



A novel approach to predicting customer lifetime value in B2B SaaS companies

Stephan Curiskis¹ · Xiaojing Dong² · Fan Jiang¹ · Mark Scarr¹

Revised: 2 March 2023 / Accepted: 10 June 2023 / Published online: 4 July 2023
© The Author(s) 2023

Abstract

In this paper, we propose a flexible machine learning framework to predict customer lifetime value (CLV) in the Business-to-Business (B2B) Software-as-a-Service (SaaS) setting. The substantive and modeling challenges that surface in this context relate to more nuanced customer relationships, highly heterogeneous populations, multiple product offerings, and temporal data constraints. To tackle these issues, we treat the CLV estimation as a lump sum prediction problem across multiple products and develop a hierarchical ensembled CLV model. Lump sum prediction enables the use of a wide range of supervised machine learning techniques, which provide additional flexibility, richer features and exhibit an improvement over more conventional forecasting methods. The hierarchical approach is well suited to constrained temporal data and a customer segment model ensembling strategy is introduced as a hyperparameter model-tuning step. The proposed model framework is implemented on data from a B2B SaaS company and empirical results demonstrate its advantages in tackling a practical CLV prediction problem over simpler heuristics and traditional CLV approaches. Finally, several business applications are described where CLV predictions are employed to optimize marketing spend, ROI, and drive critical managerial insights in this context.

Keywords Customer lifetime value · B2B · SaaS · Supervised learning · Ensembling · Hierarchical

Introduction

Customer Lifetime Value (CLV) predicts the future value each customer may generate. It serves as a foundational metric for many organizations (Venkatesan and Kumar 2004). Over the last few years there has been a rising interest among both academicians and industry practitioners in the modeling of CLV (Kanchanapoom and Chongwatpol 2023). This can be attributed to the increasing accessibility of large sets of customer data, especially for online Software-as-a-Service (SaaS) companies (Bakhshizadeh et al. 2022). In addition to the customer purchase data, SaaS companies also record how each customer utilizes the products they purchased. Such online behavioral information has proven to be valuable in predicting future customer purchases and behaviors.

For an online Business-to-Business (B2B) SaaS company, understanding CLV is critical for long-term success. B2B SaaS companies typically have longer sales cycles and higher customer acquisition costs, making it even more important to focus on customer retention. By determining the expected lifetime value of a customer, B2B SaaS businesses can prioritize their efforts toward high-value customers and tailor their services to meet their specific needs. By delivering an exceptional customer experience, companies can improve customer satisfaction and retention rates. Additionally, measuring CLV helps B2B SaaS businesses identify inefficiencies in their customer onboarding process and identify where they can optimize marketing, adjust targeting strategies, and enhance customer service and support efforts. By understanding and improving CLV, B2B SaaS companies can increase revenue from existing customers, reduce churn, and lower their cost of acquisition and improve marketing strategies (Bolton et al. 2004), helping to ensure financial success and industry leadership over the long term.

CLV predictive models are well developed for online consumer customers in the Business-to-Consumer (B2C) space where there is a well-defined direct one-to-one relationship

✉ Stephan Curiskis
scuriskis@atlassian.com

¹ Atlassian Corporation, Sydney, NSW, Australia

² Santa Clara University, Santa Clara, CA, USA



with the customer and typically an abundance of historical behavioral and purchase data. An example is for e-commerce sites such as amazon.com or walmart.com, which have millions of customers and where an individual user makes a purchase decision. By contrast the B2B SaaS use case is less well studied; the customer relationship is a lot more nuanced, highly heterogeneous, can include multiple product offerings, and may lack sufficient historic data (Horak 2017). Thus, CLV modeling poses unique challenges for B2B SaaS providers, in an industry that is highly competitive, global, and rapidly changing.

In this paper, a novel and flexible framework is proposed to predict CLV in a B2B SaaS context. The efficacy of this framework and subsequent business applications are demonstrated on data from a major provider of B2B SaaS products. The CLV estimation framework developed here addresses the following two key challenges empirically observed in B2B SaaS companies:

1. **Constrained temporal data sets:** This issue is highlighted when the available data on customer purchases and behaviors range over shorter time periods than the customer lifetime being modeled, or the data have drifted substantially during the customer data collection period.

CLV models are typically framed over the estimated lifetime of a customer base. However, for customers of B2B SaaS products these lifetimes can span over many years. This challenge arises when the available data may not span the full lifetime of the customer, due to data instrumentation, storage or system changes. For instance, in many cases new data sources that may serve as excellent model features may be instrumented and the data collected only recently, so that these data sources are missing in earlier data periods and cannot be directly incorporated into the model. Secondly, customer lifetimes may extend over several years or even decades, during which time the underlying data generating process and product offerings may change substantially. For example, the product mix may change drastically due to an increasingly competitive landscape. There may also be long-term industry or geographic changes, or events such as pandemics or war which may cause changes rapidly.

2. **Wide variation in CLV prediction drivers by customer segments:** The size and purchasing power for a particular product can vary dramatically across business customers. This leads to longtail problems that impose challenges to any prediction tasks.

This issue surfaces when B2B SaaS products are used by a wide range of heterogeneous customers. Certain features in the model may have drastically different relationships to CLV depending on the nature of the customer, or the customer segment. For instance,

large enterprise customers may be treated differently by receiving, for example, customized pricing. This may produce different CLV feature relationships than smaller customers, which will typically have default pricing per license seat. While some machine learning models may have sufficient complexity to provide a CLV prediction across all customers with reasonable performance, there may be insufficient data in certain customer segments and the model performance may suffer for these segments. Large enterprise customers are one such example, since they are often few in number and can have very different data distributions. Adopting an ensemble approach consisting of different model types may better handle these issues. For example, a tree-based ensemble model such as XGBoost may produce the best CLV predictions for the majority of the customer base, but for large customers we may find that a linear model performs better.

Considering the main CLV goal is to find the most effective ways of treating customers differently based on their potential future values, we first reframe the CLV estimation problem as a lump sum prediction for the total revenue generated across multiple products in the future. This lifts the need to develop a model that predicts a time series of cash flows and, with a well-defined label, opens up a plethora of supervised learning models that permit much richer and more flexible features. While the modeling approach is motivated by, and developed for the B2B SaaS use case the methods are more widely applicable to any situation subject to similar conditions.

The first issue above of constrained temporal data is then addressed by developing a two-step hierarchical T -period CLV model. In the first stage, n periods of revenue data are deployed for calculating the feature set. These features are then used for training a T' periods CLV model, where $T' < T$. In the next step, the T' period prediction is expanded to a T period prediction through a second model that maps the predicted T' period CLV to T period CLV. This second model is simply fit to T periods of data and relies on slowly changing features such as firmographics. For example, if we are interested in forecasting the CLV over a 5-year time horizon, the training data require at least 5 years of historical data to serve as the output label. For the features, we may require at least 2 years of prior historical data to make a reasonable prediction. This indicates data from 7 years ago will be required for training a model to forecast the customer value in the next 5 years. Due to the dynamic nature of the competitive market, it is debatable how useful data that are 7 years old will be going forward. In our approach, we may train a 2-year prediction model with 3 years of features. We then fit a second more general model to predict 5 years of revenue using the 2-year prediction. This means that we only



require 5 years of data to train our model, and we are relying more heavily on the more recent 2-year prediction.

The second issue above is handled by adopting an ensemble approach to the forecasting problem. In particular, we split our data based on the model performance over values of key features which we identify through an error diagnostic analysis. We then treat the data segments and different types of prediction models as hyperparameters, which allows for different segments of customers to be forecast more effectively by different types of models. Since the estimation problem is reformulated as a lump sum prediction, any standard supervised learning model can be employed. The final model is then an ensemble of all the prediction models over the data segments.

The proposed hierarchical ensembled CLV estimation framework was deployed at a major software company which offers Software-as-a Service to a large number of business customers. The method was implemented in multiple business use cases and has demonstrated its power in all of them.

This paper is organized as follows. “Literature review” section describes the current literature. In “Methods” section, we develop the method for the customer lifetime value modeling framework, with our hierarchical T -period model introduced in “Hierarchical T -period prediction” section and our ensembled customer model in “Ensembled Customer segment model” section. “Results and discussion” section describes the experimental results of our method compared to several baseline approaches, with a discussion. In “Business applications” section, we describe several business applications for our model, and we conclude in “Conclusion” section.

Literature review

Our work is built upon multiple streams of existing literature, including the large body of studies developing feasible empirical approaches to predict CLV, and the literature on ensemble methods to prediction models. Marketing has well accepted the concept of customer lifetime value (Dwyer 1989) and its role as a foundational metric in managing customers and resource allocations (Mulhern 1999), or in developing customer targeting schemes for marketing,

support, or other service functions (Venkatesan and Kumar 2004; Bolton et al. 2004) and its importance in building shareholder’s value (Berger et al. 2006). Reichheld and Sasser (1990) established that it is more costly to acquire new customers than keeping them. This sparked long-term research interests in evaluating the customers not just based on one-time purchase, but the value each customer brought in over the lifetime.

A long literature focusing on the transactional value and customer retention in CLV was established. This literature offered deep insights and managerial implications on the practice of valuing customers over their future purchases and its implications to the success of the brand and the company. For example, in a highly cited paper by Gupta et al. (2004), the authors demonstrated how a firm’s financial value depends on the net present value of their customers based on the future values to be generated. In particular, they found that a 1% retention rate increase will lead to a 5% improvement in the firm’s value. In the book written by Rust et al. (2000), the authors established that in order for a company to achieve its success, it is important to switch from product-focused to customer-focused view. In other words, it is critical for a company to devote their efforts and resources on enhancing the value of each customer in the long term, rather than the equity of the brand.

Various mathematical and statistical models have been proposed in the literature to evaluate the value of a customer (Berger and Nasr 1998; Gupta et al. 2004; Rust et al. 2000). The most common approach defines CLV as the Net Present Value (NPV) of the expected flow of revenue after subtracting the cost of acquiring and serving the customer, which gives the expected profit series. The NPV formulation calculates the current value in today’s value by summing the future expected profit series with an appropriate discount rate (see for example Blattberg and Deighton (1996), Lehmann and Gupta (2005), and Jain and Singh (2002)). For customer i , CLV is formulated as

$$CLV_i = \sum_{t=1}^{T_i} \frac{R_{it} - C_{it}}{(1+r)^t}. \quad (1)$$

Table 1 outlines the key notation used in this formulation. We use CLV_i to denote the Customer Lifetime Value

Table 1 Concepts used in the general formulation of Customer Lifetime Value in the literature

Variable	Definition
CLV_i	Customer Lifetime Value for customer i
NPV	Net present value, formulated as a sum over future expected value with a discount rate
T_i	Total expected lifetime for customer i
R_{it}	Expected revenue generated from a customer i in time period t
C_{it}	Expected cost from acquiring and serving customer i in time period t
r	Discount rate used to get the present value of future expected revenue less cost



for customer i . As described in Eq. (1), CLV is defined as the NPV of future, or predicted, cashflows. T_i denotes the expected lifetime of customer i . R_{it} denotes the expected revenue generated from customer i at time period t . For B2B SaaS product providers, revenue from a customer typically is received as period recurring revenue. The cost to acquire and serve customer i at time period t is represented by C_{it} . Note that the Customer Acquisition Costs (CAC) may be included in C_{it} , where $t = 1$. Lastly, r is the discount rate applied.

Built upon this basic model are a series of extensions to incorporate more complex decisions and behaviors in the predictions. The basic model considers the repurchase probability to be the main source of uncertainty, while ignoring other factors that can potentially change repurchase decisions in the future. One extension is to include additional factors into the modeling process. For example, in the book by Blattberg et al. (2001), it allows service usage and cross-buying behaviors to affect the future CLV. In another example, leveraging customer-stated data from a survey combined with their observed data from their purchase records, Kamakura et al. (2002) demonstrate that enhancing service quality will lead to higher customer retention and hence improved a firm's profit. Bolton et al. (2000) examine the impact of a loyalty rewards program and service experiences on the customers repurchase decisions and the CLVs. In a similar finding, Verhoef (2003) documented that loyalty programs with economic incentives positively affect customer retention and repurchase probabilities.

Another extension from the basic model is to relax the assumption that the customer is alive for the whole prediction period and incorporate a variable indicating the probability a customer is still active at a given time period. Such a question can be directly observed in a contractual setting but needs to be modeled in a non-contractual setting. The Pareto/NBD model developed by Schmittlein et al. (1987) offers a solution to gauge the probability a customer is still active even when we do not observe that in the data for the non-contractual setting. This approach has been applied in many studies, such as Schmittlein and Peterson (1994), Reinartz and Kumar (2003), Reinartz and Kumar (2000), Niraj et al. (2001), and Fader et al. (2005).

In the above setting, the CLV calculations focus on the retention rate of the customers at each time period, based on which the future expected profits can be calculated. Alternatively, the CLV calculations are focused on predicting the expected lifetime of each customer. In this approach, in order to predict by which time a customer is likely to churn, hazard/survival models are employed. The conditional hazard at any time period t is defined as the conditional probability a customer leaves, given that by time t he or she has not left. To model the hazard probabilities, three different approaches can be employed, depending on the formulation of the time

variable and the time-variant X variables that shift the baseline hazard.

The first approach is referred to as the proportional hazard model (Cox 1972) and defines the hazard function as a multiplication of two parts. The first part relates to a function of the continuous t variable, while the second part incorporates other time-variant X variables. This allows the probability of churn at any time to be influenced by both time and other X variables. Secondly, instead of using multiplication, the additive risks model proposed by Aalen (1980) defines the hazard as an additive function between the baseline hazard and a function of the explanatory variables. Thirdly, the accelerated failure time model proposed by Prentice and Kalbfleisch (1970) defines one of the parameters in the baseline hazard function as a function of explanatory variables. These statistical tools allow scholars to study the factors that may change customer lifetime. For example, one of the most cited marketing papers by Bolton (1998) leverages such a statistical approach and finds that higher levels of customer satisfaction increase the duration of the provider–customer relationship for a cellular telephone company.

One of the commonalities among the papers mentioned above is their focus on consumer products and measuring CLV at individual B2C consumer level. The list of studies evaluating CLV for business customers is much shorter. For example, in a study of CLV for B2B markets Horak (2017) concluded that some industry segments, such as the financial sector and telecom operators, have high-quality transaction data that can be used for revenue prediction. However, in the Information Technology market, the development of new technologies, products, and competition makes CLV prediction an ongoing challenge. Transaction data are only part of the story, since other information is also important. Venkatesan and Kumar (2004) developed a CLV model for business customers that consists of two sub-models. The first sub-model predicts the purchase frequency of each business customer using the generalized gamma specification of interpurchase timing model. The second sub-model predicts the change in the contribution margin at each time period using a panel data regression. The focus in their paper was to incorporate the impact of the marketing communications across multiple channels, and their potential cost in the future. This will help to address the resource allocation question in order to maximize CLV.

Another relevant stream of literature is related to the development of ensembling approaches to prediction models. Ever since the seminal work by Bates and Granger (1969), the power of combining multiple predictive models has demonstrated increased robustness and accuracy (Garcia-Pedrajas et al. 2005). Model ensembling has been widely adopted in a variety of prediction problems with real world applications, such as retail (Ma and Fildes 2021), weather (Gneiting and Raftery 2005), and disease prediction (Sharma et al. 2021),



among many others. They have established that the results from combining multiple models beat those from the best among the component models, see for example Stock and Watson (1998). As summarized by an excellent review paper by Wang et al. (2022), the development in this literature focuses in two areas, namely enhancing the performance of the component models and developing superior approaches in combining them. Most of the existing studies develop component models on overlapping data sets before combining them to form predictions with enhanced accuracy and reduced uncertainty. In our study, we adopted a divide-and-conquer approach, in which the data space is split into multiple subregions and a separate model is developed for each set, before combining the results.

In this section, we have identified that there is a gap in the literature regarding CLV methods specifically for companies that provide B2B products. Furthermore, the majority of the literature on CLV appears in the field of marketing, while the ensemble approaches are primarily developed from the statistics and machine learning fields. To address a practical issue in our context, and tackle the challenges, our paper contributes to both literature by leveraging the flexibility offered by an ensemble approach in predicting the CLV values across a large range of business customers. The existing literature in CLV requires explicit modeling of the time variable, including the time to churn or the time of the revenue generated. It reflects the basic structure of the CLV definition. However, it imposes restrictions regarding what factors or variables can be incorporated into the CLV prediction model. In the context of real applications of a B2B scenario, we have a large number of those factors that can potentially influence the prediction of a CLV, which includes both time-variant and time-invariant variables, and potentially their interactions. To include that many potential features into the basic CLV model imposes huge challenges in both model setting and estimation. In our approach, we propose to formulate the CLV value in the future as the outcome variable, with a fixed time frame. This definition directly addresses the business interest, and in the meantime, it allows us to leverage the large arsenal of tools from the machine learning field, especially its flexibility of incorporating a large number of factors and their possible interactions. Combining with the ensemble method, our approach enhances model flexibility even more.

Methods

This section describes the development of our model in detail. Following the literature, we define the CLV as the net present value of future expected revenue for each customer (Gupta et al. 2004; Niraj et al. 2001). The classical version (as mentioned in the previous “Literature review” section) is to frame CLV prediction as a forecasting problem of two key metrics, including (i) the renewal

rate for each active customer in each given time period; and (ii) the expected cash flow that will be generated for each active customer. This approach has demonstrated its value in many applications by offering predicted results for the future flow of revenue in a predefined period of time (Berger and Nasr 1998), before including the discount rate and calculating the net present value of such a cash flow.

In our application, however, such an approach imposes tremendous challenges. We quickly discovered that standard forecasting models delivered poor prediction performance for future revenue flows, so using such methods are not feasible in our context due to their large variations in the revenue flows in the future. This is partially due to the fact that we offer multiple products to our customers, and they are free to add new products, and hence increased revenue, at any given point in the future. Considering that our goal is to find the most effective ways of treating our business customers differently based on their potential future values, regardless of which products, we decided to redefine the CLV problem as a lump sum prediction task for all the future revenue across multiple products.

This approach presents three benefits in our context. First, it lifts our burden of trying to devise a model that can predict multiple series of cash flows in the future, one for each product, before combining them. Second, with the label well defined, this approach allows us to adopt a variety of supervised learning models that permit more flexible features and their potential interactions. Finally, given that we offer multiple products, our customers may choose to add or drop a product at any time (depending on the subscription contract, monthly or annual), which could lead to dramatic revenue changes. Forecasting the stream of revenues for each product is not feasible. Redefining the CLV into a lump sum allows us to explicitly include the adding or dropping decisions.

The second challenge facing us is that our context involves a large number of business customers that vary dramatically across the size, industry, and their usage patterns of the products. To accommodate these large variations, we adopt the divide-and-conquer type of ensemble approach, by segmenting our customers based on the error analysis.

Finally, common to all prediction problems, one challenge is how much historical data can be used to forecast the future. On one hand, according to statistics, more data are preferred to provide more stability and efficiency in prediction. On the other hand, in a high-tech world, technology changes fast and the competitive environment switches dramatically over time, relying on data from a few years back may lead to predictions that are already outdated. To strike that balance, we developed the hierarchical approach that allows us to take the best advantage of the data structure.



CLV formulation for B2B SaaS providers

Calculating CLV for customers of B2B SaaS products presents two challenges related to the product mix offered and heterogeneous sets of customers. Firstly, many providers of such software services offer several products. These products often do not serve mutually exclusive use cases, but rather complement each other and are more productive when used together; for example, one may be a primary ‘land’ product and the other products in the portfolio additional ‘expand’ products. Secondly, the customers themselves may be composed of multiple entities ranging from end users to teams with product licenses, company accounts, and corporations with multiple subsidiaries with separate licenses. An individual customer may therefore have data sources available about the usage of different products at the user level, the terms of their product license, and up to their company profile.

For reference, Table 2 outlines the key notation used throughout the rest of this paper.

As described in Eq. (1), CLV is defined as the NPV of future, or predicted, revenue. MRR denotes the Monthly Recurring Revenue, since for our data set revenue is received monthly for each customer. Monthly is a common period for revenue to be recurring for B2B SaaS product providers. MRR is used to define the CLV prediction in this study, as the net present value of future MRR. MRR is also used to calculate model features prior to the CLV prediction period.

The lifetime of a customer can be estimated on an individual per customer basis or can be treated as a constant by considering how likely an average customer is to churn. Most lifetime value models also require a churn model to inform the selection of T_i per customer as highlighted in “Literature review” section. However, a simple choice for the expected lifetime is to derive a scalar value T as one divided by the mean churn rate calculated over historic data. This fixes the lifetime for each customer and allows for lump sum regression models to be used to predict CLV. However, the average customer lifetime T may be long and we may

not have sufficient data to effectively model it. We therefore distinguish between the average customer lifetime for which we T measure CLV over, T , and the shorter timeframe for the lump sum regression prediction, T' . The timeframe for the feature set to be calculated over is denoted by n and we assume an annual discount rate of r . Lastly, since our approach is hierarchical over the two timeframes T and T' , we denote the T' period model as f and the T period model as g .

Let the set of customers be defined as C and the set of products as P . Each customer has a license associated with each product they purchase, defined by the set $L = \{L_{i,p} | i \in C, p \in P\}$. Each customer i has a set of users who have access to products through their respective licenses. Product usage data are generated for each user, which allows a range of features to be derived at the license and customer entity levels. Each customer is billed per license on a monthly basis depending on the number of enabled users during the month, the product type, and any discounting they have received. This generates a Monthly Recurring Revenue (MRR) time series for each license defined by $MRR_{l,t}$, where $l \in L$ and t represents a discrete period in the customer lifetime T_i .

For a customer i , monthly discount rate r and global estimated lifetime T , Customer Lifetime Value CLV_i is given by

$$CLV_i = \sum_{t=1}^T \sum_{p \in P} \frac{MRR_{ipt}}{(1+r)^t} \quad (2)$$

$$= \sum_{t=1}^T \frac{MRR_{it}}{(1+r)^t} \quad (3)$$

It is possible to formulate CLV at either the license level for a customer, as in Eq. (2), or for the sum of all licenses for a customer, as in Eq. (3).

CLV model features

Most B2B SaaS product providers collect a wide variety of data sources that are relevant to understand the future values of a customer, such as customer revenue stream, product license details, product feature usage information, and account-level firmographic details. However, it is common for these data sets to be stored in different systems with varying lengths of historical data. Furthermore, some sets of data may have short or quickly changing distributions. We refer to these issues as historical data constraints. Using these historical data to create feature sets for model training and testing requires careful examination of the overlapping time periods across all of the data sources. Some common types of data available for B2B SaaS providers are outlined below.

Table 2 Key notation used

Variable	Definition
CLV	Customer Lifetime Value
MRR	Monthly Recurring Revenue
T	Average customer lifetime
T'	Lump sum regression prediction timeframe
n	Timeframe for feature set to be calculated over
r	Discount rate
f	T' period CLV model
g	T period CLV model



Revenue data

In our proposed method, the CLV prediction label is constructed per customer by summing MRR data over a predefined number of time periods T . We also craft some input features using this MRR data as trend variables. Revenue data streams collected for billing purposes are generally of high quality with detailed historical coverage, so we are not constrained with this data set.

For training the CLV model, T periods of data are required for the prediction label. Features are also derived from revenue data prior to the T prediction periods, since generally a good predictor of future revenue from a customer is their more recent product invoice amount paid. Furthermore, trends and changes in revenue may provide good indications of future revenue. These features may be defined over n periods of historic data prior to the T prediction periods. This imposes the constraint of requiring $T + n$ periods of MRR data to train a model. For mature products this may be fine as revenue data collected for billing purposes are generally of high quality with detailed historical coverage, but for more recently launched products this may be a severe constraint.

Product license data

Product license information specifies the product acquired by each customer and the channel through which the product license was purchased (*e.g.*, directly purchased from a vendor's website, or through sales and marketing activity). In addition, the timing of each product acquisition and their relative sequence may offer insights on the use case the customer is seeking to fulfill and therefore can serve as meaningful features for predicting the value of the customer.

Product usage data

One of the advantages of a SaaS product, compared to a classical desktop product, is the collection of usage data from customers. These usage data are potentially the most relevant information to predict future customer product utilization, and hence the CLV. A customer with high usage of the products on a regular basis is more likely to have benefited more from the products than a less active customer, and hence is more likely to bring more revenue in the future. The type of product features utilized by each customer could indicate their level of familiarity with the products, and potential commitment to continuing paying for the product in the future. One challenge to including product usage data into a CLV prediction model lies in the way that product features could be changed in the future, new features may be added and some features can become obsolete. To tackle

that, detailed product knowledge would be a helpful guide to engineer useful features.

Firmographic data and customer segments

Firmographic data relate to attributes of a company, such as the number of employees, revenue, industry, and location. These data may be collected in a variety of ways, either by surveys or third-party sources. Product license details and firmographics will often be point in time data sets with the latest values used in prediction, but changes in the data over time may be useful to include as well.

In addition to firmographic data, B2B SaaS providers may have defined further customer segments or "personas." These data may be defined by rules or other machine learning systems. For example, customers may be classified into a high-touch sales segment based on their size, industry, and current product mix. Alternatively, customer research may be used to define personas or customer types based on product needs. These additional data may be factored into a CLV model as features.

Hierarchical T -period prediction

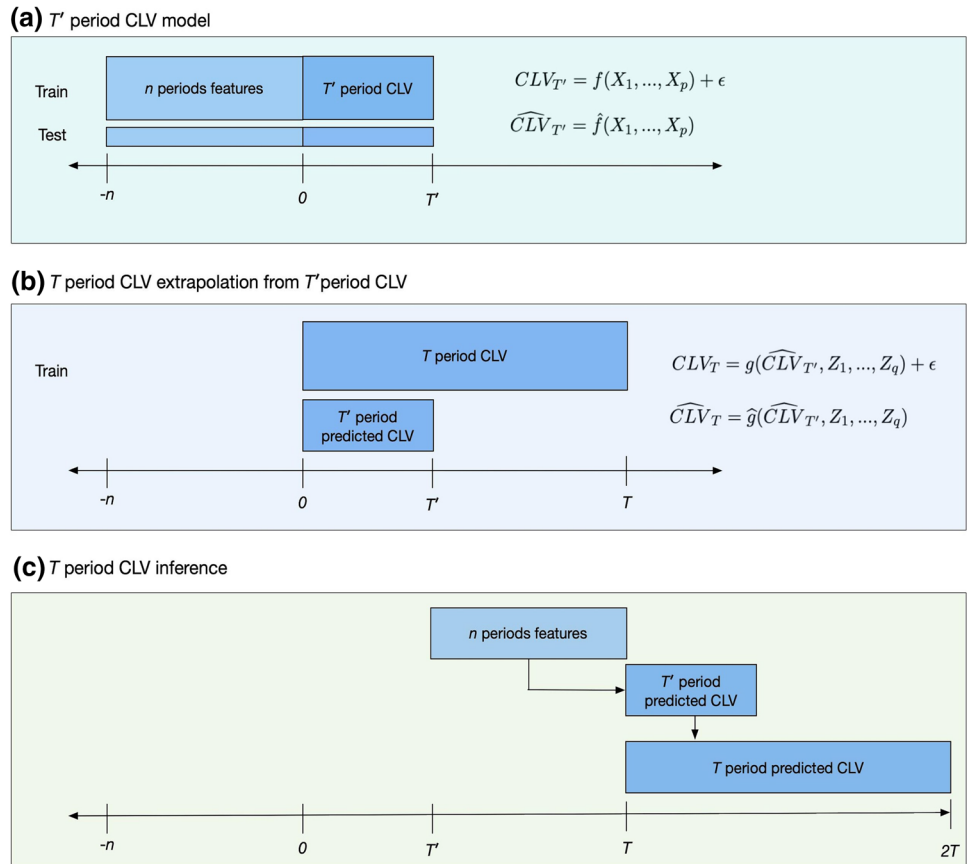
The historical data constraint presents a major challenge for CLV modeling, particularly so for B2B SaaS providers as customers are businesses and their lifetime may be long. Given the dynamic nature and competitive landscape of the technology industry, long-term historical data may exhibit significant drift in the underlying data distributions. Furthermore, new product launches may be relatively frequent, in which case there will be little historical data to train a CLV model on. Training a CLV model over a long history may be learning outdated characteristics of historical data that do not reflect future CLV and may exclude new products and changes in the customer base.

A temporal hierarchical CLV modeling approach is proposed in this paper. In this approach, the model is framed to rely more on recent data than on older data. Such an approach also addresses the limited availability of extensive historical data sets and the issue of data drift often encountered. To facilitate such a model, we define T to be the expected lifetime of our customers. The discounted lump sum value of MRR over this time period is then used as the label in the supervised learning model framework. The features utilized by this model are calculated from a certain time period before these T periods, up to n periods prior.

Figure 1 depicts the design of our hierarchical model over the training and testing data. In the first chart on the top, the data are partitioned into a feature set generated over the first n periods of data and a response set over T' periods of data.



Fig. 1 Hierarchical T period Customer Lifetime Value model for constrained data. We train a model to predict CLV on test data T' , $\widehat{CLV}_{T'}$. To produce a prediction over the full lifetime T , we then train a second model to prediction CLV_T which takes the first prediction $\widehat{CLV}_{T'}$ as input. This second model can be of simpler form and is more stable with respect to long-term drift in the data



The model prediction label is defined as the net present value of monthly recurring revenue over this period, denoted as

$$CLV_{T'} = \sum_{i=1}^{T'} \frac{MRR_i}{(1+r)^i}, \tag{4}$$

where MRR_i is the monthly recurring revenue at month i , and r is the annual discount rate. The model features are calculated over n periods of data and consist of several types of data as outlined in “CLV model features” section. These features consist of lagged MRR values, trends in MRR, and features derived from product usage, user counts, firmographics, and the customer’s product mix. The features used are denoted as X_j , where $j = 1, \dots, p$.

The T' period prediction model then takes the form of

$$CLV_{T'} = f(X_1, \dots, X_p) + \epsilon. \tag{5}$$

We train a machine learning model as the function \hat{f} , which produces the $\widehat{CLV}_{T'}$ predictions in the equation below:

$$\widehat{CLV}_{T'} = \hat{f}(X_1, \dots, X_p). \tag{6}$$

With the shorter T' period CLV prediction as defined in Eq. (6), the hierarchical T period method can be developed. We assume that only T periods of MRR data are available

to use as the response. We train a model to predict the $CLV_{T'}$ response over the same time period and evaluate the model on a random test set. This model takes as input the $\widehat{CLV}_{T'}$ estimate, so serves to map the T' period predictions to T period predictions. Since this model does not have a future ground truth for evaluation, it is framed as a linear model with aggregated customer segment level features that experience lower levels of long-term drift, such as industry, geography, and customer size. These features are denoted as Z_1, \dots, Z_q . This model is defined as

$$CLV_T = g(\widehat{CLV}_{T'}, Z_1, \dots, Z_q) + \epsilon \tag{7}$$

and our CLV_T predictions are produced with our fitted function \hat{g} as

$$\widehat{CLV}_T = \hat{g}(\widehat{CLV}_{T'}, Z_1, \dots, Z_q). \tag{8}$$

Candidate T' period CLV models

The T' period CLV model is the main focus of our methodology since it is reasonable to assume that ground truth response data is available. In comparison, due to the large time horizon of T , we assume enough data for training but



not for ground truth evaluation T periods forward. Since the T' prediction problem involves predicting the Customer Lifetime Value as the net present value of the monthly recurring revenue over T' periods, we may choose a variety of regression models. These are summarized in Table 3.

In our study, we have elected to employ a selection of tree-based ensemble regression models, a LASSO linear regression model, and a K nearest neighbors regression model for the purpose of analysis. The hyperparameters we tested involved several loss functions and some model specific settings in which we saw the largest variations in performance. For the tree-based ensemble models, we tested different loss functions (such as absolute error, squared error and quantile loss) as we hypothesized that these losses may optimize the model differently for different types of customers. In terms of hyperparameters, we tested common settings for tree-based ensemble models using the Optuna hyperparameter tuning package (Akiba et al. 2019) with 50 trials per model. We have also included a time series forecasting model, an ARIMA model with automatic selection of lag variables for both the auto-regressive and moving average components. For comparison purposes, we have trained these auto ARIMA models independently for each customer.

The performance measures used to assess the models were the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the Symmetric Mean Absolute Percentage Error (SMAPE). We found that these measures captured different types of performance relevant to our customer set.

Ensembled customer segment model

In the context of the B2B SaaS industry, customers tend to come from a wide range of businesses, from small startups to large enterprises. The drivers of CLV can vary substantially across the customer base. Building one model to fit all customers can be quite challenging, due to the nature of each customer's purchase decision process. For example, smaller businesses may purchase product licenses directly through a website on a per-user basis. On the other hand, the purchase

decisions of large enterprises may involve convoluted sales discussions and pricing negotiations.

In addition, the number of large enterprise accounts are much smaller compared to small and medium business customers. This may cause machine learning models to deliver worse performance when predicting large enterprise accounts due to the limited size of the relevant data. Considering the importance of those accounts to a SaaS provider, it may be worthwhile to train different types of models separately by key segments such as customer size. Such a segmentation-based approach may improve the overall performance of CLV predictions by ensembling multiple types of models.

For the case of CLV modeling, there are two main model ensembling strategies that can be developed. The first is to explicitly choose a customer segment to train a different type of model on. In the example given previously, a tree-based ensemble model may perform best for small- to medium-sized businesses, while a different type of model may be suitable for larger enterprise customers. Considering the small data size for the large customers, a simple linear model may be a better choice since it generally may better handle smaller sample sizes and the CLV output can scale linearly with the number of users.

The second modeling approach is to treat the model ensembling as a hyperparameter and test different models over different ranges of the input features. This approach allows for a more extensive search of the feature and model performance space but requires more computational resources. In this study, we employ a hybrid combination of these two approaches.

To accomplish our ensembling strategy, we conducted an analysis of the residuals from the T' model. We first calculated the residuals as the difference between the predicted value and the actual value in the out-of-sample test data. We then set the residual as the dependent variable in a linear regression model, and a set of firmographic data and account details as the dependent variables. The variables with high importance in predicting the residuals were then used as hyperparameters to split the data into subsets and train different models on each. If training different

Table 3 Candidate models tested for CLV over T' periods

Model	Optimal settings
LASSO linear regression	Alpha = 1
Gradient Boosted Regressor	Absolute error loss
XGBoost Regressor (Chen and Guestrin 2016)	Squared error loss
LightGBM Regressor (Ke et al. 2017)	Quantile loss
K Neighbors Regressor	Distance weights
Auto ARIMA	Between 1 and 3 lags for both the auto-regressive and moving average components



models on subsets of the data resulted in an overall performance improvement, this configuration was selected as optimal from the hyperparameter tuning step.

Empirical data

The data used to train and evaluate the CLV model presented in this paper were taken from a provider of B2B SaaS products with a large customer base. The T' time period was set to 2 years, with the prediction label calculated on the net present value of MRR from Jan 1, 2021 to December 31, 2022. All customers who had active licenses on Jan 1, 2021 were included in the sample. The average customer lifetime T for the second model was set to 5 years, which ranged from Jan 1, 2018 to December 31, 2022.

The features used in the T' model were derived from a mixture of data regarding revenue, product licenses, product usage, and firmographics, as described in “CLV model features” section. The revenue data used consisted of MRR data calculated over 3 years prior to Jan 1, 2021, and involved various measures of growth trends. Product usage data were similarly calculated over the prior 3 years. Features regarding product license details and firmographics were retrieved on Dec 31, 2020.

The training data consisted of a random 70% sample, and the test data were the remaining 30%. All model performance results were calculated on the test data set.

Results and discussion

In this section, the results of training the T' period CLV model are presented and discussed. We first describe the performance of the different models tested, and then illustrate the customer segment ensembling approach.

Figure 2 shows the performance of each candidate model for the test set across the evaluation measures of the Symmetric Mean Absolute Percentage Error (SMAPE), the Root Mean Squared Error (RMSE), and the Mean Absolute Error (MAE). For reasons of commercial sensitivity, we have indexed each measure by dividing by the value of the top performing model, which was the LightGBM regressor. This still shows the relative differences between methods, while obscuring the sensitive absolute values. For the SMAPE measure, the LightGBM regressor produced a slight performance improvement over the XGBoost regressor model but delivered more significant improvement over the other models. The worst performing model was the auto ARIMA time series model, which delivered a SMAPE value almost 1.75 times that of the LightGBM regressor.

For the RMSE measure, there was only a slight difference in performance between all of the regression models, with LightGBM still delivering the best performance. For this measure, the auto ARIMA model delivered a RMSE score nearly 5 times higher than that for the LightGBM regressor. The comparison was similar for the MAE evaluation measure, but the differences were not as pronounced. For all three measures, the LightGBM regressor delivered the best performance.

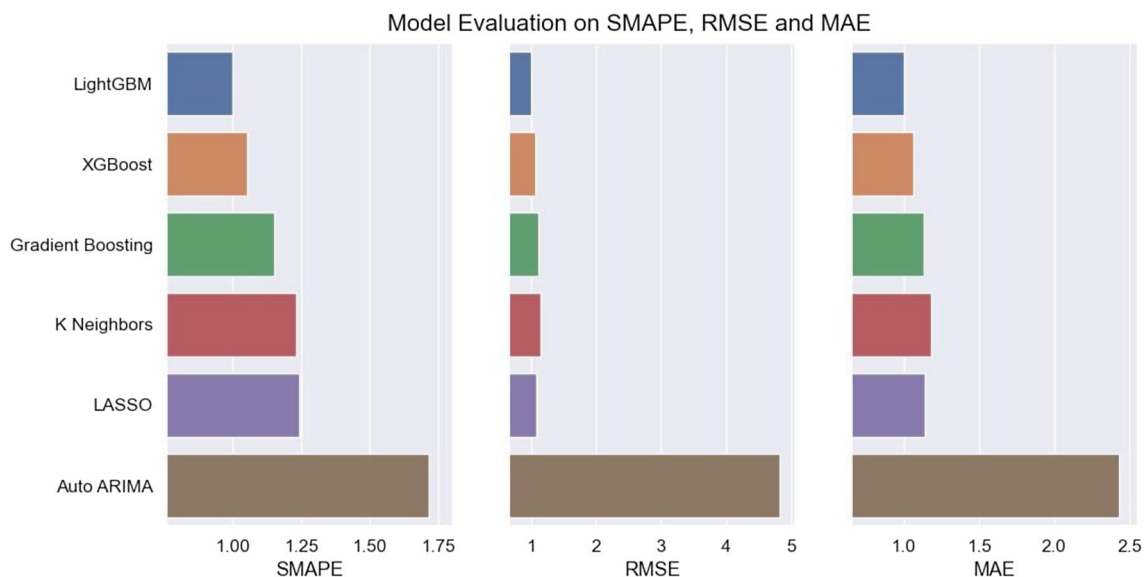


Fig. 2 Evaluation results for each $CLV_{T'}$ model on the same randomly sampled test data set. Values for each measure are indexed to the best performing model with value of 1.0



There are two main reasons for the improved performance of the lump sum regression models over the time series forecasting models. Firstly, by framing the forecasting problem as a regression prediction, we were able to include a wide variety of features calculated at the time of prediction and over the preceding periods of time. This clearly provides these models with the ability to make accurate predictions by adjusting for many other factors that may influence a customer's future revenue. The time series forecasting approach only uses the time series of revenue data to derive features for prediction, while the regression models include a similar set of time series features as well as information about the product license, usage, and firmographics.

The second reason for improved performance of the regression models is that these models are trained over all the customers in the training data. In contrast, the time series forecasting model used was trained independently on each customer's revenue series. The regression models therefore may learn generalized patterns across all the input variables, such as relationships between revenue growth and decline based on firmographics and product usage levels. The time series forecasting models simply model the trends and lags in the revenue series.

The last part of our methodology involves applying the model ensembling strategy as described in “[Ensembled customer segment model](#)” section. This is accomplished with an analysis of the prediction residuals by key model features for the LightGBM model, which gave the best overall performance. These features are then treated as hyperparameters to split the data into subsets on which different types of models are trained as part of the ensemble approach.

To illustrate this process, Fig. 3 plots the residuals by customer size, with values indexed to the ‘Very Small’ customer segment. The ‘Enterprise’ customer size segment has nearly 30 times the residual value than the ‘Very Small’ segment. On the one hand, this is not surprising since enterprise customers are generally large organizations and will have many users using the product. This implies that the revenue numbers will be large, and likewise we may expect the residuals to also be large. However, on the other hand enterprise customers may not have the same nature of relationships between revenue and the number of users, levels of product usage, and so on. In this case, our approach may treat enterprises separately with a different type of model.

Using the example in Fig. 3, based on the residuals we partitioned the data into two sets by size from Very Small to Very Large in one set, and then Enterprise in another set. We then ran a similar hyperparameter tuning step for all the models separately for each data set. The results showed that the LightGBM model performed best for the first data set, but for the Enterprise data set a LASSO model achieved the best performance. This makes sense given the nature of the Enterprise set, which had far fewer customers with

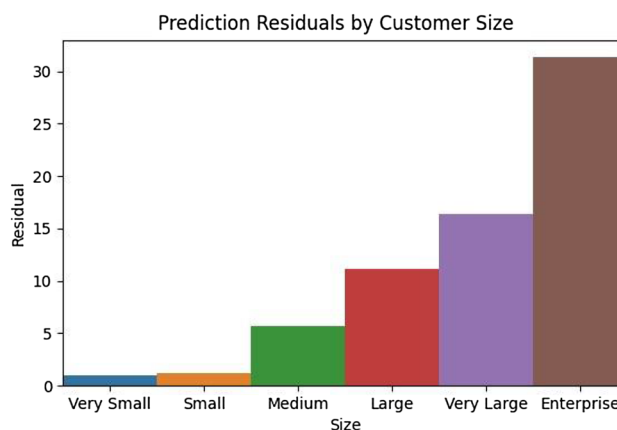


Fig. 3 Plot of residuals by customer size for the best performing LightGBM model

more variation in key features such as the number of users. A linear model like a LASSO may provide predictions that scale with the number of users for the customer, rather than a tree-based ensemble that will make prediction threshold cuts in the data. The final method was an ensemble of the two models, with the LightGBM applied for customers in the Very Small to Very Large segment, and the LASSO model applied for Enterprise customers.

Business applications

We now describe several business applications of the Customer Lifetime Value model developed in “[Methods](#)” section, based on a cloud driven B2B SaaS company. As mentioned previously, in the context of a B2B SaaS company CLV approximates the financial value of each customer. This captures the present value of future cash flows attributed to that customer during their entire relationship with the company, inclusive of expansion, cross-sell, and churn. Due to the nature of the B2B setting, the customer acquisition funnel can be complex consisting of multiple stages as illustrated in Fig. 4. In addition, a business customer may also purchase and engage with a wide selection of products with multiple editions.

These factors determine the need for multiple CLV variants across different applications. For example, CLV framed soon after a new customer starting is needed for customer acquisition campaigns. With the nature of acquisition CLV, we emphasize the power of landing, without sacrificing the short-term gain from acquiring expansion customers. On the other hand, future expected CLV is required for future growth Return Of Investment (ROI) assessments. The following sections take a closer look in more detail at some of the key applications and CLV variants.





Fig. 4 Example of a B2B acquisition funnel

Projected value

Marketing is an area where CLV can be leveraged to drive more higher-quality product evaluations (*i.e.*, ending up with a purchase). However, the challenge is that since CLV is computed at the point of purchase it is not currently defined for these pre-purchase ‘top of funnel’ use cases. To address this issue we can use CLV as the basis to define a measure we refer to as *Projected Value*. The projected value enables us to estimate the revenue of the current spending, and establish a healthy and sustainable customer acquisition flywheel and is defined as

$$\begin{aligned}
 \text{projected value} &= \# \text{ signups} \times \text{purchase rate} \\
 &\times \text{acquisition CLV}.
 \end{aligned}
 \tag{9}$$

As can be seen from Eq. (9), projected value is the estimated value of a customer signup based on a likelihood to purchase and the predicted lifetime value of a customer at the point of purchase. In Eq. (9) we employ ‘acquisition CLV,’ which takes revenue from the initial ‘land’ product and includes revenue from any additional ‘expand’ products that are added to the first land product at a later date. Acquisition CLV is computed over a *T*-period time horizon, which means additional products that are added have their lifetime values truncated to give a maximum of *T*-periods tenure for the customer. Acquisition CLV can then be combined with any relevant segmentation data, such as geography and channel.

At a macro level, projected value is used for budget planning to support future business growth. At a micro level, it provides a key input for marketing ROI optimization, which is discussed in the next section.

Return-of-investment (ROI) optimization

In a marketing context ROI is simply the net revenue that marketing campaigns generate against the amount of money



Fig. 5 Daily \$ Spend and Projected Value

a company spends on those campaigns. In other words, ROI calculations are critical to understanding expected return *vs.* marketing spend. While there are numerous approaches to measuring future expected return, CLV and projected value provide a natural mechanism to do this and furnish more optimal targeting strategies beyond straightforward product purchases. By considering projected value as a function of marketing spend, we have

$$\text{projected value} = f(\text{spend}).
 \tag{10}$$

Figure 5 shows an illustrative plot of projected value *vs.* spend, where each point on the graph represents a daily value. A simple regression model of the form given in Eq. (11) can then be fitted to the data.

$$\text{projected value} = a * \ln(\text{spend}) + c.
 \tag{11}$$

Marginal ROI is defined as the expected return for each additional dollar spent, which from Eq. (11) gives

$$\begin{aligned}
 \text{marginal roi} &= (a * \ln(\text{spend} + 1) + c) \\
 &- (a * \ln(\text{spend}) + c) \\
 &= a * \ln((\text{spend} + 1)/\text{spend})
 \end{aligned}
 \tag{12}$$

Equation (12) then provides the following decision rule:

$$\begin{aligned}
 \text{marginal roi} \geq 1 & \Rightarrow \text{increase \$ spend} \\
 \text{marginal roi} < 1 & \Rightarrow \text{decrease \$ spend}
 \end{aligned}
 \tag{13}$$

This decision rule can be applied at various levels of segmentation, for example the marketing team may define different ROI goals across different products and marketing channels to meet specific business strategies. In some instances, if the focus is primarily around brand awareness then a slightly negative ROI goal may be tolerated, however,



in other cases the team would focus on ROI neutral or positive. Overall the CLV-based ROI optimization provides a much more holistic approach to informing marketing budgeting decisions compared to other methods based, for example, on cost-per-lick (CPC) or cost-per-evaluation (CPE) type metrics. A lower CPC or CPE does not mean a higher yield return, since CPC or CPE does not consider whether the evaluation ends up with a purchase and any sales beyond the purchasing point.

Conclusion

This paper presents a flexible new machine learning framework to predict customer lifetime value in a B2B SaaS context. The methodological contributions address the key modeling challenges associated with CLV prediction in the B2B SaaS setting, such as highly heterogeneous populations, multiple product offerings, temporal data constraints, and a nuanced customer relationship. The CLV prediction is first framed as a lump sum regression problem, allowing the use of a wide range of machine learning algorithms and a much richer feature set compared to traditional time-series-based CLV forecasting methods. A two-stage temporal hierarchical regression model is then employed, which places more emphasis on recent data, mitigating the constraints associated with limited availability of extensive historical data sets and the issue of data drift. Finally, an ensemble of models is utilized within the framework to address business heterogeneity by leveraging a customer segment ensembling strategy as a hyperparameter tuning step. While the framework was motivated by and developed for the B2B SaaS space it is certainly generalizable to any application subject to similar challenges.

The hierarchical ensembled prediction framework was evaluated for various different machine learning models and segmentation ensembling strategies using sample data from a well-known B2B SaaS company across a 2-year period from Jan 1, 2021 to December 31, 2022. Several standard metrics were used to measure the model efficacy (e.g., RMSE, SMAPE, MAE), with the proposed machine learning framework demonstrating a 2–5× improvement in performance over a traditional time-series-based CLV estimation baseline. In fact there were relatively small performance improvements among the different lump sum regression models explored, indicating that in general this approach outperforms the traditional baseline irrespective of the particular machine learning model used. This is a key advantage as it allows the flexibility to employ a model based on other criteria such as simplicity, interpretability, or ‘cost-to-serve’ the predictions.

Ultimately, the value of such a novel CLV prediction framework is how it can be leveraged in practice, for example, to provide valuable insights into how best to allocate resources, refine product offerings, promote customer retention, target acquisition campaigns, or generally enhance business decisions across a company. In this study, a number of marketing applications are described where CLV predictions are employed in a B2B SaaS company to drive critical managerial insights by (i) deriving a new CLV-based metric called *projected value* that allows holistic top-of-funnel marketing spend optimization and (ii) an ROI framework that guides marketers on budget forecasting, planning, and campaign allocation spend. A direct consequence of implementing this approach has allowed marketers to allocate their budget more efficiently, target more profitable markets and channels, and establish a healthy and sustainable customer acquisition flywheel. It has also empowered them to be more data driven and optimal in day-to-day operations.

Funding Open access funding provided by SCELCC, Statewide California Electronic Library Consortium.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aalen, O.O. 1980. *A model of nonparametric regression analysis of counting processes*. New York: Springer.
- Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Bakhshizadeh, E., H. Aliasghari, R. Noorossana, and R. Ghousi. 2022. Customer clustering based on factors of customer lifetime value with data mining technique (case study: Software Industry). *International Journal of Industrial Engineering & Production Research* 33 (1): 1–16.
- Bates, J., and C. Granger. 1969. The combination of forecasts. *Journal of the Operational Research Society* 20 (4): 451–468.
- Berger, P.D., N. Eechambadi, M. George, D.R. Lehmann, R. Rizeley, and R. Venkatesan. 2006. From customer lifetime value to



- shareholder value: Theory, empirical evidence, and issues for future research. *Journal of Service Research* 9 (2): 156–167.
- Berger, Paul D., and Nada I. Nasr. 1998. Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing* 12 (1): 17–30.
- Blattberg, R.C., and J. Deighton. 1996. Manage marketing by the customer equity test. *Harvard Business Review* 74 (4): 136.
- Blattberg, R.C., G. Getz, and J.S. Thomas. 2001. *Customer equity: Building and managing relationships as valuable assets*. Brighton: Harvard Business Press.
- Bolton, R.N. 1998. A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Science* 17 (1): 45–65.
- Bolton, R.N., P.K. Kannan, and M.D. Bramlett. 2000. Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academy of Marketing Science* 28 (1): 95–108.
- Bolton, R.N., K.N. Lemon, and P.C. Verhoef. 2004. The theoretical underpinnings of customer asset management: A framework and propositions for future research. *Journal of the Academy of Marketing Science* 32 (3): 271–292.
- Chen, T., and C. Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, pp. 785–794, New York, NY, USA. ACM.
- Cox, D.R. 1972. Regression models and life tables. *Journal of the Royal Statistical Society B* 34: 187–220.
- Dwyer, F.R. 1989. Customer lifetime valuation to support marketing decision making. *Journal of Direct Marketing* 3 (4): 8–15.
- Fader, P.S., B.G. Hardie, and K.L. Lee. 2005. RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research* 42 (4): 415–430.
- García-Pedrajas, N., C. Hervás-Martínez, and D. Ortiz-Boyer. 2005. Cooperative coevolution of artificial neural network ensembles for pattern classification. *IEEE Transactions on Evolutionary Computation* 9 (3): 271–302.
- Gneiting, T., and A. Raftery. 2005. Weather forecasting with ensemble methods. *Science* 310 (5746): 248–249.
- Gupta, S., D.R. Lehmann, and J.A. Stuart. 2004. Valuing Customers. *Journal of Marketing Research* 41 (1): 7–18.
- Horak, P. 2017. Customer lifetime value in b2b markets: Theory and practice in the Czech republic. *Int. J. Bus. Manag* 12 (2): 47–55.
- Jain, D., and S.S. Singh. 2002. Customer lifetime value research in marketing: A review and future directions. *Journal of Interactive Marketing* 16 (2): 34–46.
- Kamakura, Wagner A., Vikas Mittal, Fernando de Rosa, and Jose Afonso Mazzon. 2002. Assessing the service profit chain. *Marketing Science* 21 (3): 294–317.
- Kanchanapoom, K., and J. Chongwatpol. 2023. Integrated customer lifetime value (CLV) and customer migration model to improve customer segmentation. *Journal of Marketing Analytics* 11: 172–185. <https://doi.org/10.1057/s41270-022-00158-7>.
- Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30: 3146–3154.
- Lehmann, D.R., and S. Gupta. 2005. *Managing customers as investments: The strategic value of customers in the long run*. New York: Wharton School Publishing.
- Ma, S., and R. Fildes. 2021. Retail sales forecasting with meta-learning. *European Journal of Operational Research* 288 (1): 111–128.
- Mulhern, F.J. 1999. Customer profitability analysis: Measurement, concentration, and research directions. *Journal of Interactive Marketing* 13: 25–40.
- Niraj, R., M. Gupta, and C. Narasimhan. 2001. Customer profitability in a supply chain. *Journal of Marketing* 65 (3): 1–16.
- Prentice, R.L., and J.D. Kalbfleisch. 1970. Hazard rate models with covariates. *Biometrics* 35 (1): 25–39.
- Reichheld, Fredrick, and W.Earl. Sasser Jr. 1990. Zero defections: Quality comes to services. *Harvard Business Review* 68 (5): 105–111.
- Reinartz, W.J., and V. Kumar. 2000. On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing* 64 (4): 17–35.
- Reinartz, W.J., and V. Kumar. 2003. The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing* 67 (1): 77–99.
- Rust, Roland, Valarie A. Zeithaml, and Katherine N. Lemon. 2000. *Driving customer equity, how customer lifetime value is reshaping corporate strategy*. New York: Free Press.
- Schmittlein, D.C., D.G. Morrison, and R. Colombo. 1987. Counting your customers: Who-are they and what will they do next? *Marketing Science* 33 (1): 1–24.
- Schmittlein, D.C., and R.A. Peterson. 1994. Customer base analysis: An industrial purchase process application. *Marketing Science* 13 (1): 41–67.
- Sharma, N., J. Dev, M. Mangla, V.M. Wadhwa, S.N. Mohanty, and D. Kakkar. 2021. A heterogeneous ensemble forecasting model for disease prediction. *New Generation Computing* 39: 701–715.
- Stock, J., and M. Watson. 1998. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* 23 (6): 405–430.
- Venkatesan, R., and V. Kumar. 2004. A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing* 68 (4): 106–125.
- Verhoef, P.C. 2003. Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing* 67 (4): 30–45. <https://doi.org/10.1509/jmkg.67.4.30.18685>.
- Wang, X., R.J. Hyndman, F. Li, and Y. Kang. 2022. Forecast combinations: An over 50-year review. *International Journal of Forecasting*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Stephan Curiskis is a Manager in the Machine Learning team at Atlassian Corporation, under Mark Scarr. His primary focus involves managing the development and productionization of machine learning solutions for stakeholders in Growth, Product, Marketing, and Finance. He has previously held technical and management roles in the machine learning space at several technology companies spanning the fintech, healthcare, and geospatial mapping industries. He has received his Ph.D. in machine learning from the University of Technology Sydney.

Xiaoqing Dong is a tenured Associate Professor of Marketing and Business Analytics at Santa Clara University. Her research area applies Data Analytics techniques to develop models examining how company marketing actions influence customer decisions, and provides suggestions on improving Marketing and Business decisions. Her studies span multiple industries, including social network, online retail, pharmaceutical, travel, and high tech. Her papers have appeared in top academic journals, such as *Marketing Science*, *Journal of Marketing Research*, and *Journal of Marketing Analytics*. She has received her B.E. degree from Tsinghua University in China, M.S. from MIT and PhD degree from Northwestern University.



Fan Jiang is a Manager in the Machine Learning team at Atlassian Corporation, under Mark Scarr. Her primary focus includes managing the development and productionization of machine learning solutions across business in Search, Marketing, and Product. She has previously held technical and management roles in the ML spaces at several technology companies spanning the real estate, high tech, and human resources industries. She has received her master's degree from Cornell University.

Mark Scarr has a Ph.D. in applied statistics and currently leads the Machine Learning team at Atlassian Corporation, a software company building the next generation of team collaboration and productivity software. His primary focus includes leading the development and scaling of innovative AI-infused solutions and data products across the business. He has previously held various technical and leadership positions in the ML/AI space at technology companies spanning applications from chip manufacturing to web-search and digital payments

