

Received July 9, 2020, accepted August 3, 2020, date of publication August 14, 2020, date of current version August 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3016824

An Expected Win Rate-Based Real Time Bidding Strategy for Branding Campaign by the Model-Free Reinforcement Learning Model

WEN-YUEH SHIH¹, YI-SHU LU¹, HSIAO-PING TSAI^{1,2}, (Member, IEEE),
AND JIUN-LONG HUANG¹, (Member, IEEE)

¹Department of Computer Science, National Chiao Tung University, Hsinchu 30010, Taiwan

²Innovation and Development Center of Sustainable Agriculture, Department of Electrical Engineering, National Chung Hsing University, Taichung 40227, Taiwan

Corresponding author: Hsiao-Ping Tsai (hptsai@nchu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Contract MOST 106-2221-E-009-152-MY3, Contract MOST 108-2218-E-009-049, Contract 108-2218-E-005-012, Contract 109-2218-E-009-015, and Contract


109-2221-E-009-118-MY3; and in part by the Innovation and Development Center of Sustainable Agriculture from the Featured Areas Research Center Program, Framework of the Higher Education Sprout Project, Ministry of Education (MOE), Taiwan.

ABSTRACT The bidding strategy plays the most important role to help the Demand Side Platforms (DSPs) making bidding decisions on a large number of bid requests in Real Time Bidding (RTB) to satisfy the different objectives of campaigns under the lifetime and budget constraints. In this paper, we focus on branding campaign whose objective is to obtain as many impressions as possible under the lifetime and budget constraints. To achieve the objectives of branding campaigns, we propose a novel expected win rate-based bidding strategy for branding campaign under the lifetime and budget constraints by utilizing a model-free reinforcement learning model. Specifically, to prevent missing good opportunities resulting from submitting extremely low bid prices, the concept of the base winning price is introduced to determine the lower bound of expected winning price. In addition, to obtain more impressions, the concept of the DSP-specified budget spending plan is proposed to determine the proper winning prices. The base expected win rate is then calculated based on the base winning price and the winning price determined by the DSP-specified budget spending plan. Since RTB is a dynamic environment, we propose a novel expected win rate-based bidding strategy named *EWDQN* which utilizes Deep Q Network (DQN) to dynamically determine the expected win rate according to the base expected win rate and the current status of the RTB market, and then determines the bid price according to the expected win rate. To the best of our knowledge, this is the first research applying the reinforcement learning technique on the bidding strategies for branding campaign. To measure the performance of *EWDQN*, several experiments are conducted on two real datasets. Experimental results show that *EWDQN* outperforms the-state-of-the-art bidding strategies for branding campaign in terms of the number of obtained impressions and CPM (cost per thousand impressions).

INDEX TERMS Real time bidding, online advertising, bidding strategy, reinforcement learning, demand side platform, branding campaign.

I. INTRODUCTION

In the modern world, the online advertising has become one of indispensable media in advertising delivery by displaying the ads on somewhere in websites or mobile apps. Such new

The associate editor coordinating the review of this manuscript and approving it for publication was Justin Zhang .

type of advertising delivery can help different kinds of business to create enormous economic benefit. For example, a sportswear company can obtain many display opportunities from online advertising to announce and promote their new products. Due to the popularity of Internet services such as social network and video streaming, the online advertising is able to provide a large number of display opportunities

(a display of an ad is called an *impression*), making the market size of online advertising significantly increase in recent years.¹

Although the display opportunities in online advertising are plenty, the competitors who want to display their ads are also a lot. To trade these displaying opportunities, the Real Time Bidding (RTB) is introduced to sell these displaying opportunities, called *ad slots*. In RTB, once an ad slot is created (e.g., an audience executes an app), the *publisher* (i.e., the owner of the ad slot) sends a bid request consisting of the information of the ad slot and the audience to an ad exchange (ADX) to hold an auction. The ADX then sends the bid request to all *advertisers* to notify them an auction is held. The advertisers then determine their bid prices of the bid request based on the information stored in bid request, and submit their bid prices to the ADX. The advertiser submitting the highest bid price wins the bid request. However, the advertisers usually do not have the capability to handling the bidding process, and thus, they usually outsource the bidding work to the Demand Side Platforms (DSPs) to achieve the advertisers' objectives by running the customized bidding strategies. Based on the objectives, the campaigns can be simply classified into two types: the *branding campaign* and the *performance campaign* [1].

- A branding campaign usually wants to deliver a message or promote a product. Therefore, the objective of a branding campaign is usually to acquire as many impressions as possible under the budget and lifetime constraints.
- A performance campaign is usually eager to obtain the audiences' responses, such as clicking on the ad, or performing other further actions after the click (e.g., purchasing the product after clicking the promotion ad or installing the app after watching the introduction video), called a *conversion*.

The main objective of the bidding strategy for branding campaign is to obtain as many impressions as possible under the constraint of the budget and lifetime. In addition, as mentioned in [2] advertisers usually specify their desired budget spending plans.² A popular type of the bidding strategies for branding campaign focuses on the budget control, called the pacing model [1]–[3]. The pacing model decides how to adjust the budget spending rate of current time slot based on the budget spending status of the previous time slot. Xu *et al.* proposed in [1] a smart pacing-based bidding strategy to spend budget smoothly during the entire lifetime of a campaign. Instead of using the pacing model, Shih and Huang proposed in [4] an expected win rate-based bidding strategy named *EWR* to maximize the number of obtained impressions. Experimental results show that *EWR* is of better budget control ability and obtains more impressions than the pacing model-based bidding strategies [4]. However, *EWR* is of the following drawbacks.

- 1) *EWR* may submit extremely low bid prices, and thus, may miss some good bidding opportunities. *EWR* is designed to determine the bid prices by equally assigning the remaining budget to all the bid requests in the near future. When the number of bid requests in the near future is large, the bid prices determined by *EWR* may be too low to win any bid, which may miss some good bidding opportunities.
- 2) *EWR* cannot adapt to the change of the RTB market well due to relying on the number of bid requests in the near future, which is difficult to accurately predict. *EWR* takes the number of bid requests in the near future as a feature to adjust the budget spending rate. In RTB, the number of the incoming bid requests in the near future is too dynamic to be predicted accurately. Thus, integrating the predicted number of bid requests in the near future into bidding strategies makes *EWR* not able to adapt to the change of the RTB market well.

To prevent missing good opportunities resulting from submitting extremely low bid prices (drawback 1), the concept of the *base winning price* is introduced to determine the lower bound of the expected winning price. In addition, to obtain more impressions, the concept of the *DSP-specified budget spending plan* is proposed to determine the proper winning prices. The *base expected win rate* is then calculated based on the base winning price and the DSP-specified budget spending plan.

However, the base expected win rate and the DSP-specified budget spending plan are derived according to the historical bidding records. Since the RTB market is highly dynamic, it is inappropriate to calculate the bid price by considering the base expected win rate only. In addition, the rapid change in the RTB market also makes the prediction of the future status of the RTB market difficult. Fortunately, reinforcement learning (RL) is capable of finding the optimal policies under dynamic environments, and has already been used in several areas in recent years. For example, in the electricity market, RL is not only adopted to build the bidding strategy for power trading but also to perform real-time power management [5]–[7]. In the finance market, several studies use RL to develop RL agents to help decision making in trading [8]–[11]. Such success motivates to employ RL to build the bidding strategy which is able to adapt to the rapid change of the RTB market.

To prevent using other prediction function (drawback 2), we decide to use the model-free RL model, which does not make any prediction, to build the RL agent. Therefore, we propose in this paper a novel expected win rate-based bidding strategy for branding campaign employing (1) the base expected win rate, (2) the DSP-specified budget spending plan and (3) the model-free RL model [5], [12]–[14]. Specifically, we model the bidding procedure as a Markov decision process (MDP) and propose to employ the model-free RL model to build an expected win rate-based bidding strategy named *EWDQN*. *EWDQN* utilizes Deep Q Network (DQN) to dynamically determine the expected win rate according to the

¹<https://www.emarketer.com/content/us-digital-ad-spending-2019>

²The definition of budget spending plan will be given in Section III.

base expected win rate, the remaining resource (e.g., budget) and the current status of the RTB market, and then determines the bid price according to the expected win rate. The advantages of using the model-free RL model are twofold. First, the resultant RL agent is able to adapt to the change of the RTB market by selecting the proper actions according to the current status of the RTB market. Second, since being built by the model-free RL model, the RL agent does not rely on any other prediction method. Thus, *EWDQN* is of better adaptation ability than *EWR* since *EWDQN* will not be affected by the prediction error resulting from the change of the RTB market. To the best of our knowledge, this is the first study applying the RL model on design of the bidding strategies for branding campaign. To measure the performance of *EWDQN*, several experiments are conducted on two real datasets. Experimental results show that *EWDQN* outperforms the state-of-the-art bidding strategies for branding campaign including *EWR* in terms of the number of obtained impressions and CPM (cost per thousand impressions).

The remainder of this paper is structured as follows. The related works are described in Section II. The proposed bidding strategy for branding campaign, *EWDQN*, is described in Section III. Section IV shows the experimental results. Finally, Section V gives the conclusions of this paper.

II. RELATED WORKS

A. PERFORMANCE PREDICTION

Bidding strategies usually make decisions according to the predicted performance index such as click through rate (CTR) or conversion rate (CVR). There are many types of performance prediction methods and the most common one is the logistic regression-based methods. He *et al.* [15] proposed a hybrid model by utilizing the decision tree as a pre-processing tool and taking the leaf nodes as the input of logistic regression to do CTR prediction. McMahan *et al.* [16] proposed an online learning algorithm, named Follow The (Proximally) Regularized Leader (FTRL-Proximal) algorithm, which can deal with the sparse data efficiently on the CTR prediction model. Chapelle [17] proposed the first CVR prediction model taking the delay feedback into consideration. Lee *et al.* [18] first built the data hierarchies by audiences, advertisers, and publishers information, and adopted different level data to create several weak estimators for CVR prediction. Then, Lee *et al.* proposed to take the results of these weak estimators as the inputs of logistic regression to train the final model. Zhu *et al.* [19] developed a feature selection method to find the significant features for CTR prediction and proposed a softmax-based ensemble model to do prediction by few filtered features.

In addition to the logistic regression, the factorization machine is another popular choice for performance prediction. Field-aware Factorization Machines (FFMs) [20] can combine the influences of two different features in different feature scopes. Pan *et al.* [21] proposed Sparse Factorization Machines (SFMs) which declared that the

Laplace distribution can fit better than Gaussian distribution in sparse datasets. Field-weighted Factorization Machines (FwFMs) [22] is based on the strength of interactions between different fields to improve original FFMs. Recently, the deep neural network is also widely adopted to predict CTR and CVR. Wang *et al.* [23] proposed a novel network architecture, called cross and deep network, to deal with dense and sparse features. Guo *et al.* [24] extended the wide and deep model [25] and proposed a factorization machine-based neural network to extract the feature interactions.

B. WINNING PRICE PREDICTION

The cost of each impression is also an indicator to influence the advertisers' decisions. However, due to the second price auction [26], only the winner knows the final winning price, and the others only know that their prices are not the highest. In other words, the DSPs are not able to observe the entire market situation or the distribution of winning price, thereby resulting in the censor issue. Cui *et al.* [27] proposed a bid star tree to organize the bidding history and model the price as log-normal distribution. However, it's from the perspective of SSPs. Wu *et al.* [28] first proposed the censor data problem on winning price estimation. They adopted the censor regression with the assumption that winning prices are modeled as Gaussian distribution. Wu *et al.* [29] proposed another deep learning framework, which models the winning price as Gumbel distribution, to solve the censor problem. Zhu *et al.* [30] proposed a gamma-based censored linear regression and a two-step optimization method to learn the model. Wang *et al.* [31] proposed a novel decision tree-based method and utilized the non-parametric survival models to predict winning price without making any assumption for winning price distribution. Ren *et al.* [32] developed the Deep Landscape Forecasting (DLF) model which combines the power of deep neural network and survival analysis to predict winning price.

C. THE BIDDING STRATEGIES FOR BRANDING CAMPAIGNS

To acquire more impressions or reach more audiences, the bidding strategies for branding campaigns must control the budget spending rates to fulfill advertisers' budget spending plans. The pacing models [1]–[3] are the common methods to control budget spending rates by adjusting the frequency of auction participation. To spend budget smoothly, the pacing model can depend on the budget consumption status of previous time slot, historical bid request number, and win rate to decide how often they join the auction in the current time slot [2], [3]. Xu *et al.* [1] proposed the smart pacing method. They first gathered the bid requests with similar CTR predictions into groups, and defined the group pacing rate for each group. Then, smart pacing dynamically adjusted the group pacing rate according to the current budget spending rate to control the budget spending rate to fulfill the budget spending plan. Based on psychology theory, Maehara *et al.* [33] defined a different objective for

branding campaign as maximizing the number of audiences who remember the ad after watching for a period of time. To acquire more impressions, Shih and Huang [4] introduced the concept of *expected win rate* which indicated how eager to win this bid request under the current remaining resources. They proposed a novel bidding strategy which dynamically adjusted expected win rate, instead of the pacing rate, to control the budget spending rate. Although outperforming the pacing model-based strategies, the expected win rate-based strategy proposed in [4] may submit too low bid price to win any bid request when the number of incoming bid requests is huge, thereby missing some good bidding opportunities. Such situation motivates us to introduce the concept of *base winning price* for each bid request to prevent such situation.

III. THE PROPOSED BIDDING STRATEGY FOR BRANDING CAMPAIGNS

A. PROBLEM FORMULATION

As mentioned in [1], [4], the objective of the bidding strategies for branding campaign is to obtain as many impressions as possible under the budget and lifetime constraints, which are denoted as B and T , respectively. Assume that there are n time slots in the lifetime T , say $\{t_1, t_2, \dots, t_n\}$, and the duration of each time slot is $\frac{T}{n}$.

Definition 1: There are many budget allocation methods to assign budget to each time slot t_i , such as even-based allocation and traffic-based allocation³ [1], [2]. The budget allocation specified by an advertiser is called the *budget spending plan*⁴ (or the *advertiser-specified budget spending plan*). The budgets allocated to all time slots are denoted as $\{B_1, B_2, \dots, B_n\}$, where $\sum_{i=1}^n B_i = B$.

Definition 2: The *budget spending rate* represents the budget consumption rate, denoted as $\{\frac{C_1}{B_1}, \frac{C_2}{B_2}, \dots, \frac{C_n}{B_n}\}$, where C_i is the budget spent in time slot t_i . Let C be the total budget spent during the lifetime T . C can be obtained by the summation of the spent budgets of each time slot (that is, $C = \sum_{i=1}^n C_i \leq B$).

Definition 3: To evaluate whether the budget spending rate *fulfills* the advertiser-specified budget spending plan, we define the following equation

$$\sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{B_i - C_i}{B_i}\right)^2} \leq \epsilon \text{ and } C \leq B,$$

where ϵ is a tolerance value of the average square root of the difference between the budget allocated to and the budget spent in each time slot. If the result is smaller than or equal to ϵ and the total spent budget C is smaller than and close to the

³The even-based allocation is to allocate equal budget to each time slot, while the traffic-based allocation is to allocate the budget of each time slot in proportion to the estimated number of bid requests in the time slot. Please refer to [1], [2] for details.

⁴Since the concept of the DSP-specified budget spending plan will be introduced in Section III-C, the advertiser-specified budget spending plan will be used to indicate the budget spending plan defined in [2] in the rest of this paper for better readability.

total budget B , we say that the budget spending rate fulfills the advertiser-specified budget spending plan.

As mentioned in Section I, the bidding strategy for branding campaign aims to obtain as many impressions as possible for information delivery under lifetime and budget constraints. In the meantime, the bidding strategy should also make the budget spending rate fulfill the advertiser-specified budget spending plan.

Definition 4: Similar as [1], [4], the objective function of a bidding strategy for branding campaign is formulated as below:

$$\begin{aligned} \min & \frac{1000 \times C}{Imp} \\ \text{s.t.} & \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{B_i - C_i}{B_i}\right)^2} \leq \epsilon, \\ & C \leq B, \\ & bp_j \leq \tau \quad \forall x_j, \end{aligned} \quad (1)$$

where x_j stands for the j -th bid request of m features $x_j = [x_{j,1}, x_{j,2}, \dots, x_{j,m}]$, and Imp is the number of the obtained impressions. The above objective function is to minimize the cost per thousand impressions (CPM) while the advertiser-specified budget spending plan should be fulfilled and the bid price for each bid request cannot exceed the threshold τ . In practice, setting the upper bound (i.e., τ) of the bid price is a common constraint specified by DSPs to avoid spending too much budget for one bid request.

B. DATA OBSERVATION

Shih and Huang observed in [4] that the cumulative distribution function (CDF) of winning price is nonlinear. Figure 1 shows the CDFs of the winning prices of different campaigns.⁵ As we can see, if the DSP wants to buy an impression with 100% expected win rate, the spent budget is much more than two times of the spent budget of 50% expected win rate. For example, as shown in Figure 1a, the DSPs may spend at most \$2.97 and \$0.8, respectively, to buy an impression with 100% and 50%, respectively, expected win rate in campaign 1485. This phenomenon reveals that when bidding the requests with low expected win rates, we could obtain more impressions than bidding with high expected win rates [4]. However, they did not consider the relation between the budget constraint and the distribution of winning price, so they may bid the requests with too low or too high prices. For example, the bidding strategy proposed in [4] may give an extremely low price at the beginning of a time slot whose number of incoming requests in the near future is huge, making the bidding strategy not win any auction at the beginning of the time slot. In view of this, we introduce in Section III-C1 the concept of the *base expected win rate* for each bid request to prevent missing good bid requests.

⁵Campaign 1458 and 3386 are from the iPinYou dataset, and campaign 215 is from the Tenmax dataset. The details of datasets will be given in Section IV-A.

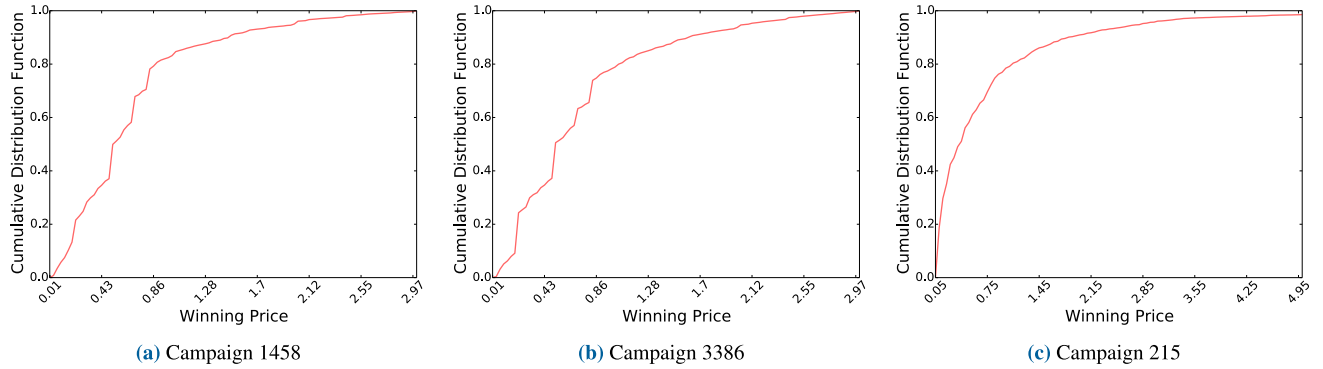


FIGURE 1. The CDF of winning price by different campaigns.

TABLE 1. Descriptions of used symbols.

Notation	Description
B	The total budget
C	The total cost
B_i, C_i	The allocated budget and cost of t_i
$B'_{i,j}$	The current budget while j -th bid request arrives in time slot t_i
T	The lifetime of campaign
t_i	The i -th time slot in T
$r_{i,j}$	The time consumption rate in time slot t_i while j -th bid request arrives
ϵ	The tolerance value of the average difference between B_i and C_i
τ	The upper bound of bid price defined by DSPs
x	The bid request which is of m features
bp	The bid price
bp^{exp}	The expected bid price
β_x	The coefficients for the bid request features
β_{bp}	The coefficient for the bid price
wp^{base}	The base winning price for each bid request
$wp^{exp}_{i,j}$	The expected winning price in time slot t_i while j -th bid request arrives
$wr^{base}_{i,j}$	The base expected win rate in time slot t_i while j -th bid request arrives
$wr^{exp}_{i,j}$	The expected win rate in time slot t_i while j -th bid request arrives
Imp	The total number of impressions
p, q	The shape parameters of beta distribution
S	The set of states
\mathcal{A}	The set of actions
\mathcal{R}	The reward function

Since the market in RTB is highly dynamic, we propose an RL-based bidding strategy to determine the expected win rate of each bid request according to the base expected win rate and the current status of the RTB market, and then determine the bid price according to the expected win rate.

C. MODELING AS A MARKOV DECISION PROCESS

MDP [8], [34] is an appropriate method to design bidding strategies for maximizing the objective under different remaining resources such as budget and lifetime. A straightforward method is to take the remaining resources as states, objective acquisitions as reward and the bid prices in the acceptable range as candidate actions. However, due to the dynamic factors such as the unknown number of incoming bid requests in the near future and uncertain winning prices of bid requests, this straightforward method does not performing well in the RTB environment.

Figure 2 shows the architecture of the proposed bidding strategy for branding campaign. Given the historical bidding records, we first apply the bid function establishment method proposed in [4] to establish the bid function, which can be used to calculate the corresponding bid price of a bid request under a given expected win rate. Then, we introduce the concept of the base expected win rate and propose a base expected win rate determination method in Section III-C1.

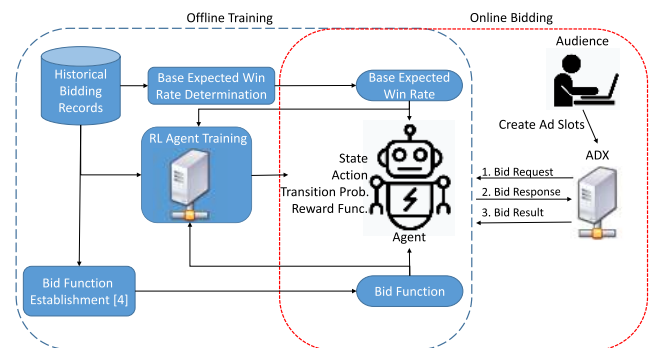


FIGURE 2. The system architecture of the proposed bidding strategy for branding campaign.

To adapt to the rapid change of the RTB market, we model the bidding process as an MDP and propose to train an RL agent to dynamically determine the bid price of each bid request according to the base expected win rate and the current status of the RTB market in Section III-C2. In the training phase, the historical bidding records are utilized

to simulate the bid requests received from the ADX. The RL agent can be trained by the simulated auctions through the RL environment and the received reward. After the offline training phase, with the aid of the established bid function and the base expected win rate, the RL agent can handle the real bid requests which are delivered from the real ADX in the online bidding phase.

1) BASE EXPECTED WIN RATE DETERMINATION

Since adjusting the expected win rate to speed up/slow down the budget spending rate considering only the remaining budget and the predicted number of incoming bid requests in the near future, *EWR* may determine extremely low expected win rates (i.e., extremely low bid prices) when the predicted number of incoming bid requests is huge, thereby missing many good opportunities in the beginning of a time slot [4]. In view of this, we propose the concept of the base expected win rate to relieve the aforementioned problem.

Definition 5: The *base winning price* is defined as the winning price which is able to obtain the most bid requests from the training data. Thus, the base winning price, denoted as wp^{base} , can be formulated as:

$$wp^{base} := \operatorname{argmax}_{y \in (0, \tau]} Imp \times \int_0^y z \times P(z) dz \leq B, \quad (2)$$

where Imp is the total number of impressions in the training data, and $P(z)$ stands for the percentage of impressions with winning price z . Note that the base winning price is constant in the online bidding phase.

In practice, the length of each time slot specified in an advertiser-specified budget spending plan is usually in the coarse-grained level (e.g., one hour or one day). Due to the rapid change of the RTB market, it is obvious that the RTB market may be not always close to the historical records. Such phenomenon may make the bidding strategy spend budget in a too slow or too fast pace, thereby missing many good bidding opportunities. In view of this, we define the concept of the *DSP-specified budget spending plan* to control the budget spending rate within each time slot to obtain more impressions.

Definition 6: Denote the time consuming rate of the current time slot t_i when the j -th bid request arrives as $r_{i,j}$. The CDF of the DSP-specified budget spending plan of the time slot t_i when the j -th bid request arrives, denoted as $DSPBP(r_{i,j})$, can be formulated as:

$$DSPBP(r_{i,j}) = \frac{r_{i,j}^{p-1} (1 - r_{i,j})^{q-1}}{\mathcal{B}(p, q)}, \quad (3)$$

where $0 \leq r_{i,j} \leq 1$ and $\mathcal{B}(\cdot)$ is the beta function. We propose to use the beta distribution to model the DSP-specified budget spending plan because the beta distribution can model different types of DSP-specified budget spending plans (as shown in Figure 3) via setting proper values of the shape parameters p and q .

With the DSP-specified budget spending plan, the expected winning price for bid request x_j in time slot t_i , denoted

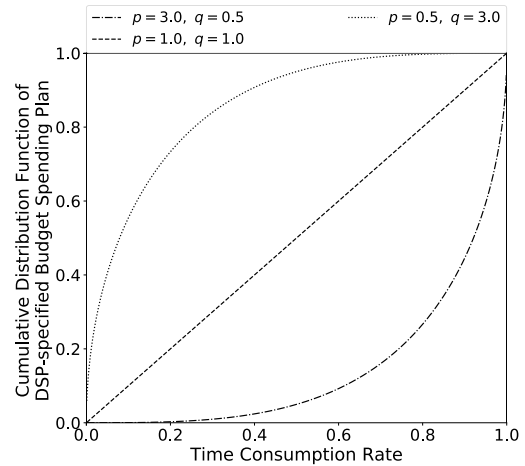


FIGURE 3. DSP-specified budget spending plan modeled by the beta distribution with different values of the shape parameters.

as $wp_{i,j}^{exp}$, can be calculated as:

$$wp_{i,j}^{exp} = \min(\max(wp^{base}, B'_{i,j} - B_i \times (1 - DSPBP(r_{i,j}))), \tau), \quad (4)$$

where $B'_{i,j}$ is the remaining budget of the current time slot t_i while bid request x_j arrives. $B_i \times (1 - DSPBP(r_{i,j}))$ shows how much budget should be remained at $t_{i,j}$, so the difference to $B'_{i,j}$ is the winning price of x_j preferred by the DSP-specified budget spending plan.

Due to the second price auction used in most RTB services, there is a gap between winning price and bid price. Thus, we adopt the method proposed in [4] to estimate the corresponding expected bid price, denoted as $bp_{i,j}^{exp}$, of the expected winning price $wp_{i,j}^{exp}$. Finally, similar to [4], the expected bid price $bp_{i,j}^{exp}$ is transformed to the *base expected win rate* for bid request x_j in time slot t_i , denoted as $wr_{i,j}^{base}$, by the following equation.

$$wr_{i,j}^{base} = p(\text{win}|x_j, bp_{i,j}^{exp}) = \frac{1}{1 + e^{-(\beta_x \times x_j + \beta_{bp} \times bp_{i,j}^{exp})}} \quad (5)$$

2) MARKOV DECISION PROCESS FORMULATION

Due to the rapid change of the RTB market, it is inappropriate to calculate the bid price by considering the base expected win rate and the bid function only and ignoring the current status of the RTB market. In addition, it is difficult to predict the future status of the RTB market such as the number of incoming bid requests in the near future.

In view of this, in this subsection, we model the bidding procedure as an MDP and propose to employ the model-free reinforcement learning model to build an expected win rate-based bidding strategy named *EWDQN*, which dynamically determines the expected win rate according to the base expected win rate, the remaining resource,⁶ and the current

⁶In *EWDQN*, a state is composed of the DSP-specified budget remaining rate, time remaining rate and hour. The details will be given latter.

status of the RTB market, and then determines the bid price according to the expected win rate.

The reason we propose to determine the bid price from the expected win rate is as follows. Since the winning price distribution of each campaign is quite different, directly determining the bid price according to the expected bid price will not perform well due the lack of consideration of the winning price distribution.

Since the RTB market usually changes rapidly, we adopt model-free reinforcement learning to prevent making any strong assumption on the status of the RTB market such as the number of bid requests in the near future. In this paper, we adopt the DQN model to find the best action for each state. DQN builds a deep neural network to model a Q-value function, $Q(s, a)$, which can estimate the Q-value of executing action a on state s . For each state, the action of the highest Q-value is the best choice. The MDP of the bidding strategy for branding campaign can be formulated as $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$, which stands for the state set, action set, transition probability, and the reward function, respectively. We then formulate the proposed bidding strategy for branding campaign as the MDP as follows.

- *State*, $\{s|s \in \mathcal{S}\}$. In order to monitor the environment and make the bidding strategy satisfying the Markov property, the content of state s is designed as below.
 - *DSP-specified budget remaining rate* is the indicator measuring the budget spending status in the current time slot t_i , defined as $\frac{B'_{i,j}}{B'_i}$, where B'_i is the sum of the residual budget of t_{i-1} and the budget allocated to t_i (i.e., B_i), and $B'_{i,j}$ is the remaining budget of t_i when the j -th bid request arrives.
 - *Time remaining rate* is the rate of the remaining time in the current time slot t_i while the j -th bid request arrives. As mentioned above, the time remaining rate of time slot t_i when the j -th bid request arrives is $1 - r_{i,j}$.
 - *Hour*, denoted as H , is the one-hot encoding of the current hour.

Therefore, the state s can be represented as $s = (\frac{B'_{i,j}}{B'_i}, 1 - r_{i,j}, H)$.

- *Action*, $\{a|a \in \mathcal{A}\}$. The agent will calculate the expected win rate for bid request x_j in time slot t_i according to the current state and the base expected win rate $wr_{i,j}^{base}$ to maximize the objective. One action represents an increment or a decrement on the base expected win rate. In other words, the expected win rate for bid request x_j in t_i , denoted as $wr_{i,j}^{exp}$, can be obtained by

$$wr_{i,j}^{exp} = wr_{i,j}^{base} + a_{i,j} \tag{6}$$

where $a_{i,j}$ is the action selected by the agent.

- *Transition probability*, $\mathcal{T} = \mathcal{P}(s'|s, a)$. Let $\mathcal{P}(s'|s, a)$ be the probability that the state will transit from state s to state s' when the agent takes action a in state s . In RTB, the transition probability will be affected by

some uncertain factors such as the winning probability of the given bid price and the distribution of the duration to next incoming bid request. Because RTB is a dynamic environment, it's hard to model these factors. Thus, the model-free RL model is adopted to learn the relationship among states and actions from historical events directly without making any assumption on transition probability [35].

- *Reward function*, $\mathcal{R}(s, a, s')$. The objective of the bidding strategy for branding campaign is to win the bid requests with the low cost and fulfill the advertiser-specified budget spending plan at the end of the lifetime. However, the intuitive method that sets the reward to one reward point when winning an impression may make the advertiser-specified budget spending plan not fulfilled, because the reward function does not contain the information on the current budget spending rate. Therefore, we propose the reward function shown in Equation (7) to take the current budget spending rate (the first component) and the expected win rate (the second component) into consideration. Regardless of the result of the current auction, the agent should get some reward since it tries to catch up with the DSP-specified budget spending plan by participating the auction. After submitting a bid price to a bid request x_j in time slot t_i , the agent gets $1 - \frac{|b(x_j, wr_{i,j}^{exp}) - bp_{i,j}^{exp}|}{\tau}$ reward point(s). The reward function should give a score indicating how the bid price determined by the agent matches the expected bid price obtained from the base winning price and the winning price preferred by the DSP-specified budget spending plan. The reward should get higher when the bid price is getting closer to expected bid price. When winning an auction, we hope that the bid price is as low as possible. Thus, the reward gets higher when the expected win rate gets lower. Thus, after winning a bid request, the agent will get extra $1 - wr_{i,j}^{exp}$ reward point(s).

$$\mathcal{R}(s, a_{i,j}, s') = \begin{cases} 1 - \frac{|b(x_j, wr_{i,j}^{exp}) - bp_{i,j}^{exp}|}{\tau} & \text{if losing} \\ 1 - \frac{|b(x_j, wr_{i,j}^{exp}) - bp_{i,j}^{exp}|}{\tau} + (1 - wr_{i,j}^{exp}) & \text{if winning} \end{cases} \tag{7}$$

After the training procedure, for each bid request x_j in time slot t_i , based on the current state, the agent selects an appropriate action, say $a_{i,j}$, and the expected win rate of the bid request x_j in time slot t_i (i.e., $wr_{i,j}^{exp}$) can be obtained by Equation (6). Once the expected win rate is obtained, we adopt the bid function proposed in [4] to calculate the corresponding bid price of the bid request x_j under the expected win rate $wr_{i,j}^{exp}$. Therefore, the bid price of bid request x_j in time slot t_i can be determined by Equation (8), which is the minimum of the result of the bid function $b(\cdot)$ proposed in [4] and the upper

TABLE 2. Dataset descriptions.

Dataset	Win Records	Average Winning Price	Standard Deviation of Winning Price	Days
iPinYou	12,236,912	0.7838	0.589	7
Tenmax	889,967	1.3591	4.451	7

TABLE 3. Campaigns descriptions.

Dataset	Campaign ID	Winning Records	Average Winning Price	Standard Deviation of Winning Price	The AUC of CTR Prediction	Days
iPinYou	1458	3,083,042	0.6889	0.5345	0.98	7
iPinYou	3386	2,847,798	0.7693	0.6125	0.73	7
Tenmax	215	272,391	0.776	1.398	0.66	7

bound of the bid price τ specified by the DSPs.

$$b(x_j, wr_{i,j}^{exp}) = \min \left(\frac{-\ln(\frac{1}{wr_{i,j}^{exp}} - 1) - \beta_x x_j}{\beta_{bp}}, \tau \right) \quad (8)$$

D. DISCUSSION

1) SUPPORT OF FIRST PRICE AUCTION

In recent years, the first price auction is getting more and more popular in RTB.⁷ Our proposed bidding strategy can support first price auction by the following minor modifications. In first price auction, there is no gap between bid price and winning price. For each bid request, the winning price is the highest received bid price. Thus, the step to obtain the corresponding expected bid price ($bp_{i,j}^{exp}$) of the expected winning price ($wr_{i,j}^{exp}$) can be skipped, and we have $bp_{i,j}^{exp} = wr_{i,j}^{exp}$. The rest procedure in our bidding strategy are then performed as usual to obtain the bid price of the bid request.

2) COOPERATION WITH FRAUD DETECTION

Among the great number of bid requests in RTB, there are many useless impressions created by bots, called *invalid ad traffic*⁸ [36]. Such invalid ad traffic cannot create any real value for the advertisers but only waste the budget. To prevent spending any budget on invalid ad traffic, there are plenty of researches to detect these fraud activities in RTB [36]–[39]. The proposed bidding strategy can cooperate with any fraud detector as follows. Each incoming bid request is first sent to the employed fraud detector. If the fraud detector reports that the bid request is invalid, our bidding strategy will drop the incoming bid request or submit the lowest bid price accepted by the ADX.⁹ Otherwise, our bidding strategy will handle the bid request by the normal procedure to determine the proper bid price.

⁷Google Ad Manager, <https://reurl.cc/b5Y3N1>

⁸Google Ads, <https://reurl.cc/3D46kX>

⁹Some ADXes will set the minimal acceptable bid price, and ask each DSP to submit an acceptable bid price for each bid request.

IV. PERFORMANCE EVALUATION

To measure the performance of *EWDQN*, several experiments are conducted in this section. The descriptions of the datasets used in given in Section IV-A, and the experimental results are given in Section IV-B.

A. DATASET DESCRIPTIONS

The experiments are conducted on two real datasets, the iPinYou dataset [40] and the Tenmax¹⁰ (An online advertising company in Taiwan) dataset. The iPinYou dataset is a seven-day dataset (from 2013/06/06 to 2013/06/12) consisting of more than 12 millions winning records. The Tenmax dataset is also a seven-day dataset (from 2016/10/01 to 2016/10/07) but of only about 900 thousands winning records. The statistics of these two datasets are listed in Table 2. As shown in Table 2, in the Tenmax dataset, the market is of wider price range than the iPinYou dataset, due to the skewed distribution of winning prices. Therefore, we can consider the iPinYou dataset is under a stable market while the Tenmax dataset is not. Because the winning price of each losing bid is unavailable, we follow [4], [28] to simulate the auctions by the winning logs, and compare our strategies with other state-of-the-art methods on these simulated auctions. However, some campaigns in the datasets are not of sufficient records. To avoid the influence of the lack of data, we select two campaigns from the iPinYou dataset and one campaign from the Tenmax dataset, which are of sufficient records for experiments. The information of these three campaigns is shown in Table 3.

B. EVALUATIONS OF BIDDING STRATEGIES FOR BRANDING CAMPAIGN

To meet the objective described in Equation (1), the number of impressions obtained is adopted as a performance metric. To understand whether the bidding strategy can fulfill the advertiser-specified budget spend plan, similar to [4], the budget spending rate is taken as the second performance metric. In addition, the CPM is also considered as the

¹⁰Tenmax, <https://www.tenmax.io/en/>

third performance metric indicating cost efficiency. The records within the first two days are used as the training data, the records within the third day are used for validation, and the rest of the records are used as the testing data. Considering the bid price constraint, τ , in Equation (1), similar to [4], we use the arithmetic and geometric means of winning price of the entire training data as the bid price constraints, which are (\$0.76 and \$0.55) in the iPinYou dataset and (\$1.3 and \$0.44) in the Tenmax dataset, respectively. To evaluate the performance under different budgets, we set the budget to three different amounts, \$5,000, \$7,500, and \$10,000, and the unit of time slot is set to one hour. Besides, similar to [35], we model the bidding strategy for branding campaign as an MDP problem with the discount factor $\lambda = 1$, because the meaning of an impression does not change over time. Experiments are conducted on a workstation equipped with an Intel E5-2683 V3 CPU, a Titan X Pascal graphics card and 64 GB main memory. The system is developed by Python with Keras and Keras-RL.¹¹ The values of the parameters of DQN are listed in Table 4.

TABLE 4. Model settings.

Model	Layers	Nodes	Output Layer	Discount Factor	Policy
Dueling DQN	3	60, 40, 30	Linear	1	ϵ from 1 to 0.1

1) PERFORMANCE COMPARISON OF DIFFERENT ACTION SETS

The action set in our proposed method is the set of candidate actions to be selected to obtain the expected win rate based on the base expected win rate by Equation (6). Since the range of the base expected win rate is [0%, 100%], the maximal range of action sets is [-100%, 100%]. To determine an appropriate action set in the following experiments, we evaluate the performance of five action sets whose ranges are [-20%, 20%], [-40%, 40%], [-60%, 60%], [-80%, 80%] and [-100%, 100%]. For each action set, the actions are set so that the range of the action set is divided into several segments with equal length 10%. Given the same budget and constraints, each action set is used to train five models following the same training process such as the number of iterations.

The experimental result is shown in Figure 4, where the average impressions of the five models of each action set is taken as the performance metric. As shown in Figure 4, the action set [-40%, 40%] is of the best performance. When the range is narrow, the model may be of too few candidate actions to optimize the budget spending. However, if the range is too wide, too many candidate actions may make the training process get too slow to have good model under limited computation resource.

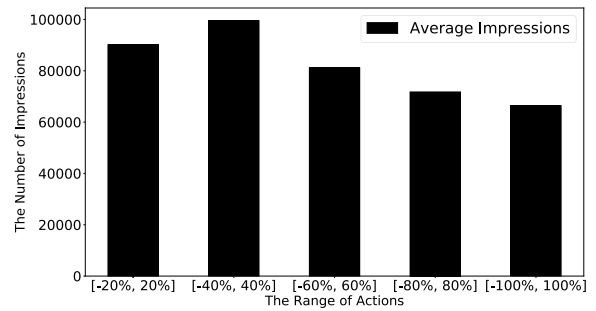


FIGURE 4. The effect of different action settings.

According to the experimental result, the action set is set to $\mathcal{A} = \{-40\%, -30\%, -20\%, -10\%, 0\%, +10\%, +20\%, +30\%, +40\%\}$ in the following experiments.

2) PERFORMANCE COMPARISON OF DIFFERENT DSP-SPECIFIED BUDGET SPENDING PLANS

Different DSP-specified budget spending plans may result in different results. To focus on the performance of the DSP-specified budget spending plans, we ignore the effect of the base winning price (i.e., w_p^{base}) when determining the expected winning price of bid request x_j in time slot t_i (i.e., $w_{p,i,j}^{exp}$) by Equation (4). As shown in Figure 3, we set three pairs of shape parameters of beta distribution to model the following three DPS-specified budget spending plans and conduct an experiment to compare their performance.

- Setting (p, q) to $(3, 0.5)$ is the plan that the budget spending rate is low in the beginning of the time slot and keeps increasing until the end of the time slot.
- Setting (p, q) to $(0.5, 3)$ is the plan that the budget spending rate is high in the beginning of the time slot and keeps decreasing until the end of the time slot.
- Setting (p, q) to $(1, 1)$ is the plan that the budget spending rate is uniform during the time slot.

The experimental results are shown in Figures 5 to 10. Since the major objective of the DSP-specified budget spending plan is to maximize the number of obtained impressions, the number of obtained impressions and CPM are key metrics to measure the performance of different DSP-specified spending plans. Both setting (p, q) to $(1, 1)$ and $(3, 0.5)$ can win 7 out of 18 cases in terms of the number of obtained impressions. On the other hand, setting (p, q) to $(3, 0.5)$ can achieve lower CPM (lower CPM is better), meaning that setting (p, q) to $(3, 0.5)$ is of better cost efficiency. In our experiments, setting (p, q) to $(3, 0.5)$ wins 11 out of 18 cases in terms of CPM. Therefore, (p, q) is set to $(3, 0.5)$ in the following experiments.

3) COMPARISON WITH OTHER BIDDING STRATEGIES FOR BRANDING CAMPAIGN

In this experiment, we compare the proposed bidding strategy for branding campaign, EWQDN, with

¹¹Keras-RL, <https://github.com/keras-rl/keras-rl>

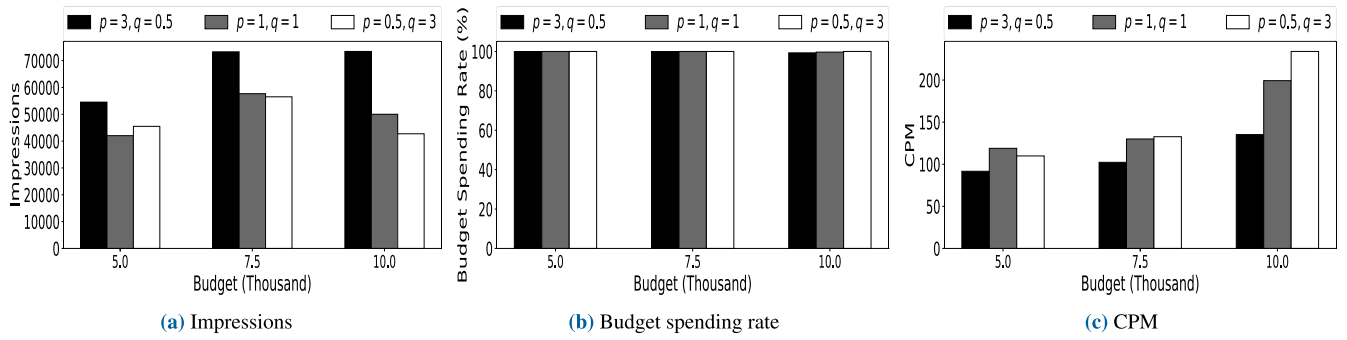


FIGURE 5. Campaign 1458 with $\tau = 0.55$.

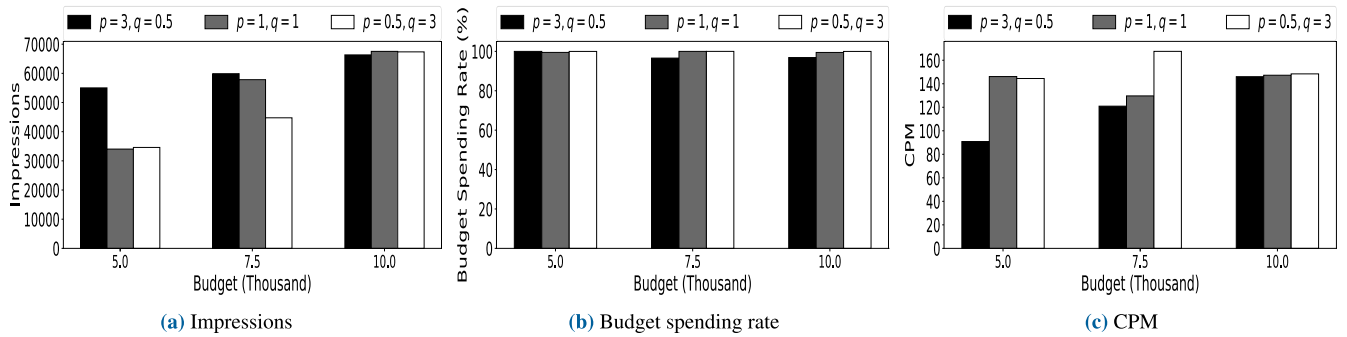


FIGURE 6. Campaign 1458 with $\tau = 0.76$.

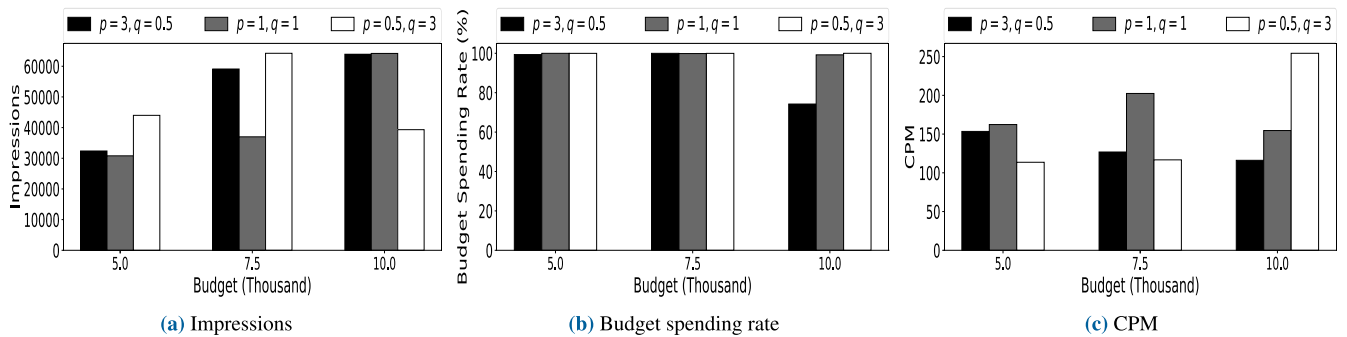


FIGURE 7. Campaign 3386 with $\tau = 0.55$.

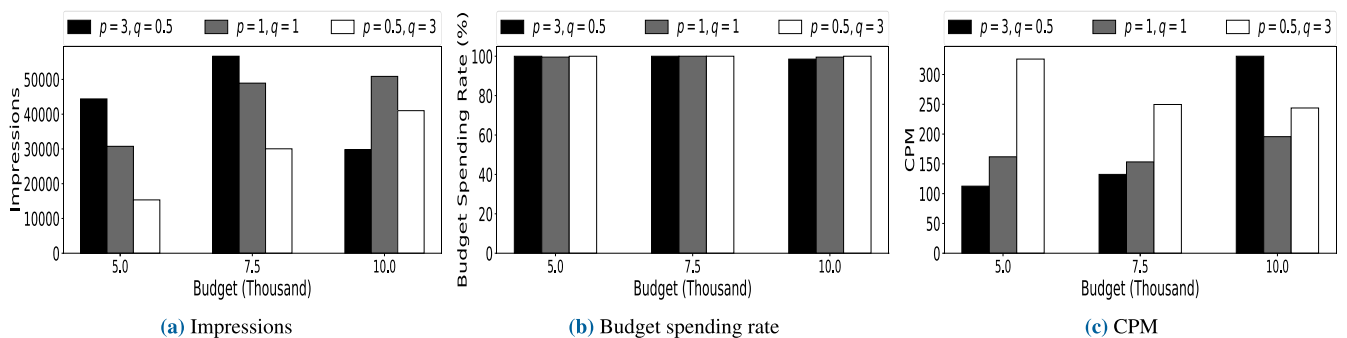


FIGURE 8. Campaign 3386 with $\tau = 0.76$.

other bidding strategies for branding campaign. The bidding strategies compared in this experiment are listed below.

- SP_{FP} is the state-of-the-art pacing model, smart pacing-based strategy [1], with fixed pricing scheme, whose bid price is set to the constant value τ .

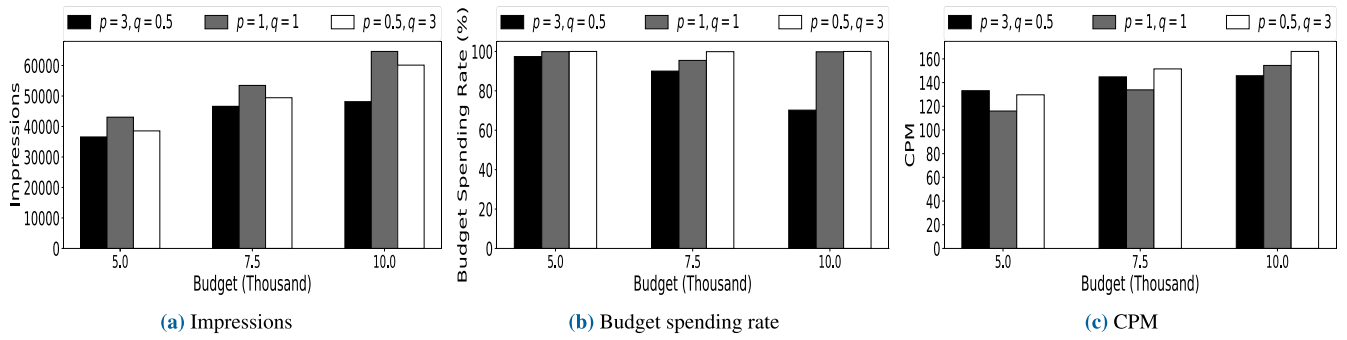


FIGURE 9. Campaign 215 with $\tau = 0.44$.

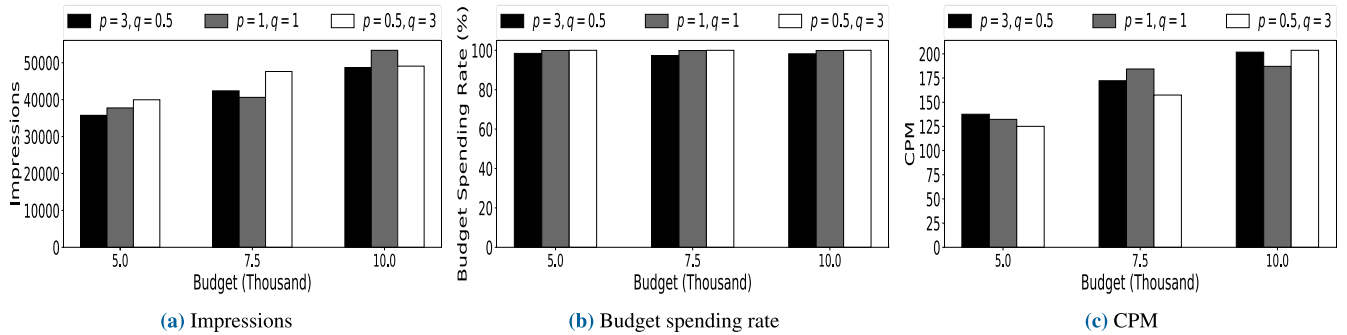


FIGURE 10. Campaign 215 with $\tau = 1.3$.

- SP_WP is another smart pacing-based method [1] adopting the winning price prediction [28] as the pricing scheme.
- EWR_CPM . The state-of-the-art expected win rate-based bidding strategy for branding campaign [4] is also adopted as one competitor.
- $EWDQN$ is our proposed expected win rate-based bidding strategy for branding campaign.
- $EWDQN$ w/o $DSPBP$ is a simplified version of $EWDQN$ disabling the DSP-specified budget spending plan. We can compare $EWDQN$ w/o $DSPBP$ and $EWDQN$ to measure the effect of the DSP-specified budget spending plan.
- $EWDQN$ w/o PRF is a simplified version of $EWDQN$ disabling the proposed reward function and applying a naive reward function that the agent will get one reward point when winning an impression. We can compare $EWDQN$ w/o PRF and $EWDQN$ to measure the effect of the proposed reward function.
- $EWDQN$ w/o BWP is a simplified version of $EWDQN$ disabling base winning price. We can compare $EWDQN$ w/o BWP and $EWDQN$ to measure the effect of base winning price.

Due to the randomness of DQN, for each DQN-based method, we train five models and select the model of the best performance in the validation set in the following experiments.

The experimental results are shown in Figures 11 to 16. Although SP_FP and SP_WP can fulfill the advertiser-specified budget spending plans, the numbers of impressions obtained by them are much less than those obtained by others strategies, making their CPM much higher than others'. Because always bidding the auctions by setting bid prices to τ , SP_FP may win several impressions of high winning prices, thereby of highest CPM in most cases. On the other hand, since bidding by the predicted winning prices, SP_WP can obtain more impressions than SP_FP in most cases. EWR_CPM also fulfills the advertiser-specified budget spending plans in all cases. With the aid of the concept of expected win rate, EWR_CPM outperforms the smart pacing-based bidding strategies in terms of the number of obtained impressions and CPM.

We now investigate the effect of the proposed reward function by comparing $EWDQN$ and $EWDQN$ w/o PRF . We can observe that $EWDQN$ outperforms $EWDQN$ w/o PRF in all cases, showing the advantage of the proposed reward function. As shown in Equation (7), the proposed reward function takes the budget spending rate and the expected win rate into account, making $EWDQN$ able to fulfill the advertiser-specified budget spending plan and obtain many impressions with low CPM. In addition, the performance of $EWDQN$ w/o PRF is unstable. In some cases, $EWDQN$ w/o PRF is even worse than EWR_CPM in impression acquisition and CPM. The reason of the instability of $EWDQN$ w/o PRF

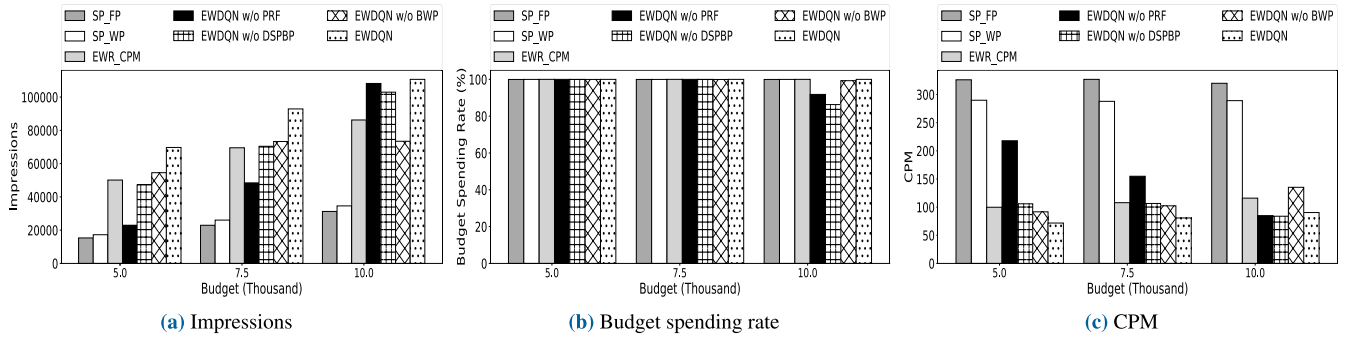


FIGURE 11. Performance comparison on campaign 1458 with $\tau = 0.55$.

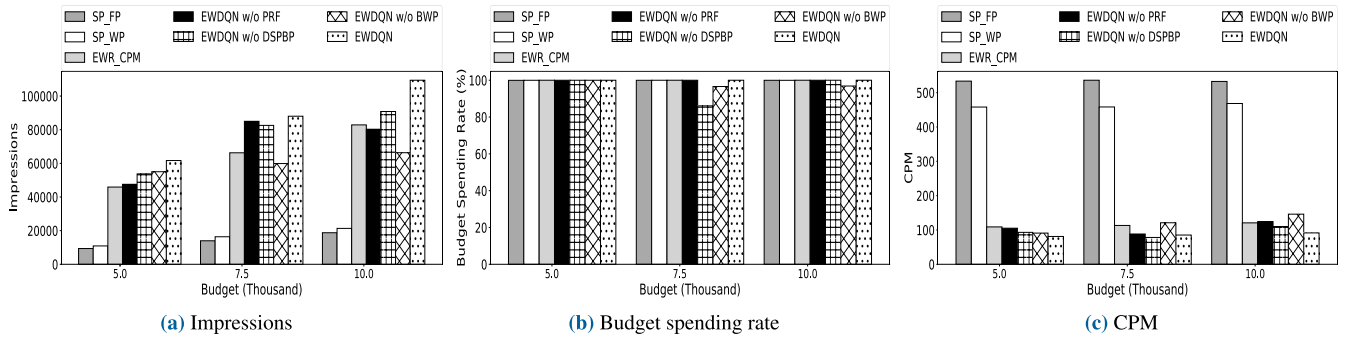


FIGURE 12. Performance comparison on campaign 1458 with $\tau = 0.76$.

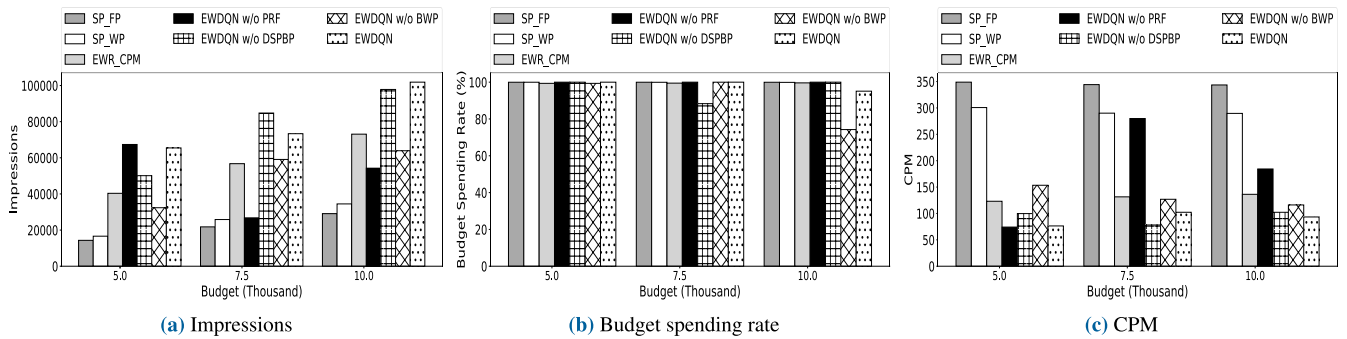


FIGURE 13. Performance comparison on campaign 3386 with $\tau = 0.55$.

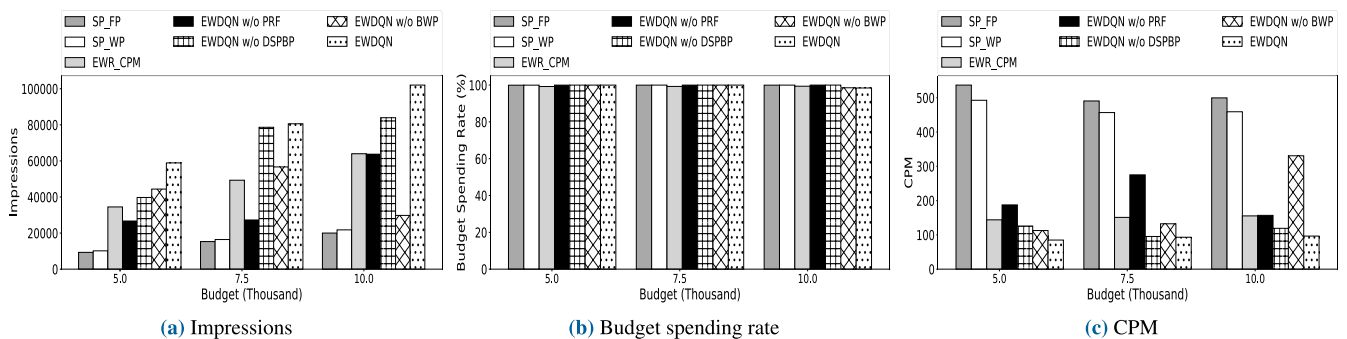


FIGURE 14. Performance comparison on campaign 3386 with $\tau = 0.76$.

is because the naive reward function applied by *EWDQN w/o PRF* does not consider the current budget spending rate and

the expected win rate. Due to not considering the budget spending rate in the reward function, *EWDQN w/o PRF*

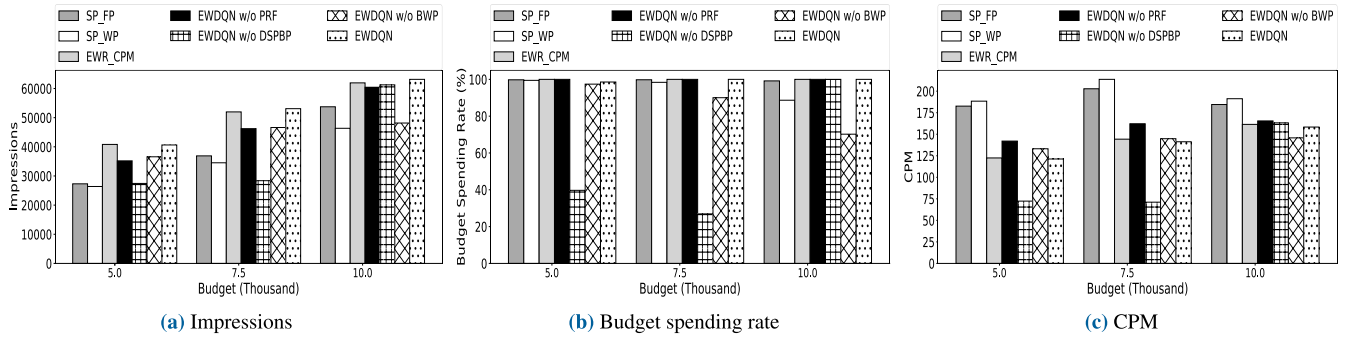


FIGURE 15. Performance comparison on campaign 215 with $\tau = 0.44$.

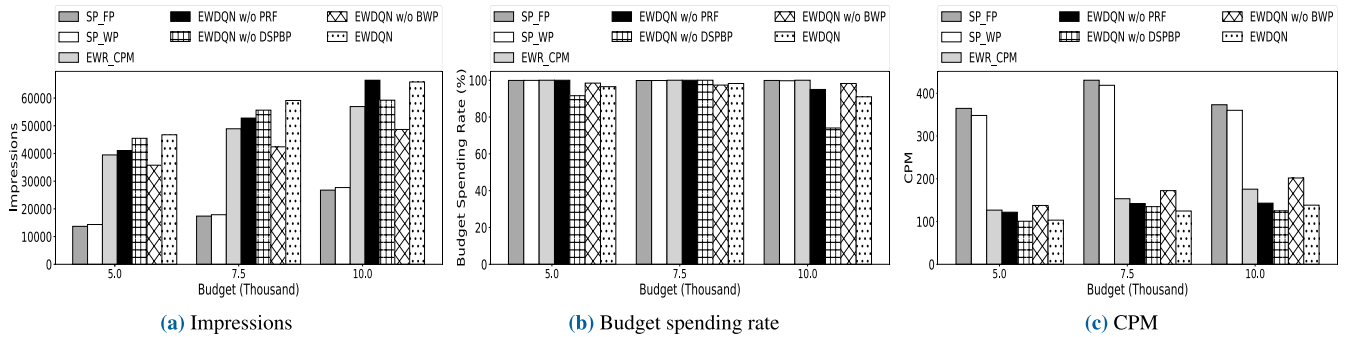


FIGURE 16. Performance comparison on campaign 215 with $\tau = 1.3$.

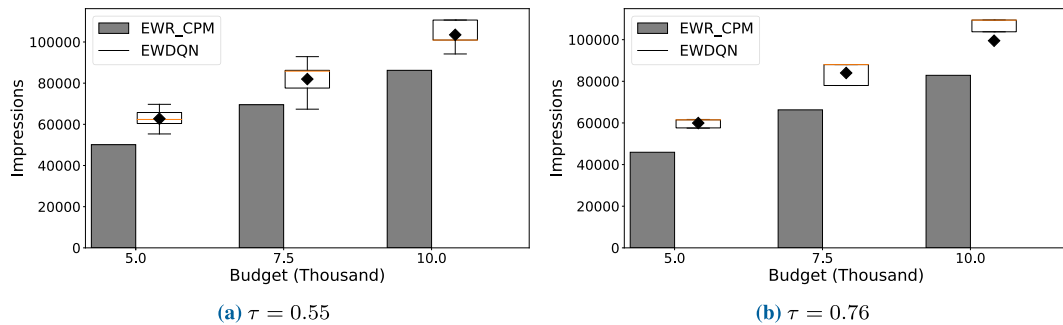


FIGURE 17. The effect of randomness on the campaign 1458.

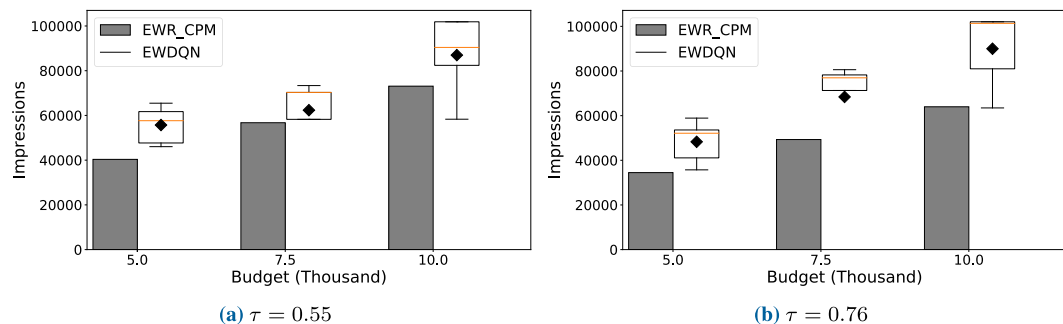


FIGURE 18. The effect of randomness on the campaign 3386.

cannot fulfill the advertiser-specified spending plan in some cases.

We then investigate the effect of the base winning price by comparing *EWDQN* and *EWDQN w/o BWP*. As mentioned

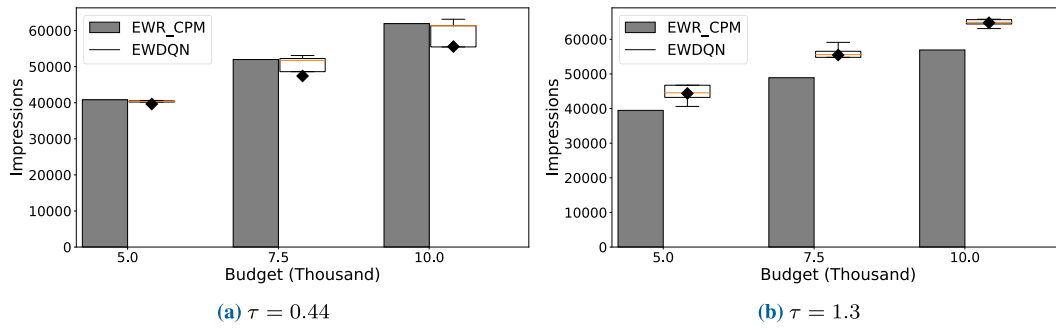


FIGURE 19. The effect of randomness on the campaign 215.

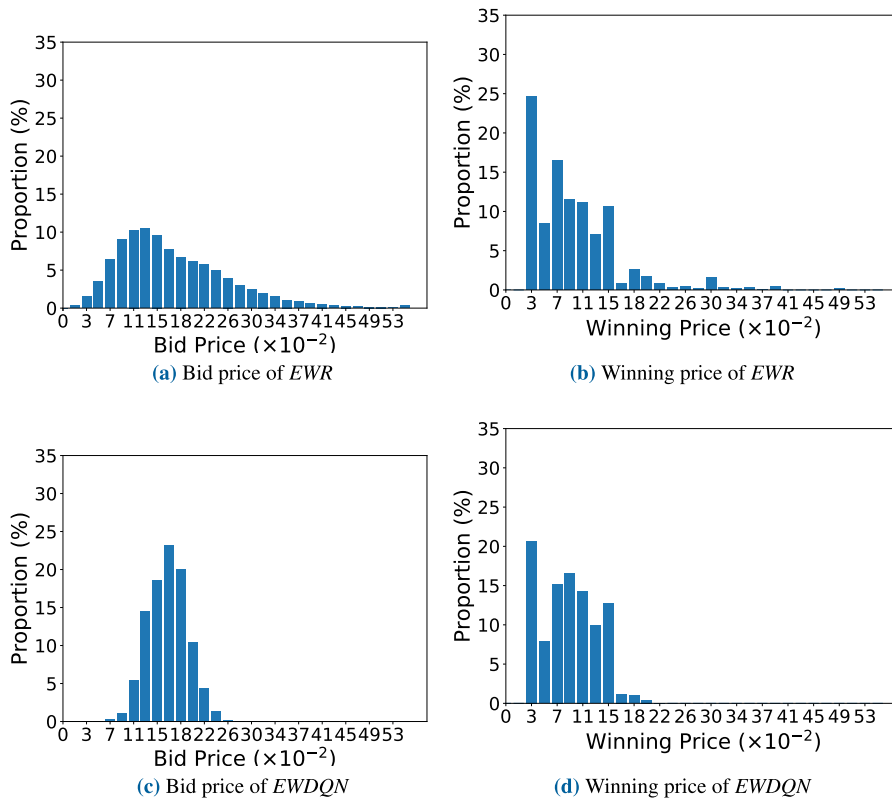


FIGURE 20. The comparisons of bidding/winning price distributions.

in Section III-C1, the objective of the base winning price is to make the bidding strategy not waste bidding opportunities by guiding the bidding strategy not to submit extremely low bid prices. Without the aid of the base winning price, the performance of *EWDQN w/o BWP* is unstable in the number of obtained impressions and CPM, and sometimes worse than *EWDQN w/o PRF*. In some cases, *EWDQN w/o BWP* even does not fulfill the advertiser-specified budget spending plan due to wasting too many bidding opportunities.

We next investigate the effect of the DSP-specified budget spending plan by comparing *EWDQN* and *EWDQN w/o DSPBP*. When the DSP-specified budget spending plan is disabled, the expected winning price equals to the base

winning price. Such change encourages the bidding strategy submitting low prices, thereby may missing some good bidding opportunities. Thus, *EWDQN w/o DSPBP* is of lower CPM than *EWDQN w/o PRF* and *EWDQN w/o BWP* in 12 out of 18 cases and even lower than *EWDQN* in 7 out of 18 cases. However, in the fulfillment of the advertiser-specified spending plan, *EWDQN w/o DSPBP* is of worse performance than *EWDQN w/o PRF*, *EWDQN w/o BWP* and *EWDQN* in most cases. It is because that *EWDQN w/o DSPBP* inclines to submit low bid prices, thereby having very low probability to win the impressions with moderate or high winning prices.

With the aid of the proposed reward function, the base winning price and the DSP-specified budget spending plan,

EWDQN outperforms the other existing bidding strategies in terms of the number of obtained impressions and CPM.

We now evaluate the effect of the randomness of DQN on the performance of *EWDQN* by comparing the performance of *EWR_CPM* and *EWDQN*. The numbers of impressions obtained by the five models of *EWDQN* are shown in box plot in Figures 17 to 19. The average performance of *EWDQN* is better than that of *EWR_CPM* in 15 out of 18 cases. *EWDQN* loses in the case with insufficient historical records and strict upper bound of bid prices. As shown in Figures 17 to 19, it is possible that we will obtain the models of worse performance. To relieve this problem, we can build multiple models and select the best model according to the results of validation. Experimental results show that with proper model selection, *EWDQN* is able to outperform *EWR_CPM* in most cases.

The reason why *EWDQN* outperforms *EWR_CPM* can be observed from the bid price and winning price distributions. We randomly select one time slot from one bidding process of campaign 1458 to observe the bid price and winning price distributions of *EWR_CPM* and *EWDQN*, as shown in Figure 20. According to Figures 20a and 20c, the range of the bid price of *EWR_CPM* is obviously wider than that of *EWDQN*. *EWR_CPM* may lose many auctions due to submitting the extremely low bid price, and remain too much budget at the end of the time slot. Therefore, *EWR_CPM* needs to offer higher bid price later for catching up with the advertiser-specified budget spend plan. Such situation results in some high bid and winning prices in the distributions of *EWR_CPM* shown in Figure 20b. On the other hand, *EWDQN* offers the bid price in a small range around the expected bid price which can avoid such unreasonable bidding.

Besides, in our experiments, the training time of *EWDQN* is about 9 hours. The average response time of *EWDQN* for each bid request is less than 5 ms, which is much less than the RTB requirement.¹² In summary, *EWDQN* is of better ability to spend the budget in a cost efficient manner as well as to control the budget spending rate to fulfill the advertiser-spending budget spending plan than *EWR_CPM*.

V. CONCLUSION

In this paper, we proposed a novel expected win rate-based bidding strategy for branding campaign named *EWDQN* by utilizing model-free RL model. We first introduced the concept of the base winning price to prevent the agent submitting extremely low bid price. We then proposed the DSP-specified budget spending plan to control the budget spending rate in each time slot for better impression acquisition. The base expected win rate was then calculated based on the base winning price and the winning price determined by the DSP-specified budget spending plan. We finally developed *EWDQN* by using DQN to dynamically determine the expected win rate according to the base expected win rate and the current status of the RTB market, and then determine the

bid price according to the expected win rate. Experimental results on real datasets showed that *EWDQN* still outperforms the-state-of-the-art bidding strategies for branding campaign in terms of the number of obtained impressions and CPM. For future works, to make the proposed bidding strategy more reliable, we plan to adopt other advanced RL models for training agents. Besides, we will create more environment variables as state descriptions for better budget control and performance.

REFERENCES

- [1] J. Xu, K.-C. Lee, W. Li, H. Qi, and Q. Lu, "Smart pacing for effective online ad campaign optimization," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2015, pp. 2217–2226.
- [2] K.-C. Lee, A. Jalali, and A. Dasdan, "Real time bid optimization with smooth budget delivery in online advertising," in *Proc. 7th Int. Workshop Data Mining Online Advertising (ADKDD)*, 2013, pp. 1:1–1:9.
- [3] D. Agarwal, S. Ghosh, K. Wei, and S. You, "Budget pacing for targeted online advertisements at LinkedIn," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 1613–1619.
- [4] W.-Y. Shih and J.-L. Huang, "An expected win rate-based real-time bidding strategy for branding campaigns on display advertising," *Knowl. Inf. Syst.*, vol. 61, no. 3, pp. 1395–1430, Dec. 2019.
- [5] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, and G. Strbac, "Deep reinforcement learning for strategic bidding in electricity markets," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1343–1355, Mar. 2020.
- [6] H. Xu, H. Sun, D. Nikovski, S. Kitamura, K. Mori, and H. Hashimoto, "Deep reinforcement learning for joint bidding and pricing of load serving entity," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6366–6375, Nov. 2019.
- [7] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, 2020.
- [8] J. Wang, M. Zhou, X. Jin, X. Guo, L. Qi, and X. Wang, "Variance minimization hedging analysis based on a time-varying Markovian DCC-GARCH model," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 2, pp. 621–632, Apr. 2020.
- [9] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, "Deep direct reinforcement learning for financial signal representation and trading," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 653–664, Mar. 2017.
- [10] T. Kim and H. Y. Kim, "Optimizing the pairs-trading strategy using deep reinforcement learning with trading and stop-loss boundaries," *Complexity*, vol. 2019, pp. 1–20, Nov. 2019.
- [11] Y. Li, W. Zheng, and Z. Zheng, "Deep robust reinforcement learning for practical algorithmic trading," *IEEE Access*, vol. 7, pp. 108014–108022, 2019.
- [12] D. P. Bertsekas, "Feature-based aggregation and deep reinforcement learning: A survey and some new implementations," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 1, pp. 1–31, Jan. 2019.
- [13] Y. Ge, F. Zhu, X. Ling, and Q. Liu, "Safe Q-learning method based on constrained Markov decision processes," *IEEE Access*, vol. 7, pp. 165007–165017, 2019.
- [14] T. Bian and Z.-P. Jiang, "Reinforcement learning for linear continuous-time systems: An incremental learning approach," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 2, pp. 433–440, Mar. 2019.
- [15] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Q. Candela, "Practical lessons from predicting clicks on ads at Facebook," in *Proc. 8th Int. Workshop Data Mining Online Advertising*, 2014, pp. 5:1–5:9.
- [16] H. B. McMahan, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, J. Kubica, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, and E. Davydov, "Ad click prediction: A view from the trenches," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2013, pp. 1222–1230.
- [17] O. Chapelle, "Modeling delayed feedback in display advertising," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 1097–1105.
- [18] K.-C. Lee, B. Orten, A. Dasdan, and W. Li, "Estimating conversion rate in display advertising from past performance data," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 768–776.

¹²Real-Time Bidding Protocol, <https://developers.google.com/authorized-buyers/rtb/start>

- [19] W.-Y. Zhu, C.-H. Wang, W.-Y. Shih, W.-C. Peng, and J.-L. Huang, "SEM: A softmax-based ensemble model for CTR estimation in real-time bidding advertising," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2017, pp. 5–12.
- [20] Y. Juan, Y. Zhuang, W.-S. Chin, and C.-J. Lin, "Field-aware factorization machines for CTR prediction," in *Proc. 10th ACM Conf. Recommender Syst.*, Sep. 2016, pp. 43–50.
- [21] Z. Pan, E. Chen, Q. Liu, T. Xu, H. Ma, and H. Lin, "Sparse factorization machines for click-through rate prediction," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 400–409.
- [22] J. Pan, J. Xu, A. L. Ruiz, W. Zhao, S. Pan, Y. Sun, and Q. Lu, "Field-weighted factorization machines for click-through rate prediction in display advertising," in *Proc. World Wide Web Conf. (WWW)*, 2018, pp. 1349–1357.
- [23] R. Wang, B. Fu, G. Fu, and M. Wang, "Deep & cross network for ad click predictions," in *Proc. ADKDD ZZZ (ADKDD)*, 2017, p. 12.
- [24] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "DeepFM: A factorization-machine based neural network for CTR prediction," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1725–1731.
- [25] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, R. Anil, Z. Haque, L. Hong, V. Jain, X. Liu, and H. Shah, "Wide & deep learning for recommender systems," in *Proc. 1st Workshop Deep Learn. Recommender Syst.*, 2016, pp. 7–10.
- [26] W. Vickrey, "Counterspeculation, auctions, and competitive sealed tenders," *J. Finance*, vol. 16, no. 1, pp. 8–37, Mar. 1961.
- [27] Y. Cui, R. Zhang, W. Li, and J. Mao, "Bid landscape forecasting in online ad exchange marketplace," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 265–273.
- [28] W. C.-H. Wu, M.-Y. Yeh, and M.-S. Chen, "Predicting winning price in real time bidding with censored data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2015, pp. 1305–1314.
- [29] W. Wu, M.-Y. Yeh, and M.-S. Chen, "Deep censored learning of the winning price in the real time bidding," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2526–2535.
- [30] W.-Y. Zhu, W.-Y. Shih, Y.-H. Lee, W.-C. Peng, and J.-L. Huang, "A gamma-based regression for winning price estimation in real-time bidding advertising," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 1610–1619.
- [31] Y. Wang, K. Ren, W. Zhang, J. Wang, and Y. Yu, "Functional bid landscape forecasting for display advertising," in *Machine Learning and Knowledge Discovery in Databases*, P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, Eds., 2016, pp. 115–131.
- [32] K. Ren, J. Qin, L. Zheng, Z. Yang, W. Zhang, and Y. Yu, "Deep landscape forecasting for real-time bidding advertising," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 363–372.
- [33] T. Maehara, A. Narita, J. Baba, and T. Kawabata, "Optimal bidding strategy for brand advertising," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 424–432.
- [34] M. K. Hanawal, H. Liu, H. Zhu, and I. C. Paschalidis, "Learning policies for Markov decision processes from data," *IEEE Trans. Autom. Control*, vol. 64, no. 6, pp. 2298–2309, Jun. 2019.
- [35] D. Wu, X. Chen, X. Yang, H. Wang, Q. Tan, X. Zhang, J. Xu, and K. Gai, "Budget constrained bidding by model-free reinforcement learning in display advertising," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 1443–1451.
- [36] A. Pastor, M. Pärssinen, P. Callejo, P. Vallina, R. Cuevas, Á. Cuevas, M. Kotila, and A. Azcorra, "Nameles: An intelligent system for real-time filtering of invalid ad traffic," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 1454–1464.
- [37] K. Springborn and P. Barford, "Impression fraud in on-line advertising via pay-per-view networks," in *Proc. 22nd USENIX Conf. Secur.* Berkeley, CA, USA: USENIX Association, 2013, pp. 211–226.
- [38] R. Oentaryo, E. Lim, M. Finegold, D. Lo, F. Zhu, C. Phua, E. Cheu, G. Yap, K. Sim, M. N. Nguyen, K. Perera, B. Neupane, M. Faisal, Z. Aung, W. L. Woon, W. Chen, D. Patel, and D. Berrar, "Detecting click fraud in online advertising: A data mining approach," *J. Mach. Learn. Res.*, vol. 15, no. 3, pp. 99–140, 2014.
- [39] S. Kaya, B. Çavdarolu, and K. S. Ensöy, "Detection of click spamming in mobile advertising," in *Advances in Operational Research in the Balkans*, N. Mladenović, A. Sifaleras, and M. Kuzmanović, Eds., 2020, pp. 251–263.

- [40] W. Zhang, S. Yuan, J. Wang, and X. Shen, "Real-time bidding benchmarking with iPinYou dataset," 2014, *arXiv:1407.7073*. [Online]. Available: <http://arxiv.org/abs/1407.7073>



WEN-YUEH SHIH received the B.S. degree in computer science from National Chiao Tung University, in 2009, and the M.S. degree in computer science and information engineering from National Cheng Kung University, in 2012. He is currently pursuing the Ph.D. degree with the Department of Computer Science, National Chiao Tung University. His research interests include wireless sensor networks, indoor localization, data mining, and machine learning.



YI-SHU LU received the B.S. degree in computer science from Chung Yuan Christian University, in 2008, and the M.S. degree in computer science and information engineering from National Taiwan Ocean University, in 2011. He is currently pursuing the Ph.D. degree with the Department of Computer Science, National Chiao Tung University. His research interests include, data mining, machine learning, and reinforcement learning.



HSIAO-PING TSAI (Member, IEEE) received the B.S. and M.S. degrees in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1996 and 1998, respectively, and the Ph.D. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in January 2009. She is currently an Assistant Professor with the Department of Electrical Engineering, National Chung Hsing University. Her research interests include trajectory data mining, robot automatic navigation, cloud computing, object tracking, and sensor data management.



JIUN-LONG HUANG (Member, IEEE) received the B.S. and M.S. degrees from the Computer Science and Information Engineering Department, National Chiao Tung University, in 1997 and 1999, respectively, and the Ph.D. degree from the Electrical Engineering Department, National Taiwan University, in 2003. He joined National Chiao Tung University, in 2005, where he is currently a Professor with the Computer Science Department. His research interests include data analytics, data mining, and blockchain.

...