UNIVERSITY OF LJUBLJANA
SCHOOL OF ECONOMICS AND BUSINESS

MASTER'S THESIS

# A COMPARISON OF THE EFFICIENCY OF MODERN MACHINE LEARNING METHODS FOR PREDICTIVE MODELLING OF CUSTOMER LIFETIME VALUE

Ljubljana, January 2020 DAVOR HAFNAR

## AUTHORSHIP STATEMENT

The undersigned **Davor Hafnar**, a student at the University of Ljubljana, School of Economics and Business, (hereafter: SEB LU), author of this written final work of studies with the title **A comparison of the efficiency of modern machine learning methods for predictive modelling of customer lifetime value**, prepared under supervision of **doc. dr. Jure Erjavec** and co-supervision of **doc. dr. Jure Demšar**

D E C L A R E

1. this written final work of studies to be based on the results of my own research;
2. the printed form of this written final work of studies to be identical to its electronic form;
3. the text of this written final work of studies to be language-edited and technically in adherence with the SEB LU's Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the SEB LU's Technical Guidelines for Written Works;
4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;
5. to be aware of the consequences a proven plagiarism charge based on the this written final work could have for my status at the SEB LU in accordance with the relevant SEB LU Rules;
6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;
7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;
8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the SEB LU Study Information System;
9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;
10. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published.

Ljubljana, ________________________ Author's signature: _________________________

(Month in words / Day / Year,
e. g. June 1st, 2012

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES



## LIST OF APPENDICES



## LIST OF ACRONYMS

**AI** – Artificial intelligence
**BTYD** – Buy ‘Till You Die

**CCLTV** – Connected customer lifetime value
**CI** – Customer intelligence
**CLTV** – Customer lifetime value
**CRM** – Customer relationship management
**CSMV** – Customer social media value
**CVM** – Customer value management
**IT** – Information technology
**ML** – Machine learning
**$R^2$** – Coefficient of determination
**RFM** – Recency, frequency, monetary value
**RMSE** – Root mean square error

# INTRODUCTION

Digital economies provide an abundance of data. In recent years, there is an increased focus on extracting business value from it. A direct way to influence business results is by building predictive models based on past customer data and making inferences from it to future customer behavior. One of the most important predictions is the customer lifetime value (CLTV, sometimes also LTV or CLV) prediction.

CLTV can play a part in customer relationship management (CRM) business strategy. CRM can help the company to increase customer retention and loyalty, provide higher customer profitability, creates value for the customer, enables customization of products and services and provide higher quality products and services.

CLTV is gaining importance as traditional metrics used in marketing (or even sales numbers) are proving as insufficient to make marketing accountable and properly show a return on marketing investment. Even financial metrics such as stock price and aggregate profit have a limited diagnostic capability when it comes to predicting marketing success (Gupta, 2006). Furthermore, the value of the customer base of the firm can be used as a proxy for the value of a high growth firm, where traditional financial methods usually have difficulties (Gupta, Lehmann & Stuart, 2004).

Typically, CLTV is defined as the present value of all future profits earned from a customer through his relationship with a company (Prasasti, Okada, Kanamori & Ohwada, 2014). It was first introduced in 1974 by Philip Kotler (Kotler, 1974). It is similar to the concept of the discounted cash flow in finance, apart from two major differences. First, it is a disaggregated metric, meaning that it is defined and estimated at the individual customer or segment level (Chen, 2018). This allows us to explicitly differentiate between more and less profitable customers - companies want to identify most valuable customers, so they can properly focus their retention resources, such as discounts, offers, bonuses and gifts (Navlani, 2018). Second, unlike in finance, CLTV can explicitly incorporate the possibility that a customer may defect to competitors in the future (Gupta, 2006).

There are multiple models that can be used to predict the lifetime value of the customer. The ones most mentioned in the literature (for example in Tsai, 2013; Gupta, 2004; Chen, 2018) are the traditional RFM models, the deterministic CLTV models, the stohastic models and various machine learning (ML) models. These models can be either used to calculate a value created in the past or to predict the value of current customers.

ML models are an exciting field as the focus on and advances in the area have been massive in recent years. ML, which can be defined as the use of computational methods using experience to improve performance or make accurate predictions (Mohri, Rostamizadeh and

Talwalkar, 2018), is the area we will be focusing in the practical part of the assignment when we will be predicting the CLTV for customers of an online casino.

The purpose of the masters' thesis is to research the theory and usage of customer lifetime value and to develop and evaluate various machine learning models for CLTV prediction that can be used to predict the monetary value of customers in an online casino.

The goal of the theoretical part of the thesis is to explore and discuss various popular approaches to calculating CLTV and establish their pros and cons. In the second part of the theoretical part, I will look at some real-life usages of CLTV on various examples from different companies.

The goal of the empirical part is to establish how good are predictions made by different machine learning models on actual customer data from an online casino.

Overall, the goal is to answer the following questions:

- What is CLTV and how does it fit in the broader context of the company?
- How is CLTV being used in different real-life contexts?
- Which of the applied machine learning methods provides the best results?
- Does machine learning provide an effective prediction for CLTV in the case of an online casino?

The key research question of the thesis:

- Which of the modern machine learning methods is the most efficient in predicting CLTV in the case of an online casino?

This master's thesis includes two parts, a theoretical and a practical one. For the first part, I will use publicly available secondary sources, namely scientific articles and online sources to research the area of CLTV and its utilization in actual examples.

In the second, practical part, I have analyzed the anonymized online casino data provided by Oryx gaming d.o.o. Using modern data analysis techniques, I will first organize the available data about the behavior of customers into the appropriate format. This data will be then used as input to various machine learning algorithms. The result will be a model that will be able to provide an estimate of monetary value for each customer. Using multiple criteria, I will then determine which model has made the best predictions which I will then use to confirm or deny the hypothesis.

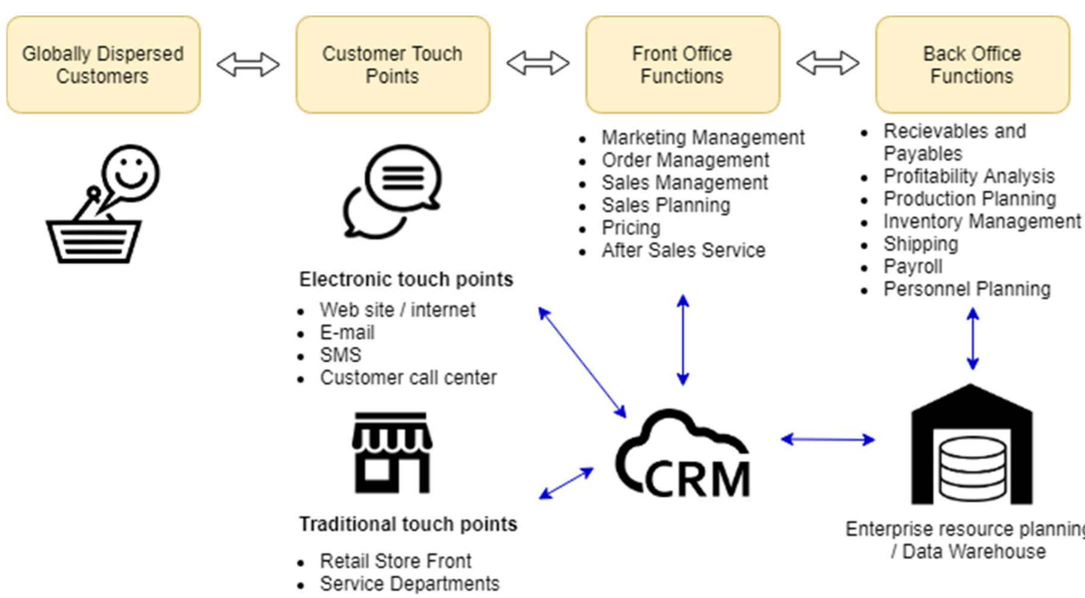
The thesis consists of nine chapters. The first chapter introduces the research topic and discusses the research problem, questions, objectives and methodology. The second chapter discusses customer relationship management. The third chapter discusses the CLTV. The fourth chapter focuses on presenting examples of implementation of CLTV on practical examples. The fifth chapter discusses the approaches to modelling of lifetime value. The sixth chapter describes the design and data preparation for a business problem we are solving. The seventh chapter presents an overview of machine learning models and scoring how well they perform. The eight chapter focuses on evaluating how the machine learning models performed. The last offers the conclusion of the thesis.

# 1 CUSTOMER RELATIONSHIP MANAGEMENT (CRM)

## 1.1 Overview

Mass production and mass marketing have increased the general product availability for customers in the mid-twentieth century (Chen and Popovich, 2003). Before that time, salesmen worked on building relationships with their customers by preparing them personalized offerings and closely followed what their customers need and want to buy. At the time, the salesmen were a part of the local community and that enabled them to know all the important information about their customers and to tailor sales offers appropriately (Rozek and Karlicek, 2014). As a result of a move towards mass marketing, the shopkeeper-customer relationship also got lost as customers got reduced to their account numbers. Today, the focus on individual customer needs is again gaining popularity, as it helps companies build long term customer loyalty. Management of the relationship with a customer, often based on or augmented by a technology solution is commonly referred to as customer relationship management or CRM (Chen and Popovich, 2003).

*Figure 1: Integration of CRM within the company*




*Source: Adapted from Chen and Popovich (2003).*

Data mining techniques and a customer-centric business strategy help the organization to get to know about the customer and offer more products and services relevant to that customer, even proactively, and build a long-lasting mutually beneficial relationship with that customer. At the same time, the data also tells the company which customers they are better off not serving – discounts and bonuses can be offered to selected customers, while withheld from others. The CRM approach, broadly speaking, represents a shift from managing product portfolios to managing portfolios of customers (Chen and Popovich, 2003). As a term, CRM has emerged in the information technology (IT) vendor and practitioner community in the mid-1990s. It is most often describing customer solutions that are technology-based, such as sales force automation (Payne and Frow, 2005). The position of a modern CRM system within a company is presented in figure 1.

It often happens that CRM means different things to different people. While for some CRM means direct e-mails, for others it means mass customization of products. For IT consultants, CRM often translates into complicated technical jargon. A high-level overview of CRM, required for a full managerial perspective, divides the concept into seven components:

- **Creation of a database** of customer activity,
- **Analysis** of that database,
- **Customer selection** based on analysis,
- **Customer targeting** with appropriate tools,
- **Relationship marketing** to build relationships with the targeted customers,
- **Privacy issues**, and
- **Metrics** for measuring the success of your CRM program (Winer, 2001).

### 1.2 Information management

An early concept of using customer data in marketing is called database marketing. It is an interactive approach to marketing that utilizes individually addressable marketing media and channels for the purposes of helping the company reach its target audience and stimulate their demand. It helps the company stay close to the customers by keeping a record of the interaction with the customer to help improve future contacts and ensure more realistic planning of all marketing activities. Database marketing relies on creating a bank of information about individual customers from, for example, records about their orders, enquiries, customer service contacts and research questionnaires (Stone and Shaw, 1988).

Customer databases have traditionally been analyzed to define customer segments. Customers with similar behavioural patterns have been grouped together with statistical methods such as cluster and discriminant analysis. The goal of the technique that was used for years was to target the most profitable prospects for emailing catalogues and to be able to tailor catalogues to different groups. In recent years, such segmentation approaches have been criticized as taking many customers to form a group focuses marketing towards an average customer in the group. Marketing tools have emerged that enable what is called one-to-one marketing and can reach customers one at a time with a tailored message. Those tools mean that usual market segmentation schemes, such as based on gender and/or age are no longer needed and the attention has moved to understand every individual customer (every individual row in the database), what can his value be, what can he deliver in the terms of profits and how to best address this customer. The idea of CLTV was born out of this focus on an individual customer. Once a lifetime value – a profit figure – can be assigned to each customer, the marketing managers can decide which customers to target (Winer, 2001).

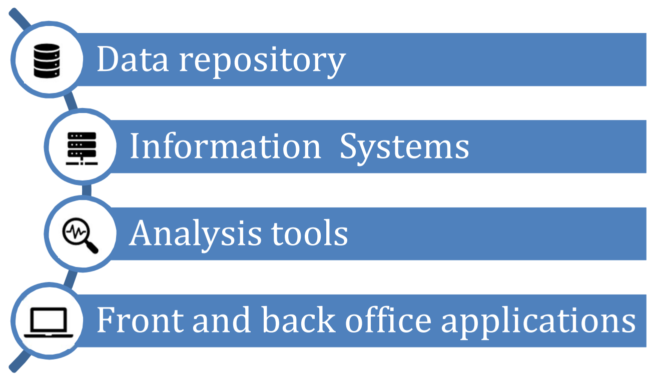
Analyses relating to customer insights can include fields such as identification of potential customers, prediction of the responses for of existing customers, calculating how much it costs to maintain a relationship with a customer and predictions about cross-selling. It is critical that there is clear communication between data analytics and marketing. Marketing's requests for analyses need to be properly understood. Analysts need to make sure that they are asking the right questions and advise marketing regarding the outcome of the analysis. Such cooperation between marketing and analysts help marketing make more fact-based decisions, while also being able to produce relevant data queries. In the end, to improve decision-making, the marketing will need to, first, understand the outcomes of the analyses and, second, be willing to use these outcomes when making decisions (Verhoef and Lemon, 2013).

The technology that emerged in recent years to enable CRM is called data warehousing. It enables CRM by consolidating and transforming data about the customer's information that enhances the company's understanding of customer behaviour. Constantly gaining knowledge about customers reduces the need for traditional, sample-based marketing research tools such as customer surveys and focus groups. Information stored by data warehouse is made available to all customer contact points in a company (Chen and Popovich, 2003).

Apart from details about customer transactions, there are many types of data generated from internal operations that can make significant contributions. There is information related to billing and account status, customer service interactions, backorders, product shipment, product returns, claims history, as well as internal operating costs that can improve the understanding of customers and their purchasing patterns. The ability of a data warehouse to store and quickly process a practically unlimited amount of data make drill-down analysis feasible as well as immediate for any department in the company that requires it (Chen and Popovich, 2003).

The process of information management collects, arranges and uses customer data from all the customer contact points with the company to generate insight into a customer and prepare appropriate marketing responses. Key material elements of the information management process (as listed in figure 2) are the data repository, which provides all the data on the customers, IT systems, which broadly include all the company's computer software and hardware, analysis tools and front office and back-office applications, which are intended to support directly reaching customers on one end and managing internal operations, administration, and supplier relationships on the other (Payne and Frow, 2005).

*Figure 2: Elements of the information management process*




*Source: Payne and Frow (2005).*

The data repository serves as a corporate memory of customers capable of relevant data analyses. In larger companies, it may include a data warehouse and related data marts and databases. IT systems refer to the computer hardware and the related software used in the organization. The historical separation between marketing and IT departments can present integration issues at the organizational level. It is critical for a company to be able to scale existing systems migrate to larger systems without disrupting business operations. For the data warehouse to be effectively used, analytical tools are required. They can be found in general data-mining packages and in specific software application packages. Data mining is an analysis of large quantities of data with the aim to discover meaningful patterns and relationships. Software applications include analytical tools that focus on specific tasks such as campaign management analysis, credit scoring, and customer profiling. To support all the activities that involve direct interaction with the customers the company uses front office applications. Those activities include sales force automation and call centre management. Back office applications are those that support internal administration activities and supplier relationships. Those include human resources, purchasing, logistics and financial processes. The key for front and back-office systems regarding CRM is that they are connected and coordinated well enough to be able to improve customer relationships and workflow (Payne and Frow, 2005).

### 1.3 Customer value management (CVM)

While CRM is a broad concept covering many aspects of managing customers, the concept of CVM focuses specifically on the value of the company's customer base and the economic values of the customer's relationship to the company. It does that by analyzing data on prospects and customers. Companies use the resulting information for the purpose of customer acquisition and retention and to drive customer behaviour with developing marketing strategies in a way that the value of all current and future customers is maximized. To increase customer value, companies focus on:

- Attracting new customers,
- Increasing customer retention,
- Facilitate the expansion of the customer base,
- Winning back old customers,
- Supporting relationship termination when needed,
- Effectively allocating resources among customers (Verhoef and Lemon, 2013).

Verhoef and Lemon (2013) suggest that for the best results, companies should use CVM to increase business performance, make sure that CVM is more customer than IT-driven, adopt CLTV as a core metric and invest in strong analytical capabilities. They should also be able to understand the key drivers of customer acquisition, retention and expansion for their business and how to manage the marketing channels to create customer value.

The fact that companies now have more data on various business processes brings more decision making that is intelligence-based, which is heavily reliant on data analytics. Having a strong analytical capability has become a strong competitive advantage. This advantage can benefit different parts of the company, be it human resources management, logistics, finance, or marketing. One of the important functions of analytics is customer intelligence (CI), which has the task to collect, store and analyze customer data. The insights into the customer need to be accurate, up to date and include as much relevant customer information as possible. The research has shown that the success of CVM in a company depends on the quality of the CI in a company – consequently, the firms with strong CI capabilities tend to have a greater firm performance (Verhoef and Lemon, 2013).

When it comes to implementing CVM on customers, it is not clear cut whether to increase or decrease the marketing investment towards loyal customers. Companies usually decide to discriminate against them and not offer them discounts and coupons because they are less price-sensitive and are less likely to buy because of the discount than customers with low levels of loyalty. Studies have shown that price promotions to existing customers can increase short term demand, but have a negative impact in the long run, while the same promotions offered to prospective customers can increase both short- and long-term demand. (Anderson and Simester, 2004). But when looked at trough prism of the CLTV, loyal customers are worth much more to the company and a loss of a loyal customer due to overcharging is extremely costly. Prioritizing switchers to loyal customers has shown to have these negative effects:

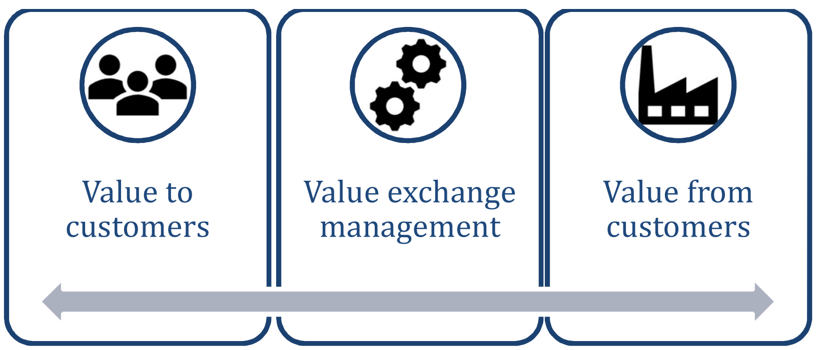
- Customers attracted by discounts lower the quality of the customer base and make it harder for the company to in the long run deliver the expected value to high-quality customers;

– Constant lowering of price cuts is teaching the customers to increase price sensibility and learn to switch;
– Loyal customers are unhappy when they realize they have been discriminated against for being loyal;
– Consumers can learn to wait and not buy until there is a discount (Villanueva and Hanssens, 2007).

The process of value creation creates programs aimed at both extracting and delivering value. The key elements (listed in figure 3) to such value creation process are:

- Determining what value the company can provide to its customer;
- Determining what value the company receives from its customers;
- Successful management of the value exchange, maximizing the lifetime value of desirable customer segments (Payne and Frow, 2005).

*Figure 3: Core elements in the value-creating process*




*Source: Payne and Frow (2005).*

The value the customer receives from the organization usually revolves around benefits that enhance the customer offer, but in recent thinking, the customer is seen also as a co-creator and co-producer. Customer's benefits can be integrated into the form of a value proposition that explains the relationship among the performance of the product, the fulfilment of the customer's needs, and the total cost to the customer over the customer relationship life cycle. To determine whether the value proposition is expected to result in superior customer experience, a company should asses the relative importance customers place on the different aspects of the product. From the perspective of the value for the organization, customer value is the outcome of the co-production of value, usage of improved acquisition and retention strategies and the utilization of effective channel management. It is important to understand how profitability varies across customers and customer segments. Also, it is important for the economics of customer acquisition and retention and also the opportunities for cross-selling and upselling must be understood. Influence of these elements, particularly retention, on customer lifetime value is key to value creation. Calculating CLTV of different segments enables companies to focus on the most profitable customers and customer segments. The value creation process is a crucial component of CRM because it translates strategies into value to be delivered to customers, and with that also explains what value can be expected by the company (Payne and Frow, 2005).

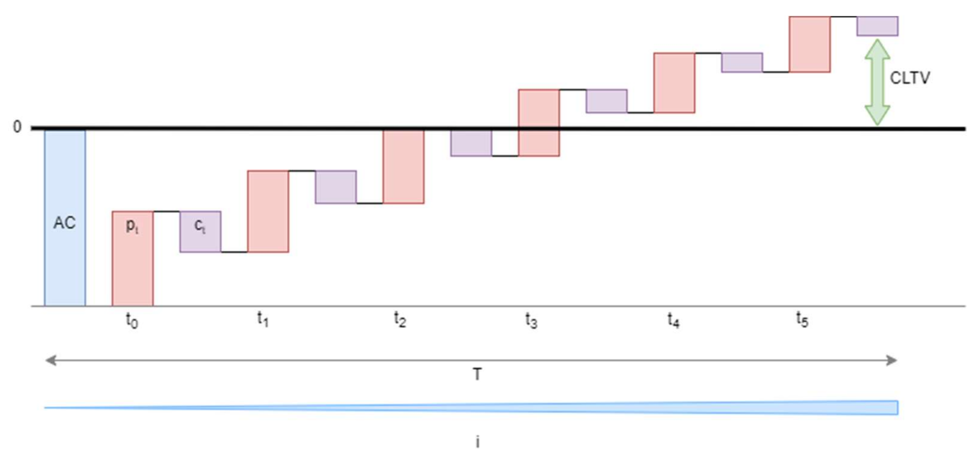
An important part of value management is performance monitoring. The problems in companies are that metrics used by companies for CRM performance measurement and control are neither well developed nor communicated. Numbers reflecting customer satisfaction or retention rarely reach the board level of the company, and even if they do, it is not clear how well they are understood or how much time is spent on them. The problem is also in the traditional performance measurement systems that don't focus on the future, but only historical financial results, which isn't appropriate to explain CRM. Standards, metrics and the key performance indicators used for CRM should ensure that activities are effective, and that feedback loop is in place to ensure continuous performance enhancements (Payne and Frow, 2005).

# 2 CLTV OVERVIEW

## 2.1 CLTV illustration

CLTV represents the present value of all future profits earned from a customer through his or his relationship with a company (Prasasti, Okada, Kanamori & Ohwada, 2014). A simplified explanation of the CLTV is presented in figure 4.

*Figure 4: A simplified explanation of CLTV*




*Source: Adapted from Bloniarz (2018).*

First, the customer needs to be acquired, which usually incurs a cost. The customer first makes a purchase $p_t$ in *t0*, and $p_t - c_t$ is the net value of that purchase. The customer pays off and starts making a profit for us at time *t3*. If the customer lifetime is represented by *T*, the CLTV for that customer is the profit the company made from it. Each future purchase has less worth to the company because of the discount rate *i* (hence the smaller columns as time periods progress).

## 2.2 Product-centric vs customer-centric metrics

Traditionally, companies put the focus on the profits from their products and use product-centric metrics, such as market share. The positive association between market share and return on investment has been proven in the past, with the logic being that firms with bigger market shares can gain economies of scale, increase market power and earn higher profits (Szymanski, Bharadwaj & Varadarajan, 1993).

The problem of product-centric metrics is that they are often be optimized incorrectly, for example by using short term sales boosters like sales discounts. The discounts only cause temporary shifts in sales, increase consumer price sensitivity and can consequently destroy customer loyalty. Product-centric metrics also only consider only one product category, while individual customers may buy from multiple categories. Purchase behaviour of customers within product categories can depend on each other and ignoring this correlation can reduce the overall value of a customer. If there is a policy of strong discounts in one product category, the same customer might be more sensitive to the price in other product categories by the same company. Being too product-centric can also mean that products within the same company can pose competition to each other as customers abandon a higher priced product and decide for a lower-priced from the same company, destroying customer value for the company in the process (Verhoef and Lemon, 2013).

*Table 1: Product-centric and customer-centric approach comparison*

| | **Product-centric approach** | **Customer-centric approach** |
|---|---|---|
| Basic philosophy | Sell products | Serve customers |
| Business orientation | Discrete transaction at a point in time | Customer life-cycle orientation |
| Product positioning | Highlight product features and advantages | Highlight product benefits in terms of meeting individual customer needs |
| Selling approach | Off-the-shelf products | Bundles that combine products, services and knowledge |
| Organizational focus | Internally focused. New product development, new account development, market share growth, and customer relations are issues for the marketing department | Externally focused. Customer relationship development, profitability through customer loyalty. Employees are customer advocates |
| Selling approach | How many customers can we sell this product to? | How many products can we sell to this customer? |

*Source: Bonnachi and Perego (2012)*

Customer-centricity addresses a growing awareness of the need to increase focus on customer-related factors such as customer satisfaction, customer service, customer loyalty, and quality as perceived by the customer. It addresses current-day shifts in management paradigms from product portfolio management to customer portfolio management, from goods-centred to service-centred dominant logic etc. Furthermore, recent research has emphasized the superiority of firms' financial performance by making marketing investments customer-centric. However, the true essence of the customer-centricity paradigm lies not in how to sell products but rather on creating value for the customer and, in the process, creating value for the firm. Customer centricity is essentially concerned with the process of dual value creation. The difference between both approaches is described in table 1.

CLTV is one of the core customer-centric metrics within the frameworks of CRM and CVM. It refers to the net present value of all future profits derived from a customer over his lifetime with the firm. CLTV can be used as:

- A metric to provide a more customer-centric culture in the company;
- A metric to evaluate marketing campaigns and investments;
- A tool to valuate the entire customer base;
- A metric for customer segmentation and resource allocation;
- An additional marketing metric in the customer database (Verhoef and Lemon, 2013).

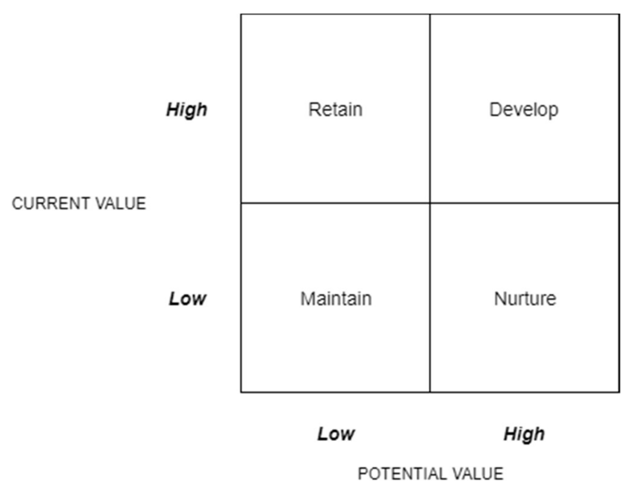
The first application of CLTV functions as a metric which goal it is to stimulate a customer-centric culture. The goal is to get everybody in the company to realize and acknowledge that the ultimate objective of the organization is to maximize CLTV. All the strategies need to be aligned on achieving this objective, consequently guaranteeing long-term perspective on customers. The second application of CLTV only focuses on increasing the accountability of marketing investments, as CLTV enables calculating return on investment (ROI) for different marketing activities. As a third possible use, CLTV can be used to assess customer equity, which can be of importance for the overall valuation of the company. This usage of CLTV underlines the importance of customers as assets, also in a financial sense, and helps companies understand whether they are exploiting the customer base to achieve short-term profits or looking longer-term (Verhoef and Lemon, 2013).

The fourth application of CLTV relates to database marketing. CLTV is calculated for current customers and then customers are usually segmented based on these calculations. Marketing resources can then be allocated across the resulting segments. A company might decide to reallocate marketing resources so that a lesser portion of the marketing budget is allocated to segments with a lower CLTV while a bigger part of the budget is allocated to segments with high CLTV. Linking CLTV to churn data for customers can also prove very beneficial.  The fifth way that doesn't create direct value to a firm is to use CLTV is as just another marketing metric present in the customer database. It is often the case that a company is still using and making decisions based on a product-centric approach, but the managers can also use CLTV to guide their decisions (Verhoef and Lemon, 2013).

## 2.3 Usage of CLTV in organizations

Usually, the first usage of CLTV is for customer acquisition, as it tells us the upper bound of what the company should spend to acquire a customer (Toccu and Fasso, 2013). Once the customer is acquired, CLTV redirects marketing focus from the mass production with one-way marketing communication back to the customer. That makes CLTV beneficial to how marketing perceives and works with customers as it shifts marketing from the transactional to a long-term relationship approach. From an organizational viewpoint, the change CLTV model brought was to view the customer as a real business partner – the customer is no longer a passive consumer, but a strategic company asset. This change is reflected in how the customer is treated and how much effort is made by the company to understand him. Once the customer is viewed as a long-term asset, relationship changes, as the company must make sure that the asset is managed correctly and that the maximum value is received from it (Rozek and Karlicek, 2014).

*Figure 5: Recommended actions based on customer value*



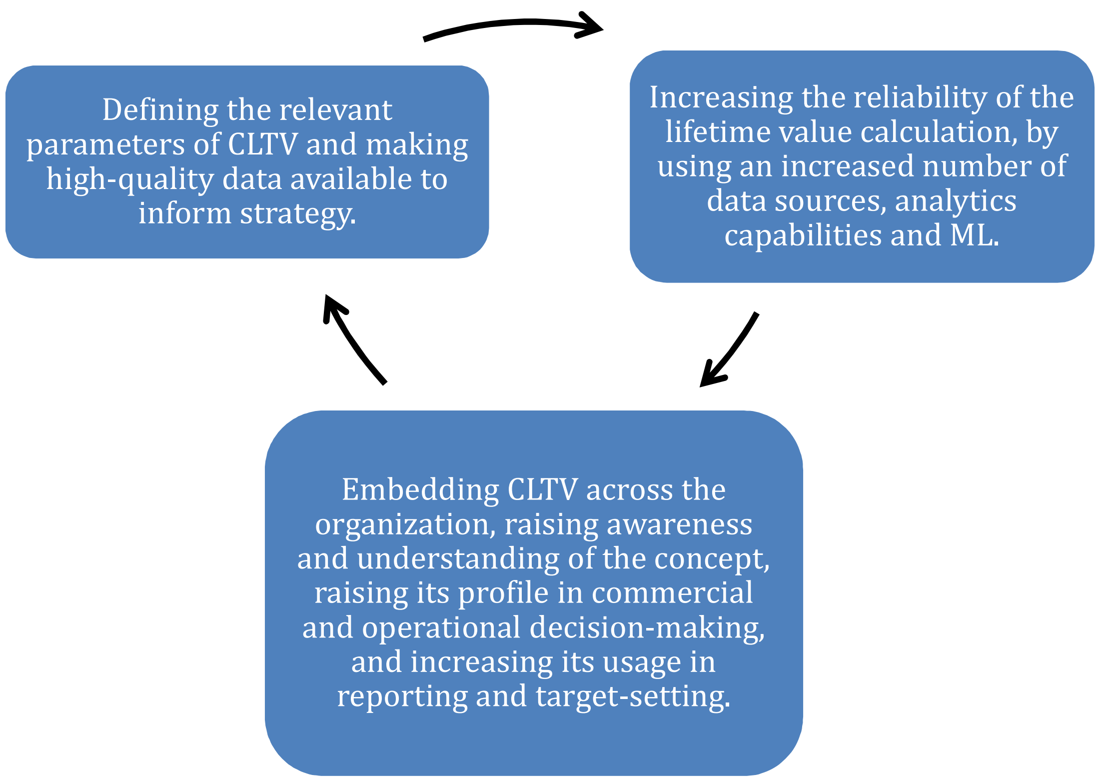

*Source: Hawkes (2000).*

An important question for any company when discussing CLTV is what better business decisions can be made by using it. What would be done differently if it was calculated? To drive strategies and plans, it could be useful to divide the value into current and potential value.

Figure 6 represents that division and each box in the matrix describes a value and indicates a strategy or action. A company clearly wants to retain customers with high current and low potential value, but the scale of future investment in them should be controlled. Also, the way in which a company will choose to manage a recently acquired high-value customer will be very different to how it will manage a customer of the same value who has already a proven to be loyal or one showing signs of possible defection (Hawkes, 2000).

The possibility to quantify the impact of each customer and calculate return on each marketing investment is a big advantage of CLTV when it comes to the financial point of view. It enables management to apply financially quantifiable key product indicators for all the marketing activities in the company and address one of the biggest griefs of marketing, namely the problem with measuring the return on investment and consequently accountability of the marketing department. The role of marketing within the company shifts towards selecting profitable target customer group, selection of the acquisition strategy and a long-term profitable retention strategy. Marketing coordinates the activities that relate to that strategy throughout the whole firm from the back-office departments like customer service and accounting to front office departments like sales. The push towards the customer has to be both external and internal to the company (Rozek and Karlicek, 2014).

*Figure 6: Constant improvement loop of CLTV in an organization*




*Source: KPMG (2019).*

Proper usage of CLTV requires constant improvement and re-checking of the basic assumptions, increasing the value of the usage of CLTV over time. The loop is represented in figure 6.

Using CLTV in combination with CRM system can increase profits via:

- increasing the number of products purchased by the customer, by cross-selling;

– increasing the price paid by the customer, by up-selling or charging higher prices;
– reducing product marginal costs;
– reducing customer acquisition costs. (Winer, 2001)

One of the major reasons that CLTV is not being implemented in small and medium-sized companies is the fact that they usually do not have analytical teams and large marketing departments. To effectively work with CLTV marketing managers need high-quality systems for aggregating customer data and financial information in the company. Similarly, managers in financial departments don't usually have the extra capacity to ensure necessary changes to set accounting systems in a way to allocate costs on customers. It can also be difficult for small companies to be able to allocate costs on customers as it can be difficult to accurately monitor the profitability of individual contracts. Accounting systems are set up to monitor corporate profitability or profitability of products, orders or projects and not cost allocations of the customer. As survival of small companies relies on focusing on the day to day of the core business, information systems and processes necessary for implementing CLTV models are usually minimized. This helps managers meet their annual targets by keeping costs under control but hurts their ability to get relevant information supporting their managerial decisions (Horák, 2017).

*Table 2: Maturity of CLTV usage in a company*

| CLTV Maturity | Sophistication of CLTV calculation | How embedded is CLTV within the organization |
|---|---|---|
| Not used | • CLTV is not calculated at all.<br>• Company relies on other metrics, such as number of contracts and total margins. | • CLTV is not used at all.<br>• The company does use several components of CLTV, but mostly in isolation. |
| Enabled | • CLTV is calculated ad hoc using a limited number of inputs and on a customer segment level.<br>• Calculations are usually on a spreadsheet. | • CLTV is used for a limited number of purposes, mostly commercial strategy.<br>• It is not a standard KPI but used on an ad hoc basis. |
| Developed | • CLTV is calculated using an increased number of parameters and updated regularly.<br>• Calculations are usually performed using analytics tooling. | • CLTV is used for key strategies such as product development, acquisition and retention.<br>• CLTV is used as one of the main metrics for steering the company. |
| Optimized | • Real-time calculation of CLTV at an individual customer level.<br>• Both internal and external data is used to calculate CLTV. | • CLTV is widely known and understood across the organization and used for strategy and target setting, but also for operational purposes. |
| Agile | • Machine learning (ML) is used to increase the accuracy of CLTV calculations.<br>• A tailored model continuously updates CLTV calculations and evaluates which inputs produce the best CLTV estimates. | • ML-based systems make decisions based on optimizing CLTV.<br>• Commercial strategies and operational actions are based on CLTV insights.<br>• CLTV is linked to executive metrics and incentives. |

*Source: KPMG (2019).*

Global accounting firm KPMG has rated companies according to their maturity when dealing with CLTV. The maturity ranges from companies who don't use the metric at all to the companies that are considered agile in their usage. The ratings are described in table 2.

### 2.4 Customer retention

Long-life customers are a benefit to the company because:

- It is cheaper to retain customers than it is to acquire them;
- Serving long-life customers is cheaper than serving new customers;
- Long-life customers help you attract new customers through word of mouth;
- Long-life customers are not as price-sensitive and pay higher prices;
- Long-life customers are more likely to buy from the company, so their value can be increased by upselling and cross-selling.

To summarize, the customer lifetime is positively related to profitability, and profits for retained customers increase over time (Villanueva and Hanssens, 2007).

*Figure 7: Ways to boost retention via customer satisfaction using CRM*


*Source: Winer (2001).*

Retention process needs to be strategically managed and should ensure that the expected CLTV is brought to the company by being strategical in building relationships with customers. The goal is to ensure repetitive buying of the acquired customers and to optimize the total customer equity by managing the optimal retention rate on the optimal customer base with a positive CLTV index. The keys to retention are customer satisfaction (the positive difference between expected and delivered quality), customer perceived value (value divided by price), product uniqueness, ease of purchase and the ease to exit (Rozek and Karlicek, 2014).

As today's customers usually have many choices, customer service should be of high priority in any company (see figure 7). Any part of the company that has direct contact with the customer has the potential either to gain or to lose repeat business. There are usually two types of programs relating to customer service; proactive and reactive programs. Reactive are the ones where they help the customer with a problem, like product failure or return and proactive are the ones where they predict a possible reactive situation and contact the customer in advance (Winer, 2001).

Loyalty programs reward customers for repeat purchasing. There are some major problems with such programs: they are expensive, mistakes are difficult to correct as customers see it as benefits being taken away from them and, most importantly, it is questionable how much they contribute to increasing loyalty and spending behavior. Loyalty programs are very common is making it hard for the company to gain a competitive advantage with such a program. Regardless, in industries like the airline industry loyalty programs can be very successful in increasing switching costs and building barriers to entry. They are also hard to avoid, as in some industries, they have become a competitive necessity (Winer, 2001).

The idea of mass customization has moved beyond just marketing to individual customers as it implies the creation of products or services for individual customers. It enables the customers to be the makers, not just takers of the product. An example of the approach is Dell with its early build-to-order computer website. It is quite easy to do with services and information goods and a harder with products, but it is possible there also, even the customization is just in variations based on the same essential product, like in ordering custom color and equipment on a car (Winer, 2001).

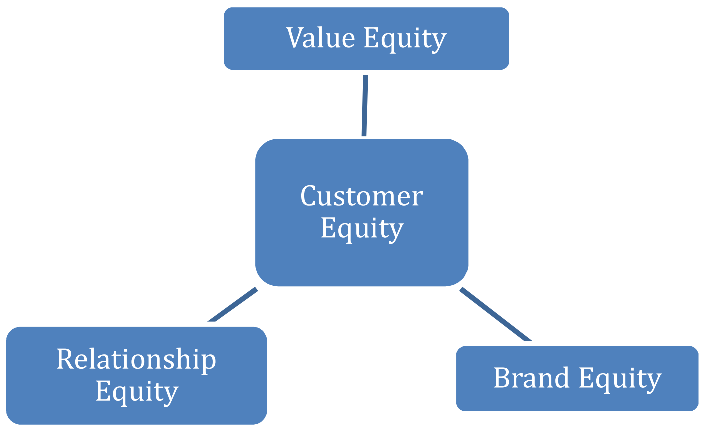
Building an online community of customers for exchanging product-related information and creating relationships between the customers and the brand is one of the more important uses the web for both online and offline companies. The goal is to take a prospective relationship and turn it into something more personal, which makes it harder for the person to leave the community of other customers. Giving customers the impression that the section of the website where tips and other information is exchanged is owned by them helps build a more personal relationship (Winer, 2001).

## 2.5 Customer equity and customer-based valuation

Long term value of the firm is in a lot of cases mostly determined by customer equity. Even if it doesn't speak about the entire value of the firm – there are physical assets, intellectual property etc. – it is the customers that provide the most reliable source of future profits. According to Lemon, Rust and Zethaml (2001), there are three drivers of customer equity – value equity, brand equity and relationship equity (see figure 8).

For the customer, value is a key component of the relationship with the company. If the company's offerings do not meet the customer's expectations or needs, the brand and relationship cannot make up for it. The value equity is an objective assessment of how useful the brand is based on how much a customer must give and how much he receives. Value equity is a combination of quality, price and convenience. Quality is about objective offerings under the firm's control, price represents what the customer must give up and convenience relates to the amount of effort required by the customer to deal with the company (Lemon, Rust and Zeithaml, 2001).

*Figure 8: Elements of customer equity*




*Source: Lemon, Rust and Zethaml (2001).*

As opposed to the value equity, which is built in relation to objective aspects of the company's offerings, value equity is built trough image and meaning. The brand serves the role to attract new customers, to remind the customers about the company's products and services and to become the customer's emotional tie to the firm. The key levers of brand equity are brand awareness, attitude toward the brand and the company's corporate ethics. Brand awareness encompasses the tools the company controls and can enhance brand awareness and marketing communications. Attitude towards the brand encompasses how close connections the firm is able to create with the customer, which is most often influenced through the specific media campaigns and can be influenced by direct marketing. Corporate ethics, company's third brand equity lever, encompasses the actions of the company that influence customer's perceptions of the organization, like sponsorships and donations or employee relations (Lemon, Rust and Zeithaml, 2001).

Relationship equity is required as the strong brand and product value by itself are no longer enough to retain a customer. It is the tendency of the customer to stick with the brand, which goes beyond just how he perceives the brand. Key levers that the company can use to enhance relationship equity are loyalty programs, special treatment, affinity programs and community or knowledge-building programs.

Loyalty programs reward customers for specific behaviors with tangible benefits, while special treatment provides intangible benefits like special statuses or memberships. The goal of affinity programs is to create strong emotional connections with customers, linking the customer's relationship with the firm with other aspects of their lives—for example, a personalized credit card. Community-building programs link the customer to a larger community of customer with similar preferences or characteristics while knowledge-building programs learn about the customer and can offer them specific benefits based on these insights, which is something the customer won't get if he switches to another company (Lemon, Rust and Zeithaml, 2001).

The value of the company's customers can be used as an evaluation tool for company value. It is in fact necessary when traditional financial approaches fail, which is often the case when business analysts try to valuate high growth and start-up companies. In early periods, those companies are investing heavily, but don't have any earnings or have negative cash flows. Startup companies also only have a limited history of cash flow and, in the case of internet-based firms tend to have very little tangible assets, especially if they are highly virtual, like tech firms. The value of those companies mostly comes from customer relations or brands (Bauer and Hammerschmidt, 2005).

A CLTV-based corporate valuation is also very relevant when the valuating marketing-related synergies conducted by acquirer companies when it comes to mergers and acquisitions. Marketing and sales synergies often have stronger effects than production, personnel or procurement aspects but their assessment has long been neglected as companies usually focus on synergies in the field of costs.

As CLTV model covers the entire operative revenue and cost in a customer-related way it can help identify synergies with regard to acquisition and marketing costs, as the economies of scale and quantity discounts in the field of advertising, promotion or direct mailing. When it comes to revenues, a customer-based model can assess whether a higher customer retention rate can be achieved through better product or service offerings by leveraging customer loyalty programs to the customers of the acquired company. The value of the customer base of an exterior company can also be used to determine how much it still makes sense to spend on acquiring a company. Similarly, the minimal lifetime value that must be achieved can be estimated in to economically justify the price that has been paid to acquire the customer (Bauer and Hammerschmidt, 2005).

## 2.6 Connected CLTV (CCLV) and customer social media value (CSMV)

Weinberg and Berger (2011) proposed an extension to CLTV that would include the value of influence associated with a customer's network of connections. They introduced CCLV as CLTV plus the present value of other customers that are due to the influence of that customer. The added value includes new customers who would not become customers otherwise, increase in purchases by existing customers, and increased retention of existing customers. They also proposed CSVM to present the value derived from social media engagement (Weinberg and Berger, 2011).

The key metric when tying together CLTV and word-of-mouth are referrals. The type of promotion known as 'refer a friend' is quite common in a variety of industries, such as web subscriptions, restaurants, wireless services and financial services. To include referrals to CLTV, the customer referral value is determined by various formulas and added to the basic CLTV engagement (Weinberg and Berger, 2011). Table 3 illustrates how CCLV can differ from CLTV.

*Table 3: Customer value considering CRV*

| Customer | CLTV | CRV | Total value | Rank by CLTV | Rank by total value |
|---|---|---|---|---|---|
| A | $ 5810,71 | $50 | $ 5860,71 | 1 | 3 |
| B | $ 5447,56 | $550 | $ 5997,56 | 2 | 1 |
| C | $ 5127,11 | $750 | $ 5877,11 | 3 | 2 |

*Source: Weinberg and Berger (2011).*

CSMV is proposed as an addition to CRV and is designed as a multiplicative model, as there is likely an interaction effect among the different social media effects on CSVM. The underlying assumption is that a person with higher CLTV will, on average, have a greater impact. It also considers what weight a social media channel has. The CCLV is calculated in equation (1).

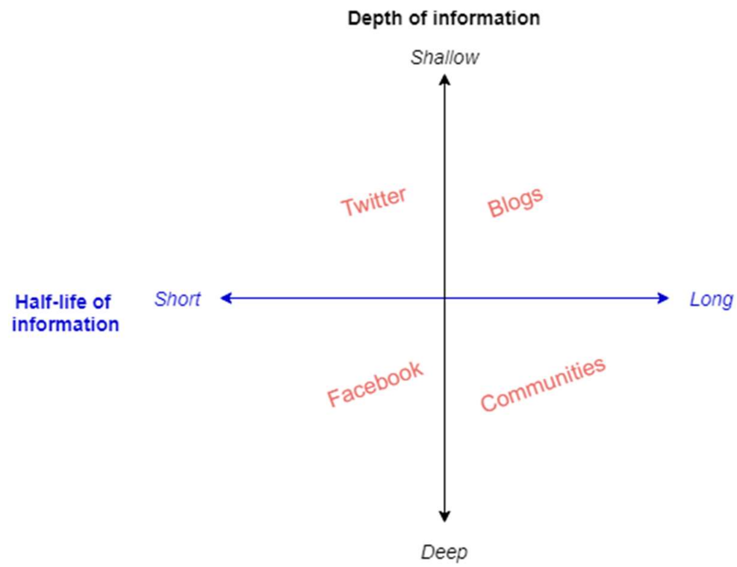
$$CCLV = CLTV + CRV + CSMV \quad (1)$$

$$CSMV_i = CLTV_i \times ([1 + SM_{i1}] \times \ldots \times [1 + SM_{ij}] \times \ldots \times [1 + SM_{iJ}] - 1) \quad (2)$$

CSMV is calculated in equation (2), where $i$ indexes individual customers from 1 to $I$ and j indexes different social media from 1 to $J$. The $SM_{ij}$ term reflects, for customer $i$, the impact of the social medium $j$. Social medium j represents any of the applications, such as blogs, communities, Twitter or Facebook (Weinberg and Berger, 2011).

The two aspects of social media that are contributing to the CSMV are the nature of social media itself and the extent to which the customer is engaged with this social medium. The dimensions of social media can be used to support marketing decisions: the depth of information and the half-life of information. The first refers to the richness of contents and the variety of perspectives —for example, on Facebook, a great amount of rich information can be found. The second, the half-life of information considers both the medium and the content and refers to how long the information appears on the screen and how big is the interest in the topic. For example, a tweet can effectively disappear from Twitter in a matter of seconds, depending on how many followers the person tweeting has (Weinberg and Berger, 2011). The relationship is illustrated in figure 9.

*Figure 9: Social media positioned regarding the depth and half-life of information*




*Source: Weinberg and Berger (2011).*

If we divide user's engagement on a social channel to high, medium and low and the depth of information and half-life to high and low, we can illustrate value assignment in table 4.

*Table 4: Value assignment example for social media channels*

| Engagement level | Facebook | Twitter | Blog | Community |
|---|---|---|---|---|
| Low | 0.04 | 0.02 | 0.04 | 0.06 |
| Medium | 0.06 | 0.04 | 0.06 | 0.08 |
| High | 0.08 | 0.06 | 0.08 | 0.10 |

*Source: Weinberg and Berger (2011).*

If we know (via a poll or direct data) that a customer is highly engaged on Facebook, has and low engagement on Twitter, blogs and community sites, we can calculate the CSMV value for him with equation (3).

$$CSMV = CLTV \times ([1 + 0{,}8] \times [1 + 0{,}2] \times [1 + 0{,}4] \times [1 + 0{,}6] - 1) \quad (3)$$

$$CSMV = CLTV \times 0.0214 \quad (4)$$

As shown in equation (4), this puts CSMV value to slightly above $^1/_5$ of the value of the user's CLTV. Having and using the CSMV can, according to the framework proposed by Weinberg and Berger (2011), have quite a big influence on the end result and ranking of the users according to their value.

### 2.7 CLTV Limitations

Usage of the CLTV also has some limitations, one of the keys being that there is no guarantee that the customer will stay consistent in his future decision-making progress, which is especially important for the more linear CLTV predictions. Another downside is a considerable investment needed to create customer databases required to make quality predictions. There is also an indirect cost which relates to the implementation of a true customer-oriented culture within the company. It is important to work with and develop customers in order to increase their value to the company and not just use CLTV as another input value for the traditional short-term marketing strategies (Rozek and Karlicek, 2014).

There are other general limits to the CLTV model. For international corporations, the models do not solve the issue of the tax burden and its impact on cash flow. CLTV by default puts the focus on customers that are more profitable and not on increasing the number of less profitable customers, which might not be an optimal strategy in all the cases. The models used also do not include factors for purchasing decisions and do not include the switch costs that the customer would incur by making future purchases elsewhere. The models also do not incorporate the risk of competition and what might happen if new competitive products are launched on the market. What also isn't included is the risk of potential customer insolvency (and unpaid invoices) and the inherent risk of not properly allocating costs to individual customers (Horák, 2017).

## 3 EXAMPLES OF CLTV IN PRACTICE

### 3.1 Summary

CLTV isn't just a theoretical concept. It has been implemented using different approaches in real-life companies. In this chapter, we briefly examined the references to the concept from the so-called big tech companies and then the ways how CLTV was tailored to some companies from different industries. To explore how to approach implementing CLTV we examined the implementation in a financial services firm, which uses CLTV to deal with customers even in real-time and example of an Italian telecommunications company, where CLTV exposed different balance between the users than when looking at sheer revenue. We also examined the case of IBM, where CLTV helped increase revenue by focusing on the right users and an example of Korean telecommunication company and the insight that CLTV brought them into their users and where to focus in the future.

### 3.2 CLTV in the context of Big Tech companies

In an article by Ciaccia (2014) a Deutsche Bank analyst is quoted that CLTV is "*obviously a metric*" that Amazon *"is focused on vs. near term margins.*" Some analysts believe that Apple, at least with some products, is starting to focus on customer lifetime value over the usual high margins. The aggressive pricing of Apple TV Plus prompted an interpretation from Macquarie Group's analyst Benjamin Schachter to write that "*while perhaps not a loss leader in the traditional sense, this signals a shift to focusing on customer lifetime value in its ecosystem*." (Savitz, 2019)

CLTV is supported by Google with Google Analytics. "*The Lifetime Value report lets you understand how valuable different users are to your business based on lifetime performance. For example, you can see lifetime value for users you acquired through email or paid search. With that information in hand, you can determine a profitable allocation of marketing resources to the acquisition of those users*," says their description (Google, 2019a).

Facebook is including CLTV as a metric that can be used by clients advertising on its platform. It defines it as "*a numeric representation of the net profit you predict will be attributable to a given customer over the duration of your relationship with them*."

They are breaking it down into the following factors:

- How often a customer makes a purchase within a typical purchase cycle,
- How much a customer spends each time they make a purchase,
- How much the company projects a customer will spend over the duration of the company's relationship with them,
- The potential length of a customer's relationship with the company (Facebook, 2019).

### 3.3 Capital One example

Capital One, credit card and financial services firm, has adopted CLTV as a core metric. They carefully segment and value its customers to understand their lifetime value. They use their value to identify the best utilization of their resources and evaluate marketing campaigns and marketing investments (Verhoef and Lemon, 2013). The company's orientation is shaped by a belief in micro-segmentation and their goal is "to deliver the right product, at the right price, to the right customer, at the right time," (Day, 2003). The micro-segmentation is done based on the customer's CLTV. In the United States, companies credit card business is segmented into several groups of customers: prime, high response, medium response, partnership, affinity, and small business. Each segment has its own manager that runs its segment as a business and is responsible for the profit and loss made by that segment (Verhoef and Lemon, 2013).

Apart from using CLTV as a tool for customer segmentation, Capital One is using it to test and evaluate new product ideas and marketing campaigns. They run tens of thousands of tests to improve customer acquisition, maximize CLTV and identify and terminate customers that are not profitable to the company (Anderson and Simester, 2011). In an example of one such test which was designed to improve customer retention Capital One compared customer responses. They tested it on three different actions over three randomized groups of customers who called in to cancel their Capital One credit card because they had received a better offer from the competition. For group A, Capital One agreed and closed the customer's account. For group B, Capital One matched the customer's alleged offer from a competitor. For group C, Capital One met the customer halfway with an offer. Later, data was collected on customers and responses to the offers, and results linked to existing customer data and CLTV. Because of these tests, when a customer calls in to cancel an account, the service representative instantly sees a recommendation for the customer based on these tests and CLTV and can make an offer instantly. The result is one of the highest retention rates in the industry and high customer satisfaction (Verhoef and Lemon, 2013).

### 3.4 Italian telecommunications company example

Toccu and Fassso (2013) have done a CLTV statistical analysis of an Italian telecommunications company. They analyzed the distribution of customers and revenues with respect to geographical data and the size of a business customer, the distribution of CLTV and used a regression model to detect the factors affecting CLTV. They analyzed a dataset of 30,379 active Italian business customers, where about 20 percent of the customers were large and on average generated about six times the amount of revenue than small customers.

*Table 5: Comparison of revenue a CLTV based valuation*

| Customer size | Of overall customers (%) | Average revenues (€) | Average CLTV (€) |
|---|---|---|---|
| Small | 80,57 | 252,36 | 265,26 |
| Large | 19,43 | 1412,28 | 833,02 |

*Source: Tocu and Fasso (2013).*

The analysis has shown that the CLTV was highly concentrated in the top 10% of the customers, who accounted for approximately 60% of all the CLTV. The comparison between small and large companies shown, that when it comes to CLTV, the difference is smaller than when looking at revenue, as large companies on average are worth, on average, in CLTV terms only about three times more than small companies (see table 5). The biggest factor contributing to CLTV for customers were mobile calls, which as an input variable influenced CLTV twice as much of fixed-line calls. The average revenues are from the data for the second half of 2010.

### 3.5 IBM example

IBM is one of the companies that explored CLTV as an indicator of customer profitability. They implemented a pilot study that included about 35,000 customers. Change from past spending history approach to CLTV approach led to the reallocation of resources for about

14% of the customers. In the pilot study, CLTV-based resource reallocation led to a tenfold increase in revenue without increasing marketing investment (Kumar, Venkatesan, Bohling and Beckmann, 2008).

At the time the study was conducted, IBM based their contacting strategy largely on a past relationship with a customer. Customers were contacted through several channels, such as a salesperson, direct mail, telesales, e-mail, and catalogues. Midmarket customers—companies with the number of employees ranging from 100 to 999—were, for example, contacted primarily through direct mail, telesales, e-mails, and catalogues (Kumar, Venkatesan, Bohling and Beckmann, 2008).

IBM sorts its customers yearly based on their score on a customer-level metric. Then they prioritize marketing contacts based on this score. Return on marketing is importantly determined by the choice of the metric used to score customers. To maximize potential value from customers, the metric is constantly refined. Through the 1990s, a customer spending score (CSS) was used to score customers. CSS was defined as the total revenue that can be expected from the customer in the next year. The customers with the highest scores were selected for future targeting. The metric was augmented yearly. It was eventually abandoned, as it disregarded the variable cost of serving the customer and therefore the bottom line. Trials of customer profitability and CLTV were proposed as an alternative for scoring customers. (Kumar, Venkatesan, Bohling and Beckmann, 2008).

*Table 6: IBM marketing resources reallocation based on CLTV*

| Decile | Not contacted until 2004 ($) | Contacted by 2004 ($) | Customer engagement |
|---|---|---|---|
| 1 | 350,471 | 2.124,48 | Super high CLTV |
| 2 | 993 | 125,46 | High CLTV |
| 3 | 669 | 43,681 | Medium CLTV |
| 4 | 638 | 23,624 | |
| 5 | 623 | 17,499 | |
| 6 | 611 | 13,675 | |
| 7 | 534 | 10,513 | |
| 8 | 444 | 8,051 | Low CLTV |
| 9 | 369 | 5,023 | |
| 10 | 80 | -35 | |

*Source: Kumar, Venkatesan, Bohling and Beckmann (2008).*

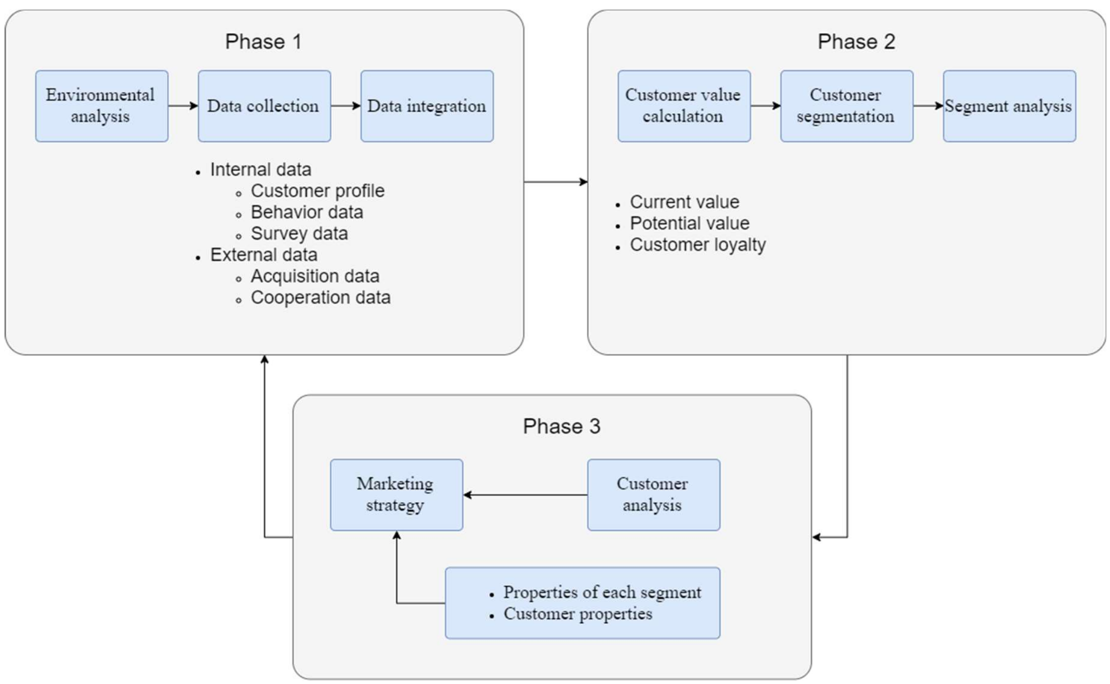
They decided for a trial run of CLTV based resource allocation on a part of the customers. The calculated the CLTV for customers and segmented them within deciles from super high to low CLTV. They looked at the customers that were contacted in previous years – so the previous system assigned them as valuable enough to allocate marketing resources on – and customers that haven't been contacted before. The calculations have shown that CLTV of the last decile of contacted customers was negative, and the reallocation of marketing resources was made from that decile to some customers that were not contacted before, but the CLTV calculation has shown they are of great value to the firm (Kumar, Venkatesan, Bohling and Beckmann, 2008).

As a result of contacting customers that weren't contacted before, the revenue from those customers went up about 10 times (see table 6). The biggest part of the increase came from customers who didn't make any purchases in the previous year but made the purchase in the next year when they were being contacted after being flagged as valuable by a CLTV calculation (Kumar, Venkatesan, Bohling and Beckmann, 2008).

### 3.6 A CLTV based strategy for a Korean telecommunication company

Kim, Jung, Suh and Hwang (2006) proposed a framework for analyzing customer value and segmenting customers based on their value and illustrated it through a case study on a wireless telecommunication company. The framework outline is presented in figure 10.

*Figure 10: A framework for analyzing customer value*




*Source: Kim, Jung, Suh and Hwang (2006).*

The data they used related to 6-month service data of a Korean wireless communication company. They calculated CLTV in three steps, by calculating current value, potential value, and customer loyalty. They calculated the current value as the average amount of service charge asked to pay for a customer and deducted average charge in arrears for a customer for the 6 months.

$$PV_i = \sum_{j=1}^{n} prob_{ij} \times profit_{ij} \quad (5)$$

They considered cross-selling and up-selling and calculated the potential value with the calculation represented by equation (5). In the equation, $prob_{ij}$ is the probability that customer *i* is going to use the service *j* among *n*-optional services. $Profit_{ij}$ stands for the profit that a company can receive from the customer *i* who uses the optional service *j*. Essentially the equation above describes expected profits from a particular customer who uses optional services provided by a wireless communication company (Kim, Jung, Suh and Hwang, 2006).

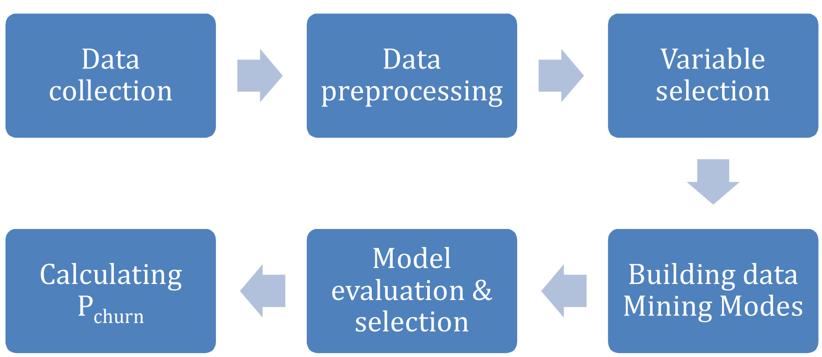
Customer loyalty is calculated as the index that customers would like to remain as customers of a company.

$$CL = 1 - CR \quad (6)$$

*CL* in equation (6) stands for customer loyalty and *CR* for churn rate. Churn predictions process is presented in figure 11.

The analysis has shown that ninety percent of customers in the low customer loyalty segment are classified by the criterion of whether the customer is using a subscription plan. Even though they pay a membership fee, customers who work as a company employee or run their own businesses also represent low customer loyalty. Based on the customer loyalty analysis, the suggestion is that the company is better of charing a small portion of the membership fee and compensate for it in some other way than to rely on the pre-paid model (Kim, Jung, Suh and Hwang, 2006).

*Figure 11: Building churn rate predictions*




*Source: Kim, Jung, Suh and Hwang (2006).*

Based on the analysis, the company should also be more attentive to customers with older devices and offer them a good plan as they are likely to pick a new provider based on what kind of device he can provide them. For customers who have high current value and high possibility to switch to other providers, a company should provide a good deal on a new phone or even offer a free upgrade to keep customers who have high current value. As customers with high potential value have been shown to be younger people from the capital area, the company needs a strategy to attract this segment by providing more attractive service plans (Kim, Jung, Suh and Hwang, 2006).

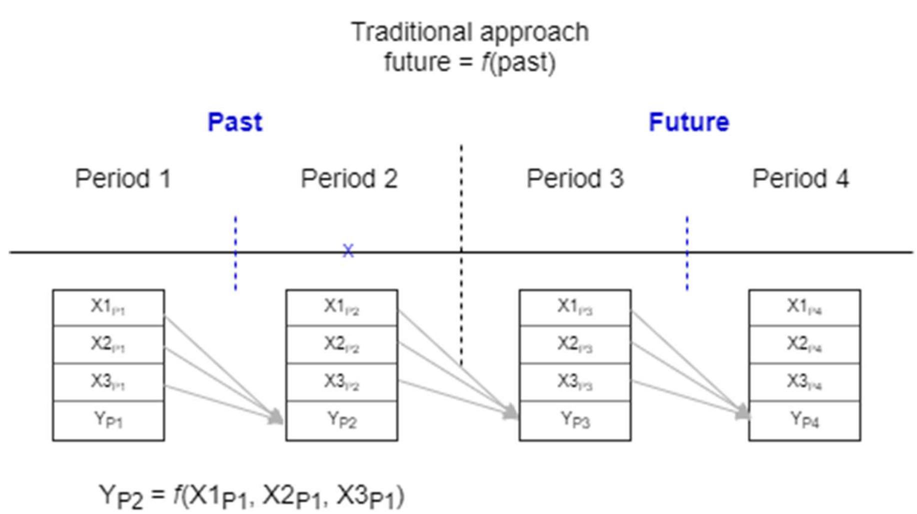
# 4 MODELLING LIFETIME VALUE

## 4.1 Traditional modelling approach – the RFM models

RFM models have been used for more than 40 years. Before the use of these models, companies used less accurate demographic profiles of customers to decide on the targeting strategy. RFM models were developed to target marketing programs such as direct mail or phone calls at specific customers with the objective to improve response rates (Gupta and others, 2006). The models got their name from the three variables they use from Period 1 to create predictions for Period 2. Those variables are:

- Recency – the time of their most recent purchase;
- Frequency – number of prior purchases;
- Monetary value – average purchase amount per transaction (Fader, Hardie and Lee, 2005).

*Figure 12: Traditional modelling approach*




*Source: Adapted from Fader and Hardie (2009).*

In a traditional modelling approach, the transaction data is split into two consecutive periods, (Period 1 and Period 2). Data from the second period is then used to create the dependent variable that we are interested in – i.e. buy / not-buy, the number of transactions or total spend. Data from the first period is then used to create the predictor variables (see figure 12). The dependent variable Y is predicted from independent variables X1, X2 and X3 once the model is established based on data for at least two periods.

An example of a simple model would be to classify customers into five groups for each of these three variables. It is also common to use weights for these cells and so create a sort of scoring system for these groups. An example of such a model is presented in figure 13. In the example, recency, frequency and monetary value scores are assigned based on ranges and weighted accordingly. The highlighted cells are used in the example of a single user later in the text.

*Figure 13: RFM calculation example*

Recency

| Last transaction within | Points |
|---|---|
| 3 months | 5 |
| 6 months | 4 |
| 9 months | 3 |
| 12 months | 2 |
| 24 months | 1 |

Frequency

| Number of purchases within 24 months | Points |
|---|---|
| 30 or more | 5 |
| 20 or more | 4 |
| 10 or more | 3 |
| 5 or more | 2 |
| 1 or more | 1 |

Monetary value

| Value of purchases within 24 months | Points |
|---|---|
| 500 $ or more | 5 |
| 200 $ or more | 4 |
| 100 $ or more | 3 |
| 50 $ or more | 2 |
| 10 $ or more | 1 |

Weights

| | | | | | |
|---|---|---|---|---|---|
| Recency | 3 | Frequency | 5 | Monetary value | 2 |

*Source: Adapted from CRMTrends (2019).*

If a customer made the last purchase 5 months ago and made 12 purchases in the total sum of 65 $ in the last 12 months, his score would be as shown in equation (7).

$$(5 \times 3) + (3 \times 5) + (2 \times 2) = 34 \tag{7}$$

Based on the frequency, recency and monetary value for the customer for Period 1 (in our case that lasted 12 months) we assigned him a score of 34 for Period 2. Once the scores are assigned to users, marketing communication programs can address the users differently based on their scores, for example, offer better customer service to those users with a better score.

There are a couple of limitations to the RFM model:

– It only outputs a score and so does not explicitly provide an exact monetary value for the customer (Gupta and others, 2006).
– RFM attempts to predict behavior in the next period, but to compute CLTV, we need to predict all the future periods for the user. Predicting the value of the user far into the future and then discounting it back to present value isn't what RFM was designed to do.
– The observed RFM variables are only imperfect indicators and fail to recognize that different slices of the data will return different values of the RFM variables and, consequently, different scoring model parameters. This has important implications when

the observed data from one period are used to make predictions of future behavior (Fader, Hardie and Lee, 2005a).

As the RFM model doesn't account for the customer's loyalty to the company, an upgraded model called LRFM (loyalty, recency, frequency and monetary value) is also frequently used. Loyalty is represented as number of days from the first to the last visit date in an observed time period, assigned a weight and included in the calculation in the same way other three parameters (Alizadeh Zoeram and Karimi Mazidi, 2018).

## 4.2 Deterministic CLTV Model

Deterministic models are designed to precisely model the outcomes as determined by parameter values and initial conditions. These models focus on inputs and outputs and don't consider possible variations. They are usually used to study firm actions such as customer acquisition, retention, customer profitability and customer cross-buying behavior (Kumar, 2019).

As CLTV is usually defined on a level of an individual customer, it allows us to differentiate individual customer in the context of CRM based on profitability for the company. The challenge with CLTV is to predict the future profits by the customer when his future transactions and frequency are unknown.

$$CLTV = \sum_{t=0}^{T} \frac{(p_t - c_t) r_t}{(1+i)^t} - AC \tag{8}$$

It is usually calculated using the formula shown in equation (8), where $p_t$ is the price paid by a consumer at time $t$, $c_t$ direct cost of servicing the customer at time $t$, $i$ the discount rate or cost of capital for the firm $r_t$ - probability of customer repeat buying at time $t$, $AC$ the acquisition cost and $T$ the time horizon for estimating CLTV (Gupta and others, 2006).

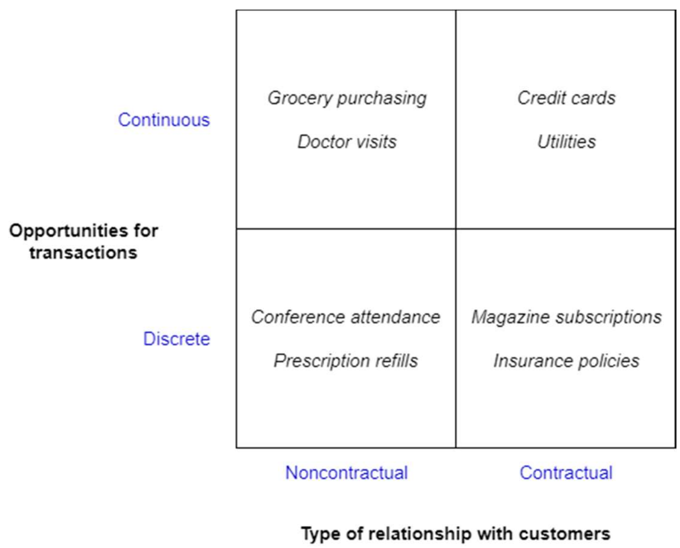
## 4.3 Buy 'Till You Die (BTYD) – Stochastic Models

In a stochastic (probabilistic) model, the observed behavior as the result of an underlying stochastic process controlled by individual characteristics that we cannot observe. The focus of this type of model is on describing and predicting the observed behavior - it assumes that the consumers' behavior varies across the population according to some probability distribution (Kumar, 2019). While deterministic models are suited for individual CLTV calculations, stochastic models should be used when computing CLTV at a customer cohort or customer base level. They are designed to integrate heterogeneity of buy and / or die probabilities among individuals (Calciu, 2009).

Models of the BTYD type are intended to capture non-contractual purchasing behavior of customers – they aim to explain the behavior of the user who are active until they drop out, i.e. die. Some of the models are Pareto/Negative Binomial Distribution (Pareto/NBD), Beta

Geometric / Negative Binomial Distribution (BG/NBD) and Beta-Geometric/Beta Binomial (BG/BB) models. They describe the scenario of the company not being able to observe the exact time at which a customer will drop out (McCarthy and Wadsworth, 2014).

*Figure 14: Classification of customer bases*




*Source: Fader and Hardie (2009).*

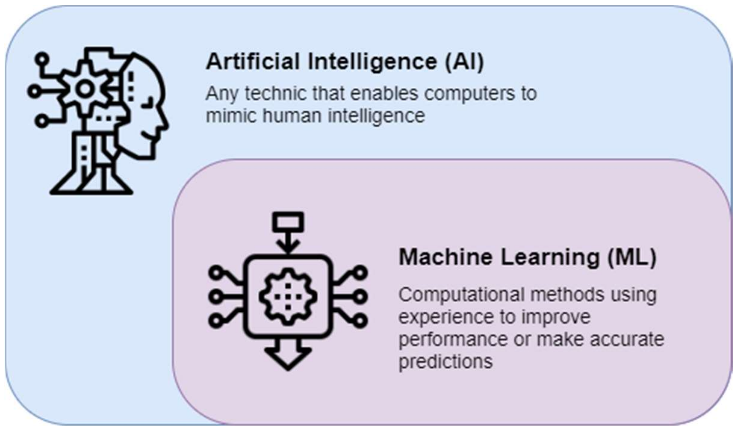
The Pareto/NBD model is used for non-contractual situations in which customers can make purchases at any time (see figure 14). It uses four parameters to describes the rate at which customers make purchases and the rate at which they drop out (McCarthy and Wadsworth, 2014).

The BG/NBD is different from Pareto/BD model only in the assumption of when the customer drops out. The Pareto model assumes that dropout can occur at any point in time and is not dependent on the occurrence of purchases. BG/NBD instead assumes that the customer drops out immediately after purchase (Fader, Hardie and Lee, 2005b). Similarly, the only difference between the Pareto/NBD and BG/BB is that the latter is used to describe situations in which customers have discrete transaction opportunities, rather than being able to make transactions at any time (McCarthy and Wadsworth, 2014).

## 4.4 Machine Learning (ML) Models

ML is broadly defined as the use of computational methods using experience to improve performance or make accurate predictions. The experience refers to past information made available to the learner and typically takes for collected electronic data. No matter the form and nature of the data, its quality and size are the most important factors when it comes to the success of the prediction. The process of ML consists of designing efficient and accurate prediction algorithms. Since the success of a learning algorithm depends on the data used, ML is related to the fields of data analytics and statistics and combines basic concepts of computer science with those from statistics, probability and optimization (Mohri, Rostamizadeh and Talwalkar, 2018).

*Figure 15: The difference between AI and machine learning*




*Adapted from Microsoft (2019).*

Instead of ML, the broader term AI (artificial intelligence) is often used to describe machine learning. The difference is explained in figure 15.

*Table 7: Some types of ML problems*

| Type of problem | Goal of the ML model | Examples |
|---|---|---|
| Classification | Assign a category to each item | • Classify customers into high value, medium value and low value,<br>• Classify texts into categories such as politics, business sports and weather,<br>• Classify images into categories such as landscape, portrait or animal. |
| Regression | Predict a real value for each item | • Assign a monetary value to each individual customer,<br>• Predict values of individual stock prices. |
| Clustering | Partition items into homogeneous regions | • Cluster uses based on their behavioral patterns,<br>• Create communities of users based only on their data. |

*Source: Mohri, Rostamizadeh and Talwalkar (2018).*

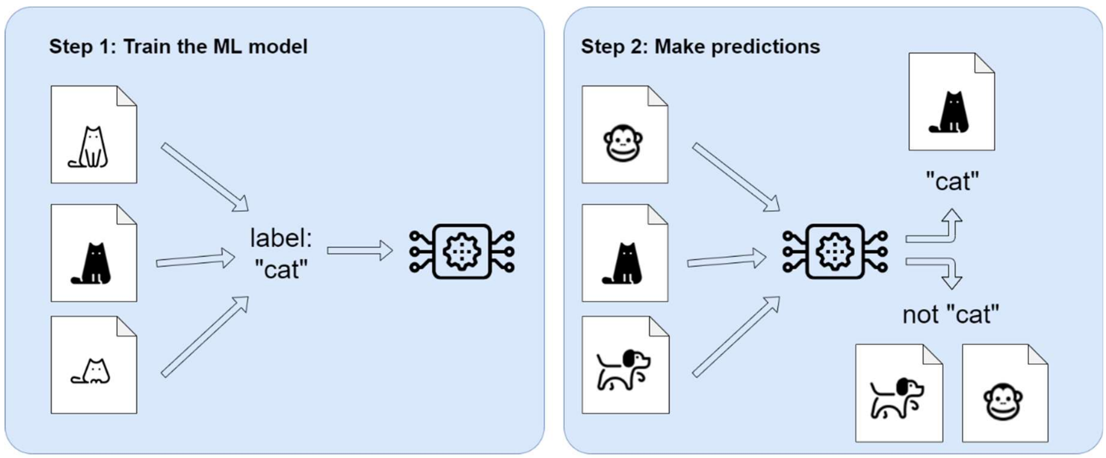
Machine learning algorithms have been used to solve a variety of problems:

- Recommendations systems, search engines,
- Fraud detection and network intrusion,
- Computer vision tasks (image recognition, face detection),
- Text document classification (spam detection),

– Natural language processing,
– Speech recognition,
– Games (chess, go, backgammon),
– Unassisted vehicle control,
– Optical character recognition,
– Medical diagnosis (Mohri, Rostamizadeh and Talwalkar, 2018).

Some major groups of learning problems that ML is used to tackle are classification problems, regression problems and clustering problems. They are explained in table 7.

*Figure 16: An example of supervised learning (classification problem)*



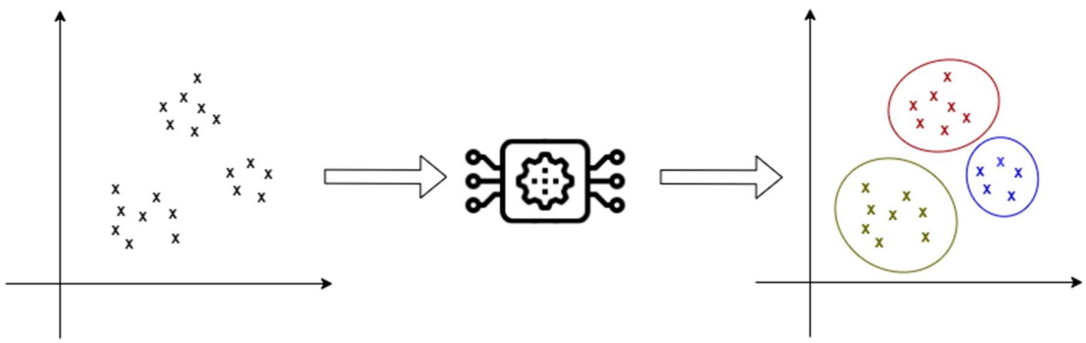

*Source: Adapted from DeBeasi (2019).*

There are multiple learning scenarios, but the two most important are supervised and unsupervised learning. In a supervised learning scenario, the learner receives a set of labelled examples as training data and makes predictions for the data points that are not observed (Mohri, Rostamizadeh and Talwalkar, 2018). This type of scenario is most often associated with classification, regression and ranking problems. See figure 16 for illustration.

In an unsupervised learning scenario, the learner receives training data that is not labelled and makes predictions for the data points that are not observed. Accuracy is harder to check for unsupervised models as there is no benchmark. Clustering is an example of an unsupervised learning scenario (Mohri, Rostamizadeh and Talwalkar, 2018). An example of unsupervised learning is in figure 17.

Traditionally, modelling in marketing literature has preferred models that are easy to interpret. In recent years computer science has developed and perfected multiple ML approaches with a great predictive ability (Gupta and others, 2006). They include neural networks models, decision tree models, random forests models, support vector machines, generalized additive models and gradient boosted trees. We will be comparing some of these models to predict CLTV in the case of actual customer data from an online casino.

*Figure 17: An example of unsupervised learning (clustering)*



*Source: Adapted from Google (2019b).*

# 5 DESIGN AND DATA PREPARATION

## 5.1 Design outline

The goal of our study was to periodically predict CLTV of live customers that recently started playing on an online casino. We made our first prediction after a customer has been playing for one week, the second prediction after two weeks and our final prediction after three weeks of playing. Note here that our methodology is flexible and could be easily adopted for making predictions on longer durations as well. Customer's duration was calculated from his first recorded event in our data base. The research question we are asking is which of the machine learning models gives us the best CLTV predictions on only a week or a couple of weeks' worth of data.

## 5.2 Data preparation

The data upon which we performed our analysis was provided by Oryx Gaming d.o.o. It represents real world data gathered by an online casino from June 2018 to October 2019. The data useful to our research relates to deposits (customers adding money to their account), withdrawals (customers withdrawing money from their account), bets (amount of money bet each time) and wins (amount of money won by the customer from a bet, zero if customer lost) by the customer. It also includes the data about the bonuses the user received, the timestamp of each transaction and whether the data belongs to a customer playing a casino slot game or making a sportsbook bet.

The starting raw data contains more than one hundred million rows. All of the data used in the thesis is hosted on a cloud data warehouse, Google BigQuery. As the raw data spans over a couple of gigabytes, we also used BigQuery as an analytical processing tool due to its extreme efficiency when working with large amounts of data.

We processed the data using a modified version of Structured Query Language (SQL) that BigQuery uses. Once the data was aggregated, we imported it in Python and modified it using the Pandas library.

To prepare the data for analysis we first had to join the transactional data (deposits, deposit bonuses, withdrawals) of each customer with his gameplay data (rounds). This data includes how much each customer bet and won in each of his actions. All the data was joined in a single table, so we were able to identify the customer and timestamp of every transaction and round. All customer IDs and timestamps were kept in the same format, which enabled us to easily group multiple data entries together.

Since customers have different lifestyles and habits, we counted the number of sessions per customer to partly account for this heterogeneity. Each customer's action (row in our data base) was assigned a session, when merging actions into sessions we used a time threshold of two hours. After adding the sessions, we grouped the data by customers. Individual values such as bets, wins, deposits and withdrawals were aggregated using summarization, averaging and counts of individual occurrences. The same process was applied to all the data partitioned by the customer and by the session of that customer. The result of grouping was a table where each row represents an individual customer.

When customers register on an online casino, they typically have to provide their demographic data. We decided to add three more variables from this data:

– Customer's age range: we divided the customers into brackets by decades;
– Customer's gender.

To ensure data quality, we decided to only include customers that placed at least one bet and have at least one deposit. Depending on the amount of time considered, the number of customers used in training and evaluation of machine learning algorithms varied from 2.407 for seven days to 1.693 for 21 days.

Constructed datasets in BigQuery were in all studied cases compact enough to enable us further analysis of the data in Python. Meaning that any additional data manipulations were conducted there. This was convenient since training of all machine learning models and their evaluation was also executed in Python.

To illustrate, for seven days, the data for the customer would look something like the data visualized in table 8.

*Table 8: An example of data after one week of playing*

| customer_id | count_sessions | sum_deposit | count_deposit | … | customer_value |
|---|---|---|---|---|---|
| ID4019 | 15 | 432 | 5 | … | 1735 |

*Source: own work.*

The calculation would be made again for 14 and 21 days. The data for the same customer after 21 days might look as in table 9.

*Table 9: An example of data after three weeks of playing*

| customer_id | count_sessions | sum_deposit | count_deposit | … | customer_value |
|---|---|---|---|---|---|
| ID4019 | 42 | 704 | 12 | … | 1322 |

*Source: own work.*

The result of the described data manipulation is a data table where each row represents the behavior of an individual customer in a chosen timeframe (7 days, 14 days or 21 days). The parameters (columns) in this table represent the independent variables (also called predictor or explanatory variables) used for building the machine learning models. Next, we appended the calculated CLTV for each customer. The CLTV parameter represents the dependent variable, also called predicted or response variable, since this is the variable which the machine learning model will predict. Since functions for calculating CLTV have different levels of complexity, we decided to try two different functions, a simple one and an aggregated one.[1]

$$CLTV = \frac{ADV \times PF}{CR} \times PM \tag{9}$$

The equation for simple CLTV calculation was as shown in equation (9). *AOV* represents the average deposit value, *PF* the purchase frequency, *CR* the churn rate and *PM* the profit margin.

$$CLTV_1 = ACV \times AL \times PM \tag{10}$$

$$CLTV_2 = AL \times E \times NV \tag{11}$$

$$CLTV_3 = \frac{SDV}{CR} \times PM \tag{12}$$

In the aggregated formula, the CLTV calculation as shown in equation (10), (11) and (12). *ACV* represents average customer value per week, *AL* average lifetime per week, E expenditure per visit, *NV* number of visits per week, *SDV* sum of the value of the deposits.

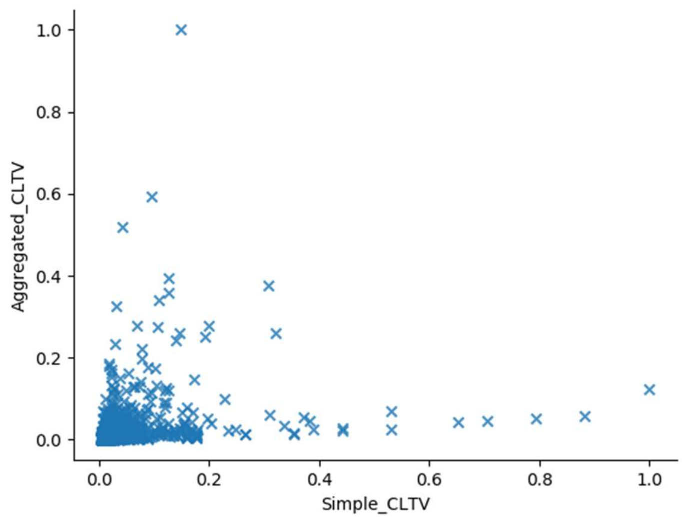
[1] For the structure of the finalized table and the description of the fields see appendix.

The three formulas are then averaged to produce a single value per customer - see equation (13).

$$CLTV = \frac{CLTV_1+CLTV_2+CL\quad_3}{3} \tag{13}$$

The two CLTV functions described above produce a similar average CLTV value – the average of simple function was at about $^3/_4$ of average for the aggregated function. Both are comparable to the estimates about the average customer lifetime value already used within the company.

*Figure 18: Results from simple and aggregated CLTV formulas*




*Source: own work.*

From the scatterplot in figure (18) - where one represents the highest value overall - we can see that the majority of the calculated CLTVs are grouped in the lower left part, with aggregated CLTV having larger outliers. Given that the estimates CLTV values derived from different formulas produce quite different results it made sense to predict the values for each formula separately.

# 6 OVERVIEW OF MODELS AND SCORING

## 6.1 Approach and methodology

The goal of our machine learning models is to provide a prediction of CLTV for each individual customer shortly after joining the casino. The longer a customer plays on a casino, more accurately we can evaluate his CLTV. In this work, we evaluated our machine learning models by comparing the customer's predicted CLTV with his actual CLTV. We made the first CLTV prediction after one week of customer's activity, the second prediction after two weeks and the final prediction after three weeks.

Table 10 visualizes this methodology: using only one week of data, we would predict the value D, using two weeks the value C and using three weeks the value B. Using all available data, we can calculate the value A, which we consider the actual CLTV of a customer.

*Table 10: Illustration of the problem and prediction goal*

| Week 1 | Week 2 | Week 3 | Week4 | Week 5 | Week 6 | ... | CLTV |
|---|---|---|---|---|---|---|---|
| | | | | | | | A |
| | | | | | | | B |
| | | | | | | | C |
| | | | | | | | D |

*Source: own work.*

We compared the efficiency of predictive modelling for two different calculations of CLTV, one derived using a simple and one derived using a more complex, aggregated function. For details about functions for calculating CLTV see chapter 5. As the goal of our model is to predict an actual value, it belongs to the regression type of problems (see chapter 5.4 and table 7).

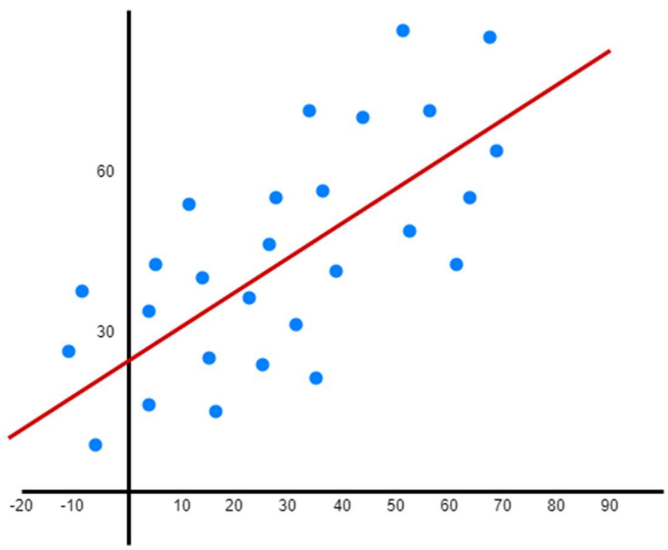
For comparison, we decided to use four popular machine learning models, one basic and three more advanced. The models we used were linear regression, neural networks, random forest and gradient boosting. The three models were chosen as they are considered among most popular for solving regression problems (Seif, 2018; Chen, 2019) and have already been compared in recent literature (Nawar and Mouazen, 2017; Krauss and Huck, 2017). We compared the performance of the models using the coefficient of determination ($R^2$) and root mean square error (RMSE).

## 6.2 Linear regression

The goal of linear regression is to model the relationship between the observed variables and a response variable by fitting a linear equation to the observed data. Every value of the independent variable X is associated with a value of the dependent variable Y. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable (or, as in our case, a vector of multiple explanatory variables) and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0) (Pires, Martins, Sousa, Ferraz and Pereira, 2008).

Figure 19 illustrates how linear regression tries to predict the value of the response variable (Y) as a linear combination (line) of one or more explanatory variables. The visualization above predicts Y from a single explanatory variable (X).

*Figure 19: Linear regression*




*Source: Adapted from Nguyen (2017).*

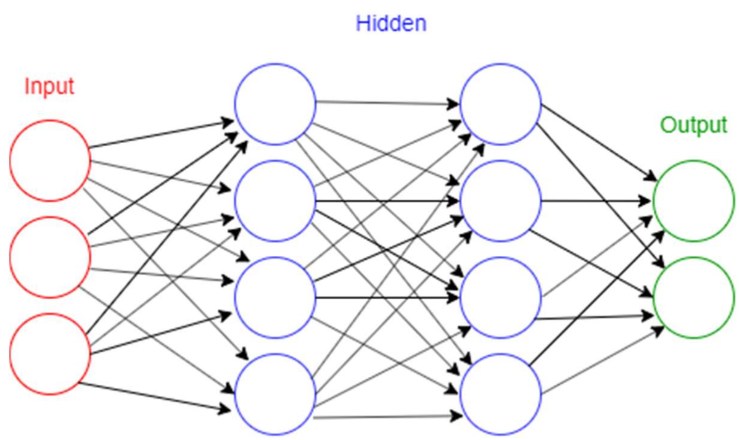
### 6.3 Neural networks

An artificial neural network consists of a number of simple processors, also called neurons, which are analogous to the biological neurons in the brain (Akhoondzadeh, 2019). Patterns are presented to the network via the input layer, which communicates to one or more hidden layers, where the actual processing is done via a system of weighted connections. The hidden layers then link to an output layer which generates the final output ('answer').

Neural networks are black boxes, as usually no satisfactory explanation of their behavior is offered (Benítez, Castro & Requena, 1997). Neural networks with more than two hidden layers are considered deep neural networks.[2] Figure 20 shows a neural network design with two hidden layers.

[2] For a more detailed description of neural networks see Da Silva, I. N., Spatti, D. H., Flauzino, R. A., Liboni, L. H. B., & dos Reis Alves, S. F. (2017). *Artificial neural networks*. Cham: Springer International Publishing.

*Figure 20: Neural networks*

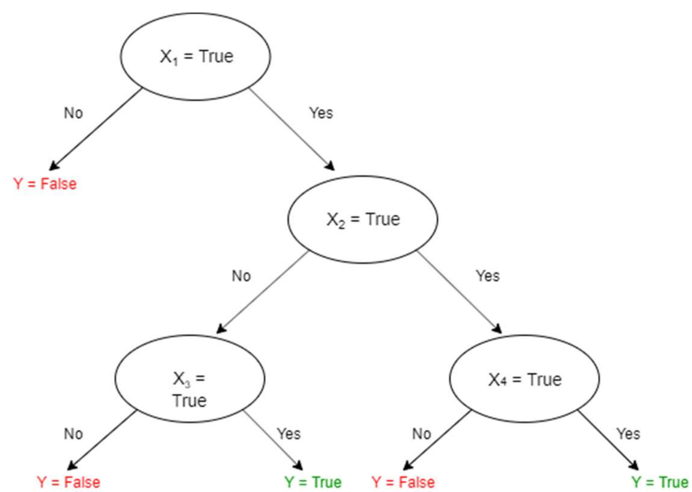



*Source: Adapted from Davydova (2017).*

## 6.4 Random forests

Random forests are a machine learning algorithm that makes its predictions by merging predictions made by multiple small decision trees. Decision trees (see figure 21) are the basis of the random forest algorithm.

*Figure 21: Decision trees (the basis for random forests)*



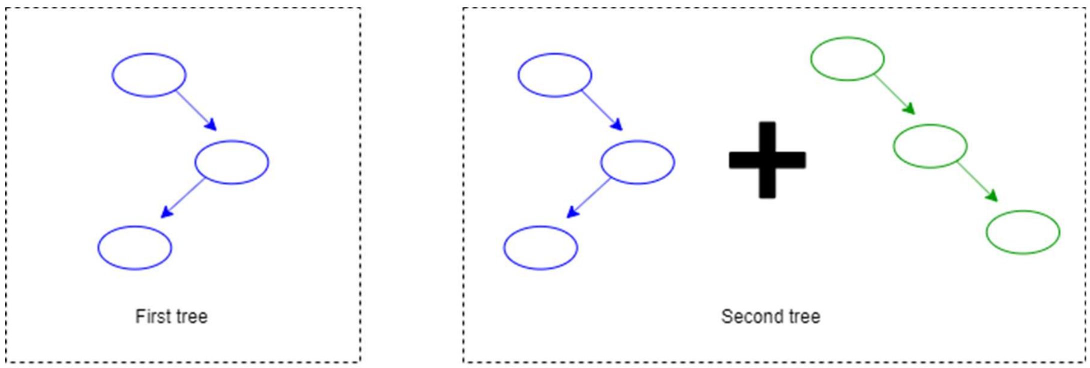

*Source: Adapted from Plapinger (2017).*

In a decision tree, we start at the top (the root) of the tree and move down, selecting the branch based on whether the condition is met or not until we finally reach the final nodes (the leaves). Essentially, decision trees make their predictions by splitting the problem into several simpler binary conditions. It consists of a series of conditions and based on whether those conditions are met, different leaves (predictions) are reached.[3]

### 6.5 Gradient boosting

While random forests utilize multiple trees in parallel and take the most often reached result as the final result, gradient boosting uses residual error from one decision tree when training another. Doing so, it connects multiple individual decision trees into a stronger learning algorithm.[4] For illustration, see figure 22.

*Figure 22: Gradient boosting*




*Source: Adapted from Rogozhnikov (2016).*

### 6.6 Development of ML models

We developed the models using Python programming language. The input data prepared with BigQuery was exported into a CSV (comma-separated values) file and imported into Python using Pandas library. We made a preliminary analysis of the data using Seaborn and Matplotlib libraries. For the development of linear regression, neural network and random forests models we used Scikit-learn Python library and for gradient boosted trees model we used XGBoost library. Scikit-learn library was also used to evaluate the performance of all models. In all the models except for linear regression hyperparameter tuning — re-running of models using different settings — was required and was performed using basic Python loops and printing the results for each iteration. Best achieved results of each model were than compared.

[3] For a more detailed description of random forest see Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

[4] For a more detailed description of gradient boosting see Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

### 6.7 Evaluation of ML models

When performing machine learning, it is a common practice to split your data into a training and a testing set. This way you can evaluate the success of your model on data that wasn't used during the training of the model, which results in the more objective evaluation. One of the more advanced approaches to splitting the data is k-fold cross-validation. Instead of using, say, 80 percent of your data to train the model and 20 percent of your data to validate it, you use all your data to both train and validate the model. It simply splits the data into training and test set, but it does this so many times that more or less all of the data is used for both training and testing. The result is usually the averaged result of all iterations. In our case, we used 10-fold cross-validation, which means our evaluation consisted of 10 iterations (see figure 23).

*Figure 23: K-fold cross-validation*


*Source: Adapted from Shaikh (2018).*

To evaluate our models we used two metrics, the coefficient of determination ($R^2$) and the root mean square error (RMSE). $R^2$ is broadly used when estimating the quality of the fit in a regression model (Renaud, Victoria-Feser: 2010). It tells us the proportion of the variance in the dependent variable that is predicted by the independent variables. It can be calculated using equation (14).

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad (14)$$

where $SS_{res}$ represents the sum of the squares of the difference between the actual values and the predicted values (sum of the squares of the residuals) and $SS_{tot}$ represents the sum of the squares of the differences between the dependent variable and its mean.

RMSE is the standard deviation of the residuals (prediction errors). It tells us how concentrated the data is around the line of best fit. In equation (15), $f$ represents predicted (forecasted) values and $o$ actual (observed) values.

$$RMSE = \sqrt{\overline{(f - o)^2}} \quad (15)$$

# 7 EVALUATION OF MODELS

## 7.1 Correlation coefficients in the data

To get a better feel for the data we first performed some elementary exploratory analysis. We calculated the correlation coefficients for the seven days of data between all independent variables and the dependent variable (CLTV). We calculated the correlation coefficients for both the simple and the aggregated CLTV functions. This exploratory analysis is visualized in figure 24 and figure 25[5].

*Figure 24: Correlation coefficients for the simple CLTV formula*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0,35 | 0,97 | 0,02 | 0,10 | 0,10 | 0,00 | 0,21 | 0,25 | -0,08 | 0,15 | 0,16 | 0,02 |
| 0,15 | 0,17 | 0,02 | 0,22 | 0,50 | -0,03 | 0,22 | 0,35 | 0,11 | 0,04 | 0,16 | 0,15 |
| 0,30 | 0,30 | 0,51 | 0,08 | 0,18 | 0,05 | -0,05 | -0,03 | 0,03 | 0,00 | 0,00 | |

*Source: own work.*

As we can see from figure 24, there is a very high correlation between the simple CLTV and one of the variables - the average deposit per customer. This is not a surprise, as the average deposit per customer is the only customer-specific data that we input into the formula to calculate the simple CLTV (the rest are casino-wide and as such constant).

*Figure 25: Correlation coefficients for the complicated CLTV formula*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0,69 | 0,25 | 0,52 | 0,36 | 0,25 | 0,21 | 0,39 | 0,16 | 0,14 | 0,54 | 0,05 | 0,34 |
| 0,53 | 0,12 | 0,32 | 0,29 | 0,13 | 0,11 | 0,28 | 0,13 | 0,17 | 0,16 | 0,45 | 0,43 |
| 0,22 | 0,19 | 0,60 | 0,21 | 0,16 | 0,04 | -0,04 | -0,04 | 0,05 | -0,02 | -0,01 | |

*Source: own work.*

The picture as presented in figure 25 is clearly different since there is no single variable that would extremely correlate with aggregated CLTV, and predictions should be more complicated when it comes to the aggregated CLTV equation.

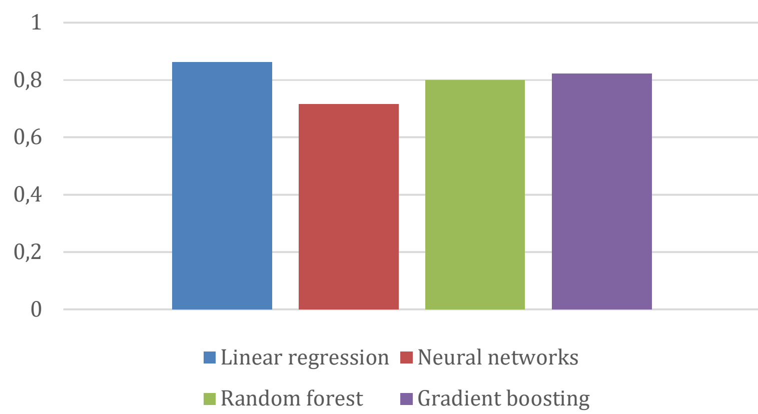
The figures indicate that the first case can probably be sufficiently explained using linear regression as the linear combination of a few parameters should suffice for a good prediction. In the second case, the predictive model is likely to be more complex.

[5] For the full list of field names, descriptions and corresponding coefficients see appendix.

## 7.2 Comparison of models for predicting the simple CLTV

According to $R^2$ scores, in our case, the simplest algorithm (the linear regression) gave the best predictions for simple CLTV. Figure 26 visualizes $R^2$ scores averaged over all three durations - one week, two weeks and three weeks. Linear regression gave the best prediction overall.

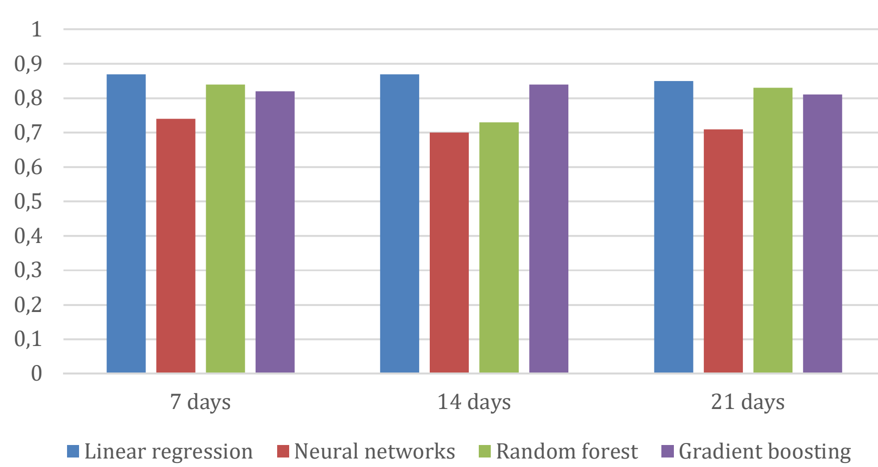
*Figure 26: Averaged $R^2$ scores for simple CLTV*




*Source: own work.*

Looking at $R^2$ scores per duration, there isn't a clear trend (figure 27). Looking at the overall performance of models, we can see that linear regression performed the best in all three cases, gradient boosting also gave good results in all three cases, while random forests performed well with 7 and 21 days of data, but interestingly performed quite poor on 14 days of data. In all cases, neural networks gave the worst predictions.

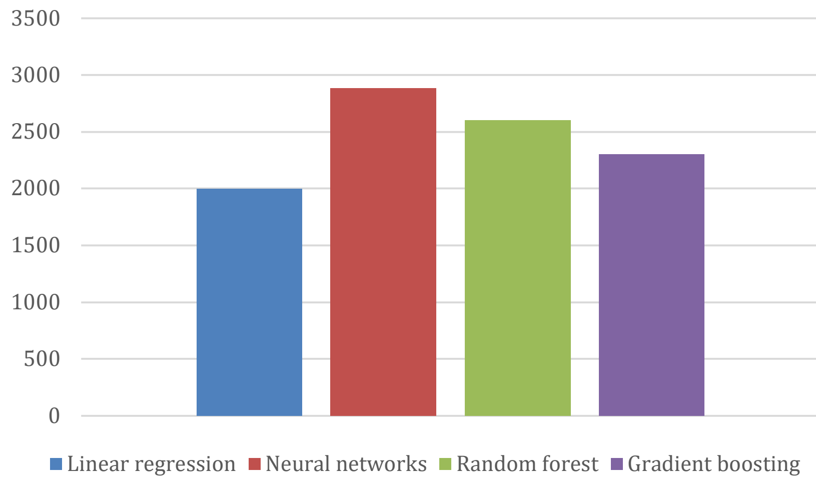
*Figure 27: $R^2$ scores for simple CLTV split by the duration of data*




*Source: own work.*

This makes sense since neural networks are probably the most complex algorithm of the four and as such require a lot of expertise and fine-tuning of parameters to achieve their optimal performance. We believe that given additional time we could configure neural networks in a way that would result in predictions comparable to other models.

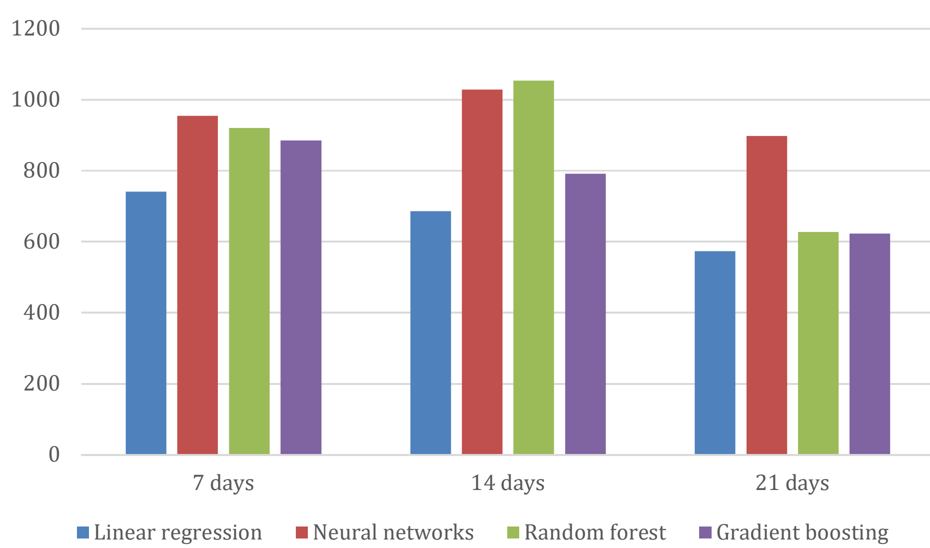
*Figure 28: Sum of RMSE scores (lower is better) for simple CLTV*




*Source: own work.*

Similarly, as with the $R^2$ scoring, the simple linear regression gave the best results when evaluated with the RMSE metric (figure 28). Again, second-best were the predictions made by gradient boosting, followed by random forest and neural networks.

*Figure 29: RMSE scores for simple CLTV split by the duration of data*



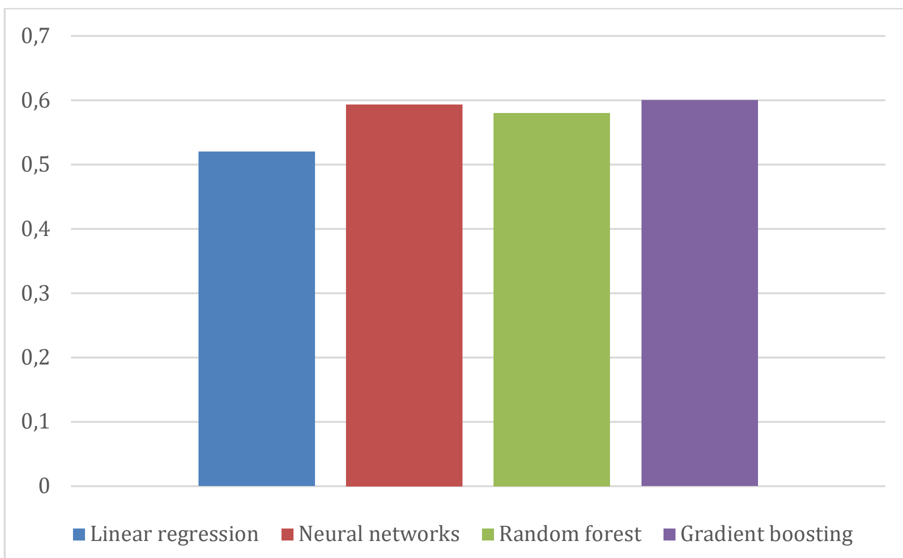

*Source: own work.*

When evaluated with RMSE, all models performed the best when predictions were based on 21 days of data (figure 29). This makes sense since the more data the prediction algorithm can learn from, the better it can predict the quality of customers. While all models performed the best when their predictions were based on 21 days of data, only two of the models (linear regression and gradient boosting) gave better results on 14 days then on 7 days of data. Interestingly, the other two (neural networks and random forest) actually performed better when learning was based on 7 days of data than when it was based on 14 days of data.

### 7.3 Comparison of models for predicting the aggregated CLTV

When the task of models is to predict aggregated CTLV, linear regression (a clear winner in the case of simple CLTV) appears to start losing steam. In our case the best algorithm for predicting aggregated CLTV was gradient boosting, closely followed by neural networks, third place went to the random forest while linear regression was the worst by a significant margin (figure 30).

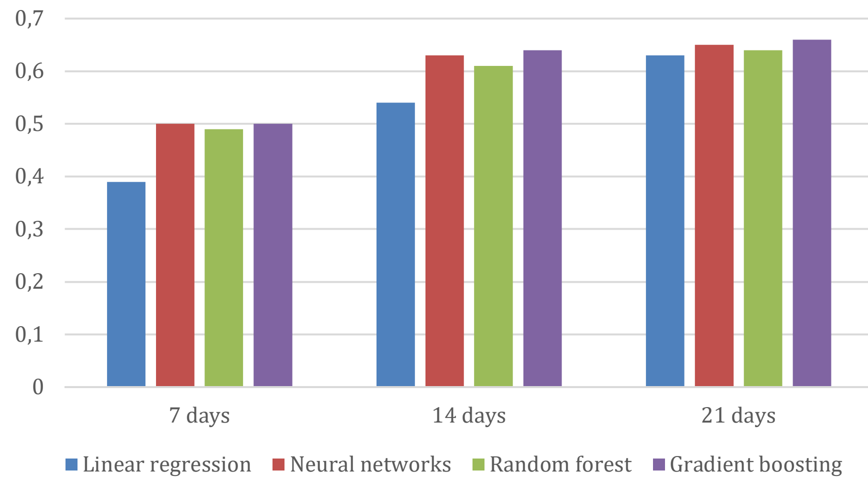
*Figure 30: Averaged $R^2$ scores for aggregated CLTV*




*Source: own work.*

We can see a clear improvement in the quality of predictions when algorithms have more data to learn from. Predictions for all four used algorithms are the best when algorithms learn from 21 days' worth of playing data, second-best when they learn from 14 days and the worst when learning from 7 days (figure 31).

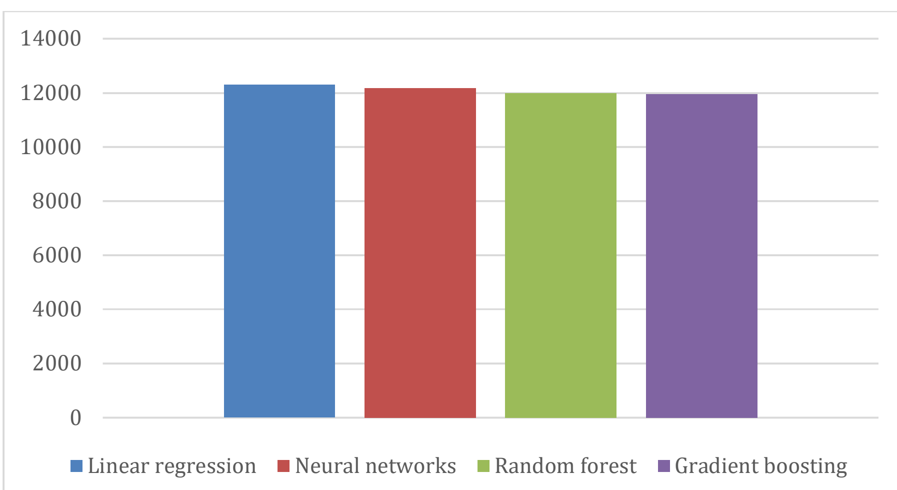
*Figure 31: $R^2$ scores for aggregated CLTV split by the duration of data*




*Source: own work.*

Looking at the summaries of RMSE score for predicting aggregated CLTV, the difference between tested approaches seems quite small, gradient boosting performed the best, but only slightly (figure 32).

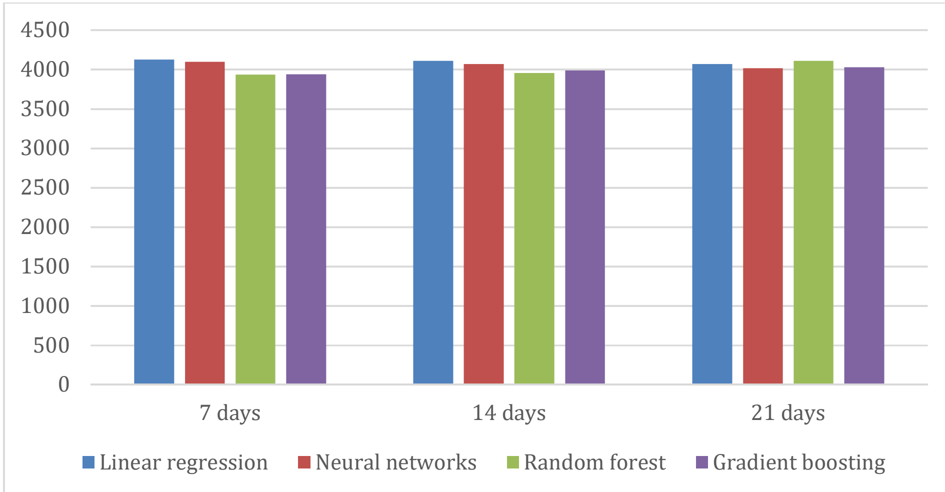
*Figure 32: Sum of RMSE scores (lower is better) for aggregated CLTV*




*Source: own work.*

Looking at the overall performance of models, we see an interesting crossover between 14 and 21 days of data, as some models performed better on 14 days of data and others on 21 days of data. Just like in the majority of previous scenarios gradient boosting seems like a good choice as it again performs very consistently. Gradient boosting actually seems to be the best method for predicting aggregated CLTV (figure 33).

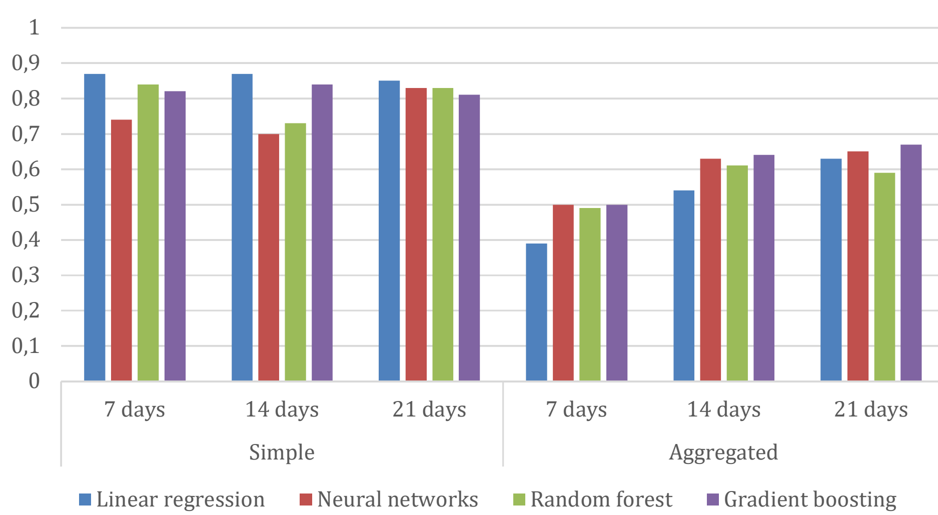
*Figure 33: Sum of RMSE scores (lower is better) for aggregated CLTV*




*Source: own work.*

Looking at $R^2$ values for predictions of both simple and aggregated CLTV (figure 34), a clear difference can be observed since the predictions for a simple CLTV are much better than those for the aggregated CLTV. The biggest difference was in the case of linear regression, a clear winner in the case of simple CLTV was the worst approach for predicting aggregated CLTV.

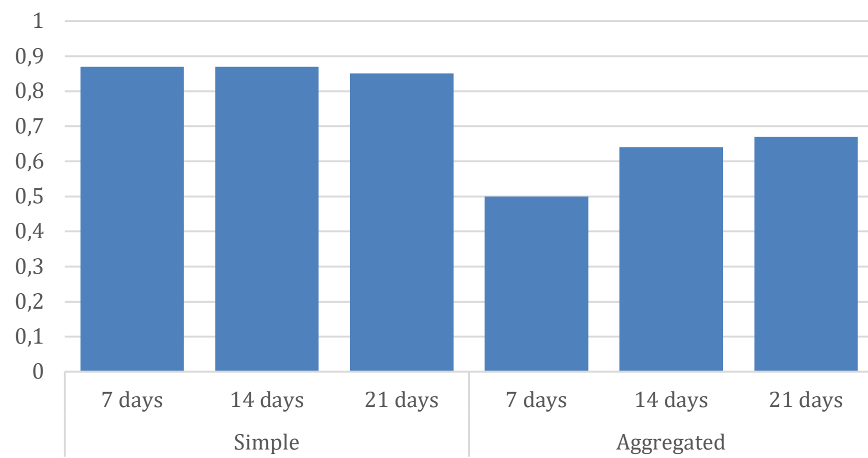
*Figure 34: $R^2$ scores: comparison of predictions for simple and aggregated CLTV*




*Source: own work.*

If we isolate just the best scores regardless of the model (figure 35), we can see that we are able to predict the simple CLTV with R2 score of over 0,8 (80%) and the aggregated CLTV from as bad as 50% (when learning on 7 days of data) to 67% (when learning on 21 days of data). That tells us there is a big difference between how well simple CLTV can be predicted compared to aggregated CLTV.

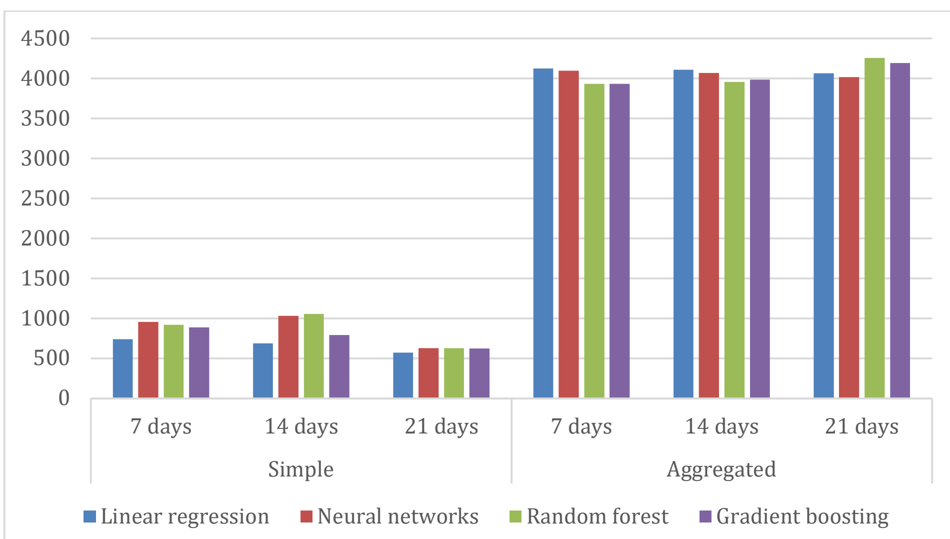
*Figure 35: Best achieved $R^2$ scores of any model*




*Source: own work.*

The look at the RMSE scoring comparison (figure 36) shows an even more glaring difference. RMSE skyrockets when predicting the aggregated formula compared to predictions of the simple formula.

*Figure 36: RMSE scores (lower is better) for predicting simple and aggregated CLTV*




*Source: own work.*

It gets clear that complex models are more suitable for complex problems, but even complex models do not perform extremely well here. This finding suggests that the parameters which online casinos gather to quantify the behavior of their players are probably not sufficient for making predictions in the case of aggregated CLTV.

## CONCLUSION

In the thesis, we used historical data of an online casino to calculate two versions of CLTV. The purpose of testing predictions for two differently calculated CLTVs was that a simpler calculation is suitable for companies that do not have a lot of customer data or that only recently started using CLTV as a metric for evaluating the quality of their customers. The aggregated CLTV is an example of a calculation that might be used in a company that has been using and fine-tuning CLTV for a while. The main goal of this thesis was the evaluation of the performance of four machine learning algorithms when predicting CLTV. The algorithms were tested on the case of an online casino, where the quality of customers has to be calculated from data about their behavior. In our case, machine learning algorithms were tested when training on 7, 14, and 21 days worth of data. The four machine learning algorithms that we tested were linear regression, random forest, neural networks and gradient boosting. Linear regression is one of the simplest machine learning algorithms, while the rest that we tested are more complicated.

Our research has shown us that the complexity of the function for calculating CLTV plays a crucial role in prediction's accuracy. When the CLTV function is relatively simple and it does not use many different parameters of the data, simple models usually make quite good predictions. This was the same in our case where the simplest of the models (linear regression) actually made the most accurate predictions for the value of simple CLTV. This has many benefits and is very practical as initial CLTV formulas are usually simple. Simple models usually also offer good performance and are quite easy to interpret, meaning that the time needed for building them is low and the reasoning behind the model's decisions is easy to understand.

When using a more elaborate CLTV function for calculating aggregated CLTV, the quality of predictions dropped significantly. This is most likely because the data we used to train the models can't provide a good enough prediction for aggregated CLTV. The results would perhaps improve if we combined the parameters together to form new ones to use for training or if we had more details about customer's behavior. During the testing, it did become clear that more complex prediction algorithms are best suited for more complex problems, such as the prediction of aggregated CLTV from just seven days' worth of data.

Overall, the differences between tested models weren't that big. When considering building a model to use for production with the purpose of providing business value, in our case linear regression offers the most bang for the buck. In the case of more complex models, the hyperparameter tuning of the models - tweaking different preferences and retraining the models - required hours, if not days of our time, even when we automated the whole process.

Linear regression, on the other hand, works with no tuning and returns the results almost instantly. When the performance of models is so close, implementing anything other than linear regression would make little sense.

The rise in value for complex problems for more advanced models does seem interesting and we believe those models would be worth revisiting in cases when:

- the business value of predicting CLTV in the company is clear and established, and the precision of prediction is of great value;
- the size of the training set is significantly higher.

The fact is that the amount of data fed into the models was relatively low, and if we would repeat the experiment with data of 10.000 or 100.000 customers, more complex models would probably perform better.

Unsurprisingly, we established that the complexity of the problem has the biggest bearing on the quality of the solution, in this case, the efficiency of the predictive model. Complex modern machine learning methods didn't prove to be the silver bullet and couldn't offset the difference in problem complexity. They did, however, signal that they are not to be discarded and that they have the potential to provide added value over linear regression if the conditions were to change.

## REFERENCE LIST


1. Akhoondzadeh, M. (2019). *Artificial Neural Network (ANN).* Retrieved November 16, 2019, from https://www.researchgate.net/publication/336999943_Artificial_Neural_Network_ANN
2. Alizadeh Zoeram, A. & Karimi Mazidi, A. R. (2018). New Approach for Customer Clustering by Integrating the LRFM Model and Fuzzy Inference System. *Iranian Journal of Management Studies*, *11*(2), 351–378.
3. Anderson, E. T. & Simester, D. I. (2004). Long-run effects of promotion depth on new versus established customers: Three field studies. *Marketing Science*, *23*(1), 4–20.
4. Anderson, E. T. & Simester, D. I. (2011). A step-by-step guide to smart business experiments. *Harvard Business Review*, *89*(3), 98–105.
5. Bauer, H. H. & Hammerschmidt, M. (2005). Customer-based corporate valuation: Integrating the concepts of customer equity and shareholder value. *Management Decision*, *43*(3), 331–348.
6. Benítez, J. M., Castro, J. L. & Requena, I. (1997). Are artificial neural networks black boxes? *IEEE Transactions on neural networks*, *8*(5), 1156–1164.

7. Bloniarz, B. (2018). *Customer Lifetime Value: Models, Metrics and a Multitude of Uses*. Retrieved November 20, 2019, from https://pyvideo.org/pydata-la-2018/customer-lifetime-value-models-metrics-and-a-multitude-of-uses.html
8. Bonacchi, M. & Perego, P. (2012). Improving profitability with customer-centric strategies: The case of a mobile content provider. *Strategic Change*, *20*(7–8), 253–267.
9. Calciu, M. (2009). Deterministic and stochastic Customer Lifetime Value models. Evaluating the impact of ignored heterogeneity in non-contractual contexts. Journal of Targeting, *Measurement and Analysis for Marketing*, *17*(4), 257–271.
10. Chen, I. J. & Popovich, K. (2003). Understanding customer relationship management (CRM) People, process and technology. *Business process management journal*, *9*(5), 672–688.
11. Chen, L. (2019). *Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting) - Step by Step Explained*. Retrieved December 22, 2019, from https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725
12. Ciaccia, C. (2014). *Why Amazon (AMZN) Prime Is More Important to Amazon's Future Than You Think*. Retrieved October 20, 2019, from https://www.nasdaq.com/articles/why-amazon-amzn-prime-more-important-amazons-future-you-think-2014-07-24
13. CRMTrends (2019). *Marketing Analysis*. Retrieved December 19, 2019, from http://www.crmtrends.com/analytics.html
14. Davydova, O. (2017). *7 Types of Artificial Neural Networks for Natural Language Processing*. Retrieved November 10, 2019, from https://www.kdnuggets.com/2017/10/7-types-artificial-neural-networks-natural-language-processing.html
15. Day, G. S. (2003). Creating a superior customer-relating capability. *MIT Sloan Management Review*, *44*(3), 77.
16. DeBeasi, P. (2019). *Training versus Inference*. Retrieved December 19, 2019, from https://blogs.gartner.com/paul-debeasi/2019/02/14/training-versus-inference/
17. Facebook (2019). *About Customer Lifetime Value*. Retrieved November 25, 2019, from https://www.facebook.com/business/help/1730784113851988
18. Fader, P. S. & Hardie, B. G. (2009). Probability models for customer-base analysis. *Journal of interactive marketing*, *23*(1), 61–69.
19. Google (2019a). *Lifetime Value*. Retrieved November 25, 2019, from https://support.google.com/analytics/answer/6182550?hl=en
20. Google (2019b). *Clustering Algorithms*. Retrieved December 19, 2019, from https://developers.google.com/machine-learning/clustering/clustering-algorithms
21. Gupta, S., Lehmann, D. R. & Stuart, J. A. (2004). Valuing customers. *Journal of marketing research*, *41*(1), 7–18.
22. Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., ... & Sriram, S. (2006). Modeling customer lifetime value. *Journal of service research*, *9*(2), 139–155.

23. Hawkes, V. A. (2000, January). The heart of the matter: The challenge of customer lifetime value. *CRM Forum Resources,* 13(1), 2–10.
24. Horák, P. (2017). Customer lifetime value in B2B markets: theory and practice in the Czech Republic. *Int. J. Bus. Manag*, *12*(2), 47–55.
25. Kim, S. Y., Jung, T. S., Suh, E. H. & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert systems with applications*, *31*(1), 101–107.
26. KPMG (2019). *Customer lifetime value in retail energy markets*. Retrieved December 12, 2019, from https://assets.kpmg/content/dam/kpmg/xx/pdf/2019/06/customer-lifetime-value-in-retail-energy-markets.pdf
27. Krauss, C., Do, X. A. & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689–702.
28. Kumar, V., Venkatesan, R., Bohling, T. & Beckmann, D. (2008). Practice Prize Report—The power of CLV: Managing customer lifetime value at IBM. *Marketing science*, *27*(4), 585–599.
29. Lemon, K. N., Rust, R. T. & Zeithaml, V. A. (2001). What drives customer equity? *Marketing management*, *10*(1), 20–25.
30. McCarthy, D. & Wadsworth, E. (2014). *Buy'Til You Die-A Walkthrough*. Retrieved December 22, 2019, https://cran.r-project.org/web/packages/BTYD/vignettes/BTYD-walkthrough.pdf
31. Microsoft (2019). *Deep learning vs. machine learning*. Retrieved December 22, 2019, from https://docs.microsoft.com/en-us/azure/machine-learning/service/concept-deep-learning-vs-machine-learning
32. Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.
33. Nawar, S. & Mouazen, A. (2017). Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. *Sensors*, *17*(10), 2428.
34. Nguyen, K. (2017) *Optimize Linear Regression Model with Intel® DAAL*. Retrieved November 10, 2019, from https://software.intel.com/en-us/articles/optimizing-linear-regression-method-with-intel-data-analytics-acceleration-library
35. Payne, A. & Frow, P. (2005). A strategic framework for customer relationship management. *Journal of marketing*, *69*(4), 167–176.
36. Pires, J. C. M., Martins, F. G., Sousa, S. I. V., Ferraz, M. C. M. A. & Pereira, M. C. (2008). Prediction of the daily mean PM10 concentrations using linear models. *American Journal of Environmental Sciences*, *4*(5), 445.
37. Plapinger, T (2017). *What is a Decision Tree?* Retrieved November 10, 2019, from https://towardsdatascience.com/what-is-a-decision-tree-22975f00f3e1
38. Prasasti, N., Okada, M., Kanamori, K. & Ohwada, H. (2014). Customer lifetime value and defection possibility prediction model using machine learning: An

application to a cloud-based software company. In *Asian Conference on Intelligent Information and Database Systems* (pp. 62-71). Cham: Springer.
39. Rogozhnikov, A. (2016). *Gradient Boosting explained*. Retrieved November 10, 2019, from http://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html
40. Rozek, J. & Karlicek, M. (2014). Customer lifetime value as the 21st century marketing strategy approach. *Central European business review*, *3*(2).
41. Savitz, Eric J. (2019). *Apple Stock Is Rising Because the Latest Big News Was More Exciting Than Expected*. Retrieved December 6, 2019, from https://www.barrons.com/articles/apple-stock-is-up-as-analysts-weigh-in-on-iphone-11-news-51568219415
42. Seif, G (2018). *Selecting the Best Machine Learning Algorithm for Your Regression Problem*. Retrieved December 22, 2019, from https://www.kdnuggets.com/2018/08/selecting-best-machine-learning-algorithm-regression-problem.html
43. Shaikh R. (2019). *Cross Validation Explained: Evaluating estimator performance*. Retrieved November 17, 2019, from https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff8
44. Stone, M. & Shaw, R. (1988). *Database marketing*. Aldershot: Gower.
45. Szymanski, D. M., Bharadwaj, S. G. & Varadarajan, P. R. (1993). An analysis of the market share-profitability relationship. *Journal of marketing*, *57*(3), 1–18.
46. Toccu, M. & Fasso, A. (2013). Statistical analysis of customer lifetime value: a case study on telecommunications data using SAS. *Statistica & Applicazioni*, *11*(2), Vita e Pensiero.
47. Tsai, C. F., Hu, Y. H., Hung, C. S. & Hsu, Y. F. (2013). A comparative study of hybrid machine learning techniques for customer lifetime value prediction. *Kybernetes*, *42*(3), 357-370.
48. Verhoef, P. C. & Lemon, K. N. (2013). Successful customer value management: Key lessons and emerging trends. *European Management Journal*, *31*(1), 1–15.
49. Villanueva, J. & Hanssens, D. M. (2007). Customer equity: Measurement, management and research opportunities. *Foundations and Trends® in Marketing*, *1*(1), 1–95.
50. Weinberg, B. D. & Berger, P. D. (2011). Connected customer lifetime value: The impact of social media. *Journal of Direct, Data and Digital Marketing Practice*, *12*(4), 328–344.
51. Winer, R. S. (2001). A framework for customer relationship management. *California management review*, *43*(4), 89–105.

# APPENDICES

## Appendix 1: Povzetek (Summary in Slovene language)

Sodobna digitalna ekonomija nam nudi izobilje podatkov. Zato ni čudno, da je zadnja leta vse večji poudarek na pridobivanju vrednosti iz teh podatkov. Eden od bolj neposrednih načinov, na katerega lahko analiza podatkov vpliva na poslovne rezultate je izdelava napovednih modelov, ki temeljijo na zgodovinskih podatkih uporabnikov in nam omogočajo napovedi obnašanja teh uporabnikov v prihodnosti.

Osredotočanje na posameznega kupca je bilo pravilo nekje do sredine 20. stoletja, ko je manjše trgovine začela nadomeščati množična proizvodnja in prodaja. Premik nazaj h kupcu se je zgodil v zadnjih desetletjih z razvojem tehnologije, katera nam ponovno omogoča do kupca ne le pristopiti na različne načine, ampak mu tudi omogočiti produkte, prirejene njegovim željam in okusu. To nam omogoča zbiranje, hranjenje in obdelava podatkov o kupcu ter izdelava novih metrik, ki so uporabne pri kategorizaciji in razumevanju tega kupca.

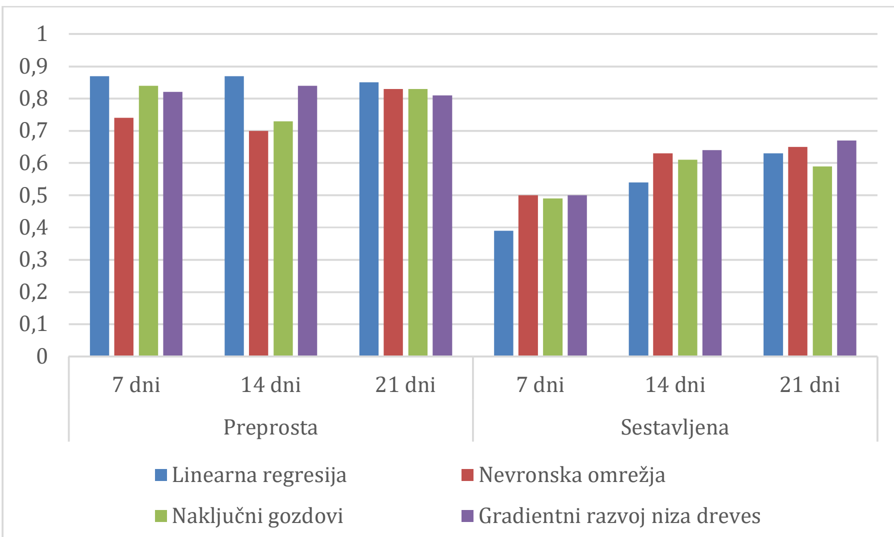
Ena bolj uporabnih in pomembnih metrik se nanaša na življenjsko vrednost kupca (angleška kratica CLTV). V osnovi nam CLTV pove sedanjo vrednost vseh bodočih dobičkov, pridobljenih s strani posameznega kupca skozi celotno življenjsko dobo njegovega sodelovanja/odnosa s podjetjem. Ta podatek o kupcu je podjetju v pomoč pri ocenjevanju, koliko se mu splača vlagati v posameznega kupca. Konkretno, podjetju pove, kateremu kupcu najbolj splača ponuditi popuste, bonuse in druge ugodnosti in pa kolikšna je še razumna vrednost teh ugodnosti za posameznega kupca. Agregirani podatki o vrednosti kupcev podjetja so se izkazali tudi koristni za ocenjevanje vrednosti podjetij – sploh hitro rastočih start-upov – in pa za ocene o tem, koliko je podjetju še smiselno zapraviti za to, da pridobi novo stranko.

Ker se CLTV nanaša na dobičke skozi vso življenjsko dobo kupca, je kvaliteta napovedi ključna za njegovo uporabnost. Za napovedovanje se uporabljajo različni modeli, od enostavnih, ki temeljijo na nekaj spremenljivkah do kompleksnejših matematičnih modelov, ki se ukvarjajo predvsem s tem, koliko časa bo kupec še ostal zvest podjetju. V zadnjih letih so zaradi hitre rasti na področju stojnega učenja za napovedovanje vse pogosteje uporabljani tudi tovrstni modeli.

V praktičnem delu naloge smo se osredotočili na primerjavo uspešnosti napovedovanja CLTV nekaterih naprednejših modelov strojnega učenja: naključni gozdovi, nevronske mreže in gradientni razvoj niza dreves. V primerjavo smo kot merilo uspešnosti dodali tudi enega najpreprostejšim modelov strojnega učenja, linearno regresijo.

Namen primerjave je bil napovedati CLTV uporabnika na podlagi 7, 14 in 21 dni podatkov za posameznega uporabnika. Vrednost CLTV, ki so jo modeli morali napovedati, smo izračunali na podlagi vseh razpoložljivih podatkov. Za izračun smo uporabili dve različni formuli, enostavno in sestavljeno. Namen različnih formul je simulacija različnih stopenj kompleksnosti izračuna CLTV v različnih podjetjih oz. podjetjih na različni stopnji zrelosti uporabe CLTV – ob predpostavki, da kompleksnost izračuna in posledično njegova točnost raste z zrelostjo uporabe metrike. Za praktični del smo uporabili in obdelali prave anonimizirane podatke, ki nam jih je priskrbelo podjetje Oryx Gaming d.o.o. Po obdelavi smo iz dobrih 100 milijonov vrstic prišli na agregirane in prečiščene podatke za od 2.407 uporabnikov s podatki za 7 dni do 1.693 uporabnikov s podatki za 21 dni.

Rezultati, primerjani s pomočjo determinacijskega koeficienta ($R^2$) so prikazani v tabeli XY. Za napovedovanje CLTV, izračunane s preprosto formulo, se je kot najboljša izkazala linearna regresija, naš najpreprostejši model. Kvaliteta napovedi je vidno padla, ko smo želeli napovedati CLTV izračunan s sestavljeno formulo. Tu se je najbolje odrezala metoda gradientnega razvoja niza dreves (slika XY).




*Source: own work.*

Iz primerjave lahko zaključimo, da je za preprosto CLTV formulo najboljša preprosta rešitev, to je linearna regresija. Ko je problem postal težji in je bilo potrebno napovedati sestavljeno CLTV formulo, so boljše rezultate pokazali kompleksnejši modeli. Vseeno pa sodobne metode strojnega učenja niso odtehtale kompleksnosti težave in v splošnem so bile napovedi za preprosto formulo veliko boljše kot za sestavljeno. Do nadalnjih izboljšav teh modelov bi verjetno prišlo s povečevanjem števila vhodnih podatkov (naš vzorec za učenje je bil relativno majhen) in z nadaljnim umerjanjem hiperparametrov, čeprav pri zadnjem bistvenih izboljšav verjetno ni razumno pričakovati.

## Appendix 2: Description of input data

| Field name | Field description |
|---|---|
| customer_id | Unique identifier for the customer |
| sum_deposit | Sum of value of all the deposits made by the customer |
| avg_deposit | Average value of a deposit by the customer |
| count_deposit | Number of individual deposits made by the customer |
| sum_withdrawal | Sum of value of all the withdrawals made by the customer |
| avg_withdrawal | Average value of a withdrawal by the customer |
| count_withdrawal | Number of individual withdrawals made by the customer |
| sum_bonus | Sum of value of all the bonuses given to the customer |
| avg_bonus | Average value of a bonus given to the customer |
| count_bonus | Number of individual bonuses given to the customer |
| sum_bets | Sum of value of all the bets on the slots made by the customer |
| avg_bets | Average value of a bet on the slots made by the customer |
| count_bets | Number of individual bets on the slots made by the customer |
| sum_wins | Sum of value of all the wins on the slots made by the customer |
| avg_wins | Average value of a win on the slots for the customer |
| count_wins | Number of individual wins on the slots by the customer |
| sum_sbook_bets | Sum of value of all the bets on the sportsbook made by the customer |
| avg_sbook_bets | Average value of a bets on the sportsbook made by the customer |
| count_sbook_bets | Number of individual bets on the sportsbook made by the customer |
| sum_sbook_wins | Sum of value of all the wins on the sportsbook made by the customer |
| avg_sbook_wins | Average value of a wins on the sportsbook made by the customer |
| count_sbook_wins | Number of individual wins on the sportsbook made by the customer |
| count_sessions | Number of sessions by the customer |
| avg_bets_per_session | Average value of a bet on a slot per session for the customer |
| avg_wins_per_session | Average value of a win on a slot per session for the customer |
| avg_sbook_bets_per_session | Average value of a bet on sportsbook per session for the customer |
| avg_sbook_wins_per_session | Average value of a win on sportsbook per session for the customer |
| avg_deposit_per_session | Average value of a deposit per session for the customer |
| avg_withdrawal_per_session | Average value of a withdrawal per session for the customer |
| avg_bonus_per_session | Average value of a bonus per session for the customer |
| gender_male | Customer is male (true/false) |
| gender_female | Customer is female (true/false) |
| age_group_20s | Customer is under 30 (true/false) |
| age_group_30s | Customer is over 30 and under 40 (true/false) |
| age_group_40s | Customer is over 40 and under 50 (true/false) |
| age_group_over50s | Customer is over 50 (true/false) |
| customer_value | CLTV for the customer |

## Appendix 3: Correlations between inputs and CLTV

| Field name[6] | 7 days simple CLTV | 7 days aggregated CLTV |
|---|---|---|
| customer_id | / | / |
| sum_deposit | 0,35 | 0,69 |
| avg_deposit | 0,97 | 0,25 |
| count_deposit | 0,02 | 0,52 |
| sum_withdrawal | 0,10 | 0,36 |
| avg_withdrawal | 0,10 | 0,25 |
| count_withdrawal | 0,00 | 0,21 |
| sum_bonus | 0,21 | 0,39 |
| avg_bonus | 0,25 | 0,16 |
| count_bonus | -0,08 | 0,14 |
| sum_bets | 0,15 | 0,54 |
| avg_bets | 0,16 | 0,05 |
| count_bets | 0,02 | 0,34 |
| sum_wins | 0,15 | 0,53 |
| avg_wins | 0,17 | 0,12 |
| count_wins | 0,02 | 0,32 |
| sum_sbook_bets | 0,22 | 0,29 |
| avg_sbook_bets | 0,50 | 0,13 |
| count_sbook_bets | -0,03 | 0,11 |
| sum_sbook_wins | 0,22 | 0,28 |
| avg_sbook_wins | 0,35 | 0,13 |
| count_sbook_wins | 0,11 | 0,17 |
| count_sessions | 0,04 | 0,16 |
| avg_bets_per_session | 0,16 | 0,45 |
| avg_wins_per_session | 0,15 | 0,43 |
| avg_sbook_bets_per_session | 0,30 | 0,22 |
| avg_sbook_wins_per_session | 0,30 | 0,19 |
| avg_deposit_per_session | 0,51 | 0,60 |
| avg_withdrawal_per_session | 0,08 | 0,21 |
| avg_bonus_per_session | 0,18 | 0,16 |
| gender_male | -0,05 | 0,04 |
| gender_female | 0,05 | -0,04 |
| age_group_20s | 0,03 | -0,04 |
| age_group_30s | 0,03 | 0,05 |
| age_group_40s | 0,00 | -0,02 |
| age_group_over50s | 0,00 | -0,01 |
| customer_value | 1,00 | 1,00 |

[6] See chapter 2 of the appendix for the description of the fields.