Neural Network –an artificial neural network used to analyse visual images –how to drive.

Less than 100 hours of driving video were “fed” into a CNN, that computes a proposed steering command: the proposed command is compared to the desired command and the weight are adjusted in order to reduce the error.

After the training, the system was able to interpret pixels from a single front-facing camera to provide a steering command.

The experiment showed surprising results: with minimum training, the system was able to learn how to drive in different situations and even in areas with unclear visual such as parking lots. The system automatically learnt to recognize the importance of some features, necessary to provide the desired steering command, without being explicitly being trained to do it.

The experiment highlights one of the advantages of Deep Learning. In fact, the Neural Network was able to learn the entire task of driving without the need for manually indicating the specific tasks to perform, such as detecting the lane or the path planning

Deep Learning would deserve a deeper explanation. Nevertheless, even though it is not possible to do it in this thesis, it deserved at least to be mentioned, in order to provide a comprehensive overview of Machine Learning.

3. Hypotheses Development

The research has the aim to investigate the influence of projects' features on the Customer Experience and individuals' preferences.

Being a relative new domain, not many studies have already explored the problem from a data mining and Machine Learning perspective (Rakesh et al., 2015).

Furthermore, Kickstarter is a highly heterogeneous online platform where a customer may be influenced by various features of different nature.

Mollick (2014) analyses the nature of project features: in particular, he discovered that the significant variables can be summarized in two categories. The first category is the *preparedness* (e.g.: video, photo, spelling check, constant updates), and the second category is the *social capital* (Number of Facebook friends of the founder). Another study (Greenberg, 2013), related the success of a project strongly with the use of specific phrases in the description.

In the following study, the variables that are considered are only those that have two main features:

- Are visible on the webpage of the project
- Do not change but remain constant during the whole campaign

The hypotheses are developed in order to investigate the research question. Customers might perceive some project features as symbols for more quality, they might prefer to invest in a project rather than in another because its features are more appealing. Therefore, the hypotheses are considered taking into account four project characteristics:

- Category
- Country
- Goal
- Currency

to determine whether there is a relation between the features of a project and the individuals' willingness to invest and therefore help a project to reach its goal.

3.1 Category: what kind of projects do customers prefer?

On Kickstarter, there are 15 categories in which the projects are divided. Those categories are:

1. Art
2. Comics
3. Crafts
4. Dance
5. Design
6. Fashion
7. Film and Video
8. Food
9. Games
10. Journalism
11. Music
12. Photography
13. Publishing
14. Technology
15. Theater

These 15 categories cover a wide range of individuals' personal interests.

Usually, 20% - 40% of *initial* investment come from friends and family. For those groups, it is rather important to support the project because of the relationship with the founder rather that for the features of the project itself (An et al., 2014). This situation changes when considering *active investors* who are more interested in funding a project that match their own interest. (An et al., 2014). Furthermore, not all categories are treated the same on Kickstarter. In fact, Design and Technology project, because of their nature, deliver *concrete* products as a reward (Mollick, 2013).

In a later study (Mollick, 2016), the categories are clustered in two general groups:

- *Product oriented,* generally created in the attempt of building an organization and offering products that have a more commercial nature

- *Art oriented*, generally created by one-time or informal groups, with the main aim of delivering purely creative projects

Given these differences in the very nature of the product categories, it is almost automatic to expect some differences in the campaign results of projects that belong to different categories. Kickstarter is born with the aim of helping *to bring creative projects to life*, hence it is possible to expect projects that offer something artistic and intangible to be more successful than projects that offer something merely physical.
Therefore, the first hypothesis is the one that follows:

> *H1: Individuals prefer to become "backers" of campaigns that are mostly about creative projects*

3.2 Goal: do customers prefer realistic of ambitious goals?

The goal is the amount of funding that the founder wants to rise from the *crowd*.

Kickstarter is based on an *all-or-nothing* model: this means, that if a project fails to achieve the initial goal, it does not receive neither the collected funds. Therefore, founders tend to set realistic goals in order to reduce the risk of missing the needed collected amount (An et al., 2014). Yet, a higher goal may be an indicator of an ambitious project, and attract expert investors (Sahlmann, 1997).

Current literature reveals some contrasting findings about the influence of the goal on the willingness of individuals of investing in a project.

A study found a substitution effects in contribution consistent with altruism models: that is, individuals are less likely to decide to invest in projects that have already met their goal, instead, they prefer to help a founder that is still striving to achieve the needed amount of investments. Furthermore, a higher goal might be perceived by investors as an opportunity to really help someone who needs their contribution; on the other hand, it might appear as an impossible target and therefore let the project appear as a bad investment (Burtch, 2013).

For another study, the goal is not important for backers. Rather, backer do invest until the goal is not reached (Rakesh, 2015).

The previous studies do not provide a clear path through which an assumption might be formulated. The contrasting results seem to signal that the goal has some relevance, even though previous studies have contrasting opinions whether this influence is positive or negative. In fact, as discovered by previous studies:

- A low goal is easier to achieve but does not make investors feel as if they fail to contribute the project may not be able to achieve the needed funding
- A higher goal can be interpreted as a symbol for an ambitious project, but is more difficult to achieve

Hence, the second hypothesis is the following:

> *H2: individuals are influenced in their investing decision by the value of the goal of a project*

3.3 Country: are there boundaries on the internet?

In Kickstarter, founders and investors tend to reveal the place in which they are located, and their city compares below of the project's photo (An et al., 2014).

In a study made on musicians seeking crowdfunding, Agrawal (2011) tried to understand whether geographic proximity could influence crowdfunding as happens with Venture Capital firms. The research shows that investment patterns over time do not depend on the distance between founder and backer. Therefore, crowdfunding may reduce the friction related with geographic boundaries, as happening until now.

In fact, Crowdfunding may be able to interrupt the habit for investors of relating their investment decision to the location of the target. For instance, Sorenson and Stuart (2005) report that on average target firms are no more than 70 apart from the VC, while 50% of angel investors chose to finance firms that are within half a day of travel (Sohl, 1999).

New findings (Agrawal, 2011) highlight how the geographic dispersion of crowdfunding investment breaks any limit related to geographic distance. This result is consistent with previous research in retail and advertising that show how online channels allow people to overcome offline barriers (Choi et al., 2010; Brynjolfsson et al., 2009; Goldfarb et al., 2010).

The main findings of previous work and research papers seem to provide a clear evidence of the fact that customer do not express preferences in terms of geographical location of the project funders. Hence, the third hypothesis is the as follows:

H3: individuals' investment decisions are not influenced by geographic boundaries

3.4 Currency: might it be a problem?

Currency has never been one of the main variables considered when analysing the customer experience on Kickstarter and the success factors of a project. therefore, no main theoretical basis is provided. Yet, some insight may be given by Kickstarter and its user itself.
A project's currency depends on the country whether the founder meets Kickstarter's *eligibility requirements* and launches the project.
Thanks to a feature introduced on Kickstarter in 2017, backers are given the opportunity to see an approximate conversion of the "original" goal in one of the above-mentioned currencies. Yet, creators may not change the project's currency.

On Kickstarter's FAQs (Frequently asked questions), it is possible to find the following questions:

> ***What currency will my pledge be collected in?***
> *Your pledge will be collected in the project's native currency.*
> *If you have questions about foreign currency transactions, we'd recommend reaching out to your financial institution for further clarification.*
>
> ***I pledged to a project based in a different country and was charged an amount different from my initial pledge. What happened?***
> *Your final pledge amount may differ based on the exchange rate value at the time a successfully funded project ends. We use the service Open Exchange Rates to determine the final value.*

The existence of explanation as the one above, lead to the assumption that there might be an impact of the currency on a project.

On the internet, a divergence of opinions emerges. On one hand, some project founders are concerned about the lower familiarity that people might have with a currency different from the one used in their country; on the other hand, some project founders believe that the conversion displayed automatically on the webpage is enough. It is a matter of fact that even though the conversion is displayed, some problems remain unsolved:

- The conversion fails to consider also the additional costs that may occur for a backer if he wants to pledge in a currency different from his own

- The conversion might not be enough to prevent confusion between users (e.g.: Sangjin Goodridge wrote: *"I pledged for a product that said $38. What I didn't realise was that was in UDS rather than AUD. When I checked my bank transactions I was a little surprised to see that nearly $50 AUD had been taken out")*[1]

This scenario leads to the belief that there might be some differences in individuals' preferences given by the currency. If an individual sees a project that asks funding in a currency different from the one he uses, he might be afraid to incur in some hidden costs and in missing to understand the exact amount of money he is investing.
Hence, the fourth hypotheses is:

H4: Individuals prefer to invest in their own currency and therefore projects that ask for diffused currencies are more successful

[1] Kickstarter, Campus Questions, *How important is USD currency and shipping cost for US backers?*

4. Descriptive statistics

The previous chapter focussed on the features of a project that might influence the individuals' investing decision.

In order to prove whether those hypotheses are verified or not, it is necessary to use a database that is significant and that offers the asked features for each project.

The purpose of this chapter is to provide a detailed insight of the database on which the subsequent analysis is based on. Hence, the following pages will illustrate the source and structure of the database, the main features and a deep examination obtained through descriptive statistics.

4.1 Dataset

The dataset contains 378.661 rows (projects). With a missing rate below 1%, all projects ever launched on Kickstarter until January 2018 where the starting point for this analysis.

The dataset is publicly available on Kaggle, a platform for predictive modelling and analytics competitions in which companies and users upload datasets to stimulate statisticians and data miners to compete and produce the best predictive models.

For each, project, the information listed below were available:

Variable	**Description**
Name	Name of the project
Category	Detailed categories
Main Category	15 main categories
Currency	Currencies used to fund projects
Deadline	Last day to fund a project
Goal	Amount that founders are willing to pledge
Launched	Day in which the project was launched on the platform
Pledged	Pledged amount
State	Outcome of the campaign
Backers	People who funded
Country	Country of origin of the project

Usd_pledged	Conversion in usd of the pledged column (Kickstarter)
Usd_pledged_real	Conversion in usd of the pledged column (Fixer.io API)
Usd_goal_real	Conversion in usd of the goal column (Fixer.io API)

One particular mention has to be done regarding the *State* of a project.

In fact, the aim of this research is to provide some insight on the influence of customer experience on the willingness to fund a project. Because the more customers want to invest in the project, the more the project might be successful, the success / fail outcome of projects seemed to be a good indicator for individuals' investment propension.

Hence, the dependent variable is the state of the project taken under a binomial form: success or fail. Nevertheless, the possible outputs of a project may be several:

- Success
- Fail
- Suspended
- Cancelled
- Live
- Undefined

Therefore, the dataset was "cleaned" in order to not consider projects that have a *State* different from Success or Fail. Hence, the final dataset contains 331.675 projects.

Now, it is possible to obtain the *success rate:* 60% of project launched on Kickstarter fail, while only 40% of projects is capable of achieving the pledging goal.

Therefore, the projects that are unsuccessful are 50% more than the projects that are successful.

The output of a project can be influenced by many variables.

The study will continue with a deeper descriptive analysis of those variables that are supposed to have some influence on the investment decision of customers.

4.2 Category

In Kickstarter, the projects are divided in 15 categories, that in the dataset figure as *main category.*

These categories are:

- *Art* for projects in visual art, dance, and performance
- *Comics,* for comics creators
- *Crafts,* for products made or to make by hand
- *Dance,* for dance events and performances
- *Design,* for decorative and innovative products
- *Fashion,* for stylish cloths and accessories
- *Film & Video,* for movies and videos
- *Food,* for foods, drinks, and places where to eat and drink
- *Games,* for any kind of game
- *Journalism,* for articles, blogs and magazines
- *Music,* for tours and CDs
- *Photography,* for photobooks and exhibitions
- *Publishing,* for books
- *Technology,* for technological gadget and devices
- *Theater,* for performances and shows

The above-mentioned “main” categories comprehend some “minor” categories:

- 13 for Art (Ceramics, Conceptual Art, Digital Art, Illustration, Installation, Mixed Media, Painting, Performance Art, Public Art, Sculpture, Textiles and Video Art)
- 5 for Comics (Anthologies, Comic Books, Events, Graphic Novels and Webcomics)
- 13 for Crafts Candles, Crochet, DIY, Embroidery, Glass, Knitting, Pottery, Printing, Quilts, Stationery, Taxidermy, Weaving, and Woodworking)
- 4 for Dance (Performances, Residencies, Spaces, and Workshops)
- 6 for Design (Architecture, Civic Design, Graphic Design, Interactive Design, Product Design, and Typography)
- 8 for Fashion (Accessories, Apparel, Childrenswear, Couture, Footwear, Jewelry, Pet Fashion, and Ready-to-wear)

- 19 for Film & Video (Action, Animation, Comedy, Documentary, Drama, Experimental, Family, Fantasy, Festivals, Horror, Movie Theaters, Music Videos, Narrative Film, Romance, Science Fiction, Shorts, Television, Thrillers, and Webseries)
- 12 for Food (Bacon, Community Gardens, Cookbooks, Drinks, Events, Farmer's Markets, Farms, Food Trucks, Restaurants, Small Batch, Spaces, and Vegan)
- 7 for Games (Gaming Hardware, Live Games, Mobile Games, Playing Cards, Puzzles, Tabletop Games, and Video Games)
- 5 for Journalism (Audio, Photo, Print, Video and Web)
- 18 for Music (Blues, Chiptune, Classical Music, Comedy, Country & Folk, Electronic Music, Faith, Hip-Hop, Indie Rock, Jazz, Kids, Latin, Metal, Pop, Punk, R&B, Rock, and World Music)
- 6 for Photography (Animals, Fine Art, Nature, People, Photobooks, and Places)
- 17 for Publishing (Academic, Anthologies, Art Books, Calendars, Children's Books, Comedy, Fiction, Letterpress, Literary Journals, Literary Spaces, Nonfiction, Periodicals, Poetry, Radio & Podcasts, Translations, Young Adult, and Zines)
- 15 for Technology (3D Printing, Apps, Camera Equipment, DIY Electronics, Fabrication Tools, Flight, Gadgets, Hardware, Makerspaces, Robots, Software, Sound, Space Exploration, Wearables, and Web)
- 7 for Theater (Comedy, Experimental, Festivals, Immersive, Musical, Plays, and Spaces)

For the study, the main categories were chosen as a variable, in order to provide a clearer division between the nature of the projects.

Category	Failed	Successful	Total	% Projects	% failed	% successful
Dance	1.235	2.338	3.573	1%	35%	65%
Theater	3.708	6.534	10.242	3%	36%	64%
Comics	4.036	5.842	9.878	3%	41%	59%
Music	21.752	24.197	45.949	14%	47%	53%
Art	14.131	11.510	25.641	8%	55%	45%
Games	16.003	12.518	28.521	9%	56%	44%
Film & Video	32.904	23.623	56.527	17%	58%	42%
Design	14.814	10.550	25.364	8%	58%	42%
Publishing	23.145	12.300	35.445	11%	65%	35%
Photography	6.384	3.305	9.689	3%	66%	34%
Fashion	14.182	5.593	19.775	6%	72%	28%
Food	15.969	6.085	22.054	7%	72%	28%
Crafts	5.703	2.115	7.818	2%	73%	27%
Journalism	3.137	1.012	4.149	1%	76%	24%
Technology	20.616	6.434	27.050	8%	76%	24%
Total	**197.719**	**133.956**	**331.675**	**100%**	**60%**	**40%**

Table 2: Successful projects by category

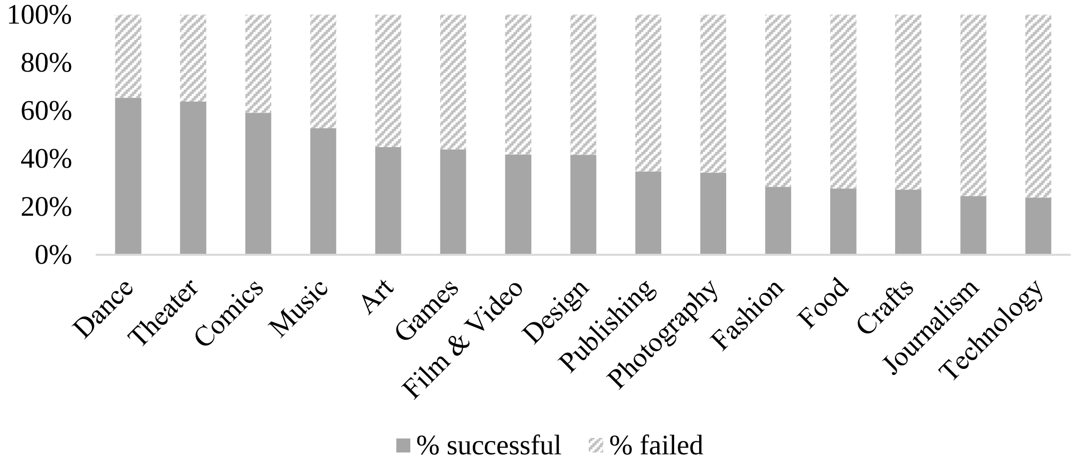

Figure 8: Graph of Successful projects by category

The table (

Table 2: Successful projects by category) and the graph (*Figure 8: Graph of Successful projects by category*) provide a deep insight on the distribution and the *performance* of Kickstarter projects.

The highest relative number of successful projects is in Dance (65%), Theater (64%), Comics (59%) and Music (53%). These categories are the only ones where the success rate is higher than the fail rate (see *Figure 9: Graph of number of projects by Category*).

Instead, the categories that perform the worst are Technology (24%), Journalism (24%) and Crafts (27%).

One interesting fact to highlight is that the categories that perform the best are not those with the highest relative number of projects. In fact, as shown by *Figure 9: Graph of number of projects by Category*, the widest categories for number of projects are Film & Video (17%), Music (14%) and Publishing (11%), that have a success rate of 42%, 53% and 35% respectively, not that different from the average rate of 40%.

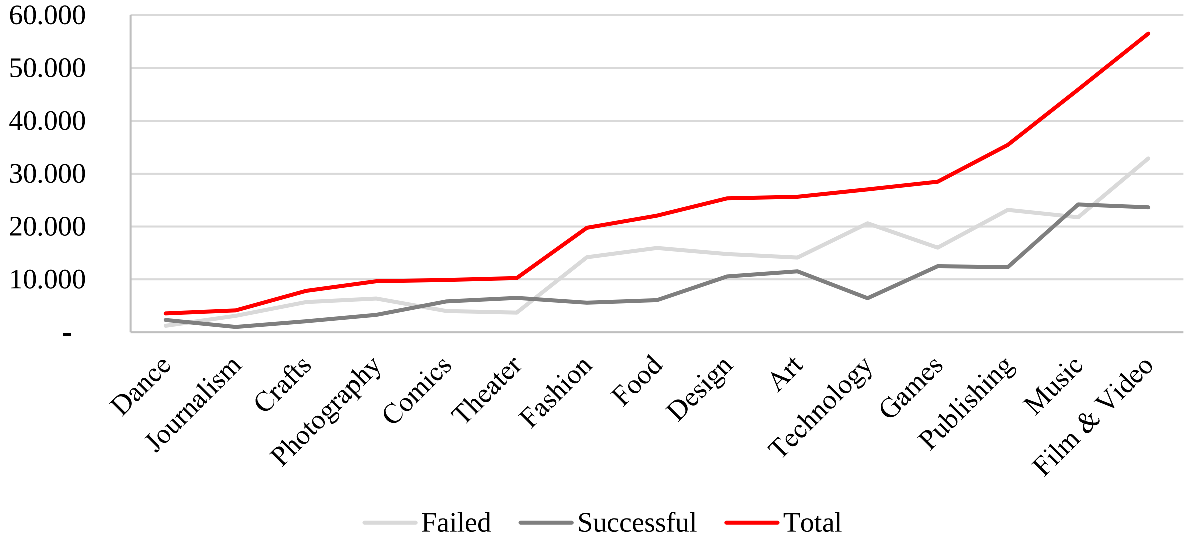

Figure 9: Graph of number of projects by Category

4.3 Goal

The goal of the project is the total amount that the founder wants to collect from backers. The creators of the campaign have the freedom to choose the value of the goal.

A higher goal is more difficult to achieve, but it attracts more backers because the project seems more ambitious (Sahlmann, 1997). On the other hand, setting a lower goal may give to creators the feeling that they will be able to achieve the needed amount more easily.

Goal range	Projects	% Projects	Failed	Successful	% failed	% successful
0-50	27.588	8,3%	12.135	15.453	44%	56%
50-100	27.893	8,4%	13.499	14.394	48%	52%
100-250	48.257	14,5%	24.219	24.038	50%	50%
250-500	64.181	19,4%	35.816	28.365	56%	44%
500+	163.756	49,4%	112.050	51.706	68%	32%
Total	**331.675**	**100,0%**	**197.719**	**133.956**	**60%**	**40%**

Table 3: Successful projects by goal range

As shown in *Table 3: Successful projects by goal range*, the projects have been divided into 5 categories, depending on the value of the goal. In order to make the comparison effective, all the currencies different from USD have been converted into USD.

The table above shows the distribution of projects accordingly to their goal. What is important to notice is that the success rate decreases with the increasing of the value of the goal. The projects with the higher goals are those that perform the worst (32% of success rate), while those with the lowest goal have the highest success rate (56%).

This seems to highlight a contrast between the creators and customer’s preferences.

With a wider classification, dividing the goal in three groups:

- *Low*: 0-100.000 USD
- *Medium*: 100.000-500 USD
- *High*: +500.000 USD

It appears clear how the frequency of projects differs:

- *Low*: 16,7%
- *Medium*: 33,9%
- *High*: 49,4%

Therefore, while creators prefer to ask for higher goals, customers seem to prefer to back projects that have a more realistic goal.

4.4 Country

The countries that figure in the database are the following:

- Australia (AU)
- Austria (AT)
- Belgium (BE)
- Canada (CA)
- Denmark (DK)
- France (FR)
- Germany (DE)
- Great Britain (GB)
- Ireland (IE)
- Hong Kong (HK)
- Italy (IT)
- Japan (JP)
- Luxembourg (LU)
- Mexico (MX)
- Netherlands (NL)
- New Zeeland (NZ)
- Norway (NO)
- Singapore (SG)
- Spain (ES)
- Sweden (SE)
- Switzerland (CH)
- United States (US)

Hence, 22 countries are involved in the study.

Country	Failed	Successful	Total	% projects	% failed	% successful
HK	261	217	478	0,1%	54,6%	45,4%
US	152132	109379	261511	78,8%	58,2%	41,8%
GB	17394	12081	29475	8,9%	59,0%	41,0%
SG	276	178	454	0,1%	60,8%	39,2%
DK	567	362	929	0,3%	61,0%	39,0%
FR	1615	998	2524	0,8%	64,0%	36,0%
NZ	826	448	1274	0,4%	64,8%	35,2%
SE	1002	509	1511	0,5%	66,3%	33,7%
CA	8238	4137	12375	3,7%	66,6%	33,4%
LU	38	19	57	0,0%	66,7%	33,3%
JP	16	7	23	0,0%	69,6%	30,4%
AU	4610	2011	6621	2,0%	69,6%	30,4%
IE	479	207	686	0,2%	69,8%	30,2%
BE	371	152	523	0,2%	70,9%	29,1%
CH	465	187	652	0,2%	71,3%	28,7%
MX	1015	396	1411	0,4%	71,9%	28,1%
NO	421	163	584	0,2%	72,1%	27,9%
DE	2505	937	3442	1,0%	72,8%	27,2%
ES	1383	492	1875	0,6%	73,8%	26,2%
NL	1795	619	2414	0,7%	74,4%	25,6%
AT	378	107	485	0,1%	77,9%	22,1%
IT	1932	439	2371	0,7%	81,5%	18,5%
Total	**197719**	**133956**	**331675**	**100,0%**	**59,6%**	**40,4%**

Table 4: Successful projects by Country

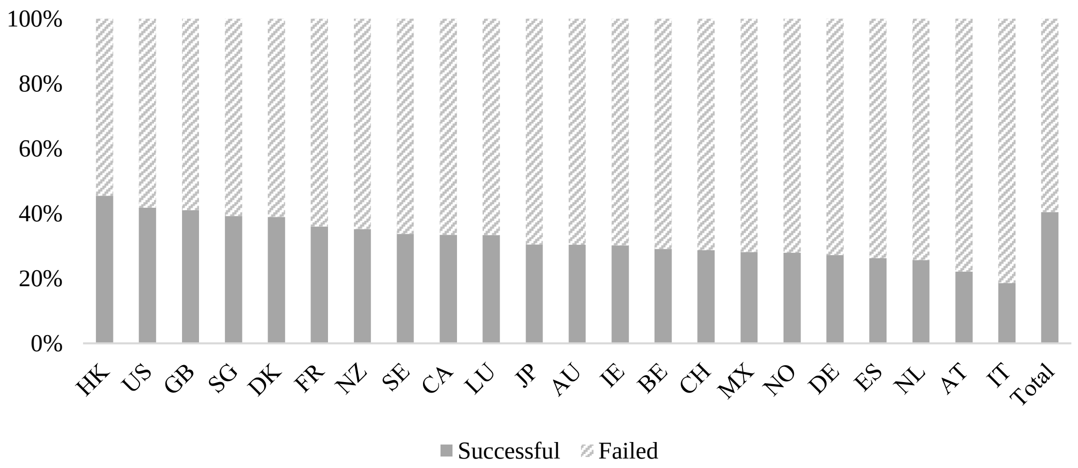

Figure 10: Graph of successful projects by currency

The table (
Table 4: Successful projects by Country) and the graph (*Figure 10: Graph of successful projects by currency*) above provide an insight on how the success rate changes accordingly with the country.

The countries that perform the best are Hong Kong, United States and Great Britain, with a success rate of 45,4%, 41,8%, and 41,0% respectively.

United States and Great Britain are also the countries where the largest amount of projects is launched. Of the considerer dataset of 331.675 projects, 261.511 come from US – accounting for 78,85% of projects launched on Kickstarter worldwide, and 29.475 come from Great Britain – accounting for 8,9% of projects worldwide. Instead, Hong Kong faces a different situation. In fact, "only" 478 projects were launched there. Yet, it has to be taken into account that Hong Kong had less time compared to other countries: in fact, Kickstarter opened the platform to this region only in 2016 (Lockett, 2016).

The countries that perform the poorest are Italy (18,5%), Austria (22,1%) and the Netherlands (25,6%).

One interesting fact to notice is that US, GB and HK are also the only three countries that perform above the average of 40%.

4.5 Currency

Currently, Kickstarter supports the following currencies:

- Australian Dollar (AUD)
- British Pound (GBP)
- Canadian Dollar (CAD)
- Danish Krone (DKK)
- Euro (EUR)
- Hong Kong Dollar (HKD)
- Japanese Yen (JPY)
- Mexican Peso (MXN)
- New Zealand Dollar (NZD)
- Norwegian Krone (NOK)

- Singapore Dollar (SGD)
- Swedish Krona (SEK)
- Swiss Franc (CHF)
- United States Dollar (USD)

Totally, 14 currencies are available.

Currency	Failed	Successful	Total	% projects	% failed	% successful
HKD	261	216	477	0,1%	55%	45%
USD	152132	109379	261511	78,8%	58%	42%
GBP	17395	12081	29476	8,9%	59%	41%
SGD	276	178	454	0,1%	61%	39%
DKK	567	362	929	0,3%	61%	39%
NZD	826	448	1274	0,4%	65%	35%
SEK	1001	509	1510	0,5%	66%	34%
CAD	8238	4137	12375	3,7%	67%	33%
JPY	16	7	23	0,0%	70%	30%
AUD	4610	2011	6621	2,0%	70%	30%
CHF	465	187	652	0,2%	71%	29%
MXN	1015	396	1411	0,4%	72%	28%
NOK	421	163	584	0,2%	72%	28%
EUR	10496	3882	14378	4,3%	73%	27%
Total	**197719**	**133956**	**331675**	**100,0%**	**60%**	**40%**

Table 5: Successful projects by Currency

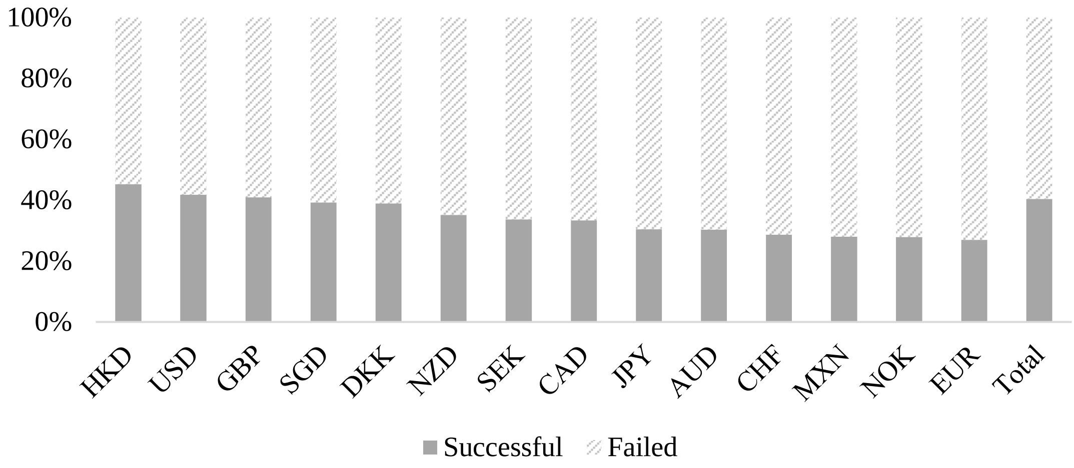

Figure 11: Graph of successful projects by Currency

The table (*Table 5: Successful projects by Currency*) and the graph (*Figure 11: Graph of successful projects by Currency*) above provide an insight on the success rate of Kickstarter projects based on Currency. The currencies with the highest success rate are Hong Kong Dollar (45%), US Dollar (42%) and GBP (41%). US Dollar and GBP are also the two currencies that are used the most, with a frequency of 78,8% and 8,9% respectively. These three currencies are also the only ones that perform above the 40% average

The Kickstarter projects that ask for funds in the following currencies are those that perform the worst: Euro (27%), Norwegian Krone (27%) and Mexican Peso (27%).

It appears clearly how the data for Currency and Country are similar. This is given by the fact that – exception made for Europe – there is a perfect correspondence between the two variables (see: *Table 6: Projects by Country and Currency*).

	AUD	**CAD**	**CHF**	**DKK**	**EUR**	**GBP**	**HKD**	**JPY**	**MXN**	**NOK**	**NZD**	**SEK**	**SGD**	**USD**	**Total**
AT					485										**485**
AU	6621														**6621**
BE					523										**523**
CA		12375													**12375**
CH			652												**652**
DE					3442										**3442**
DK				929											**929**
ES					1875										**1875**
FR					2524										**2524**
GB						29475									**29475**
HK					1		477								**478**

IE					686										**686**
IT					2371										**2371**
JP								23							**23**
LU					57										**57**
MX									1411						**1411**
NL					2414										**2414**
NO										584					**584**
NZ											1274				**1274**
SE												1511			**1511**
SG													454		**454**
US														261511	**261511**
Total	**6621**	**12375**	**652**	**929**	**14378**	**29476**	**477**	**23**	**1411**	**584**	**1274**	**1510**	**454**	**261511**	**331675**

Table 6: Projects by Country and Currency

In fact, excluding the one project launched in Hong Kong and selecting Euro as currency, the only currency where there emerge some differences in the dataset between countries, is Euro. In fact, Euro is used in 9 countries:

- Austria
- Belgium
- Germany
- Spain
- France
- Ireland
- Italy
- Luxembourg
- Netherlands

	Projects	Frequency
DE	3442	23,9%
FR	2524	17,6%
NL	2414	16,8%
IT	2371	16,5%
ES	1875	13,0%
IE	686	4,8%
BE	523	3,6%
AT	485	3,4%
LU	57	0,4%

Total	**14377**	**100%**

Table 7: Frequency of projects in European Countries

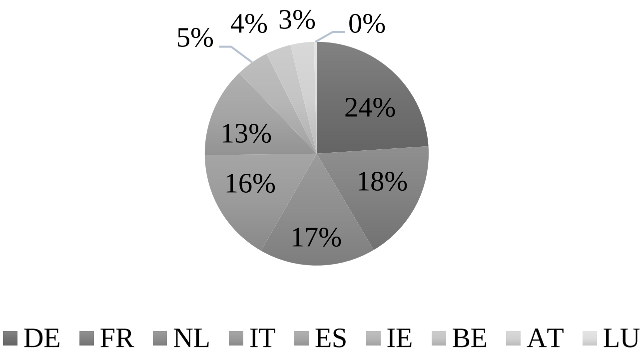

Figure 12: Graph of frequency of projects in European Countries

As shown by *Table 7: Frequency of projects in European Countries* and *Figure 12: Graph of frequency of projects in European Countries*, European countries that use Euro as a Currency and that launch the highest number of projects are Germany (23,9%), France (17,6%) and the Netherlands (16,8%). Instead, the European Countries with the lowest number of launched projects are Luxembourg (0,4%), Austria (3,4%) and Belgium (3,6%).

5. Methodology and Model

This chapter focusses on the methodology used to develop the model.

This is the core of the study. The literature review and the data analysis have provided the basis for developing the model presented and explained in this chapter.

The first part provides an overview of the preparation and cleaning of the dataset and of the used algorithm.

Afterwards, the second part analyses the dependent variable and the independent variables.

The third part provides an insight of the process of the model building

The considered variables are five: four independent variables and one dependent variable. The dependent variable is the fail / success outcome of a project, while the four dependent variables are (1) Category, (2) Goal, (3) Country and (4) Currency.

The predictive ML Model was tested in order to find the right combination of variables. Therefore, the best model happens to be the one built only on three variables: (1) Category, (2) Goal and (3) Country.

5.1 Preparing the Dataset

The dataset contains information about the features of the projects launched on Kickstarter until January 2018.

The aim of the research is to understand which projects customer prefer to invest in. In order to measure these preference, the success or fail outcome of a project has been chosen as *dependent variable.*

The dataset was provided with a further column with the following values:

- *1,* if the project was successful
- *0,* if the project failed

The considered independent variables are four:

- *Category*, categorical variable
- *Goal*, discrete variable
- *Currency*, categorical variable

- *Country*, categorical variable

Hence, three variables are categorical (category currency, and country), and one variable is discrete (goal).

Since the *dependent variable* is *dichotomic*, using a linear regression did not appear as the optimal solution.

The *logistic regression* is more appropriate. In fact, it permits to relate both discrete and categorical variables with a dichotomic variable.

The formula of the logistic regression is:

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

with

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

$\beta_0 =$ intercept

$\beta_1 =$ coefficient of x_1

$x_1 =$ dependent variable 1

$\beta_k =$ coefficient of x_k

$x_k =$ dependent variable k

The coefficients provide the log of the odds of each variable.

The aim of the research is to provide a predictive model in order to forecast whether a given project will have success based on customer experience and preferences.

Therefore, the research was developed with the implementation of Machine Learning through R, a software environment mainly used for statistical computing and graphics.

First, the dataset was “cleaned”, removing all the projects that had a *state* different from *successful* or *failed:* hence, projects that were *undefined, live, cancelled* or *suspended* were not considered. The figure below provides a sample of the dataset in R (see: *Figure 13: Kickstarter Dataset in R*).

	row.names	ID	name	category	main_category	currency	state	backers	country	usd.pledged	usd_pledged_real	usd_goal_real	SUCCFAILED
1	1	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	failed	0	GB	0.00	0.00	1533.95	0
2	2	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	failed	15	US	100.00	2421.00	30000.00	0
3	3	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	failed	3	US	220.00	220.00	45000.00	0
4	4	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	failed	1	US	1.00	1.00	5000.00	0
5	6	1000014025	Monarch Espresso Bar	Restaurants	Food	USD	successful	224	US	52375.00	52375.00	50000.00	1
6	7	1000023410	Support Solar Roasted Coffee & Green Energy! Sol>	Food	Food	USD	successful	16	US	1205.00	1205.00	1000.00	1
7	8	1000030581	Chaser Strips. Our Strips make Shots their B*tch!	Drinks	Food	USD	failed	40	US	453.00	453.00	25000.00	0
8	11	100004721	Of Jesus and Madmen	Nonfiction	Publishing	CAD	failed	0	CA	0.00	0.00	2406.39	0
9	12	100005484	Lisa Lim New CD!	Indie Rock	Music	USD	successful	100	US	12700.00	12700.00	12500.00	1
10	13	1000055792	The Cottage Market	Crafts	Crafts	USD	failed	0	US	0.00	0.00	5000.00	0
11	14	1000056157	G-Spot Place for Gamers to connect with eachother>	Games	Games	USD	failed	0	US	0.00	0.00	200000.00	0
12	15	1000057089	Tombstone: Old West tabletop game and miniatures >	Tabletop Games	Games	GBP	successful	761	GB	57763.78	121857.33	6469.73	1
13	16	1000064368	Survival Rings	Design	Design	USD	failed	11	US	664.00	664.00	2500.00	0
14	17	1000064918	The Beard	Comic Books	Comics	USD	failed	16	US	395.00	395.00	1500.00	0
15	18	1000068480	Notes From London: Above & Below	Art Books	Publishing	USD	failed	20	US	789.00	789.00	3000.00	0
16	19	1000070642	Mike Corey's Darkness & Light Album	Music	Music	USD	successful	7	US	250.00	250.00	250.00	1
17	20	1000071625	Boco Tea	Food	Food	USD	failed	40	US	1781.00	1781.00	5000.00	0
18	21	1000072011	CMUK. Shoes: Take on Life Feet First.	Fashion	Fashion	USD	successful	624	US	34268.00	34268.00	20000.00	1
19	22	1000081649	MikeyJ clothing brand fundraiser	Childrenswear	Fashion	AUD	failed	1	AU	0.00	0.81	2026.10	0
20	23	1000082254	Alice in Wonderland in G Minor	Theater	Theater	USD	failed	12	US	650.00	650.00	3500.00	0
21	24	1000087442	Mountain brew: A quest for alcohol sustainability	Drinks	Food	NOK	failed	3	NO	6.18	6.29	65.55	0
22	25	1000091520	The Book Zoo - A Mini-Comic	Comics	Comics	USD	successful	66	US	701.66	701.66	175.00	1
23	26	1000102741	Matt Cavenaugh & Jenny Powers make their 1st albu>	Music	Music	USD	successful	147	US	15827.00	15827.00	10000.00	1
24	27	1000103948	Superhero Teddy Bear	DIY	Crafts	GBP	failed	0	GB	0.00	0.00	17489.65	0
25	28	1000104688	Permaculture Skills	Webseries	Film & Video	CAD	successful	571	CA	43203.25	42174.03	15313.04	1
26	29	1000104953	Rebel Army Origins: The Heroic Story Of Major Gri>	Comics	Comics	GBP	successful	27	GB	167.70	160.60	142.91	1
27	30	100011318	My Moon - Animated Short Film	Animation	Film & Video	USD	successful	840	US	10120.00	57577.31	50000.00	1

Figure 13: Kickstarter Dataset in R

Machine Learning works because the computer “learns” from past experience and applies to new data the knowledge acquired in the past. Therefore, the dataset is divided in two sets of data, a *training set* and a *test set.*

- *Training set*

  It consists of 70% of the total rows in the dataset. The ML model is trained on the train data, to find relevant relationships between the independent variables and the dependent variable and to compute the coefficients

- *Test set*

  It consists of the remaining 30% of the dataset. The Machine Learning Model is tested with this data, in order to prove the validity of the developed model

Both the train and the test set are labelled: this means, the value of the dependent variable is known.
Therefore, it is possible to make a comparison between the output predicted by the Machine Learning Model and the real output.

Now – using only the training set – each variable is tested alone, in order to gain some deeper understanding of the relationship between the considered variable and the *independent variable*. But first, it is necessary to provide some further information about the dependent variables and why it has been chosen.

5.2 Variables Description

5.2.1 Dependent Variable

In order to choose the best dependent variable capable of reflecting the Customer Experience and preferences, the success or failure of the project seemed to be the best solution.

In fact, using the number of backers as dependent variable would not give a general perspective on the customer experience. In fact, a project could have many backers but with a low average investment. On the other hand, using the total pledged amount as dependent variable would not give any indications about the number of backers.

Instead, knowing if a project was successful or not, is more consistent with the aim of this research. In fact, neither a large number of backers or a high funded amount taken alone are comprehensive indicators of the Customer Experience. Being successful on Kickstarter, instead, reflects both the above-mentioned phenomenon, indicating individuals' engagement with the project.

5.2.2 Independent Variables

The four dependent variables are tested with a Logistic Regression in order to provide some deeper understanding of the analysed phenomenon and to investigate the validity of the hypotheses.

Main Category

The first independent variable to be tested is the category.
It is a categorical variable that assumes value 1 for the category of the project, 0 for any other category.
On R, the logistic regression

glm(formula = Y ~ X, family = binomial(link = "logit"))

where

Y = Dependent Variable

X = Main Category

provides the following results:

	Estimate	Std. Error	Z value	Pr(>|z|)	
(Intercept)	-0.20960	0.01500	-13.971	< 2e-16	***
Comics	0.57434	0.02867	20.036	< 2e-16	***
Crafts	-0.77472	0.03412	-22.707	< 2e-16	***
Dance	0.87390	0.04494	19.446	< 2e-16	***
Design	-0.11110	0.02135	-5.205	1.94e-07	***
Fashion	-0.71652	0.02409	29.739	< 2e-16	***
Film & Video	-0.71652	0.01815	-6.760	1.38e-11	***
Food	-0.76493	0.02344	-32.630	< 2e-16	***
Games	-0.76493	0.02071	-1.756	0.0792	.
Journalism	-0.91059	0.04552	20.004	< 2e-16	***
Music	0.31503	0.01870	16.846	< 2e-16	***
Photography	-0.43216	0.02978	-14.510	< 2e-16	***
Publishing	-0.41868	0.02005	-20.878	< 2e-16	***
Technology	-0.95436	0.02272	-42.014	< 2e-16	***
Theater	0.78237	0.02893	27.047	< 2e-16	***
Signif.codes	*0 '***'*	*0.001 '**'*	*0.01 '*'*	*0.05 '.'*	

Table 8: Coefficients of Logistic Regression with Main Category as Independent Variable

(Reference Group: Art)

The variables are significant, meaning that there is a relationship between the category and the success of a project. Only a few categories have positive log odds, and they are, in decreasing order: Dance, Theater, Comics and Music.

The boxplot in *Figure 14* has the main categories on the horizontal axis and a success indicator (% of pledged amount on the goal, if >100% then successful) on the vertical axis.

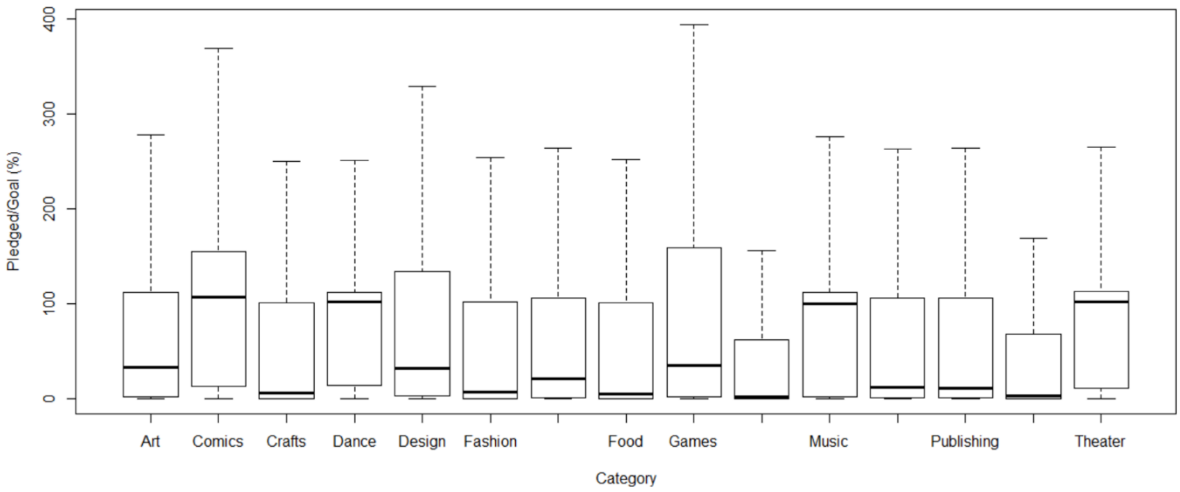

Figure 14: Boxplot of the performance of Kickstarter projects by Category

The boxplots in *Figure 14* do not consider outliers and is therefore a good graphic representation of the projects' performance.

What is interesting, is the strong negative relationship between Technology and performance. How may it be that Technology may not perform as well as other technologies on a platform that exists because of the technological and digital progress itself?

This might be because lately and on a global scale Governments have become more receptive towards technological start-ups.

For example, in Italy the Decree *D.L. 179/2012* introduced the figure of the *innovative start up*, in order to promote the technological development of the country. the decree sets some important advantages – especially of fiscal nature – that could be sufficient for the best start-ups and projects to grow without the help of a crowdfunding platform. Furthermore, those people that use an advanced and alternative platform as Kickstarter are maybe the most sensitive to the quality of technology, and thus prefer more advanced products of well-known companies. Nevertheless, the available data is not enough to continue with further investigation in this direction.

Goal

The second independent variable to be tested is the goal.

On R, the logistic regression

glm(formula = Y ~ X, family = binomial(link = "logit"))

where

Y = Dependent Variable

X = Goal

provides the following results:

	Estimate	Std. Error	Z value	Pr(>|z|)	
(Intercept)	-0.1469	0.005087	-28.88	<2e-16	***
Goal	-0.00001731	2.465e-07	-70.19	<2e-16	***
Signif.codes	*0 '***'*	*0.001 '**'*	*0.01 '*'*	*0.05 '.'*	

Figure 15: Coefficients of Logistic Regression with Goal as Independent Variable

The coefficients in *Figure 15* highlight that, the higher the goal, the lower the probability of success.

A possible explanation is that *people prefer winners, not losers.*

When people have to choose between supporting a new project that has not achieved the pledging goal yet and a project that has already passed the goal, people chose the latter.

Providing a numerical example:

- Project A: has reached 400.000 $ with a funding goal of 100.000 $
- Project B: has reached 400.000$ with a funding goal of 500.000 $

Based on the numbers above, on the Kickstarter webpage the following percentages would appear:

- For Project A: 400% funded
- For Project B: 80% funded

The difference in the proportion between pledged and needed is impressive, even though the pledged amount is exactly the same.

Some support to this theory is provided by the fact that there are projects that are able of raising investments far above their goal, such as *Exploding Kittens,* that raised 8,782,571 $ with a 10.000$ goal (87.825%), and *Pebble Time - Awesome Smartwatch, No Compromises,* that raised 20,338,986 $ with a 500.000 $ goal (4.067%).Anyway, investors might perceive high goals are a symptom rather of presumptuousness than of ambitiousness, and hesitate in front of something that seems too risky.

So, is there any correlation between the goal and the possibility of a successful campaign?

Country

The third independent variable to be tested is the Country.

It is a categorical variable that assumes value 1 for the country of the project, 0 for any other country.

On R, the logistic regression

glm(formula = Y ~ X, family = binomial(link = "logit"))

where

Y = Dependent Variable

$X =$ Country

provides the following results:

	Estimate	Std. Error	Z value	Pr(>|z|)	
(Intercept)	-1.15336	0.13071	-8.824	< 2e-16	***
AU	0.32289	0.13459	2.399	0.016437	*
BE	0.14937	0.17697	0.844	0.398660	
CA	0.47694	0.13267	3.595	0.000324	***
CH	0.20454	0.16736	1.222	0.221647	
DE	0.17064	0.13847	1.232	0.217839	
DK	0.74012	0.15354	4.820	1.43e-06	***
ES	0.06391	0.14542	0.439	0.660308	
FR	0.62383	0.13980	4.462	8.11e-06	***
GB	0.79446	0.13148	6.043	1.52e-09	***
HK	1.13394	0.17330	6.543	6.01e-11	***
IE	0.32371	0.16388	1.975	0.048229	*
IT	-0.29762	0.14480	-2.055	0.039843	*
JP	0.19785	0.54223	0.365	0.715195	
LU	0.53432	0.35634	1.499	0.133746	
MX	0.19022	0.14894	1.277	0.201561	
NL	0.12696	0.14191	0.895	0.294166	
NO	0.18052	0.17208	1.049	0.294166	
NZ	0.51185	0.14846	3.448	0.000565	***
SE	0.43871	0.14630	2.999	0.002711	**
SG	0.56941	0.17468	3.260	0.001115	**
US	0.82369	0.13080	6.297	3.03e-10	***
Signif. codes	*0 '***'*	*0.001 '**'*	*0.01 '*'*	*0.05 '.'*	

Figure 16: Coefficients of Logistic Regression with Country as Independent Variable

(Reference Group: AT)

From *Figure 16* it seems as only few countries have a relationship with the success of the project.

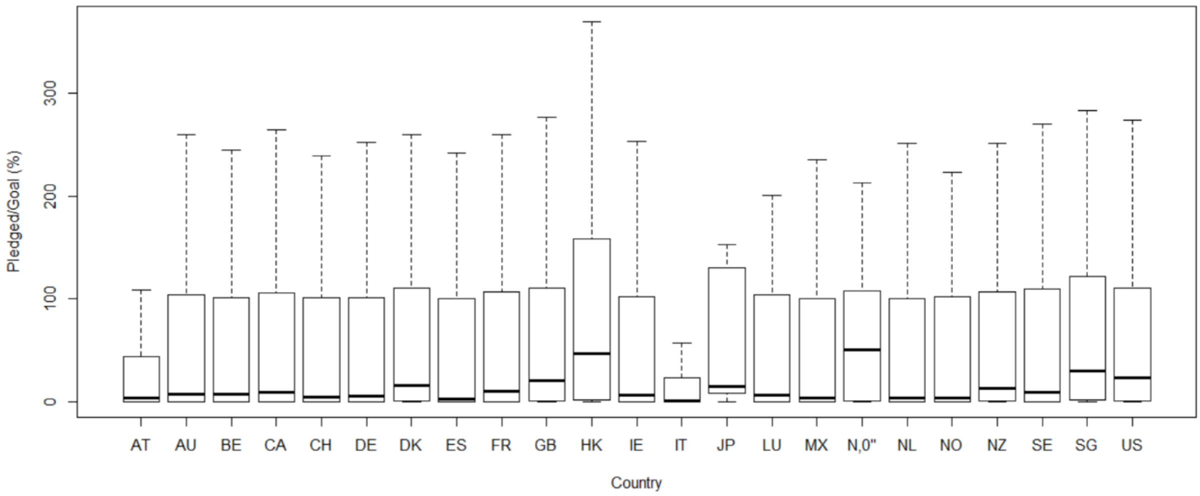

Figure 17: Boxplot of the performance of Kickstarter projects by Country

The boxplot in *Figure 17* relates the country with the percentage if funding on the initial goal (if > 100, then the project was successful).

The boxplot shows more precisely the gap that exists between Italy and all the other countries.

It appears that there are differences in preferences between countries. This might be related to the perception that customers have about different countries. For instance, a significant sample is given by Italy's situation, the country that performs the wost.

It is possible to suppose that this difference depends also on the situation in which Italy is living: far beyond other European Countries in terms of economic strength and technological revolution.

In international Marketing, the phenomenon of the *country of origin effect* has been discussed widely (De Nisco et al., 2016), with the main idea that the country of origin may have tremendous influence on the acceptance and success of products (Dichter, 1962).

Italy has an international appeal in *food, shoes, leather goods, wines and liquors, clothing* and *home furniture* (De Nisco et al., 2016). And technology? Technology in not one of the sectors for which the "Made in Italy" is known. This means, before changing the customer experience on Kickstarter, it might be necessary to change the perception of Italy abroad.

The European country that performs the best is UK.

A possible explanation for this phenomenon might be the *maturity of the UK Market*. In fact, UK is a much more mature market compared to other European countries, and this is reflected by the fact that approximately 10% of all Kickstarter projects are launched in this country (78% of projects are launched in the US). The Alternative Finance Maturity Index developed by CrowdfundingHub (2016) in the *Current State of Crowdfunding in Europe Report,* shows how UK is the only European Country that, exception made for *Consumer Interests,* distinguishes itself in a positive manner in the other fourteen research areas (*Degree of Organization, Volumes Diversity in Platforms, Level of Activity, Cross Border Activity, Approach Banking Industry, Approach Government, Donation Based Crowdfunding, Reward Based Crowdfunding, P2P Lending, Crowdfunding, Equity Based Crowdfunding, Access to Finance SME's, Registration Obligations, and Tax Reliefs*). This means that it might be plausible to assume that UK attracts not only more projects but also more backers.

From an International perspective, it is interesting to how the boxplot highlights the outstanding performance of project launched in Hong Kong. This might be linked to fact that Hong Kong is trying to boost the economy through programs and initiatives linked with innovation, new projects and start-ups. For instance, the *Hong Kong Federation of Youth Groups (HKFYG)* is a service organization established in 1960 that provides *opportunities and facilities for the social, educational, cultural, emotional and physical development of young people*[2].

Currency

The fourth independent variable to be tested is the currency.
It is a categorical variable that assumes value 1 for the currency of the project, 0 for any other currency.
On R, the logistic regression

glm(formula = Y ~ X, family = binomial(link = "logit"))

where
Y = Dependent Variable
X = Currency

provides the following results:

[2] HKFYG, https://hkfyg.org.hk

	Estimate	Std. Error	Z value	Pr(>|z|)	
(Intercept)	-0.82510	0.03177	-25.972	< 2e-16	***
CAD	0.12886	0.03906	3.299	0.00097	***
CHF	-0.05861	0.10862	-0.540	0.58949	
DKK	0.37240	0.08617	4.322	1.55e-05	***
EUR	-0.17902	0.03891	-4.600	4.22e-06	***
GBP	0.45933	0.03479	13.202	< 2e-16	***
HKD	0.74555	0.11516	6.474	9.53e-11	***
JPY	0.13195	0.46400	0.284	0.77612	
MXN	-0.08282	0.07669	-1.080	0.28018	
NOK	-0.04445	0.11038	-0.403	0.68718	
NZD	0.14870	0.07759	1.916	0.05531	.
SEK	0.13623	0.07272	1.873	0.06103	.
SGD	0.39223	0.12144	3.230	0.00124	**
USD	0.49358	0.03212	15.367	< 2e-16	***
Signif.codes	*0 '***'*	*0.001 '**'*	*0.01 '*'*	*0.05 '.'*	

Table 9: Coefficients of Logistic Regression with Currency as Independent Variable

(Reference group: AUD)

From *Table 10* it emerges that not every currency seems to have a relationship with the success rate of a project.

Surprisingly, using Euro as currency lowers the log of the odds of having a successful project, while GBP increases it.

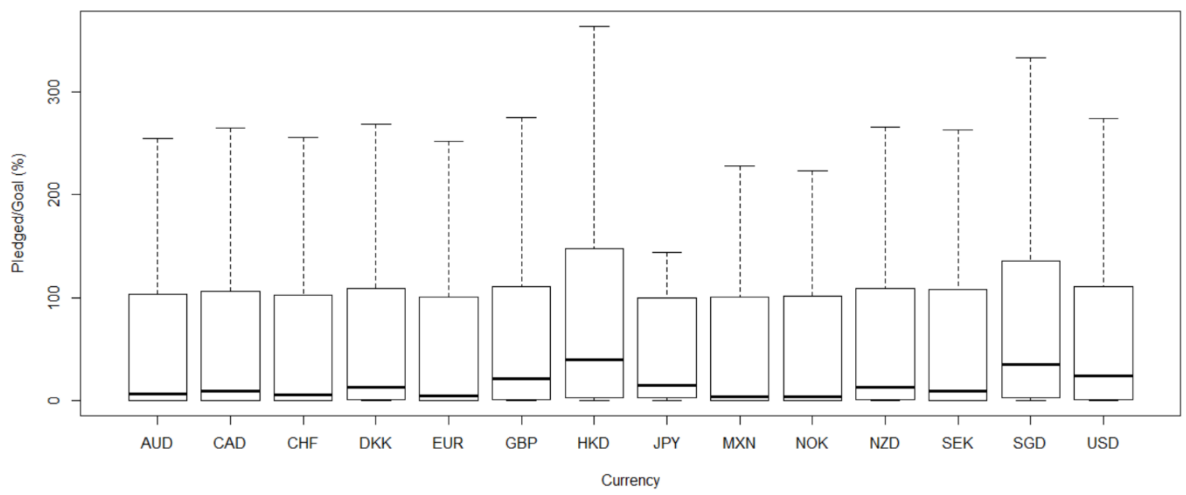

Figure 18: Boxplot of the performance of Kickstarter projects by Currency

The currency that performs the worst is Euro, yet this does not provide specific information of Italy, because Euro is used in 9 of the analysed countries.

The results are quite similar to the ones obtained drawing a logistic regression with *Country* as independent variable: GDP and HKD, the currencies of UK and Hong Kong respectively, perform the best. Instead, Euro is the currency that performs the worst.

5.3 Building the Model

Now that each Independent Variable has been analyzed individually, it is possible to run the ANOVA test to compare the consistency of different models built through the combination of variables.

The ANOVA test compares the models, providing useful indication about "added value" that a new independent variable brings to the model.

First, the following two models are compared:

- Model 1: $logit(p) = \beta_0 + \beta_1 Category$
- Model 2: $logit(p) = \beta_0 + \beta_1 Category + \beta_2 Goal$

The ANOVA test provides the following results:

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	
Model 1	232357	302738				
Model 2	232356	295059	1	7679.4	< 2.2e-16	***
Signif. codes	*0 '***'*	*0.001 '**'*	*0.01 '*'*	*0.05 '.'*		

Table 10: ANOVA Test Model 1 and Model 2

Table 10 shows that by adding *Goal* to the Model the deviance drops by 7679,4. Therefore, Model 2 fits to the data significantly better than the Model containing only the Category

Now, it is possible to do a further step, adding another independent variable, Country, to build a new Model, that will be called Model 3. The comparison between Model 2 and Model 3, where

- Model 2: $logit(p) = \beta_0 + \beta_1 Category + \beta_2 Goal$
- Model 3: $logit(p) = \beta_0 + \beta_1 Category + \beta_2 Goal + \beta_3 Country$

provides the following results:

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	
Model 2	232356	295059				
Model 3	232334	294098	22	960.98	< 2.2e-16	***
Signif. codes	*0 '***'*	*0.001 '**'*	*0.01 '*'*	*0.05 '.'*		

Table 11: ANOVA test Model 2 and Model 3

Adding *Country* as an independent variable (see *Table 11*), the deviance is reduced by a value of 960,98. Model 3 is significant, and it fits to the data better than Model 2.

The last independent variable to test in the Model is the Currency. Therefore, the comparison between Model 3 and Model 4, where

- Model 3: $logit(p) = \beta_0 + \beta_1 Category + \beta_2 Goal + \beta_3 Country$

- Model 4: $logit(p) = \beta_0 + \beta_1 Category + \beta_2 Goal + \beta_3 Country +$
$+ \beta_4 Currency$

provides the following results

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
Model 3	232334	294098			
Model 4	232328	294095	6	2.6701	0.849
Signif. codes	*0 '***'*	*0.001 '**'*	*0.01 '*'*	*0.05 '.'*	

Table 12: ANOVA test Model 3 and Model 4

Table 12 shows that adding the fourth independent variable (*Currency*), the deviance is not reduced significantly and that Model 4 does not fit the data.

Therefore, the best and final Model is Model 3.

Hence, the final predictive Machine Learning Model consists of three independent variables and is as follows:

$$logit(p) = \beta_0 + \beta_1 Category + \beta_2 Goal + \beta_3 Country$$

Where *Category* and *Country* are Categorical Variables, and *Goal* is a discrete variable.

The coefficients of the predictive Model figure in *Table 13: Predictive Model*.

	Estimate	Std. Error	Z value	Pr(>|z|)	
(Intercept)	-0.6089	0.1331	-4.575	4.76e-06	***
Comics *(x_1)*	0.5804	0.02904	19.985	< 2e-16	***
Crafts *(x_1)*	-0.8133	0.03405	-23.888	< 2e-16	***
Dance *(x_1)*	0.8062	0.04476	18.011	< 2e-16	***
Design *(x_1)*	0.06186	0.0 2194	2.819	0.004817	**
Fashion *(x_1)*	-0.6502	0.02433	-26.724	< 2e-16	***
Film & Video *(x_1)*	0.005492	0.01846	0.298	0.766075	
Food *(x_1)*	-0.5937	0.02384	-24.904	< 2e-16	***
Games *(x_1)*	0.1229	0.02114	5.813	6.15e-09	***
Journalism *(x_1)*	-0.8214	0.04575	-17.956	< 2e-16	***
Music *(x_1)*	0.2968	0.01887	15.724	< 2e-16	***
Photography *(x_1)*	-0.4556	0.02990	-15.236	< 2e-16	***
Publishing *(x_1)*	-0.4146	0.02020	-20.528	< 2e-16	***
Technology *(x_1)*	-0.5802	0.02350	-24.692	< 2e-16	***
Theater *(x_1)*	0.7880	0.02932	26.881	< 2e-16	***
Goal *(x_2)*	-0.00001539	0.0000002490	-61.810	< 2e-16	***
AU *(x_3)*	0.01714	0.1363	1.257	0.208577	
BE*(x_3)*	0.1555	0.1773	0.877	0.380443	
CA*(x_3)*	0.2789	0.1343	2.077	0.037788	*
CH*(x_3)*	0.2560	0.1708	1.499	0.133995	
DE*(x_3)*	0.09292	0.1404	0.662	0.508065	
DK*(x_3)*	0.4930	0.1567	3.147	0.001650	**
ES*(x_3)*	0.02347	0.1471	0.160	0.873228	
FR*(x_3)*	0.4908	0.1422	3.453	0.000555	***
GB*(x_3)*	0.5467	0.1331	4.109	3.98e-05	***
HK*(x_3)*	0.9364	0.1742	5.376	7.63e-08	***
IE*(x_3)*	0.2752	0.1667	1.651	0.098773	.
IT*(x_3)*	-0.3321	0.1473	-2.254	0.024176	*
JP*(x_3)*	-0.1655	0.5987	-0.277	0.782156	
LU*(x_3)*	0.4492	0.3926	1.144	0.252568	
MX*(x_3)*	-0.1084	0.1507	-0.719	0.472260	
NL*(x_3)*	0.09182	0.1438	0.639	0.523006	
NO*(x_3)*	-0.1039	0.1753	-0.593	0.553445	
NZ*(x_3)*	0.4083	0.1505	2.713	0.006674	**
SE*(x_3)*	0.2689	0.1483	1.813	0.069824	.
SG*(x_3)*	0.4510	0.1787	2.525	0.011581	*
US*(x_3)*	0.5721	0.1323	4.323	1.54e-05	***
Sign. codes	*0 '***'*	*0.001 '**'*	*0.01 '*'*	*0.05 '.'*	

Table 13: Predictive Model

Deviance Residuals

Min	1Q	Median	3Q	Max
-1.5766	-1.0093	-0.7643	1.1970	8.4904

Table 14: Deviance Residuals of the Predictive Model

5.4 Testing the Model

Now that the model has been developed, two further steps have to be done:

- Testing the Model on the test set
- Measuring the accuracy of the prediction made on the test set

The prediction has a value between 0 and 1 and expresses the probability of success of a given project in the dataset. For instance, if a project has a *Predicted value* of 0,3, it means that it has 30% of success of achieving the funding goal.

Eventually, a new dichotomic variable, *Result*, is added to the dataset.
The column *Result* has value:

- *1,* if the probability of success is > 0,5
- *0,* if the probability of success is < 0,5

In order to calculate whether the predicted result matches the true outcome of the project, the new dichotomic variable *Match* is added to the dataset.
It has value:

- *1,* if the value in the column *Result* matches with the real outcome of the project
- *0,* if the value in the column *Result* does not match with the real outcome of the project

For instance, the project *Wheeling, Careening, and Wandering through the EMP* had a goal of $ 850. The model predicted that it had almost 90% probability of being successful, and therefore the value *1* was displayed in column *Result.*

The project collected $ 875, hence it was successful. The value in the *Result* column matched with the real outcome, hence the value *1* figures in the column *Match.*

	row.names	ID	name	SUCCFAILED	percentage_of_funding	Pred	Result	match
1	14	1000056157	G-Spot Place for Gamers to connect with eachother>	0	0.000000e+00	-3.098831e+00	0	1
2	16	1000064368	Survival Rings	0	2.656000e+01	5.927155e-03	0	1
3	17	1000064918	The Beard	0	2.633333e+01	5.044934e-01	1	0
4	18	1000068480	Notes From London: Above & Below	0	2.630000e+01	-4.937655e-01	0	1
5	19	1000070642	Mike Corey's Darkness & Light Album	1	1.000000e+02	2.723362e-01	0	0
6	27	1000103948	Superhero Teddy Bear	0	0.000000e+00	-1.154996e+00	0	1
7	33	1000120151	Feather Cast Furled Fly Fishing Leaders	1	1.000000e+02	-3.588046e-01	0	0
8	36	1000129669	Squatch Watchers	0	1.373370e+01	-1.819942e-01	0	1
9	37	1000131947	Arrows & Sound Debut Album	1	2.160335e+02	2.122881e-01	0	0
10	44	1000170964	Penny Bingo Playing Card Game fun for the whole f>	0	5.706667e+01	7.971153e-02	0	1
11	48	1000184224	Star and the Snowman	1	1.175125e+02	-1.499686e-01	0	0
12	52	100020143	H2O The Underwater Portraits	0	6.315789e-02	-6.354282e-01	0	1
13	54	1000212914	A Teacher's Travels in China	1	1.216543e+02	-5.473578e-01	0	0
14	58	1000227361	UNDERWATER	1	1.016842e+02	-1.739878e-01	0	0
15	64	1000235643	HIIT Bottleâ„¢	1	8.333200e+02	-8.544335e-01	0	0
16	68	1000256230	Shreddit - Privacy on Reddit	0	0.000000e+00	-6.767990e-01	0	1
17	70	1000257563	"Flying" Carpets from Azerbaijan, Iran and Turkey>	0	4.004453e-02	-9.984985e-01	0	1
18	79	1000320473	Uncommon Rhythm - Season One	0	3.505051e+01	-4.974466e-01	0	1
19	83	1000332383	Road to the Shire	1	1.011250e+02	-8.591738e-02	0	0
20	84	1000333671	Spiral Electric Skylab Recording	1	3.080000e+02	2.683330e-01	0	0
21	93	1000354338	"Little Shop of Horrors" at the Browncoat Theatre!	1	1.037500e+02	7.111046e-01	1	1
22	96	1000386601	Los Angeles International Student Film Festival	1	1.288660e+02	-4.588535e-02	0	0
23	98	1000392220	The Girls Bathroom	1	1.126495e+02	-5.389176e-02	0	0
24	100	1000399155	New Lasagna	0	0.000000e+00	-6.943054e-01	0	1
25	101	1000405957	Just Another Reality Show: Season 3	1	1.512000e+02	-2.987254e-02	0	0
26	107	10004373	Bad Example (A Southern Trailer Park Fairy Tale) >	0	1.222222e+00	-5.068919e-02	0	1
27	119	1000491830	The Legendary YD 2016 Tour	0	1.581028e-01	1.750584e-01	0	1

Figure 19: Sample of the predictions of the ML Model

Now, to determine the *Accuracy* of the model, that means, the probability that the model predicts correctly the success or failure of a project, we compute the ratio between the number of correct prediction and the total number of projects in the test set: the model has 61% of probability of predicting correctly the outcome of a Kickstarter campaign.

6. Discussion

In this section, the data and analyses provided in the previous chapters are the basis to develop conclusions and to test the validity of the initial four hypotheses.

The results obtained through the deep analysis of the dataset have been enough to develop important conclusions

Eventually, the chapter does a focus on the limitations of the research in order to draw guidelines for further investigation.

6.1 Main Findings

In the accuracy test, the Machine Learning Model performs 61%: this means, the model is able to predict 6 times out of 10 if the project will be successful or not. Considering that the prediction power of the model depends solely on three independent variables, it is a significant output for the experiment. In fact, those three variables that remain constant during the whole campaign, are demonstrated to influence whether the project will be successful or not.

In its 2017 Security Review Report, Android shares the performance results of its Machine Learning algorithms: they were able to successfully detect 60,3% of Potentially Harmful Applications (PHAs) identified by Google Play Protect in 2017.

Comparing the results of the experiments to those of Android, it becomes clearer how tough it is to perform predictions with Machine Learning, even for a tech giant.

This happens because there are multiple features and aspects that may influence and determine an output, and considering and interpreting properly those variables requires a huge work.

The results of the analysis on the variables and the Predictive Model provide enough evidences to draw conclusions on the four hypotheses developed and suggested in this Master Thesis.

- *H1: Individuals prefer to become "backers" of campaigns that are mostly about creative projects*

Hypothesis 1 is confirmed by the fact the categories that perform the best are those that offer intangible and artistic products.

This categorical variable is significant when analysed in the logistic regression and also in the Model.

The relevance of this variables is highlighted also from the high improvement that it offers to the Model, reducing the Deviance in a consistent way.

There are no notable differences in the coefficients of this categorical variable between the logistic regression *(*

Table 2: Successful projects by category*)* and the Predictive Model.

The categories that perform the best remain *Dance*, *Theater*, *Comics* and *Music*. A notable difference is that *Film & Video* is no more significant in the Predictive Model.

This may happen because this variable suffers from the influence of the other categorical variable, *Country*. In fact, there might be an evaluation of aspects related to the geography of the project, especially culture and language. In fact, a customer might consider to invest in those film and video production that are made in a language he understands.

- *H2: individuals are influenced in their investing decision by the value of the goal of a project*

Hypothesis 2 is tested with the Model and is true.

Differently from the findings of other studies (Sahlmann, 1997; Rakesh, 2015), it emerged that not only creators prefer realistic goals, even backers do. A possible explanation might be that lower goals are just easier to achieve because funding from friends and family are enough.

Yet, the relationship that occurs between *Goal* and success of a project appears to be only slightly negative: in fact, on average for each one dollar increase in the goal, the log of the odds is reduces by -0.00001539 on average given all the other variables as constant.

Therefore, it is possible to suppose that the results provided by the model depends on to opposite phenomenon happening. On the one side, it is statistically true that a lower goal is easier to achieve. On the other hand, customers prefer to invest in ambitious projects. Hence, the final result is that even though the relationship between the dependent and the independent variable is negative, the value of the coefficient is very low.

- *H3: individuals' investment decisions are influenced by geographic boundaries*

Hypothesis 3 is verified by the Model but is not true.

In fact, relevant differences occur between projects that are launched in different countries. The influence of the *Country* on the Customer Experience is very strong in the Logistic Regression, and remain strong also in the final Predictive Model.

This might depend on the fact that customer “trust” other countries more than other. Furthermore, it may be that previous successful projects of a country attract customers to invest again for a project of the same creator or of the same country at least.

The model offers no insight on this argument, but leaves space for further investigation.

- *H4: Individuals prefer to invest in their own currency and therefore projects that ask for diffused currencies are more successful*

Hypothesis 4 is verified and is not true.

Even though it seems to have a strong relationship with customer preferences, the categorical variable has no significance when put in a Model with other variables (as happens in *Table 12: ANOVA test Model 3 and Model* 4). This means that people do not care anymore in which currency their investment will be collected and converted in. Probably, the approximate value of the investment in their own currency is enough to permit them to choose whether to pledge or not, without being a determinant element for their final choice.

6.2 Limitations and Recommendations for further studies

Overall, it is clear that the customer experience on Kickstarter depends on many factors that have not been taken in consideration during the last part of this dissertation and in the model.

Navigating on the internet, approaching an online Crowdfunding platform and deciding whether or not to trust a project and funding is a long process where many factors can concur in influencing the final result. Yet, even though the model is built upon “only” three variables, it is worthful enough to reduce deviance consistently.

Obviously, elements such as presentation video, photo, description, name of the project and other visual elements are part of the customer experience of potential investor from the moment they land on the webpage of a specific project. Nevertheless, it is often necessary to have some time in order to gain further information on the project through the interaction with those elements, because their message does not arrive as immediately as it happens with the numeric and concise information displayed on the webpage of the project.

Hence, the result of the experiments highlights two limits.

On one hand, the validity of the model could increase with the inclusion of further independent variables. In fact, the process that happens from the moment in which a customer opens a crowdfunding webpage to the final investment is long, complex and with many facets, not considered exhaustively by the four independent variables analysed for the model building.

On the other hand, implementing advanced algorithms and machine learning techniques may help to improve the quality of the prediction.

In fact, Machine Learning is experimenting a continuous and fast development, offering new and sophisticated tools whose implementation may bring a positive impact in the accuracy of the model. For instance, sentiment analysis or NLP might increase the understanding of the perception of title and description for the potential investors.

Even though the potential improvement margin is consistent, the model is still satisfying. Yet, the simultaneous development of the two above mentioned possibilities – new variables and advanced Machine Learning techniques – might lead to a model that helps founders to draw and design their campaign in the way that mostly increases their probability of success.

Conclusion

The results of the study provide some findings that might be useful for creators' in order to know in advance whether the project they are going to launch will deliver an appealing Customer Experience and attract individuals' investments.

In fact, from the Predictive Model it emerged that some features deliver a higher Customer Experience, hence increasing individuals' willingness to invest in a project.

Previous researches were more focussed on one specific feature, such as the description (Mitra et al., 2014), the frequency of updates (Xu et al., 2014), the appearance in the "Projects We Love" Category (Gutsche et al., 2018), the founders' friends on Twitter and Facebook (Greenberg et at., 2013) and publicly available Tweets (Etter et al., 2013), their non-profit nature (Lambert et al., 2010).

The study tried to fill a gap left open by previous research: trying to analyse the success of a project not from a product and campaign but from a Customer Experience perspective, highlighting which project's features have a positive impact on individuals' willingness to invest.

The study takes into account only those variables that remain constant during the whole campaign and that are visible on the main page of each Kickstarter project.

The findings of the research prove that customers are more attracted towards projects that offer artistic contents, with realistic goals, and that they are influenced by the creators' country of origin.

Yet, the model presents some limitations, leaving space for further studies that could improve the quality of the Model by adding further variables and implementing more sophisticated Machine Learning algorithms.

Bibliography

1. Abbott, L. (1955), *Quality and competition; an essay in economic theory*, New York. Columbia University Press, p. 39

2. Alpaydin, E. (2010), *Introduction to Machine Learning*, The MIT Press, Second Edition

3. Adroid (2018), *Android Security 2017 Year in Review* https://source.android.com/security/reports/Google_Android_Security_2017_Report_Final.pdf

4. Agrawal, A., Catalini, C., Goldfarb, A. (2010), *The geography of crowdfunding*, SSRN Electronic Journal

5. An, J., Quercia D., Crowcroft J. (2014), *Recommending Investors for Crowdfunding Projects*, Cornell University Library

6. Avery, J., Steenburgh, T., Deighton J., Caravella, M. (2012), *Adding Bricks to Clicks: Predicting the Patterns of Cross-channel Elasticities Over Time*, Journal of Marketing, 76 (3), 96–111

7. Bass, F. M. (1974), *The Theory of Stochastic Preference and Brand Switching*, Journal of Marketing Research, Vol. 11, No. 1, pp. 1-20

   https://www.jstor.org/stable/i358097

8. Belkahla, W., Triki A. (2011), *Customer Knowledge Enabled Innovation Capability: Proposing a Measurement Scale*, Journal of Knowledge Management, 15 (4), 648-674

9. Belleflamme, P., Lambert, T., Schwienbacher, A. (2013), *Crowdfunding: Tapping the right crowd*, Journal of Business Venturing

10. Berger, P., Magliozzi, T. (1992), *The effect of sample size and proportion of buyers in the sample on the performance of list segmentation equations generated by regression analysis*, Direct Marketing, 6: 13-22

11. Berry, L. L. (1995), *Relationship Marketing of Services: Growing Interest, Emerging, Perspectives*, Journal of the Academy of Marketing Science, 23 (4), pp. 236–45

    https://www.academia.edu/2005612/Relationship_marketing_of_services_growing_interest_emerging_perspectives

12. Bhattacharyya, D. K. (1999*), On the Economic Rationale of Estimating the Hidden Economy,* The Economic Journal, 109: 348-359

13. Bitner, M. J., Brown, S.W., Meuter, M.L. (2000), *Technology infusion in Service Encounters*, Journal of the Academy of Marketing Science, 28 (1), 138-149

14. Bitran, G., Mondschein, S (1996), *Mailing Decisions in the Catalog Sales Industry,* Management Science, 1996, vol. 42, issue 9, 1364-1381

15. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., Zieba, K. (2016), *End to End Learning for Self-Driving Cars*

16. Bose, I., Mahapatra, M. K. (2001), *Business Data Mining – A Machine Learning Perspective,* Infromationa and Management, Numer 39, pp. 211-225

17. Brodie, R. J., Hollebeek, L. D., Juric, B., Ilic, A. (2011), *Customer Engagement: Conceptual Domain, Fundamental Propositions, and Implications for Research*, Journal of Service Research 14(3), pp. 252-271

18. Brynjolfsson, E., Hu, Y.J., Rahman, M.S. (2009), *Battle of the Retail Channels: How Product Selection and Geography Drive Cross-Channel Competition*, Management Science, 55 (11), 1755–1765

19. Brynjolfsson, E., Hu, Y.J., Rahman, M.S. (2013), *Competing in the Age of Omnichannel Retailing*, MIT Sloan Management Review, 54 (4), 23–9

20. Burtch, G., Ghose, A., Wattal, S. (2013), *An Empirical Examination of the Antecedents and Consequences of Contribution Patterns in CrowdFunded Markets,* SSRN, 1–46

21. Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., Xie, Z. (2018), *Deep Learning and Its Applications in Biomedicine*, Genomics, Proteomics & Bioinformatics, Volume 16, Issue 1, Pages 17-32

22. Chan, C.S.R., Parhankangas, A. (2017), *Crowdfunding Innovative Ideas: How Incremental and Radical Innovativeness Influence Funding Outcomes,* Entrepreneurship Theory and Practice, March 2017 pp. 237-263

23. Chandrayan, P. (2017), *Machine Learning Part 3: Logistic Regression* https://towardsdatascience.com/machine-learning-part-3-logistics-regression-9d890928680f

24. Choi, J., Bell, D. (2010), *Preference Minorities and the Internet*, Working paper, Wharton

25. Christensen, C. M., Cook, S., Hall, T. (2005), *Marketing Malpractice: The Cause and the Cure*, Harvard Business Review, 83 (12), pp. 74–83

26. Christensen, C. M. Raynor, M.E. (2003), *The Innovator's Solution*, Boston: Harvard Business Press

27. Cook, G. (2014), *Customer experience in the omni-channel world and the challenges and opportunities this presents*, Journal of Direct, Data and Digital Marketing Practice,Vol. 15, Number 4. pp. 262–266

28. Crawford, M., Khoshgoftaar, T.M., Prusa, J.D.,. Richter, A.N., Najada, H.A. (2015), *Survey of review spam detection using machine learning techniques*, Journal of Big Data, 2, 23

29. Crone, S., Lessmann, S., Stahlbock, R. (2006), *The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing*, European Journal of Operational Research

30. CrowdfundingHub (2016), *Current State of Crowdfunding in Europe* https://link.springer.com/content/pdf/10.1057%2Fdddmp.2014.16.pdf

31. Cui, G., Wong, M.L., Lui, H. (2006), *Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming*, Management Science, 52, 597-612

32. Day, G.S. (1969), *A two-dimensional concept of brand loyalty*, Journal of Advertising Research, Vol. 9. pp. 29-35

33. De Nisco, A., Mainolfi, G. (2016), *Competitiveness and Foreign Perception of Italy and Made In Italy on the Emerging Markets*, Rivista Italiana di Economia Demografia e Statistica, Vol. LXX, n.3

34. Deleersnyder, B., Geyskens, I., Gielens, K., Dekimpe, M.G. (2002), *How Cannibalistic Is the Internet Channel? A Study of the Newspaper Industry in the United Kingdom and the Netherlands*, International Journal of Research in Marketing, 19 (4), 337–48

35. Dichter, E. (1962), *The world customer*, Harvard Business Review, Vol. 40, No.4, pp. 113-122

36. Duncan, E. (2017), *Customer experience. New capabilities, new audiences, new opportunities.* Number 2, McKinsey&Co

37. Edelman Insights (2013), *Brandshare*

38. Etter, V., Grossglauser, M., Thiran, P. (2013), *Launch Hard or Go Home!*, COSN'13, October 7–8, Boston, Massachusetts http://dx.doi.org/10.1145/2512938.2512957

39. The Financial Brand (2014), *Omnichannel Banking: More Than a Buzzword* https://thefinancialbrand.com/37996/bank-channel-usage-research/

40. Gentile, C., Spiller. N., Noci, G. (2007), *How to Sustain the Customer Experience: An Overview of Experience Components that Co-create Value with the Customer*, European Management Journal Vol. 25, No. 5, pp. 395–410

41. Gerber, E.M., Hui, J.S., Kuo, P. (2012), *Crowdfunding: Why People Are Motivated to Post and Fund Projects on Crowdfunding Platforms* https://pdfs.semanticscholar.org/c1e2/a1068f0af1c3120c62be5943340518860ecb.pdf

42. Goldfarb, A., Tucker, C. (2010), *Advertising Bans and the Substitutability of Online and Offline Advertising*, Journal of Marketing Research

43. Goodfellow, I., Bengio, Y., Courville A. (2016), *Deep Learning*, MIT Press

44. Green, P., Frank, E., Robinson, P. (1967), *Cluster Analysis in test Market Selection*, Management Science, Management Science, 13, B. pp. 387-400

45. Greenberg, M. D., Pardo, B., Hariharan, K., Gerber, E. (2013), *Crowdfunding Support Tools: Predicting Success & Failure*, CHI'13 Extended Abstracts on Human Factors in Computing Systems. pages 1815–1820, ACM https://egerber.mech.northwestern.edu/wp-content/uploads/2012/11/Crowdfunding-Support-Tools_Predicting-Success-Failure_2013.pdf

46. Gutsche, J., Sylla (2018), *S. Success Factors of Crowdfunding Projects on the Kickstarter Platform*, German-Turkish Perspectives on IT and Innovation Management, FOM-Edition (FOM Hochschule für Oekonomie & Management), Springer Gabler, Wiesbaden

47. Hirschman E. C., Holbrook, M. B. (1982), *Hedonic Consumption: Emerging Concepts, Methods and Propositions*, Journal of Marketing, Vol. 46, No. 3., pp. 92-101

http://criticalmanagement.uniud.it/fileadmin/user_upload/Hirschman__Holbrook_1982.pdf

48. Holbrook, M.B., Hirschman, E.C. (1982) *The experiential aspects of consumption: Consumer fantasy, feelings and fun,* Journal of Consumer Research 9(2), 132–140

49. Howard, John A., Sheth, J. (1969), *The Theory of Buyer Behavior*, John Wiley & Sons

50. Janasik, N., Timo H., Bruun, T. (2009), *Text Mining in Qualitative Research: Application of an Unsupervised Learning Method*, Organizational Research Methods, 12 (3), 436-460

51. Jenkinson, A. (2007), *Evolutionary implications for touchpoint planning as a result of neuroscience: A practical fusion of database marketing and advertising*, Database Marketing & Customer Strategy Management Vol. 14, 3, 164–185 https://link.springer.com/article/10.1057/palgrave.dbm.3250054

52. Josephat, P. (2012), *A Logistic Regression Model of Customer Satisfaction of Airline*, International Journal of Human Resource Studies, Vol. 2, No. 4

53. Karim, M., Rahman, R.M. (2013)**,** *Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing*, Journal of Software Engineering and Applications, Number 6, pp. 196-206 http://file.scirp.org/pdf/JSEA_2013042913162682.pdf

54. Koch, J., Siering, M. (2015), *Crowdfunding Success Factors: The Characteristics of Successfully Funded Projects on Crowdfunding Platforms*, Twenty-Third European Conference on Information Systems (ECIS), Münster, Germany

55. Kotsiantis, S. B. (2007), *Supervised Machine Learning: A Review of Classification Techniques*, Informatica 31, pp. 249-268 https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

56. Kusinitz, S. (2014), *The Definition of a Buyer Persona [in Under 100 Words]*. Hubspot

57. Lambert T., Schwienbacher, A. (2010), *An Empirical Analysis of Crowdfunding*

58. Laura, Luigi (2017), *Machine Learning and Object Driven Marketing*. Cap. 5

59. Learned-Miller, Erik G. (2014), *Introduction to Supervised Learning*. University of Massachusetts

60. Lemon, K.N., Verhoef, P.C. (2016), *Understanding Customer Experience Throughout the Customer Journey*, Journal of Marketing, Vol. 80, 69–96

https://www.ama.org/publications/JournalOfMarketing/Pages/understanding-customer-experience-customer-journey.aspx

61. Levy, S. J. (1959), *Symbols for Sale,* Harvard Business Review, July-August. Number 37, p. 117-119 https://www.uibk.ac.at/smt/marketing/brandresearchlab2/files/symbolsforsale_levy.pdf

62. Libai,B., Bolton, R., Bügel, M. S., Ruyter, K., Götz, O., Risselada H., Stephen, A. T. (2010), *Customer-to-Customer Interactions: Broadening the Scope of Word of Mouth Research,* Journal of Service Research, 13: 267

63. Linden, G., Smith B., York, J. (2003), *Amazon.com Recommendations: Item-to-Item Collaborative Filtering,* IEEE Computer Society

64. Lingqvist, O., Plotkin, C., & Stanley, J. (2015), *Do you really understand how your business customers buy?* McKinsey Quarterly, 74—85

65. Lockett, H. (2016), *Kickstarter Opens in Hong Kong, Singapore,* Financial Times

66. MacQueen, J.B. (1967), *Some Methods for Classification and Analysis of Multivariate Observations,* Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkley, University of California Press, 1, pp. 281-297

67. https://pdfs.semanticscholar.org/a718/b85520bea702533ca9a5954c33576fd162b0.pdf

68. Metters, R., Queenan, C., Ferguson, M., Harrison, L., Higbie, J., Ward, S., Barfield, B., Farley, T., Kuyumcu, H. A., Duggasani, A. (2008), *The "Killer Application" of Revenue Management: Harrah's Cherokee Casino & Hotel,* Interfaces, 38(3),161-175

69. Meyer, C., Schwager, A. (2007), *Understanding Customer Experience,* Harvard Business Review

70. Mikroyannidis, A., Babis, T. (2006), *Heraclitus ii: A Framework for Ontology Management and Evolution,* in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Toyoaki Nishida, (Chair), Hong Kong, China: IEEE Computer Society, 514-521

71. Mithcell, Tom M. (1997), *Machine Learning,* McGraw Hill https://www.cs.ubbcluj.ro/~gabis/ml/ml-books/McGrawHill%20%20Machine%20Learning%20-Tom%20Mitchell.pdf

72. Mitra T. Gilbert, E. (2014), *The Language that Gets People to Give: Phrases that Predict Success on Kickstarter*, CSCW'14, February 15–19, Baltimore, Maryland, USA

73. Mollick, E. (2013), *Delivery Rates on Kickstarter,* The Wharton School University of Pennsylvania

74. Mollick, E. (2014), *The dynamics of crowdfunding: An exploratory study*, Journal of Business Venturing. 29 (1), 1–16

75. Mollick, E. (2016), *Containing Multitudes: The Many Impacts of Kickstarter Funding*, The Wharton School University of Pennsylvania

76. Montgomery D.B., Silk, A.J. (1971), *Cluster of Consumer Interests and Opinion Leaders' Sphere of Influence*, Journal of Marketing Research, Number 8, pp. 317-321

77. Morgan, R. M., Hunt, S.D. (1994), *The Commitment–Trust Theory of Relationship Marketing*. Journal of Marketing, Number 58. pp. 20–38 http://www.jstor.org/stable/1252308

78. Murthy, S.K. (1998), *Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey*, Data Mining and Knowledge Discovery, 2, pp. 345–389

79. Myers, J. B., Nicosia, F.M. (1968), *On the Study of Consumer Typologies*, Journal of Marketing Research, Number 5, pp. 182-183

80. Nasoi, R. (2017), *6 Customer Journey Mapping Examples: How UX Pros Do It*, https://conversionxl.com/blog/customer-journey-mapping-examples/

81. Neslin, S.A., Grewal, D., Leghorn, R., Shankar, V., Teerling, M. L., Thomas, J. S., Verhoef, P. C. (2006), *Challenges and Opportunities in Multichannel Customer Management.* Journal of Service Research. Volume 9. No. 2, pp. 95-112

82. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.317.9521&rep=rep1&type=pdf

83. Ordanini, A., L. Miceli, M. Pizzetti and A. Parasuraman (2011), *Crowdfunding: Transforming customers into investors through innovative service platforms*, Journal of Service Management 22(4): 443-470

84. Ordenes, F.V., Theodoulidis, B., Burton, J., Gruber, T., , Zaki, M. (2014), *Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach*, Journal of Service Research, Vol 17, Issue 3, pp. 278 - 295

85. Parasuraman, P., Zeithaml, Valarie A., Berry, Leonard L. (1988), *SERVQUAL: A multiple- Item Scale for measuring consumer perceptions of service quality*. Journal of Retailing, Volume 64, Number 1

86. Payne, A., Frow, P. (2005), *A Strategic Framework for Customer Relationship Management*, Journal of Marketing, Number 69, pp. 167–76

87. Punj, G., Stewart, D. W. (1983), *Cluster Analysis in Marketing Research: Review and Suggestions for Application.* Journal of Marketing Research, Vol. 20. pp. 134-148 https://www.researchgate.net/profile/Girish_Punj/publication/215666129_Cluster_analysis_in_marketing_research_Review_and_suggestions_for_application/links/0deec532b6d6edc01d000000/Cluster-analysis-in-marketing-research-Review-and-suggestions-for-application.pdf

88. Pearl, L., Steyvers, M. (2013), *"C'mon – You Should Read This": Automatic Identification of Tone from Language Text*, International Journal of Computational Linguistics (IJCL), Volume 4, Issue 1

89. Qiu, C. (2013), *Issues in Crowdfunding: theoretical and empirical investigation on Kickstarter,*

90. Pednault E., Abe, N., Zadrozny, B., Wang, H., Fan, W. and Apte (2002), C. *Sequential Cost-Sensitive Decision-Making with Reinforcement Learning*, KDD '02 Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 259-268

91. PwC (2015), *Making omnichannel work: The "to do" list for banks*

92. Rakesh, V., Jaegul Choo, Chandan K. Reddy (2015), *Project Recommendation Using Heterogeneous Traits in Crowdfunding*, Association for the Advancement of Artificial Intelligence http://dmkd.cs.wayne.edu/Papers/ICWSM15.pdf

93. Rathmell, John M. (1966), *What Is Meant by Services?* Journal of Marketing, Vol. 30, No. 4, pp. 32-36

94. Rawal, P. (2013), *AIDA Marketing Communication Model: Stimulating a purchase decision in the minds of the consumers through a linear progression of steps,* IRS's International Journal of Multi-Disciplinary Research in Social & Management Sciences, Issue 1, Volume 1

95. Reichheld, F. E., Sasser, W. Earl Jr. (1990), *Zero Defections: Quality Comes to Services*, Harvard Business Review, Number 68, pp. 105-111

96. Reinartz, W., Kumar, V. (2000), *On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing*, Journal of Marketing, Number 64, pp. 17-35

97. Reinders, M. J., Dabholkar, P. A., & Frambach, R. T. (2008), *Consequences of forcing consumers to use technology-based self-service*, Journal of Service Research, 11(2), 107–123

98. Rigby, Dan (2011), *The Future of Shopping*, Harvard Business Review, 89 (12), 65–76

99. Rosenbaum, M., Otalora, M.L., Ramirez, G.C. (2016), *How to create a realistic customer journey map,* Kelley School of Business, Indiana University

100. Roux, M. (2015), *A comparative study of divisive hierarchical clustering algorithms*, arXiv

101. https://arxiv.org/ftp/arxiv/papers/1506/1506.08977.pdf

102. G. A. Rummery and M. Niranjan (1994), *On-Line Q-Learning Using Connectionist Systems*

103. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.17.2539&rep=rep1&type=pdf

104. Rust, R. T., Lemon, K. N., Zeithaml, V. A. (2004), *Return on Marketing: Using Customer Equity to Focus Marketing Strategy*, Journal of Marketing, Vol. 68, pp. 109–127

105. Sahlman. W.A. (1997), *How to write a great business plan*, Harvard Business Review

106. Sands, S., Ferraro, C., & Luxton, S. (2010), *Does the online channel pay? A comparison of online versus offline information search on physical store spend*, The International Review of Retail, Distribution and Consumer Research,Vol. 20 No. 4, pp. 397–410

107. Sarstedt, M. (2008), *Market segmentation with mixture regression models: Understanding measures that guide model selection*, Journal of Targeting, Measurement and Analysis for Marketing, Vol. 16, 3, pp. 228–246

108. Schmidhuber, J, (2015), *Deep learning in neural networks: An overview,* Neural Networks 61, pp. 85–117

109. Serban, I.V., Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Rajeswar, Alexandre de Brebisson, Jose M. R. Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, Yoshua Bengio (2018), *A Deep Reinforcement Learning Chatbot,*

110. Shalev-Shwartz, Shai. Ben-David, Shai (2014), *Understanding Machine Learning*, Cambridge University Press

111. Sheth, J. N., Sisodia, R. S., Sharma, A. (2000), *The Antecedents and Consequences of Customer-Centric Marketing,* Journal of the Academy of Marketing Science, Volume 28 No. 1. pp. 55-66
https://link.springer.com/article/10.1177/0092070300281006

112. Sohl, J. E. (1999), *The early-stage equity market in the USA*, Venture Capital: An International Journal of Entrepreneurial Finance, 1, 101–120

113. Sorenson, O., and T. E. Stuart (2001), *Syndication Networks and the Spatial Distribution of Venture Capital Investments,* American Journal of Sociology, 106(6), 1546– 1588.

114. Sterne, J. (2017), *Artificial Intelligence for Marketing: Practical Applications*, Wiley

115. Strong Jr, E.K. (1925), *The Psychology of Selling and Advertising*, McGraw Hill, p. 349 and p. 9

116. Sun, Baohong. (2018), *Technology Innovation and Implications for Customer Relationship Management*, Marketing Science, 25, pp. 594-597

117. Sundsøy, P., Bjelland, J., Iqbal, A.M., Pentland, A.S., De Montjoye, Y. (2014), *Big Data-Driven Marketing: How machine learning outperforms marketers' gut feeling,* Computer Science, Volume 8393, pp 367-374

118. Sutton, R. Barto, A. (2017), *Reinforcement Learning*, Second Edition, MIT Press

119. http://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

120. Triggs, R. (2018), *Artificial Intelligence vs Machine Learning: what's the difference?,* Android Authority

121. Tschimmel, K. (2012), *Design thinking as an effective toolkit for innovation.,* In Proceedings from the XXIII ISPIM Conference: Action for innovation: Innovating from experience, 2012

122. Verhoef, P. C., Donkers, B. (2001), *Predicting Customer Potential Value*, Decision Support Systems, Number 32. pp. 189–199

123. Verhoef, P.C., Kannan, P.K., Inman; J.J. (2015), *From Multi-Channel Retailing to Omni-Channel Retailing: Introduction to the Special Issue on Multi-Channel Retailing*, Journal of Retailing 91 (2) 174–181

124. Verhoef, P.C., Neslin, S.A., Vroomen, B. (2007), *Multichannel Customer Management: Understanding the Research-Shopper Phenomenon*, International Journal of Research in Marketing, 24 (2), pp. 129–48.

125. Wang, C. (2016), *The Promise of Kickstarter: Extents to Which Social Networks Enable Alternate Avenues of Economic Viability for Independent Musicians Through Crowdfunding*, Social Media + Society http://journals.sagepub.com/doi/pdf/10.1177/2056305116662394

126. Wash, R. (2013), *The Value of Completing Crowdfunding Projects*, Association for the Advancement of Artificial Intelligence https://pdfs.semanticscholar.org/c4fd/4477920c1f0b78d4c6198bb3333c7d9a6035.pdf

127. Watkins, C.J.C.H. (1989), *Learning from delayed rewards*

128. Watkins, C.J.C.H. (1992), *Q-learning*, Machine Learning. 8, pp. 279-292

129. Webster Jr. F.E., and Wind Y. (1972), *A General Model for Understanding Organizational Buying Behavior*, Journal of Marketing

130. Wells, W.D. (1975), *Psychographics: A Critical Review*, Journal of Marketing Research, Number 12, pp. 196-213

131. Wetzels, M., Ruyter, K., Bloemer, J. (1997), *On the relationship between perceived service quality, service loyalty and switching costs.* International Journal of Strategic Innovative Marketing https://pdfs.semanticscholar.org/105c/4accbf582eb902cd65621210615f9995a6a0.pdf

132. Wheat, R.E., Yiwei Wang, Jarrett E. Byrnes, and Jai Ranganathan (2013), *Raising money for scientific research through crowdfunding*, Trends in Ecology & Evolution, February 2013, Vol. 28, No. 2

133. Xiao, S., Xue Tan, Ming Dong, Jiayin Qi (2014), *How to Design Your Project in the Online Crowdfunding Market? Evidence from Kickstarter*, Thirty Fifth International Conference on Information Systems, Auckland

134. Xu, A., Yang, X., Rao, H., Fu, W. T., Huang, S. W., & Bailey, B. P. (2014), *Show me the money! An analysis of project updates during crowdfunding campaigns*, In CHI 2014: One of a CHInd - Conference Proceedings, 32nd Annual ACM Conference on Human Factors in Computing Systems (pp. 591-600). Association for Computing Machinery https://pdfs.semanticscholar.org/105c/4accbf582eb902cd65621210615f9995a6a0.pdf

135. Yuan H., Y.K. Lau, Wei Xu (2016), *The determinants of crowdfunding success: A semantic text analytics approach*, Decision Support Systems 91, pp. 67–76

136. Zhang, S., Zhang, C. (2003), *Data Preparation for Data Mining*, Applied Artificial Intelligence 17, pp. 375–381

137. Zheng, H., Yi-Chia Wang, & Piero Molino (2018), *COTA: Improving Uber Customer Care with NLP & Machine Learning*, Uber https://www.cse.ust.hk/~qyang/Docs/2003/Data_Preparation_for_Data_Mining_ZZY.pdf

138. Zvilichovsky, D. Inbar, Y. Barzilay, O. (2014), *Playing Both Sides of the Market - Success and Reciprocity on Crowdfunding Platforms*, Working Paper

Executive Summary

The Master Thesis had the main aim to analyse whether and how the features of Kickstarter projects influence the Customer Experience and individuals' preferences.

First, the Thesis provided a Literature Review that illustrates the development of Marketing theories on Customer Experience, previous research of Machine Learning applications in Marketing and Customer Experience in Kickstarter. Then, an introduction on Machine Learning provided the basic knowledge to understand what Machine Learning is and the value that its implementation may bring to companies and customers to improve the quality of *touchpoints.*

Relying on this introduction, the study examined how four features of Kickstarter projects – (1) *Main Category,* (2) G*oal,* (3) *Country* and (4) *Currency* – influence the Customer Experience, analysing how combinations of these variables affect projects' success rate.

First, each dependent variable was analysed separately and then used to build step-by-step a predictive model, made with three of the four independent variables – (1) *Main Category,* (2) G*oal,* and (3) *Country* – and based on a multivariate logistic regression. Then, the model was trained on 70% of the dataset, and then tested on the remaining 30%.

The Predictive Model has an accuracy of 61%, meaning that it is capable of predicting the outcome of a campaign 6 times out of 10.

The findings of the research proved that customers are more attracted towards projects that offer artistic contents, with realistic goals, and that they are influenced by the creators' country of origin.

1. Literature Review

The first chapter of the Master Thesis has the aim to provide the necessary knowledge basis through a deep insight of the Existing Literature.

The chapter is divided in three parts, that gradually move from a more general framework to the narrower field related with Kickstarter.

The first part explains the evolution of the Marketing theories related with the Customer Experience, from the 1920s until nowadays. What emerges is how the customer became increasingly meaningful for Marketing strategies.

The second part takes a more specific look on previous studies and Research made in the Customer Experience field through the implementation of Machine Learning techniques.

The third and last part of the chapter introduces studies and discoveries regarding Customer Experience on Kickstarter from different perspectives. The sum of these three parts, together, provides the fundament for the development of the hypothesis that help responding to the research question: whether Kickstarter's projects features influence Customer Experience and individuals' preferences.

1.1 Customer Experience theories

Customer experience has broadly been defined by scholars and marketers as a *multidimensional construct that involves cognitive, emotional, behavioural, sensorial, and social components* (Schmitt, 1999).

Recently, with the increasing number of contact points between a company and its customers, the attention towards customers became visible with the monitoring of the experiences originated by those contact points (Gentile et al., 2007). Thus, it gains more and more importance to consider aspects that are part of the emotional and irrational side of the customer behaviour (Holbrook et al., 1982).

The importance of customer experience lies in the fact that it encompasses every aspect of a company's offering (Meyer et al., 2007). Anyhow, the path that lead to this current of thought has been long, and multiple theories have been developed in the last decades, that can be divided into three big research areas: (1) customer experience and the customer journey, (2) customer experience measurement, and (3) customer experience management. Customer experience has clearly gained a central position in Marketing strategies: it depends on the interactions between customers and firms through multiple channels and media, making it more complex for businesses to ensure to their customers a high quality end-to-end experience.

Researchers have been interested in the individualization and analysis of those interactions, also known as *touchpoints*. One commonly diffused believe (Jim Sterne, 2017) is that through the implementation of the right strategy during a specific touch point, it might be possible to show the right message in front of the right person at the right time in the right context on the right device and figuring out whether any of the work that was done had an impact on the buying decision.

1.2 Customer Experience and Machine Learning

Until recently, traditional statistical models have dominated the analysis of customer preferences and responses to direct Marketing (Berger et al., 1992). Yet, statistical models make several stringent assumptions on the types of data and their distribution that may limit the potentiality of the analysis (Cui et al., 2006). Furthermore, when applied to real data, the model incurs in some issues because the key assumption of the research methods are often violated (Bhattacharyya 1999).

Machine Learning evolved with the aim of eliminating the dispendious and time-consuming processes needed to develop knowledge-based system (Bose et al., 2001). In fact, an accurate analysis of this information with traditional tools requires a significant time contribution (Janasik et al., 2009).

In Customer Experience, Machine Learning is used to identify purchasing intentions for different products and services (Crone et al., 2006), or to analyse customers' responses to direct marketing in order to build a model capable of predicting individuals' purchases (Cui, 2006).

Another application of Machine Learning permits to analyse texts to provide a holistic perspective of textual and nontextual information (Mikroyannidis et al., 2006), for instance to analyse customer experience feedbacks (Ordenes et al., 2014)

Machine Learning is also used to daily life tasks, such as spam detection (Crawford, 2015), and chatbots (Serban et al., 2018).

What becomes clear from the above-mentioned studies, is that ML techniques are applied across different industries and for multiple purposes.

1.3 Customer Experience on Kickstarter

Crowdfunding platforms, such as Kickstarter, offer project founders the possibility to ask for funding for their idea from all internet users active on these open online services (Mollick, 2014).

Many studies and researches have already tried to analyse and understand the relation between the variables of a campaign and the success of a project.

Depending on the funded amount, the backers may receive a "reward". Small amounts (generally below $10), are without any type of reward. Higher amounts generally permit to receive something that goes from a customized "thank you", to one or more prototypes of the backed product, that can come in different models.

From a general perspective, some studies compare profit and non-profit crowdfunding initiatives (Lambert et al., 2010).

Statistic variables of project and founders have also been widely analysed. Elements such as founders' friends on Facebook and followers on Twitter were insert in the model (Greenberg et at., 2013), while other more comprehensive studies focussed on project-specific and founder-specific variables (Koch et al., 2015).

The question that drove other researches was whether information about the founder influences the funding success on future projects (Zvilichovsky et al., 2013). Kickstarter itself recommends backers to check the founders of a project, to know their history and eventually their previous successfully launched projects, that might be seen as a sign of their capability of transforming an idea into something real.

Furthermore, through a semantic text analytics approach, a team of researchers started to build a model to analyse which features of the description were more likely to help projects to obtain funding (Yuan et al., 2016).

From a data mining and Machine Learning perspective, only a few analyses have already been made.

A research (Mitra et al., 2014) uses Natural Language Processing techniques in order to analyse the textual content of Kickstarter projects, while another study leverages the updates of projects to determine their success rate (Xu et al., 2014).

Etter et al. (2013) build prediction models relating the probability if success of a project with the number and quality of Tweets publicly available.

2. Introduction to Machine Learning

This chapter has the aim to provide an introduction on Machine Learning and on its main features, types and applications.

First, it is necessary to understand what Machine Learning is, and what is the meaning behind *learning.* The chapter provides an overview of those tasks where Machine Learning provides a higher efficiency compared to normal computer programs, and an insight of the typologies of Machine Learning: Supervised Learning – when the computer learns from labelled data, Unsupervised Learning – when the computer learns from unlabelled data, and Reinforcement Learning – when the computer learns to maximize a reward signal. The chapter provides also an insight of Deep Learning, a subset of Machine Learning, that deserves to be mentioned for its wide implementations in Marketing.

2.1 Features of Machine Learning algorithms

The new digital devices are recording a huge amount of reliable data, while computer technologies make it possible to store it and to access it from physically distant locations over a computer network (Alpaydin, 2010). Experts believe that there are some common patterns that explain the data, and that they do not have a completely random nature. *Consumer behaviour* offers a good example: usually, people buy less ice cream in winter than in summer and when they buy a beer they could buy also chips.

Developing *a good and useful approximation* may help in achieving some interesting discoveries, such as patterns or regularities.

This is exactly what *Machine Learning* is all about: the automated detection of meaningful patterns in data (Shalev-Shwartz, 2014). In fact, machine learning is making computers build a model and its parameters using the training data or past experience. The belief is that, behind the large amount of data, there are simple patterns that can be discovered and used by machine learning to build up a model**.**

The model may be *predictive* - if it is used to make prediction on future outcomes, or *descriptive* – if it is used to gain better understanding on the data, or both.

Eventually, Machine Learning may be preferred over normal programs in those situations where there is:

- Complexity
- Need for adaptivity

The first relevant choice is about the type of training experience (Mitchell, 1997).

The training experience has the following attributes:

- Direct or indirect feedback
- Degree of control of the sequence of the training examples
- Similarly of training set with the test set

Then, it is necessary to establish which kind of knowledge will be used by the program. The aim is to create a very expressive representation as close as possible to the ideal target function. However, the more expressive the function, the more data will be required in order to make the program capable of choosing among the alternative hypothesis.

2.2 Types of Machine Learning

There are three types if Machine Learning: Supervised Machine Learning, Unsupervised Machine Learning and Reinforced Machine Learning.

A key point that deserves a short focus is the importance of *high-quality data*. This is the reason why data preparation is a fundamental stage of data analysis. A lot of raw data is available in various data sources and on the Web, and companies and firms are interested in "cleaning" data to shape it into a high-quality data set that can be analysed and used to generate profits (Zhang et al., 2003).

In fact, *real world data* may be (Zhang et al., 2013):

- Incomplete
- Noisy
- Inconsistent

Preparing those data to generate a quality dataset comprehends actions as selecting only relevant data, removing anomalies filling the missing value. At the end of the whole process, the quality of the data will provide a higher quality of the pattern.

2.2.1 Supervised Learning

Supervised Learning is based on the idea of learning from experience.

Usually, the computer is provided with two sets of data, a *training set* and a *test set.*

In the training set, the data are already labelled. In this way, the program can learn from the training set and try to label as correctly as possible the data in the test set. The aim of the program is to understand how to classify properly the unlabelled data through the example it has been provided with.

Supervised learning algorithms can be deployed for *Classification* and *Regression* problems. In classification problems, the aim is to predict whether the output will be in a certain way or not (the labels of the data are categorical). Instead, in regression problems, the aim is to predict the value of the output (the labels of the data are numerical).

2.2.2 Unsupervised Learning

While in supervised learning the labels are given, in the unsupervised learning the only available information are the data. The aim is to find some patterns in the data, analysing the input space and searching for some structures that happen to be present more often than others. One of the most common ways to find patterns is using *clustering***.**

Clustering is one method to find cluster or groups of data with similar characteristics. The aim is to create groups such that the objects of one group are similar (or related) to one another and different from (or unrelated to) the objects in other groups. The goal is to minimize the *intra-cluster distance* (the distance between the elements in the same cluster) and to maximize the *inter-cluster distance* (the distance between elements in different clusters.

The customer segmentation may also allow to identify *outliers* – customers that differ strongly among the others, that could indicate the existence of a niche market with possibility of exploitation for the company (Alpaydin, 2010).

Clustering can be *partitional* or *hierarchical.*

2.2.3 Reinforcement Learning

Reinforcement Learning is about analysing the environment, in order to understand which actions to take in order to maximize a reward signal.

It is at the same time the problem, the class of solution methods that work well on the class of problems, and the field that studies these problems and their possible solutions (Sutton et al., 2017).

It is possible to identify six main components in a reinforcement learning system:

- Agent
- Environment
- Policy
- Reward signal
- Value function
- Model of the environment (optional)

2.2.4 Deep Learning

Even though Deep Learning is not properly a *type* of Machine Learning but more a *subset* of it, it deserves to be mentioned.

There are some tasks that appear natural and immediate for the human brain: recognizing a face, understanding whether a person is angry, knowing the difference between the picture of a cat and of a dog. If people were asked to explain how they could were able to solve those tasks, they would not know which process they followed: it just happened.

Deep Learning is a subset of Machine Learning that helps computers to deal with these kinds of tasks, easier to perform than to describe. The computer understands the world in terms of a hierarchy of concepts, allowing the computer to learn complicated concepts by building them out of simpler ones (Goodfellow et al., 2016).

Deep Learning works through the implementation of Artificial Neural Networks, structures that have been inspired by the human brain. An artificial neural network (ANN) is made by many simple, connected processors (neurons), each producing a sequence of real-valued activation. Input neurons get activated through sensors perceiving the environment, other neurons get activated through weighted connections from previously active neurons (Schmidhuber, 2015).

3. Hypotheses Development

The research has the aim to investigate the influence of projects' features on the Customer Experience and individuals' preferences.

In the following study, the variables that are considered are only those that have two main features:

- Are visible on the webpage of the project
- Do not change but remain constant during the whole campaign

The hypotheses are developed in order to investigate the research question. Customers might perceive some project features as symbols for more quality, they might prefer to invest in a project rather than in another because its features are more appealing. Therefore, the hypotheses are considered taking into account four project characteristics:

- Category
- Country
- Goal
- Currency

to determine whether there is a relation between the features of a project and the individuals' willingness to invest and therefore help a project to reach its goal.

The four hypotheses are the following:

- *H1: Individuals prefer to become "backers" of campaigns that are mostly about creative projects*

- *H2: individuals are influenced in their investing decision by the value of the goal of a project*

- *H3: individuals' investment decisions are not influenced by geographic boundaries*

- *H4: Individuals prefer to invest in their own currency and therefore projects that ask for diffused currencies are more successful*

4. Descriptive statistics

The previous chapter focussed on the features of a project that might influence the individuals' investing decision.

In order to prove whether those hypotheses are verified or not, it is necessary to use a database that is significant and that offers the asked features for each project.

The purpose of this chapter is to provide a detailed insight of the database on which the subsequent analysis is based on. Hence, the following pages will illustrate the source and structure of the database, the main features and a deep examination obtained through descriptive statistics.

5. Methodology and Model

This chapter focusses on the methodology used to develop the model.

The first part provides an overview of the preparation and cleaning of the dataset and of the used algorithm.

Afterwards, the second part analyses the dependent variable and the independent variables.

The third part provides an insight of the process of the model building

The considered variables are five: four independent variables and one dependent variable. The dependent variable is the fail / success outcome of a project, while the four dependent variables are (1) Category, (2) Goal, (3) Country and (4) Currency.
The predictive ML Model was tested in order to find the right combination of variables. Therefore, the best model happens to be the one built only on three variables: (1) Category, (2) Goal and (3) Country.

5.1 Preparing the Dataset

The dataset contains information about the features of the projects launched on Kickstarter until January 2018.

The aim of the research is to understand which projects customer prefer to invest in. In order to measure these preference, the success or fail outcome of a project has been chosen as *dependent variable.*

The dataset was provided with a further column with the following values:

- *1,* if the project was successful
- *0,* if the project failed

The considered independent variables are four:

- *Category,* categorical variable
- *Goal,* discrete variable
- *Currency,* categorical variable
- *Country,* categorical variable

Hence, three variables are categorical (category currency, and country), and one variable is discrete (goal).

Since the *dependent variable* is *dichotomic*, using a linear regression did not appear as the optimal solution.

The *logistic regression* is more appropriate. In fact, it permits to relate both discrete and categorical variables with a dichotomic variable.

The formula of the logistic regression is:

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

With

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

β_0 = intercept

β_1 = coefficient of x_1

x_1 = dependent variable 1

β_k = coefficient of x_k

x_k = dependent variable k

The coefficients provide the log of the odds of each variable.

The aim of the research is to provide a predictive model in order to forecast whether a given project will have success based on customer experience and preferences.

Therefore, the research was developed with the implementation of Machine Learning through R, a software environment mainly used for statistical computing and graphics.

First, the dataset was "cleaned", removing all the projects that had a *state* different from *successful* or *failed:* hence, projects that were *undefined, live, cancelled* or *suspended* were not considered. The figure below provides a sample of the dataset in.

5.2 Variables Description

5.2.1 Dependent Variable

In order to choose the best dependent variable capable of reflecting the Customer Experience and preferences, the success or failure of the project seemed to be the best solution.

In fact, using the number of backers as dependent variable would not give a general perspective on the customer experience. In fact, a project could have many backers but with a low average investment. On the other hand, using the total pledged amount as dependent variable would not give any indications about the number of backers.

Instead, knowing if a project was successful or not, is more consistent with the aim of this research. In fact, neither a large number of backers or a high funded amount taken alone are comprehensive indicators of the Customer Experience. Being successful on Kickstarter, instead, reflects both the above-mentioned phenomenon, indicating individuals' engagement with the project.

5.2.2 Independent Variables

The four dependent variables are tested with a Logistic Regression in order to provide some deeper understanding of the analysed phenomenon and to investigate the validity of the hypotheses.

- *Main Category*

  The first independent variable to be tested is the category. It is a categorical variable that assumes value 1 for the category of the project, 0 for any other category.

  The variables are significant, meaning that there is a relationship between the category and the success of a project. Only a few categories have positive log odds, and they are, in decreasing order: Dance, Theater, Comics and Music

- *Goal*

  The second independent variable to be tested is the goal. It is a discrete variable.

  The variable is significant. The coefficients in *Figure 15* highlight that, the higher the goal, the lower the probability of success

- *Country*

  The third independent variable to be tested is the Country.

  It is a categorical variable that assumes value 1 for the country of the project, 0 for any other country.

  It appears that there are differences in preferences between countries. This might be related to the perception that customers have about different countries

- *Currency*

  The fourth independent variable to be tested is the currency.

  It is a categorical variable that assumes value 1 for the currency of the project, 0 for any other currency. Not every currency seems to have a relationship with the success rate of a project. Using Euro as currency lowers the log of the odds of having a successful project, while GBP increases it.

5.3 Building the Model

Now that each Independent Variable has been analyzed individually, it is possible to run the ANOVA test to compare the consistency of different models built through the combination of variables.

The ANOVA test compares the models, providing useful indication about "added value" that a new independent variable brings to the model.

The following models are compared:

- Model 1: $logit(p) = \beta_0 + \beta_1 Category$
- Model 2: $logit(p) = \beta_0 + \beta_1 Category + \beta_2 Goal$
- Model 3: $logit(p) = \beta_0 + \beta_1 Category + \beta_2 Goal + \beta_3 Country$
- Model 4: $logit(p) = \beta_0 + \beta_1 Category + \beta_2 Goal + \beta_3 Country + \beta_4 Currency$

The final predictive Machine Learning Model consists of three independent variables and is as follows:

$$logit(p) = \beta_0 + \beta_1 Category + \beta_2 Goal + \beta_3 Country$$

Where *Category* and *Country* are Categorical Variables, and *Goal* is a discrete variable.

5.5 Testing the Model

Now that the model has been developed, two further steps have to be done:

- Testing the Model on the test set
- Measuring the accuracy of the prediction made on the test set

The prediction has a value between 0 and 1 and expresses the probability of success of a given project in the dataset. For instance, if a project has a *Predicted value* of 0,3, it means that it has 30% of success of achieving the funding goal.

Eventually, a new dichotomic variable, *Result*, is added to the dataset.
The column *Result* has value:

- *1,* if the probability of success is > 0,5
- *0,* if the probability of success is < 0,5

In order to calculate whether the predicted result matches the true outcome of the project, the new dichotomic variable *Match* is added to the dataset.

It has value:

- *1,* if the value in the column *Result* matches with the real outcome of the project
- *0,* if the value in the column *Result* does not match with the real outcome of the project

Now, to determine the *Accuracy* of the model, that means, the probability that the model predicts correctly the success or failure of a project, we compute the ratio between the number of correct prediction and the total number of projects in the test set: the model has 61% of probability of predicting correctly the outcome of a Kickstarter campaign.

6. Discussion

The results of the analysis on the variables and the Predictive Model provide enough evidences to draw conclusions on the four hypotheses developed and suggested in this Master Thesis.

> *H1: Individuals prefer to become "backers" of campaigns that are mostly about creative projects*

Hypothesis 1 is confirmed by the fact the categories that perform the best are those that offer intangible and artistic products.
This categorical variable is significant when analysed in the logistic regression and also in the Model.

> *H2: individuals are influenced in their investing decision by the value of the goal of a project*

Hypothesis 2 is tested with the Model and is true.
Differently from the findings of other studies (Sahlmann, 1997; Rakesh, 2015), it emerged that not only creators prefer realistic goals, even backers do.

H3: individuals' investment decisions are influenced by geographic boundaries

Hypothesis 3 is verified by the Model but is not true.
In fact, relevant differences occur between projects that are launched in different countries.

H4: Individuals prefer to invest in their own currency and therefore projects that ask for diffused currencies are more successful

Hypothesis 4 is verified and is not true.
Even though it seems to have a strong relationship with customer preferences, the categorical variable has no significance when put in a Model with other variables.

The findings of the research prove that customers are more attracted towards projects that offer artistic contents, with realistic goals, and that they are influenced by the creators' country of origin.

Yet, the model presents some limitations, leaving space for further studies that could improve the quality of the Model by adding further variables and implementing more sophisticated Machine Learning algorithms.