MAXIMIZING CUSTOMER LIFETIME VALUE USING DYNAMIC PROGRAMMING: THEORETICAL AND PRACTICAL IMPLICATIONS

Eman AboElHamd, Department of Operations Research and Decision Support, Cairo University
Hamed M. Shamma, School of Business, The American University in Cairo
Mohamed Saleh, Department of Operations Research and Decision Support, Cairo University

ABSTRACT

Dynamic programming models play a significant role in maximizing customer lifetime value (CLV), in different market types including B2B, B2C, C2B, C2C and B2B2C. This paper highlights the main contributions of applying dynamic programming models in CLV as an effective direct marketing measure. It mainly focuses on Markov Decision Process, Approximate Dynamic Programming (i.e. Reinforcement Learning (RL)), Deep RL, Double Deep RL, finally Deep Quality Value (DQV) and Rainbow models. It presents the theoretical and practical implications of each of the market types. DQV and Rainbow models outperform the traditional dynamic programming models and generate reliable results without overestimating the action values or generating unrealistic actions. Meanwhile, neither DQV nor Rainbow has been applied in the area of direct marketing to maximize CLV in any of the market types. Hence, it is a recommended research direction.

Keywords: Approximate Dynamic Programming (ADP), Customer Lifetime Value (CLV), Deep Q Network (DQN), Double Deep Q Network (DDQN), Markov Decision Process (MDP), Reinforcement Learning (RL), Business to Business (B2B), Business to Consumer (B2C), Consumer to Business (C2B), Consumer to Consumer (C2C), Business to Business to Consumer (B2B2C).

INTRODUCTION

Customer lifetime value (CLV) is defined as the present value of all future profits that are obtained from the customers over their life of relationship with a firm. Cannon et al. (2014) mentioned that identifying the most profitable customers and investing marketing resources in them, would help the firm to gain more profits, steady market growth, and increase its market share. Thus, it is not surprising for the firm to treat customers differently according to their level of profitability.

CLV is tightly related to the market types of Business to Business (B2B), Business to Consumer (B2C), Consumer to Business (C2B), Consumer to Consumer (C2C), and Business to Business to Consumer (B2B2C). Starting from B2B that refers to any company sells products or services to another company or organization (Cortez et al. 2019). Meanwhile, B2C is considered as a traditional and more popular market type, that controls the relationship between the firm and its customers (Lin et al. 2019). Meanwhile, C2B refers to the consumer who offer product or service to a business entity, C2C is the way that allows customers to interact with each other to buy and sell products or services (Rachbini et al. 2019). Combing B2B to B2C results in B2B2C, that might be considered as the most interesting marketing type. It is used usually in Food and Beverage companies; it refers to the process that when the consumer buys certain products from two different interacted companies (i.e. when the consumer buys Burger and Coca, the former is his need and the latter is an add-on), Baidokhti et al. (2019). CLV plays a significant role in measuring the value of the second party (i.e. the customer) in each of these market types, and in even B2B2C newly evolved market type (Lau et al. 2019). Figure 1 presents the relationship between CLV and each of these market types.

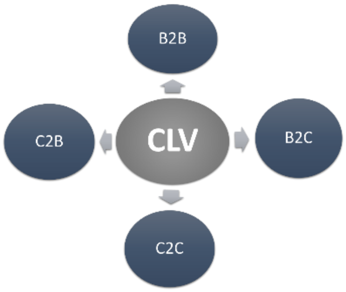

FIGURE 1
RELATIONSHIP BETWEEN CLV AND B2B, B2C AND C2C

This paper presents and analyzes the latest contributions of dynamic programming models in the area of maximizing CLV. It categorizes these contributions according to their market type and application area. It mainly focuses on B2B, B2C and C2C markets and briefly touches the contributions related to B2B2C market type. The rest of the paper is organized as follows; Section-2 introduces the dynamic programming models. It presents their main ideas and the algorithm of each. Section-3 reviews the literature contributions in the area of maximizing CLV, focusing on the theoretical models. While, Section-4 presents the practical applications of each of the dynamic programming models. It is divided into subsections according to the market types of interest (i.e. B2B, B2C, C2B and C2C). Section-5 summarizes and discusses the paper. Finally, Section-6 concludes the paper and highlights the future research directions.

BACKGROUND

This section presents the dynamic programming models used to maximize CLV. Briefly introduces the main idea behind each of those algorithms and its corresponding steps. It is constructed in a way that shows the evolution of the dynamic programming algorithms in the area of maximizing CLV; such that every algorithm overcomes the limitations of its previous one. Subsection-2.1 presents Markov Decision Process (MDP) as a basic and traditional model, then Approximate Dynamic Programming (RL) model that outperforms MDP in complex problems and overcomes most of its limitations. Section-2.3 highlights the main idea and the algorithm of major deep reinforcement learning approaches (i.e. Deep Q Networks, and Double Deep Q Networks).

Markov Decision Process

Markov Decision Process (MDP), belongs to a dynamic programming umbrella. It is the designed in a way that allows the agents to take actions in a grid-world environment, based on their current state and a taken action. Hence, MDP depends mainly on the agent's current state (s), next state (s'), an action (a) that invokes the transition from the current to the next state, and finally, the discounted accumulated reward value (r) as demonstrated in Figure 2. These factors are combined in a tuple (s, a, s', r, γ), where γ is a discount factor ($0 < \gamma <= 1$) and related to each other mathematically in Eqs. [1, 2], where Eq. [1] describes the transition probability function, and Eq. [2] presents the reward function.

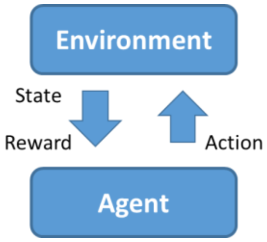

FIGURE 2
MDP KEY ELEMENTS

In MDP, the mapping between states and actions is called a *"Policy"*, and the goal is to find the optimal policy that would maximize the long term accumulated rewards (Ekinci et al. 2014). This optimal solution is achieved either by (value iteration algorithm), or (policy iteration algorithm). These are the two main solution approaches of finding optimal policy in MDP. The steps of each algorithm are listed in Algorithm-1, and Algorithm-2 respectively (Van et al. 2009, AboElHamd et al. 2019).

$$P_a{}^{ss'} = P[S_{t+1} = s' \,| S_t = s, A_t = a] \quad (1)$$

$$R_s{}^a = E[R_{t+1} \,| S_t = s, A_t = a] \quad (2)$$

Algorithm-1: Value Iteration

```
1: initialize V arbitrarily (i.e. $V(x) = 0, \forall s \in S$)
2: repeat
        loop to all $s \in S$
        set $\gamma$ = V(s)
        loop to all $a \in A(s)$ and do
        $Q(s,a) = \sum_{s'} T(s,a,s')(R(s,a,s') + \gamma V(s'))$
        $\Delta = max_a Q(s,a) V(s)|)$
3: until $\Delta < \sigma$
```

- Adapted from AboElHamd et al. (2019)

Approximate Dynamic Programming

This subsection presents Approximate dynamic programming (ADP) as a more advanced algorithm than MDP. It is a modeling approach that solves large and complex problems. It's proved to outperform MDP in many cases especially in complex problems, as its mechanism allows it to overcome the curses of dimensionality. In fact, it is capable of overcoming the three types of dimensionality problems (i.e. in state space, action space or outcome space). Algorithm-3 illustrates the steps of generic ADP as mentioned by Powell et al. (2008) and already presented in the work of AboElHamd et al. (2019).

Algorithm-2: Policy Iteration

```
1: $V(s) \in <$ and $\pi(s) \in A(s)$ arbitrarily $\forall s \in S$)
        [POLICY EVALUATION]
2: repeat
        $\Delta = 0$
        Loop to all $s \in S$ and do
        set $\gamma = V^{\pi}(s)$
        $V(s) = \sum_{s'} T(s,\pi(s),s')(R(s,\pi(s),s') + \gamma V(s'))$
        $\Delta = max(\Delta, |\gamma - V(s)|)$
3: until $\Delta < \sigma$
        [POLICY IMPROVEMENT]
4: policy-stable = True
5: loop to all $s \in S$ and do
        b = $\pi(s)$
        $\pi(s) = argmax_a \sum_{s'} T(s,a,s')(R(s,a,s') + \gamma V(s'))$
        if $b \neq \pi(s)$ then policy-stable = False
6: if policy-stable then stop; else go to step-2
```

- Adapted from AboElHamd et al. (2019)

Approximate dynamic programming (also called Reinforcement learning) especially in computer science's context. While, ADP term is mainly used in the context of Operations Research. Q-learning algorithms play a significant role in ADP/RL, because of being model free algorithm. The main idea of Q-learning is to learn the action-value function $Q(s, a)$. During the learning process, at every state, action is evaluated and the goal is to select the action that maximizes the long term rewards. Traditionally, Q-learning was implemented using lookup tables; where, Q values were stored for all possible combinations of states and actions. Eventually, machine learning algorithms help in learning the Q-values. Q-learning in its basic form is mentioned in Algorithm-4 (AboElHamd et al. 2019).

Algorithm-3: Approximate Dynamic Programming

```
1: initialization
        initialize $V_t^0(S_t)\forall$ state $S_t$ choose an initial state $S_0^{\,1}$
        set n = 1
2: choose a sample path $\omega^n$
3: for $t = 0$, $t$++, while $t < T$ do
        assuming a maximization problem, solve

            $\gamma_t^{\,n} = max_{xt}(C_t(S_t^{\,n}, x_t) + \gamma E[V_{t+1}^{\,n-1}(S_{t+1})|S_t])$
        update $V_t^{n-1}(S_t)$ using

        $V_t^n(S_t) = \begin{cases} (1 - a_{n-1})V_t^{n-1}(S_t^n) + a_{n-1}\gamma_t^n, & if\ S_t = S_t^n \\ V_t^{n-1}(S_t), & otherwise \end{cases}$

        Compute $S_{t+1}^n = S^M(S_t^n, x_t^n, W_{t+1}(\omega^n))$
4: let n = n+1. If n < N, go to step 1.
```

- Adapted from AboElHamd et al. (2019)

Algorithm-4: Q-Learning

```
1: start with $Q_0(s, a)\ \forall s, a$
2: get initial state s
3: **for** $k$ = 1, $k$++, while not converged do
        sample action a, get next state $s$
4:          **if** $s'$ is terminal **then**:
                target = $R(s, a, s')$
                sample new initial state $s'$
        **else** $target = R(s, a, s') + \gamma max_{a'} Q_k(s', a')$
5:      $Q_{k+1}(s, a) \leftarrow (1 - \alpha(Q_k(s, a) + \alpha[target])$
6:      $s \leftarrow s'$
```

- Adapted from AboElHamd et al. (2019)

Deep Reinforcement Learning

The goal of dynamic programming is to find a solution for the problem at hand. In many cases, an optimal solution is obtained. Meanwhile, in complex and large problems, finding the exact and optimal solution is almost impossible. Hence, searching for an approximate solution is needed. This was the main motivation behind utilizing ADP at first and eventually, deep reinforcement learning (Or Deep Q Networks (DQN)), Algorithm-5. DQN combines Q Learning and neural networks algorithms. The basic idea of DQN is that the action-value function Q is approximated using neural network (or more precisely, deep learning) algorithm, instead of using lookup tables (Li et al. 2015). Figure-5 lists the main steps of DQN algorithm (AboElHamd et al. 2019).

Algorithm-5: DQN with experience replay

```
1: initialize replay memory $D$ to capacity $N$
2: initialize action-value function $Q$ with random weights $\theta$
3: initialize target action-value function $Q^0$ with weights $\theta^- = \theta$
4: **for** $episode$ = 1, while $episode < M$ **do**
        initialize sequence $s_1 = [x_1]$ and pre-processed sequence $\varphi_1 = \varphi(s_1)$
5:      **for** $t$ = 1, while $t < T$ do
            select a random action $a_t$ with probability $\varepsilon$; otherwise, select $a_t$ =
            $argmax_a Q(\emptyset(s_t), a, \theta)$
```

```
execute action $a_t$ and observe a reward $r_t$ and $x_{t+1}$ set $s_{t+1} = s_t, a_t, x_{t+1}$ and pre-process $\varphi_{t+1} = \varphi(s_{t+1})$
store transition $(\emptyset_t, a_t, r_t, \emptyset_{t+1})$ in $D$
sample random mini-batch of transitions $(\emptyset_t, a_t, r_t, \emptyset_{t+1})$ from $D$
$set\ y_j = \begin{cases} r_j, & if\ episode\ terminates\ at\ step\ j+1 \\ r_j + \gamma max_{a'} Q'(\emptyset_{j+1}, a';\ \theta^-), & otherwise \end{cases}$
optimize $\theta$ using gradient descent for $(y_j - Q(\emptyset_j, a_j;\ \theta))^2$
set $Q' = Q$ every C steps
```

- Adapted from AboElHamd et al. (2019)

This sections briefly presents the main implementation steps of each dynamic programming models, mentioned in Figure 3. The upcoming two sections analyze their corresponding theoretical and practical applications. The analyzed contributions are classified according to their market types, whether being related to B2B, B2C or C2C markets. Similar to the work in this paper, is the one done by AboElHamd et al. (2019) who conducted a survey on the most significant contributions in the area of CLV whether by developing basic models, for calculating the value of CLV, analyzing it, segmenting the customer base upon its value, predicting it or even maximizing it. Meanwhile, there are two differences between their work and the work in this paper. First, this work focuses mainly on the contributions that maximize CLV, analyzing its theoretical and practical implications. Second, this work classified the theoretical and practical work to market types (i.e. B2B, B2C and C2C).

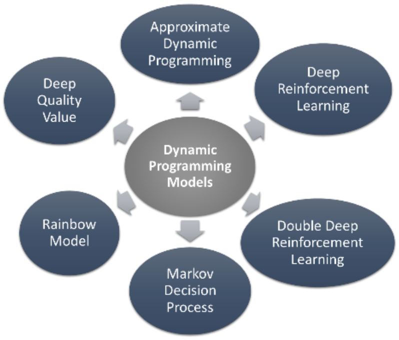
Approximate Dynamic Programming
Deep Reinforcement Learning
Deep Quality Value
Dynamic Programming Models
Rainbow Model
Double Deep Reinforcement Learning
Markov Decision Process

FIGURE 3
CLV MODELS

THEORETICAL MODELS OF CLV

Researchers competed in developing different models to maximize CLV, due to its significance in direct marketing. This section focuses on analyzing the theoretical contributions, while Section-4 presents the practical contributions. As mentioned in the previous section, states and actions are two of the main MDP's elements. These might take either discrete or continuous values, and accordingly the dynamic programming model is defined as being discrete or continuous. Table 1, lists the type of states and actions in the list of publications analyzed in this paper. Table 1 shows that the majority of the researchers assume discrete states/actions and only few of them work on problems with continuous states or actions. Simester et al. (2006) already reviewed the history of field experiments in marketing over 20 years, in their book chapter. They grouped set of papers into topics and reviewed them per topic.

Table 1
TYPE OF PUBLICATIONS' STATES, ACTIONS

		States		**Actions**	
	Publication	**Discrete**	**Continuous**	**Discrete**	**Continuous**
Markov Decision Process	Ching et al. (2004)	√	-	√	-
	Haenlein et al. (2007)	√	√	√	-
	Labbi et al. (2007)	√	-	√	-
	Mannor et al. (2007)	-	√	√	-
	Ma et al. (2008)	√	-	-	-
	Cheng et al. (2012)	-	√	√	-
	Ekinci et al. (2014)	-	√	√	-
	Wang et al. (2014)	√	-	√	-
	Zhang et al. (2014)	-	√	-	√
	Ekinci et al. (2014)	√	-	√	-
	Klein et al. (2015)	√	-	√	-
	Permana et al. (2017)	√	-	-	-
	Hwang et al. (2016)	√	-	√	-
Approximate Dynamic Programming	Simester et al. (2006)	√	-	√	-
	Bertsimas et al. (2007)	√	-	√	-
	Chen et al. (2012)	√	-	√	-
	Bravo et al. (2014)	-	√	-	√
	Wei et al. (2014)	-	√	-	√
	Jiang et al. (2015)	-	√	-	√
	Jiang et al. (2015)	-	√	-	√
	James et al. (2016)	√	-	√	-
Deep Reinforcement Learning	Hasselt et al. (2010)	√	-	√	-
	Silver et al. (2013)	√	-	√	-
	Tkachenko et al. (2015)	-	√	√	√
	Theocharous et al. (2016)	√	-	√	-
	Li et al. (2015)	-	√	√	-
	Shaohui et al. (2016)	-	√	√	-
	Hessel et al. (2017)	√	-	√	-
	Zhao et al. (2018)	-	√	-	√
	Kalashnikov et al. (2018)	√	-	√	-
	Sabatelli et al. (2018)	√	-	√	-

Markov Decision Process Models

Over many years, MDP proved to have significant results in the area of maximizing CLV. In fact, MDP best fit the CLV maximization problem, due to the model assumptions and mechanism that is illustrated in the previous section. This encourages many researchers to compete in this area of research, and some of them is mentioned in this subsection. Starting by bben et al. (2008) who presented the arguments of the academics and practitioners, as in favor of or against the usage of stochastic models in practice for analyzing the customer activities. They added a comparison between the quality of the results from these models and a simple heuristic algorithm. As a conclusion, they confirmed that the heuristic model performed at least as well as the stochastic model expect for the prediction of the future purchases. Hence, they recommended the usage of stochastic customer base analysis models at the end. In the same year, Ma et al. (2008) constructed a Markov decision framework to capture the relationship marketing.

Ekinci et al. (2014) designed a methodology to help managers in determining the optimal promotion campaigns that maximize CLV. For this, they utilized classification and regression tree (CART) and stochastic dynamic programming algorithms. In the same year, Zhang et al. (2014) built a Markov decision process model with multivariate interrelated state-dependent behavior. The uniqueness in their model lied in its ability to capture the effect of firm pricing decisions in customer purchasing behavior and consequently, in customer lifetime profitability in a Business to Business (B2B) market. Clempner et al. (2014) presented a method for finding local optimal policy for CLV, developed for a class of ergodic controllable finite Markov chains for a non-converging state value function. They validated their method by simulated credit-card marketing experiments.

Vaeztehrani et al. (2015) utilized stochastic dynamic programming model that was optimized using two deterministic linear programming algorithms. They concluded with set of insights including that the loyalty might decrease in short term net revenue. They also compared different loyalty programs. The most exiting finding is that their analytical long term evaluation of loyalty programs was capable of determining the most appropriate loyalty factors. In the same year, Klein et al. (2015) have had two contributions, initially, they studied how to balance the tradeoff between short term attainable revenues and long term customer relationship using Markov decision process model. Then they investigated the impact of limited capacity on CLV by introducing an opportunity cost-based approach that understood customer profitability as a customer's contribution to customer equity.

In the last two years, the researchers tried to add to what have been done as well. Taking Gilbert et al. (2017) as example. They proposed an offline algorithm that computed an optimal policy for the quantile criterion. Their algorithm could be applied both for finite and infinite horizons. As a future work they mentioned the aim to upgrade their algorithm to a reinforcement learning where the dynamics of the problem is unknown and has to be learned. Table 2 summarizes these contributions with their corresponding authors list and publication year.

Table 2
RESEARCHERS CONTRIBUTIONS IN MARKOV DECISION PROCESS

Authors	**Publication Title**
bben et al. (2008)	Instant customer base analysis: Managerial heuristics often get it right
Zhang et al. (2014)	Dynamic targeted pricing in B2B relationships
Clempner et al. (2014)	Simple computing of the customer lifetime value: A fixed local-optimal policy approach
Vaeztehrani et al. (2015)	Developing an integrated revenue management and customer relationship management approach in the hotel industry
Klein et al. (2015)	Maximizing customer equity subject to capacity constraints
Gilbert et al. (2017)	Optimizing Quantiles in Preference Based Markov Decision Processes

Approximate Dynamic Programming

Approximate Dynamic Programming (ADP), or Reinforcement Learning (RL), is considered as a complex Markov decision process, that is very powerful in solving large scale problems. Although ADP and RL are used interchangeably and present the same concept. ADP is more popular in the context of Operations Research (OR) while, RL is widely used in the context of computer science. It is considered as a branch of machine learning, that focuses on how the agent will perform in an environment when being in certain state, and has to take action to be able to move to another state while maximizing a commutative reward function. On another hand, reinforcement learning is very close to Markov Decision Process (MDP). However, the former don't assume knowledge about the mathematical model and is mainly used when having a model for the problem is infeasible. This is why machine learning algorithms (i.e. neural network) is used to approximate the solution function of the reinforcement learning. ADP is able to overcome the limitations of MDP including the curse of dimensionality, whether it is related to state space, action space or even outcome space (Powell et al. 2008). The solution of ADP is an approximate one, especially in the complex problems where finding exact or optimal solution is almost impossible.

The motivation behind utilizing ADP mentioned by Powell et al. (2018) who listed set of limitations for MDP including its inability to completely capture the dynamism in the relationship management, and suffers from the curse of dimensionality that arises from increasing the customer segments and the organization's actions related to each marketing strategy. They also noticed that MDP depends on the having a model and a transition probability matrix (TPM) for the problem at hand, and in some cases, TPM could not be constructed. Also, MDP needs a lot of feature engineering prior to the modeling phase. All the mentioned limitations of MDP motivated the researchers to utilize ADP to maximize CLV. Starting from Bertsimas et al. (2007) who presenting a Bayesian formulation for the exploration, exploitation tradeoff. They applied the multi-armed bandit problem in marketing context, where the decisions were made for batches of customers and the decisions may vary within each batch. Their proposal integrated Lagrangian decomposition-based ADP with a heuristic model based on a known asymptotic approximation to the multi-armed bandit. Their proposed model showed outperforming results.

On another hand, the banking sector attracted many researchers to enhance the bank and customer's relationship using different approximate dynamic programming approaches. For example, Chen et al. (2012) utilized dynamic programming to solve the problem of maximizing bank's profit. The optimization problem is formulated as a discrete, multi-stage decision process. The obtained solution is globally optimal and numerically stable. Mirrokni et al. (2016) explored a restricted family of dynamic auctions to check the possibility of implementing them in online fashion and without too much commitment from the seller in a space of single shot auctions that they called (bank account). While, James et al. (2016) developed a single server queuing model that determined a service policy to maximize the long term reward from serving the customers. Their model excluded the holding costs and penalties obtaining from the waited customers who left before receiving the service. Also, they estimated the bias function used in a dynamic programming recursion using ADP. Different than this, Ohno et al. (2016) proposed an approximate dynamic programming algorithm called simulated-based modified policy iteration, for large scale undiscounted Markov decision processes. Their algorithm overcame the curse of dimensionality when tested on set f numerical examples.

Leike et al. (2017) focused on illustrating the deep reinforcement learning environment components, and the properties of intelligent agents. They listed eight RL related problems including safe interruptibility, avoiding side effects, absent supervisor, reward gaming, safe exploration, as well as robustness to self-modification, distributional shift, and adversaries. These problems were categorized into two categories (i.e. robustness and specification problems) according to whether the performance function was corresponding to the observed reward function or not. They built two models to deal with these problems (i.e. A2C and Rainbow).

Arulkumaran et al. (2017) wrote a survey that covered central algorithms in deep reinforcement learning including deep Q-network, trust region policy optimization, and asynchronous advantage actor-critic. They also highlighted the unique advantages of neural networks and tried to focus on a visual understanding of RL. Geert et al. (2017), their application was medical, however their writing style is perfect and their demonstrative charts are awesome. Garcia et al. (2015) wrote a survey paper presenting a concept of safe reinforcement learning. They classified the papers in light of this concept to two approaches, optimization criterion and exploration process. They concluded by highlighting the effect of preventing the risk situations from the early steps in the learning process. Table 3 lists the mentioned contributions.

Table 3
RESEARCHERS CONTRIBUTIONS IN APPROXIMATE DYNAMIC PROGRAMMING

Authors	Publication Title
Simester et al. (2006)	Dynamic catalog mailing policies
Bertsimas et al. (2007)	A learning approach for interactive marketing to a customer segment
Chen et al. (2012)	Robust model-based fault diagnosis for dynamic systems
Garcia et al. (2015)	A comprehensive survey on safe reinforcement learning
James et al. (2016)	Developing effective service policies for multiclass queues with abandonment: asymptotic optimality and approximate policy improvement
Ohno et al. (2016)	New approximate dynamic programming algorithms for large-scale undiscounted Markov decision processes and their application to optimize a production and distribution system
Mirrokni et al. (2016)	Dynamic Auctions with Bank Accounts
Leike et al. (2017)	Ai safety gridworlds
Arulkumaran et al. (2017)	A brief survey of deep reinforcement learning

Deep Reinforcement Learning

Reinforcement learning utilizes Q learning as a model free algorithm, to maximize CLV, especially if the problem doesn’t have exact solution. In RL, Q-learning is trained using Neural networks algorithm, and an approximated solution is generated. In Deep reinforcement learning, Q value is approximated using deep learning model, instead of multi-layer perceptron neural network. Meanwhile, there are many advantages for using deep reinforcement learning. It overcomes the curse of dimensionality problem caused by the huge number of states, actions or outcome spaces; it also minimizes the complexity of the problem and save a lot of feature engineering steps. Finally, it approximates the Q value in case of absence of transition probability matrix with the help of Q-learning as a model free reinforcement learning technique.

PRACTICAL MODELS OF CLV

This section introduces the relationship between CLV and different marketing types, including B2B, B2C and C2C markets. It focuses on the practical contributions of CLV in each of these markets. The upcoming three sections present the practical applications in different business sectors, while the last section is devoted to relating dynamic programming approaches to marketing types (i.e. B2B, B2C and C2C) for maximizing CLV in each of these markets. Figure-3 demonstrates the number of publications for each market type. It might indicate the interest and competence of the researchers in the area of B2C, as it has the maximum number of publications. Meanwhile, very few researchers contributed in the area of C2B by proposing theatrical contributions. Figure 4 does not include the number of publications related to B2B2C. In fact, there are about six contributions in this area presented in this section. Meanwhile, adding B2B2C converts Figure 4 to 3x3 matrix instead of 2x2. Hence, it is left as future work to be analyzed in details alongside with its applications on CLV.

	Consumer	Business
Consumer	C2C **3**	C2B **2**
Business	B2C **35**	B2B **5**

FIGURE 4
NUMBER OF PUBLICATIONS FOR EACH MARKET TYPE

CLV in B2B Market

Business to Business (B2B) refers to any firm that sells its products or services to any other firm. Many researchers contributed in analyzing, studying B2B markets and related them to CLV. Cortez et al. (2019) conducted a study to predict the marketing capabilities of B2B in a short future period, i.e. from three to five years ahead. Meanwhile, tackling the situation only from supplier's perspective was one of their limitations. Nyadzayo et al. (2019) studied the effect of engagement on B2B marketing. Their study was a cross sectional one, and it would be more generalizable if it would be longitudinal. McCarthy et al. (2020) developed a conceptual framework to analyze customer based corporate valuation. Studying the effect of customer equity and CLV on the overall valuation and engagement of the customer. Their framework was for both B2B and B2C markets. Nenonen et al. (2016) developed a conceptual framework that analyzed the customer relationships in B2B markets. Depending on four research propositions. They had many findings including that firms should develop customer portfolio models and treated their customer upon. On top of their limitations was its small scale due to the nature of empirical scope. While, Horák et al. (2017) analyzed the role of CLV in B2B markets, focusing on Czech Republic in small and medium size companies (SMEs). Their major limitation was the lake of presenting CLV in practice using more concrete examples and demonstrative figures. Utami et al. (2019) tried to analyze and fill the gap in value co-creation concept moving from being conceptual to applying it a practical context. They relied upon vegetable market in their analysis in Indonesia.

Finally, Hallikainen et al. (2019) studied the relationship between customer big data analytics and customer relationship performance. They found that the former improved customer relationship performance in B2B market. Yet, their findings need to be generalized to be applied on different industries. Memarpour et al. (2019) tried to allocate the limited promotion budget and maximize the engagement of the customer using MDP. Their model was applied on real dataset meanwhile the results were not generalized to different application domains. ASLAN et al. (2018) utilized many algorithms to analyze CLV. Their model focus on B2B market and was applied on of the IT companies. Ekinci et al. (2014) designed a two-step methodology study that aimed to maximize CLV via determining optimal promotion campaigns; Based on stochastic dynamic programming and regression tree. Their model is applied in banking sector and they also tried to determine the states of the customers according to their values using CART technique. Also, in the same year, Ekinci et al. (2014) tackled the problem from another perspective. They developed a simple, industrial specific, easily measurable model with objective indicators to predict CLV. They injected the predicted CLV as states in Markov decision process model. Their proposed model is tested in the banking sector. One of the strengths of their model is that they conducted set of in-depth interviews to collect the most effective indicators for CLV. Chan et al. (2011) added a comprehensive decision support system to the contributions in the literature. They predicted customer purchasing behavior given set of factors including product, customer, and marketing influencing factors. Their model made use of the predicted customer purchasing behavior to estimate the customer's net CLV for a specific product.

Table 4
APPLICATIONS' AREAS OF B2B RESEARCH CONTRIBUTIONS

Authors List	Publication Title	Application Area	Source of Data	Dynamic Programming Model Type
Ekinci et al. (2014)	Analysis of customer lifetime value and marketing expenditure decisions through a Markovian-based model	Banking	Collected from one of the Banks	MDP
ASLAN et al. (2018)	Comparing Customer Segmentation With CLV Using Data Mining and Statistics: A Case Study	IT	Anonymous	MDP
Cortez et al. (2019)	Marketing role in B2B settings: evidence from advanced, emerging and developing markets	Economy	Data gathered in South America by the "Centro de Marketing Industrial", Universidad de Chile (CMI) and in the USA by the "Center for Business and Industrial Marketing (CBiM)" and the "Institute for the Study of Business Markets (ISBM)"	NA
Memarpour et al. (2019)	Dynamic allocation of promotional budgets based on maximizing customer equity	Telecom	Ario Payam Iranian Company	MDP

CLV in B2C Market

Ching et al. (2004) developed a stochastic dynamic programming model for both finite and in finite time horizons to optimize CLV. Their model is tested using practical data of computer service company. Wu et al. (2018) developed a framework to analyze the value co-creation of the customers and suppliers, taking the mobile hotel booking as an application. On top of their findings was relating CLV and customer Influencer Value (CIV). They found that CIV moderated CLV. While, Tarokh et al. (2017) contributed in the area of CLV by building two models. Initially, they group the customers based on their behavior and then utilized MDP to predict future behavior of the customers. In the same year, Tarokh et al. (2017) published another paper with a bit similar contribution that highlighted the effect on Stochastic approaches in the area of CLV. The third contribution for Tarokh et al. (2019) confirmed the effectiveness of MDP in CLV, by applying it on a medical industry. Yet, their research had some limitations including being conducted under certain conditions that might limit its applicability in real life. In their book chapter, Burelli et al. (2019) presented an overview of CLV modeling and prediction in different application fields; focusing on the free-to-play games. In fact, the most interesting part in their contribution is the ability of predicting revenue from a player that did not made any purchase yet.

Sekhar et al. (2019) studied set of CRM practices, reflecting them customer loyalty, using bivariate correlation. Their model that has been applied on telecom sector, proved the strong relationship and impact of CRM on shaping customer loyalty. In their book chapter, Bonacchi et al. (2019) analyzed the way of empowering decision making in light of performing customer behavior analytics. Their analysis is a comprehensive and might be cornerstone for effective research ideas. Altinay et al. (2019) reviewed and analyzed the recent studies in sharing economy that mainly focused on hospitality and tourism industries. The power of their analysis lied in the ability to introduce both theoretical and practical implications. Lin et al. (2019) utilized Spark MapReduce to proposed a large sum submatrix bi-clustering algorithm. The aim was to identify the most profitable customers, then segment those customer according to their purchasing behavior. Their model was applied on a real world telecom dataset based on the data consumption of the users. Haenlein et al. (2007) who developed a customer valuation model that combined first order Markov decision model with classification and regression tree (CART). The power of their model was in its ability to deal with both the discrete and continuous transitions. Their model was tested on a real life data from a leading German bank with a 6.2 million datasets. Labbi et al. (2007) proposed a comprehensive framework that combined customer equity with lifetime management. Their framework maximized ROI as it helped in the optimal planning and budgeting of targeted marketing campaigns. Their proposed model combined advanced Markov decision process models with Monte Carlo simulation, and portfolio optimization. Their model was tested on the Finnair case study. The contribution of Mannor et al. (2007) was a bit different. They proposed a model using a finite-state, finite-action, in finite-horizon and discounted rewards Markov decision process. Their model was tested on a large scale mailing catalog dataset. Wang et al. (2014) estimated the lifetime duration of the customers and their discounted expected transactions, using a simple and a standard discrete-time based transaction data. They also identified the relational and demographical factors that might cause the variance in customer's CLV. they concluded by set of insights to the marketing managers and decision makers.

Bravo et al. (2014) contributed in the area of marketing budget allocation for the sake of optimizing CLV, by introducing a decomposition algorithm that overcame the curse of dimensionality in stochastic dynamic programming problems. Wei et al. (2014) utilized iterative adaptive dynamic programming to establish a data-based iterative optimal learning control scheme, that is used for a discrete-time nonlinear systems. Their model is used to solve a coal gasification optimal tracking control problem. Neural networks were used to represent the dynamical process of coal gasification, coal quality and reference control. finally, iterative ADP was mainly used to obtain the optimal control laws for the transformed system. Jiang et al. has many contributes in this context. In (2015) they proposed Monotone-ADP, a provably convergent algorithm that exploited the value function monotonicity to increase the convergence rate. Their algorithm was applied on a finite horizon problem and show numerical results for three application domains including optimal stopping, energy storage (allocation), and glycemic control for diabetes patients. The same researchers and within the same year, published a paper that formulated the problem of the real-time placement of the battery storage operators while simultaneously accounting for the leftover energy value. Their algorithm exploited value function monotonicity to be able to find the revenue generating bidding policy. They also proposed a distribution free variant of the ADP algorithm. Their algorithm is tested on New York Independent System and recommended that a policy trained on historical real time price data using their proposed algorithm was indeed effective.

Keropyan et al. (2012) tried to differentiate the journey of the low-value and high-value customer within the firm. For this, they utilized MDP and heuristic models, and applied them on real life data. One of their limitations were in their assumption for defining high value customers as the ones who spend long time in one purchase. Meanwhile, it might not usually the case, but the ones who had more purchases. While, Cheng et al. (2012) developed a framework consisted of three groups of techniques to compute CLV based on its predicted value, identified its critical variables and finally predicted the profit of the customers under different purchasing behavior using ANN. Their model is tested on a dataset related to a company in Taiwan. Hwang et al. (2016) proposed a method to calculate CLV dynamically for the sake of adopting personalized CRM activities. They also applied data mining techniques to predict CLV and their model is tested on a wireless Telecom industry in Korea. Zhang et al. (2018) designed a prediction probabilistic model to measure the lifetime profitability of customers. They were interested in the customers whose purchasing behavior followed purchasing cycles. Their model is measured using the inter purchase time of the customers and was assumed to follow Poisson distribution. They also measured the customer lifetime profitability based on a proposed a customer’s probability scoring model. Their model was applied on dataset for 529 customers from catalog firm and showed outperforming results. While, Jasek et al. (2019) also contributed in the area of CLV, but with a predictive model using MDP, and many other models. Although their model has been applied on large amount and real dataset, it is still limited to the prediction task not to upgrade it to a maximization task. Besides some limitations in their assumptions for to transition probability matrix of MDP model.

Jasek et al. (2019) conducted a comparison between eleven models based on set of online stores datasets. All of the proposed models were probabilistic ones that achieved outperforming results. Although their work is perfectly presented and their proposed models had outperforming results; these models could not capture the seasonal purchasing behavior of the customers. Also Dahana et al. (2019) were interested in the online marketing. They studied the effect of lifetime on CLV through a segmentation model on online fashion retailer. On top of the main limitations of their interesting study was being limited on fashion products only and not generalized. Dachyar et al. (2019) also performed a segmentation task to measure CLV. Their model aimed to identify the level of loyalty of the online customer based on CLV. Meanwhile, their proposed model lakes the generalizability as well. Theocharous et al. (2013) who explored marketing recommendations through reinforcement learning and Markov decision process, and evaluated their performance through an offline evaluation method using a well-crafted simulator. They tested their proposal on several real world datasets from automotive and banking industry. Also, Theocharous et al. (2015), built a comprehensive framework to utilize reinforcement learning with off-policy techniques to optimize CLV for personalized ads recommendation systems. They compared the performance of life time value matric to the click through rate for evaluating the performance of personalized ads recommendation systems. Their framework was tested on a real data and proved its outperformance. While, Silver et al. (2013) proposed a framework for concurrent reinforcement learning using temporal difference learning, that captured the parallel interaction between the company and its customers. They tested their proposal on large scale dataset for online and email interactions. Tripathi et al. (2018) utilized reinforcement learning in a bit different way. They proposed a model that combined RL with Recurrent Neural Network (RNN) for personalized video recommendation. Their model that has been tested on real user segments for a month, showed outperforming results. Their interesting model was well presented, but combining audio and text of the customer’s feedback is still needed. Barto et al. (2017) stated five applications for reinforcement learning, including personalized web services. This included recommending the best content for each particular user based on his profile interests inferred from his history of online activity.

Deep reinforcement learning is a rich research area, where many researchers competed in utilizing it for the sake of maximizing CLV. Their models are being built on different industries, including banking, retail, direct mailing campaigns, and many more. However, one of the most significant ways of interaction between the firm and its customers in direct marketing is through mailing campaigns. It attracted many researchers and motivated them to analyze the dynamic implications of mailing decisions. Li et al. (2015) focused on deep learning and built a deep reinforcement learning model (i.e. RNN and LSTM) on a partially observable state space and discrete actions. Their model has been applied on a KDD Cup 1998 mailing donation dataset. While, Tkachenko et al. (2015) proposed a framework that captured the autonomous control for customer relationship management system. In their model, they utilized Q learning to train deep neural network, assuming two assumptions. First was that the customer’s states represented by recency, frequency, and monetary values. Second, the actions are both discrete and continuous. They assumed that the CLV of each customer was represented by the estimated value function. Their model was run on a KDD Cup 1998 mailing donation dataset. Shaohui et al. (2016) built an approach that aimed to optimize the mailing decisions by maximizing the customer’s CLV. Their proposal was a two-step approach started from a non-homogenous MDP that captured the dynamics of customers, mailings interactions, then determined the optimal mailing decisions upon using partially observable MDP.

On another hand, Li et al. (2017) summarized the achievements of deep reinforcement learning. They analyze its core elements, mechanisms and applications, including its applicability in Business management in different industries including ads, recommendation, marketing, finance and health care. They also mentioned topics that were not reviewed until the time they wrote their manuscript and listed set of resources and tutorials for deep reinforcement learning. Meanwhile, Lang et al. (2017) focused on e-commerce and built a model that understood the customer behavior in e-commerce using recurrent neural networks. The power of their model was in its ability to capture the customers’ actions sequences; hence, it overcame the Logistic Regression model as a traditional vector based method. Also, Zhao et al. (2018) focused on e-commerce transactions as well; however, they combined Markov decision process with unbounded action space with deep learning, to build a deep reinforcement learning platform that overcame the limitations of the other fraud detection mechanisms. Their model not only maximized the e-commerce profit but also reduced the fraudulent behavior. It was applied on a real world dataset. The main drawback of their model was that it applied only in e-commerce dataset not generalized or applied on many industries to test its robustness. Chen et al. (2018) applied DQN on video games industry, hence tackled CLV from a different perspective. These researchers recommended Convolutional Neural Network (CNN) as the most efficient among all neural network structures. It can predict the economic value of the individual players, and also best suited the nature of video games’ large datasets.

Although deep reinforcement learning has many significant contributions in the area of direct marketing to maximize CLV. It has many limitations; on top of them is that might overestimate the action values and hence, generates unrealistic actions, as it always shows the action that maximizes Q values, as mentioned in Algorithm-5 and demonstrated in Eq. (3); where Q represents the CLV, r is the rewards, γ is the discount factor, s is the next states and, a is the next actions that maximize Q values. These limitations motivated few researchers to develop modified versions of DQN, on top of them was (double deep reinforcement learning). DDQN overcame the disadvantages of DQN by generating reliable and robust actions values. DDQN has been designed to have two decoupled (i.e. separate) networks, one for the selection of the optimal action that maximizes Q and one for the evaluation of this selected action. The decoupling process helps DDQN to generate reliable actions. Eq. (4) demonstrates DDQN model, where θ represents the weights of the first network and θ represents the weights of the second network (Hasselt et al. 2010).

$$Q(s,a) = r + \gamma \max(Q(s',a',\theta) \qquad (3)$$

$$Q(s,a) = r + \ \gamma Q(s', argmax_a((Q(s',a';\ \theta), \theta') \qquad (4)$$

As mentioned in Table 5, Hasselt et al. (2010), and Hasselt et al. (2016) developed many algorithms that utilized DDQN. Meanwhile, their papers are theoretical and lake the applicability, as these only contained the theory of the model and without being applied on real life case studies. While, Kalashnikov et al. (2018) proposed a scalable reinforcement learning approach. Their model was applied on robotic manipulations. However, it still far from the direct marketing area. Hence, applying double deep reinforcement learning in the context of direct marketing to maximize CLV still unreached research area. However, empirical studies proved some limitations for DDQN and this what was mentioned by (Hessel et al. 2017) who proposed a model called *"Rainbow*". It combined six extensions of DQN and applied on Atari 2600 benchmark. The main drawback or their model is it was not applied on marketing area, although it was proved to be very robust and reliable. Sabatelli et al. (2018) contributed to the literature by introducing deep quality value (DQV), that used a model called (value neural network) to estimate the temporal-difference errors. The later are used by a second quality network for directly learning the state-action values. DQV model proved its effectiveness in four games of Atari Arcade Learning. DQV was not applied on the context of direct marketing yet.

Table 5
APPLICATIONS' AREAS OF B2C RESEARCH CONTRIBUTIONS

Authors List	Publication Title	Application Area	Source of Data	Dynamic Programming Model Type
Ching et al. (2004)	Customer lifetime value: stochastic optimization approach	IT	Computer Service Company	MDP
Haenlein et al. (2007)	A model to determine customer lifetime value in a retail banking context	Banking	German Bank	MDP
Labbi et al. (2007)	Customer Equity and Lifetime Management (CELM)	Tourism	Finnair's frequent-flyer program	MDP

Mannor et al. (2007)	Bias and variance approximation in value function estimates	Mailing Catalog	Mail-Order Catalog Firm	MDP
Chan et al. (2011)	A dynamic decision support system to predict the value of customer for new product development	Electrical	Anonymous	MDP
Keropyan et al. (2012)	Customer loyalty programs to sustain consumer fidelity in mobile telecommunication market	Telecom	Mobile Operator	MDP
Cheng et al. (2012)	Customer lifetime value prediction by a Markov chain based data mining model: Application to an auto repair and maintenance company in Taiwan	Automotive	Auto Repair and Maintenance Company	MDP
Silver et al. (2013)	Concurrent reinforcement learning from customer interactions	Mailing	Anonymous	Reinforcement Learning
Theocharous et al. (2013)	Lifetime value marketing using reinforcement learning	Automotive and Banking	Anonymous	Reinforcement Learning
Wang et al. (2014)	The antecedents of customer lifetime duration and discounted expected transactions: Discrete-time based transaction data analysis	Telecom	Italian cellphone network firm	MDP
Bravo et al. (2014)	Valuing customer portfolios with endogenous mass and direct marketing interventions using a stochastic dynamic programming decomposition	Manufacturer	Manufacturing Company	ADP
Wei et al. (2014)	Adaptive dynamic programming for optimal tracking control of unknown nonlinear systems with application to coal gasification	Oil and Gas	Collected by real-world industrial processes	ADP
Jiang et al. (2015)	An Approximate Dynamic Programming Algorithm for Monotone Value Functions	Healthcare	Diabetes patients	Monotone ADP
Jiang et al. (2015)	Optimal hour-ahead bidding in the real-time electricity market with battery storage using approximate dynamic	Electricity	New York Independent System Operator	ADP

	programming			
Li et al. (2015)	Recurrent reinforcement learning: a hybrid approach	Mailing	KDD1998	Deep Reinforcement Learning
Theocharous et al. (2015)	Personalized Ad Recommendation Systems for Life-Time Value Optimization with Guarantees	Banking	Anonymous	Reinforcement Learning
Tkachenko et al. (2015)	Autonomous CRM control via CLV approximation with deep reinforcement learning in discrete and continuous action space	Mailing	KDD1998	Deep Reinforcement Learning
Hwang et al. (2016)	A Stochastic Approach for Valuing Customers: A Case Study	Telecom	Telecom Wireless Company, Korea	MDP
Tkachenko et al. (2016)	Customer simulation for direct marketing experiments	Mailing	KDD1998	Deep Reinforcement Learning
Hasselt et al. (2016)	Deep Reinforcement Learning with Double Q-Learning	Gaming	Atari 2600 games	Deep Reinforcement Learning
Shaohui et al. (2016)	A nonhomogeneous hidden Markov model of response dynamics and mailing optimization in direct marketing	Mailing	KDD1998	Deep Reinforcement Learning
Li et al. (2017)	Deep reinforcement learning: An overview	Mailing	KDD1998	Deep Reinforcement Learning
Hessel et al. (2017)	Rainbow: Combining improvements in deep reinforcement learning	Gaming	Atari 2600 games	Deep Reinforcement Learning
Lang et al. (2017)	Understanding consumer behavior with recurrent neural networks	E-Commerce	Zalando	Deep Reinforcement Learning
Tarokh et al. (2017)	A new model to speculate CLV based on Markov chain model	Manufacturing	Manufacturing company in Iran	MDP
Tripathi et al. (2018)	A reinforcement learning and recurrent neural network based dynamic user modeling system	E-Commerce	LIRIS-ACCEDE and Proprietary Public Datasets	Reinforcement Learning
Zhang et al. (2018)	Assessing lifetime profitability of customers with purchasing cycles	Manufacturing	Catalog Firm	MDP
Zhao et al. (2018)	Impression Allocation for Combating Fraud in E-commerce Via Deep Reinforcement Learning with Action Norm Penalty	E-Commerce	Anonymous	Deep Reinforcement Learning

Chen et al. (2018)	Customer Lifetime Value in Video Games Using Deep Learning and Parametric Models	Gaming	“Age of Ishtaria” Mobile Game	Deep Reinforcement Learning
Sabatelli et al. (2018)	Deep Quality-Value (DQV) Learning	Gaming	Atari 2600 games	Deep Reinforcement Learning
Jasek et al. (2019)	Predictive Performance of Customer Lifetime Value Models in E-commerce and the Use of Non-Financial Data	E-Commerce	Online Stores	MDP
Jasek et al. (2019)	Comparative analysis of selected probabilistic customer lifetime value models in online shopping	Retail	Online Store	ADP
Dahana et al. (2019)	Linking lifestyle to customer lifetime value: An exploratory study in an online fashion retail market	Retail	Online Store	ADP
Dachyar et al. (2019)	Loyalty Improvement of Indonesian Local Brand Fashion Customer Based on Customer Lifetime Value (CLV) Segmentation	E-Commerce	Local Brand Fashion	ADP
Tarokh et al. (2019)	Modeling patient's value using a stochastic approach: An empirical study in the medical industry	Medical	Dental clinic in Tehran	MDP

CLV in C2C Market

This section is devoted to list the contributions of dynamic programming in C2C market, to maximize CLV. Liang et al. (2011) introduced a social commerce concept by merging data from Facebook social media to e-commerce. They divide their social commerce concept to online, offline and mobile social commerce. In their thesis, Meire et al. (2018) presented the social media usefulness from a marketing perspective. They relied upon social media datasets for commerce and web logs. Rachbini et al. (2019) tried to relate brand equity, and value equity to relationship equity and to reflect this on customer loyalty. Their study was effective, yet, the generalization of their analysis on other areas and industries was one of its limitations. Rihova et al. (2019) explored and evaluated the practice-based segmentation and compared it to the conceptual segmentation. Focusing on C2C markets. Their major limitation was limiting their analysis to certain conceptual settings, hence, limits its generalization. The list of contributions in C2C market is stated in Table 6.

Table 6
APPLICATIONS' AREAS OF C2C RESEARCH CONTRIBUTIONS

Authors List	Publication Title	Application Area	Source of Data
Liang et al. (2011)	Social Commerce: A New Electronic Commerce	Commerce	Social Media
Meire et al. (2018)	A Marketing Perspective on Social Media Usefulness	Commerce	Different sources for commercial and web data
Rachbini et al. (2019)	Determinants of Trust and Customer Loyalty on C2C E-marketplace in Indonesia	Commerce	E-marketplace in Indonesia

CLV in C2B

This section presents the contribution of the researchers in the area of C2B market. Meanwhile, these contributions are noticed to be very few compared to the contributions in other mentioned marketing types. Hence, this section presents the major contributions in both theoretical and practical aspects, because there is almost no contribution in the area of dynamic programming in these two market types. Hence, this section presents the published contributions regardless whether these belong to dynamic programming or not. As mentioned in Section-1, C2B assumes the user has the power and ability to initiate and lead the transaction process between him and the firm. O'Hern et al. (2013) analyzed C2B market type through a term of user generated content (UGC). They highlighted the various types of UGC, benefits and challenges generated by these types. They also projected the implications of UCG on marketing decisions. Holm et al. (2012) analysis CLV in light of customer probability analysis. Also, they implemented some models for determining customer profitability and build a framework for customer profitability and guide the managers who need to implement it in their firms. Yet, these contributions are still theoretical in nature as summarized in Table 7. Meanwhile, up to the knowledge of the researchers of this manuscript and according to their conducted research, it is even hard to find recent contributions this area of research.

Table 7
APPLICATIONS' AREAS OF C2B RESEARCH CONTRIBUTIONS

Authors List	Publication Title
Holm et al. (2012)	Measuring customer profitability in complex environments: an interdisciplinary contingency framework
O'Hern et al. (2013)	The empowered customer: User-generated content and the future of marketing

SUMMARY AND DISCUSSION

This section is devoted to discuss and summarize the contribution of dynamic programming models in different market types, including B2B, B2C and C2C, for the sake of maximizing CLV. These contributions were classified to theoretical and practical ones. Listing these models happened in a way such that each of these algorithms built on its previous one and tried to overcome its drawbacks. Starting from MDP as a simple and easy to implement algorithm that has many limitations including its dependency of the existence of TPM, the necessity of having a model for the problem at hand. Besides that, it suffers from the curse of dimensionality. The evolution of MDP is ADP algorithm that overcame the former's drawbacks and could find an approximate solution for a given problem instead of an exact one; but it could not look ahead and this restricted its ability to learn. Bunch of researchers utilized DQN for solving the mentioned issue of APD. Meanwhile, DQN is proved to overcome the action values but generated unrealistic results, although it had got higher convergence than the traditional Q learning algorithms. The overestimation of the actions' values mentioned in case of DQN, was treated after proposing DDQN. The latter utilized two separate networks, one of them for the action selection and the other was for its evaluation. Meanwhile, it is proved to overestimate the action values as well in some cases and consequently, to generate inaccurate customer value. Recently, many researchers utilized DQV and rainbow for many applications, but it was not applied to the area of maximizing CLV up until this moment.

CONCLUSION AND FUTURE WORK

This paper presented a review and analysis for the theoretical and practical contributions of dynamic programming models, in different market types, including B2B, B2C, C2B, C2C and briefly B2B2C, to maximize CLV. It started from MDP as a basic dynamic programming model that had many contributions in the different market types, especially B2C. The paper listed its limitations, including, its dependency on having a model formulated for the problem, the necessity of having a transition probability matrix, and its suffer from the curse of dimensionality, especially when was applied on complex problems with large datasets. These limitations encouraged the researchers to utilize ADP (or RL). The paper also presented the effectiveness of Q learning algorithm in generating outperforming results as a model free algorithm. It was capable of finding approximate solution for the problem that might not had an exact or optimal one, especially when the learning process for its values occurred by neural network or deep learning algorithms. Although deep reinforcement learning model outperformed many other algorithms, it had a main drawback of overestimating the action values, hence generating unrealistic results. Meanwhile, double deep reinforcement learning proved to overcome these issues and generated robust actions, but this might not always be the case. The paper presented the effectiveness of many other algorithms that outperformed DDQN, including Rainbow, and DQV models. However, none of the latter models was applied on the area of direct marketing. This area of research is still very rich and there are many gaps for future research directions. For instance, DQV and Rainbow models, might be applied and tested on different application areas. Also, other algorithms might be integrated with Q learning to help it in generating more robust results. Another research direction is to focus on other environmental factors that contribute in CLV including governmental regulations, the competitors' decisions, the changes in effect of the price changes, and stock market. These factors might be analyzed and taken into consideration when determining the CLV for each customer. Also, other engagement indicators might be integrated with CLV, including customer referral value and customer influencer value. Finally, more interest in analyzing CLV in the market type of C2B and B2B2C might be given.

REFERENCES

AboElHamd, Eman, Hamed M. Shamma, & Mohamed Saleh (2019). Dynamic Programming Models for Maximizing Customer Lifetime Value: An Overview. Proceedings of SAI Intelligent Systems Conference. Springer, Cham, 419-445.

Agrawal, M.L. (2003). Customer relationship management (CRM) and corporate renaissance. *Journal of Services Research, 3*(2), 149.

Ahmadalinejad, Mehrpooya, & Seyed Navid Nabavi (2016). Social CRM: A solution for realization of virtual banking. *International Journal of Mechatronics Electrical and Computer Technology, 6*(22), 3134-3141.

Altinay, Levent, & Babak Taheri (2019). Emerging themes and theories in the sharing economy: a critical note for hospitality and tourism. *International Journal of Contemporary Hospitality Management*, *1*(31), 180-193.

Arulkumaran, Kai, et al. (2017). A brief survey of deep reinforcement learning. arXiv preprint arXiv:1708.05866

ASLAN, Damla, & Metehan TOLON (2018). Comparing Customer Segmentation with CLV Using Data Mining and Statistics: A Case Study.

Baidokhti, Farhoodeh Zamani (2019). An Exploration of how Social Media Data is Used in Companies: Evidence from the Telecom Industry. Diss. Trinity College Dublin.

Barto, Andrew G., P.S. Thomas, & Richard S. Sutton (2017). Some Recent Applications of Reinforcement Learning. Proceedings of the Eighteenth Yale Workshop on Adaptive and Learning Systems.

Bertsimas, Dimitris, and Adam J. Mersereau (2007). A learning approach for interactive marketing to a customer segment. Operations Research 55(6), 1120-1135.

Bonacchi, Massimiliano, and Paolo Perego (2019). Customer Analytics for Internal Decision-Making and Control. Customer Accounting. Springer, Cham. 37-66.

Bulsara, M., and Thakkar, H. (2015). Customer Feedback-based Product Improvement: A Case Study. Productivity, 56(1), 107.

Burelli, Paolo (2019). Predicting Customer Lifetime Value in Free-to-Play Games. Data Analytics Applications in Gaming and Entertainment, 11-79.

Cannon, J. N., and Cannon, H. M. (2014). Modeling Strategic Opportunities in Product-Mix Strategy: A Customer-Versus Product-Oriented Perspective. Developments in Business Simulation and Experiential Learning, (35).

Chan, S. L., and W. H. Ip (2011). A dynamic decision support system to predict the value of customer for new product development. Decision Support Systems 52(1), 178-188.

Chen, Jie, and Ron J. Patton (2012). Robust model-based fault diagnosis for dynamic systems. Springer Science and Business Media, (3).

Chen, Pei Pei, et al (2018). Customer Lifetime Value in Video Games Using Deep Learning and Parametric Models. arXiv preprint arXiv:1811.12799.

Cheng, C-J., et al (2012). Customer lifetime value prediction by a Markov chain based data mining model: Application to an auto repair and maintenance company in Taiwan. Scientifica Iranica19(3), 849-855.

Ching, Wai-Ki, et al (2004). Customer lifetime value: stochastic optimization approach. Journal of the Operational Research Society 55(8), 860-868.

Clempner, Julio B., and Alexander S. Poznyak (2014). Simple computing of the customer lifetime value: A fixed local-optimal policy approach. Journal of Systems Science and Systems Engineering 23(4), 439-459.

Cortez, Roberto Mora, and Wesley J. Johnston (2019). Marketing role in B2B settings: evidence from advanced, emerging and developing markets. Journal of Business & Industrial Marketing 3(1), 36-45.

Dachyar, M., F. M. Esperanca, and R. Nurcahyo (2019). Loyalty Improvement of Indonesian Local Brand Fashion Customer Based on Customer Lifetime Value (CLV) Segmentation. IOP Conference Series: Materials Science and Engineering. IOP Publishing. 598(1).

Dahana, Wirawan Dony, Yukihiro Miwa, and Makoto Morisada (2019). Linking lifestyle to customer lifetime value: An exploratory study in an online fashion retail market. Journal of Business Research 99, 319-331.

Ekinci, Yeliz, et al (2014). Analysis of customer lifetime value and marketing expenditure decisions through a Markovian-based model. European Journal of Operational Research 237(1), 278-288.

Ertz, Myriam, et al (2019). Made to break? A taxonomy of business models on product lifetime extension. Journal of Cleaner Production 598(1), 867-880.

Esteban-Bravo, Mercedes, Jose M. Vidal-Sanz, and Goekhan Yildirim (2014). Valuing customer portfolios with endogenous mass and direct marketing interventions using a stochastic dynamic programming decomposition. Marketing Science 33(5), 621-640.

Ford, Caroline (2013). Smartphone apps on the mobile web: an exploratory case study of business models. Third Annual International Conference on Engaged Management Scholarship, Atlanta, Georgia, 3(1).

Fotiadis, Anestis K., and Chris Vassiliadis (2017). Being customer-centric through CRM metrics in the B2B market: the case of maritime shipping. Journal of Business & Industrial Marketing, 32(3), 347-356.

Frow, Pennie, et al (2015). Managing co-creation design: A strategic approach to innovation. British Journal of Management 26(3), 463-483.

Garcia, Javier, and Fernando FernÆndez (2015). A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 16(1), 1437-1480.

Gilbert, Hugo, Paul Weng, and Yan Xu (2017). Optimizing Quantiles in Preference Based Markov Decision Processes. AAAI.

Haenlein, Michael, Andreas M. Kaplan, and Anemone J. Beeser, (2007). A model to determine customer lifetime value in a retail banking context. European Management Journal 25(3), 221-234.

Hallikainen, Heli, Emma Savimäki, and Tommi Laukkanen (2019). Fostering B2B sales with customer big data analytics. Industrial Marketing Management.

Hasselt, Hado V (2010). Double Q-learning. In Advances in Neural Information Processing Systems, 2613-2621.

Hessel, Matteo, et al (2017). Rainbow: Combining improvements in deep reinforcement learning. arXiv preprint arXiv:1710.02298.

Holm, Morten, V. Kumar, and Carsten Rohde (2012). Measuring customer profitability in complex environments: an interdisciplinary contingency framework. Journal of the Academy of Marketing Science 40(3), 387-401.

Horák, Pavel (2017). Customer Lifetime Value in B2B Markets: Theory and Practice in the Czech Republic. Int. J. Bus. Manag 12(2), 47-55.

Hwang, H (2016). A Stochastic Approach for Valuing Customers: A Case Study. Int. J. Softw. Eng. Its Appl 10(3), 67-82.

James, Terry, Kevin Glazebrook, and Kyle Lin (2016). Developing e ective service policies for multiclass queues with abandonment: asymptotic optimality and approximate policy improvement. INFORMS Journal on Computing 28(2), 251-264.

Jasek, Pavel, et al (2019). Comparative analysis of selected probabilistic customer lifetime value models in online shopping. Journal of Business Economics and Management 20(3): 398-423.

Jasek, Pavel, et al (2019). Predictive Performance of Customer Lifetime Value Models in E-Commerce and the Use of Non-Financial Data. Prague Economic Papers, (6), 648-669.

Jiang, Daniel R., and Warren B. Powell (2015). An approximate dynamic programming algorithm for monotone value functions. Operations Research, 63(6), 1489-1511.

Jiang, Daniel R., and Warren B. Powell (2015). Optimal hour-ahead bidding in the real-time electricity market with battery storage using approximate dynamic programming. INFORMS Journal on Computing 27(3), 525-543.

Kaelbling, Leslie Pack, Michael L. Littman, and Andrew W. Moore (1996). Reinforcement learning: A survey. Journal of artificial intelligence research (4), 237-285.

Kakalejčík, Lukáš, Jozef Bucko, and Martin Vejačka (2019). Differences in buyer journey between high-and low-value customers of e-commerce business. Journal of Theoretical and Applied Electronic Commerce Research, 14(2), 47-58.

Kalashnikov, Dmitry, et al (2018). Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. arXiv preprint arXiv:1806.10293.

Keropyan, Aras, and Ana Maria Gil-Lafuente (2012). Customer loyalty programs to sustain consumer fidelity in mobile telecommunication market. Expert Systems with Applications, 39(12), 11269-11275.

Klein, Robert, and Johannes Kolb (2015). Maximizing customer equity subject to capacity constraints. Omega (55), 111-125.

Kumar, V., Girish Ramani, and Timothy Bohling (2004). Customer lifetime value approaches and best practice applications. Journal of Interactive marketing 18(3), 60-72.

Labbi, Abderrahim, et al (2007). Customer Equity and Lifetime Management (CELM). Marketing Science, 26(4), 553-565.

Lang, Tobias, and Matthias Rettenmeier (2017). Understanding consumer behavior with recurrent neural networks. International Workshop on Machine Learning Methods for Recommender Systems.

Lau, Eric Kin Wai, Abel Zhao, and Anthony Ko (2019). Hybrid Artificial Intelligence B2B2C Business Application–Online Travel Services. International Conference on Knowledge Management in Organizations. Springer, Cham, 515-524.

LEI, Shen-hong, et al (2017). B2B2C Platform of National Featured Agricultural Products Design Based on iOS. DEStech Transactions on Computer Science and Engineering itme.

Leike, Jan, et al (2017). Ai safety gridworlds. arXiv preprint arXiv:1711.09883.

Li, Xiujun, et al (2015). Recurrent reinforcement learning: a hybrid approach. arXiv preprint arXiv:1509.03044.

Li, Yuxi (2017). Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274.

Liang, Ting-Peng, and Efraim Turban (2011). "Introduction to the special issue social commerce: a research framework for social commerce." International Journal of electronic commerce 16(2), 5-14.

Lin, Qin, et al (2019). A Novel Parallel Biclustering Approach and Its Application to Identify and Segment Highly Profitable Telecom Customers. IEEE Access (7), 28696-28711.

Litjens, Geert, et al (2017). A survey on deep learning in medical image analysis. Medical image analysis (42), 60-88.

Liu, Derong, Ding Wang, and H. Ichibushi (2012). Adaptive dynamic programming and reinforcement learning. UNESCO Encyclopedia of Life Support Systems.

Ma, Ming, Zehui Li, and Jinyuan Chen (2008). Phase-type distribution of customer relationship with Markovian response and marketing expenditure decision on the customer lifetime value. European Journal of Operational Research 187(1), 313-326.

Ma, Shaohui, et al (2016). A nonhomogeneous hidden Markov model of response dynamics and mailing optimization in direct marketing. European Journal of Operational Research 253(2), 514-523.

Mannor, Shie, et al (2007). Bias and variance approximation in value function estimates. Management Science 53(2), 308-322.

McCarthy, Daniel, and Fernando Pereda (2020). Assessing the Role of Customer Equity in Corporate Valuation: A Review and a Path Forward. Available at SSRN.

Meire, Matthijs (2018). A Marketing perspective on social media usefulness. Diss. Ghent University.

Memarpour, Mehrdad, et al (2019). Dynamic allocation of promotional budgets based on maximizing customer equity." Operational Research, 1-25.

Mirrokni, Vahab S., et al (2016). Dynamic Auctions with Bank Accounts. IJCAI, 16(1), 387-393

Myllymäki, Markus (2019). A Dynamic Allocation Mechanism for Network Slicing as-a-Service.

Nenonen, Suvi, and Kaj Storbacka, (2016). Driving shareholder value with customer asset management: Moving beyond customer lifetime value. Industrial Marketing Management 52(1), 140-150.

Nyadzayo, Munyaradzi W., Riza Casidy, and Park Thaichon (2019). B2B purchase engagement: Examining the key drivers and outcomes in professional services. Industrial Marketing Management.

O'Hern, Matthew S., and Lynn R. Kahle, (2013). The empowered customer: User-generated content and the future of marketing. Global Economics and Management Review, 18(1), 22-30.

Ohno, Katsuhisa, et al (2016). New approximate dynamic programming algorithms for large-scale undiscounted Markov decision processes and their application to optimize a production and distribution system. European Journal of Operational Research, 249(1), 22-31.

Permana, Dony, Udjianna S. Pasaribu, and Sapto W. Indratno (2017). Classification of customer lifetime value models using Markov chain. Journal of Physics: Conference Series. IOP Publishing, 893(1)

Polychronakis, Yiannis, and Xiang Li (2008). Barriers to international supply chain collaboration for small Chinese costume jewellery suppliers. The Cyprus Journal of Sciences, 6(1), 181-201.

Powell, Warren B (2008). Approximate dynamic programming: lessons from the field. Winter Simulation Conference. IEEE, 205-214.

Powell, Warren B (2009). What you should know about approximate dynamic programming. Naval Research Logistics (NRL), 56(3), 239-249.

Rachbini, Widarto, Iha Haryani Hatta, and Tiolina Evi (2019). Determinants of Trust and Customer Loyalty on C2C E-marketplace in Indonesia. International Journal of Civil Engineering and Technology, 10(3).

Rihova, Ivana, et al (2019). Practice-based segmentation: taxonomy of C2C co-creation practice segments. International Journal of Contemporary Hospitality Management., 31(9), 3799-3818.

Sabatelli, Matthia, et al (2018). Deep Quality-Value (DQV) Learning. arXiv preprint arXiv:1810.00368.

Sekhar, K. Chandra, and P. Malyadri (2019). An Empirical Study on CRM Practices and Customer Loyalty with Reference to BSNL. Journal of Commerce 7(4), 13-19.

Silver, David, et al (2013). Concurrent reinforcement learning from customer interactions. International Conference on Machine Learning, 924-932.

Simester, Duncan (2017). Field experiments in marketing. Handbook of Economic Field Experiments. North-Holland 1(1), 465-497.

Simester, Duncan I., Peng Sun, and John N. Tsitsiklis (2006). Dynamic catalog mailing policies. Management science 52(5), 683-696.

Tarokh, Mohammad Jafar, and Mahsa EsmaeiliGookeh (2019). Modeling patient's value using a stochastic approach: An empirical study in the medical industry. Computer methods and programs in biomedicine, 176(1), 51-59.

Tarokh, MohammadJafar, and Mahsa EsmaeiliGookeh (2017). A new model to speculate CLV based on Markov chain model. Journal of Industrial Engineering and Management Studies 4(2), 85-102.

Tarokh, Mohammadjafar, and Mahsa EsmaeiliGookeh (2017). A Stochastic Approach for Valuing Customers. International Journal of Information and Communication Technology Research 9(3), 59-66.

Theocharous, Georgios, and Assaf Hallak (2013). Lifetime value marketing using reinforcement learning. RLDM, 19(1).

Theocharous, Georgios, Philip S. Thomas, and Mohammad Ghavamzadeh (2015). Personalized Ad Recommendation Systems for Life-Time Value Optimization with Guarantees. IJCAI.

Tkachenko, Yegor (2015). Autonomous CRM control via CLV approximation with deep reinforcement learning in discrete and continuous action space. arXiv preprint arXiv:1504.01840.

Tkachenko, Yegor, Mykel J. Kochenderfer, and Krzysztof Kluza (2016). Customer simulation for direct marketing experiments. Data Science and Advanced Analytics (DSAA), IEEE International Conference on. IEEE, 478-487.

Tripathi, Abhishek, T. S. Ashwin, and Ram Mohana Reddy Guddeti (2018). A reinforcement learning and recurrent neural network based dynamic user modeling system. IEEE 18th International Conference on Advanced Learning Technologies (ICALT). IEEE, 411-415.

Utami, Hesty Nurul, Eleftherios Alamanos, and Sharron Kuznesof (2019). How Have Things Changed? Value Co-Creation Reinvents Agribusiness–a Multiple B2B Stakeholder Perspective. PROCEEDINGS BOOK, 28(1).

Vaeztehrani, Amirhossein, Mohammad Modarres, and Samin Aref (2015). Developing an integrated revenue management and customer relationship management approach in the hotel industry. Journal of Revenue and Pricing Management 14(2), 97-119.

Van Hasselt, Hado, Arthur Guez, and David Silver (2016). Deep Reinforcement Learning with Double Q-Learning. AAAI. 2(1).

Van Otterlo, Martijn (2009). Markov Decision Processes: Concepts and Algorithms. Course on ‘Learning and Reasoning.

W bben, Markus, and Florian V. Wangenheim (2008). "Instant customer base analysis: Managerial heuristics often get it right." Journal of Marketing, 72(3): 82-93.

Wang, Chao, and Ilaria Dalla Pozza (2014). The antecedents of customer lifetime duration and discounted expected transactions: Discrete-time based transaction data analysis, 2014-203.

Wei, Qinglai, and Derong Liu (2014). Adaptive dynamic programming for optimal tracking control of unknown nonlinear systems with application to coal gasification. IEEE Transactions on Automation Science and Engineering 11(4): 1020-1036.

Wu, Jialin Snow, Rob Law, and Jingyan Liu (2018). Co-creating value with customers: a study of mobile hotel bookings in China. International Journal of Contemporary Hospitality Management.

Zhang, Jonathan Z., Oded Netzer, and Asim Ansari (2014). Dynamic targeted pricing in B2B relationships. Marketing Science 33(3), 317-337.

Zhang, Qin, & P. B. Seetharaman (2018). Assessing lifetime profitability of customers with purchasing cycles. Marketing Intelligence and Planning 36(2), 276-289.

Zhao, Mengchen, et al (2018). Impression Allocation for Combating Fraud in E-commerce Via Deep Reinforcement Learning with Action Norm Penalty. IJCAI, 3940-3946