Multi-arm Bandits Techniques In Online Advertising

***Abstract*—This research paper explores the application of multi-armed bandit algorithms in optimizing content recommendations on online media platforms. The study aims to investigate the use of different bandit algorithms for dynamically selecting and adjusting content recommendations to maximize advertising revenue. A comprehensive experimental analysis is conducted, comparing the proposed methods against random approaches. The paper discusses the significance and potential benefits of utilizing multi-armed bandit algorithms in real-world recommendation systems. According to the experimental results, the ϵ-greedy algorithm demonstrates superior optimization capabilities compared to other Agent's decision functions.**

***Index Terms*—multi-armed bandit algorithms, epsilon greedy algorithm, Upper Confidence Bound algorithm, Uncertainty-Based Confidence-Bound Reinforcement Learning, Thompson Sampling**

I. Introduction

In response to the rapid growth of online media platforms, there has been a need for effective content recommendation systems that can personalize user experiences by suggesting relevant and engaging content. To address the exploration-exploitation trade-off in recommendation systems, multi-armed bandit algorithms have emerged as a promising approach (Cazzolla Gatti, 2021) [1].

Multi-armed bandit problems can be viewed as a simplified form of reinforcement learning, characterized by independent states where learning occurs from individual outcomes of either winning or losing. The focus is solely on assessable feedback. The outcome of each action is only dependent on the current state and is not influenced by the historical outcomes of previous actions. The problem also can be defined as a Markov decision process with a single state.

The main research objectives include evaluating the performance of different bandit algorithms in improving content recommendations on an online media platform, analyzing their impact on user engagement metrics, and discussing the advantages and challenges associated with deploying these techniques in real-world scenarios.

II. Related Work

Various multi-armed bandit (MAB) algorithms have been developed to tackle the exploration-exploitation trade-off and enhance content recommendations across different domains.

Silva et al. (2022) [2] offers a comprehensive explanation of the principles behind the MAB problem. The MAB problem is a classic sequential decision-making problem in which an agent selects actions from a set of options to maximize cumulative rewards, similar to a gambler choosing slot machines for maximum winnings. The agent must find a balance between exploring different options and exploiting the option with the highest expected reward.

The MAB problem is represented by (N, A, R), where N is the number of trials, A is the set of actions, and R is the cumulative reward. By addressing the exploration-exploitation trade-off, the MAB problem is relevant in recommendation systems, where the objective is to maximize user satisfaction by dynamically selecting actions (e.g., recommending items) while learning and optimizing the recommendations. This facilitates adaptive and personalized content recommendations that strike a balance between exploration and exploitation.

The multi-armed bandit (MAB) algorithm has been widely applied in various practical domains, addressing the exploration-exploitation trade-off and optimizing decision-making processes. One prominent application area is online advertising, where MAB algorithms dynamically adjust ad strategies to maximize revenue by experimenting with different options.

According to Issa Mattos et al. (2019) [3], MAB based experiments have emerged as a novel practice in the industry to provide faster value to customers through online experiments. These experiments have the potential to deliver quicker results and achieve improved resource allocation compared to traditional A/B experiments. By leveraging MAB algorithms, organizations can dynamically allocate resources based on real-time feedback, enabling more efficient exploration of different options and faster identification of optimal strategies.

In addition, Yan et al. (2022) [4] highlighted the trade-off faced by recommender systems between exploring new items to maximize user satisfaction and exploiting already interacted items to match user interests. This challenge, known as the exploration/exploitation (EE) dilemma, has been effectively addressed by the multi-armed bandit (MAB) algorithm.

III. Methodology

The goal of this study is to explore and evaluate different MAB algorithms for optimizing content recommendations in an online shopping context. The Methodology section includes the collection of data and the application of agent's decision function.

A. Data collection

The Online Shoppers Purchasing Intention Dataset provided by UCL [5] is utilized as the data source for this study. This dataset contains information from online shopping sessions, with a total of 12,330 instances. Among these sessions, 84.5% (10,422) are negative class samples where shopping did not occur, while the remaining 15.5% (1,908) are positive class samples ending with a purchase.

Due to the unavailability of reliable advertising data in public datasets, the "region" attribute in the dataset is assumed to represent advertisements and treated as arms, resulting in a total of 9 arms. The "revenue" attribute is considered as the reward, and 1000 trials are conducted. By calculating the conversion rates of the 9 advertisements using different multi-armed bandit algorithms, this study aims to formulate optimal advertising placement strategies for future dynamic ad campaigns.

B. Agent's decision function

There are various agent's decision functions in the MAB model. The methods employed in this study include the ϵ-greedy algorithm, Upper Confidence Bound (UCB), Uncertainty-Based Confidence-Bound Reinforcement Learning (UCBRL), and Thompson Sampling.

1) ϵ-greedy algorithm: The ϵ-greedy algorithm, which is widely used in multi-armed bandit problems, offers a straightforward yet powerful strategy. It strikes a balance between exploration and exploitation by employing a probabilistic approach when selecting arms. Specifically, with a probability of $(1-\epsilon)$, it chooses the arm with the highest estimated reward, aiming to exploit the perceived best option. Meanwhile, with a probability of ϵ, it opts for a random arm, enabling exploration to uncover potentially superior alternatives.

To summarize, the ϵ-greedy algorithm can be expressed as follows based on Cazzolla Gatti (2021) [1]:

- Select a random number, r, between 0 and 1.
- If $r \leq \epsilon$, choose a random arm uniformly from all available arms.
- If $r > \epsilon$, select the arm with the highest estimated reward based on previous observations.

2) Upper Confidence Bound (UCB) algorithm: The Upper Confidence Bound (UCB) algorithm is also a well-known approach utilized in multi-armed bandit problems. It relies on confidence intervals to estimate the upper bound of the expected reward associated with each arm. By selecting the arm with the highest upper confidence bound, the algorithm promotes the exploration of arms that possess uncertain reward estimates, leading to a more comprehensive exploration-exploitation trade-off (Radović & Erceg, 2021) [6].

Mathematically, the UCB algorithm can be expressed as follows:

For each arm i, calculate the upper confidence bound (UCB_i) using the formula:

$$\text{UCB}_i = \bar{X}_i + \sqrt{\frac{2\ln(N)}{n_i}}$$

where $\bar{X}_i$ is the average reward obtained from arm i so far, N is the total number of rounds played, and n_i is the number of times arm i has been selected. Then, select the arm with the highest UCB_i value.

By employing the UCB algorithm, the bandit problem solver leverages confidence intervals to estimate upper bounds, prioritizing arms that exhibit higher uncertainty in their reward estimates.

3) Uncertainty-Based Confidence-Bound Reinforcement Learning (UCBRL): Uncertainty-Based Confidence-Bound Reinforcement Learning (UCBRL) is a reinforcement learning approach that combines the principles of Upper Confidence Bound (UCB) algorithms with uncertainty estimation techniques. It aims to balance exploration and exploitation in reinforcement learning problems by incorporating uncertainty measures into the decision-making process. UCBRL estimates the upper confidence bound of the expected value of each action and selects actions based on these bounds, taking into account the uncertainty associated with each estimate [7].

The UCBRL algorithm can be summarized as follows:

For each action i, calculate the upper confidence bound (UCB_i) using the formula:

$$UCB_i = Q_i + c \cdot \sigma_i,$$

where Q_i is the estimated value of action i, σ_i is the uncertainty or standard deviation associated with the estimate, and c is a tunable exploration parameter. Then, select the action with the highest UCB_i value.

By incorporating uncertainty measures in the form of σ_i, UCBRL encourages exploration of actions with higher uncertainty, allowing the agent to gather more information and refine its estimates. The exploration parameter, c, controls the balance between exploration and exploitation, enabling the agent to adapt its behavior based on the specific problem and requirements.

4) Thompson Sampling: Thompson Sampling is a widely used probabilistic algorithm in multi-armed bandit problems, which effectively balances exploration and exploitation by leveraging Bayesian inference. This algorithm assigns a probability distribution to each arm based on prior beliefs and observed rewards. The selection of an arm is then determined by drawing samples from these probability distributions, with arms having higher probabilities being more likely to be chosen [8].

The Thompson Sampling algorithm can be summarized as follows:

1) For each arm i, maintain a probability distribution P_i, representing the belief or uncertainty about the true reward distribution of arm i.
2) Sample a reward parameter θ_i from the probability distribution P_i for each arm i.
3) Select the arm with the highest sampled reward parameter θ_i.
4) Observe the actual reward obtained by playing the selected arm.
5) Update the probability distribution P_i for the selected arm i based on the observed reward, using Bayesian inference techniques.
6) Repeat steps 2-5 for subsequent rounds.

By sampling reward parameters from the probability distributions, Thompson Sampling incorporates uncertainty into the decision-making process. This allows for exploration of arms with higher uncertainty while exploiting arms that are believed to have higher expected rewards based on the probability distributions. Over time, as more observations are made, the

algorithm adapts its beliefs and tends to converge towards selecting the arms with the highest true rewards.

C. Modeling

The conversion rates of different advertising arms is calculated by grouping the dataset by the 'Region' column and taking the mean of the 'Revenue' column. TABLE I displays the results of calculating the conversion rates, showing that Ad9 has the highest conversion rate. Ideally, the agent's decision functions should allocate a majority of opportunities to Ad9 to maximize the revenue. Subsequently, this criterion will be used to evaluate the effectiveness of the agent's decision functions.

TABLE I: The conversion rates of Ads

Bandits	Conversion Rates	Reward Ranking
Ad1	0.1613	4
Ad2	0.1655	2
Ad3	0.1452	7
Ad4	0.1481	6
Ad5	0.1635	3
Ad6	0.1391	8
Ad7	0.1564	5
Ad8	0.1290	9
Ad9	0.1683	1

IV. Results/Analysis

The allocation of arms is observed using different agent's decision functions. These decision functions determine how the agent selects which arm to pull based on the available information. By applying various decision functions, we can examine how the arms are allocated and the impact on the overall performance of the multi-armed bandit system.

Figure 1 illustrates the selection frequencies of different arms by various agent's decision functions. As mentioned earlier, it can be observed that the ϵ-greedy algorithm yields the best results, allocating the highest number of selections to Ad9.

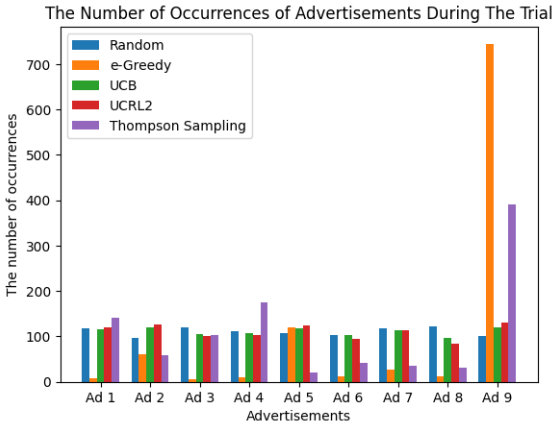

Fig. 1: The Number of Occurences of Advertisements During The Trial

The regret of different algorithms can be calculated by comparing the cumulative reward obtained by each algorithm with the cumulative reward that could have been achieved by always selecting the best arm. A lower regret indicates better performance of the algorithm in terms of maximizing rewards. By analyzing the regret of different algorithms, we can assess their effectiveness in balancing exploration and exploitation and making optimal decisions.

According to Figure 2, when comparing the regret of different algorithms, it was observed that the ϵ-greedy algorithm consistently exhibited lower regret compared to the other algorithms. This indicates that the ϵ-greedy algorithm was more effective in minimizing the difference between the cumulative rewards obtained and the optimal rewards that could have been achieved. Thompson Sampling showed relatively lower regret values, suggesting their ability to make better decisions and explore more rewarding options. On the other hand, the UCB algorithm and UCBRL had higher regret, indicating that they may have been more cautious or conservative in its decision-making process, resulting in missed opportunities for higher rewards.

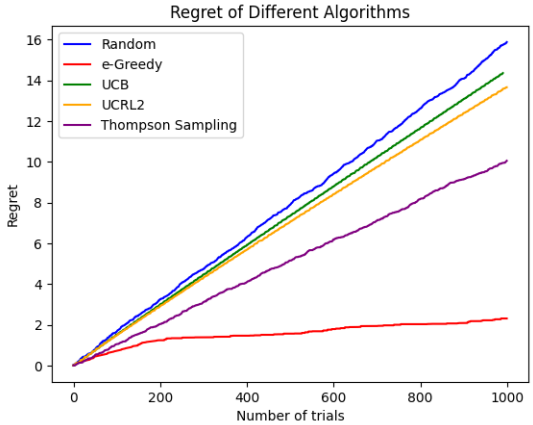

Fig. 2: Regret of Different Algorithms

The total reward achieved by different algorithms refers to the cumulative sum of rewards obtained by each algorithm throughout the experimentation process. It provides a measure of the overall performance of the algorithms in terms of maximizing rewards. By comparing the total rewards of different algorithms, we can assess their effectiveness in selecting arms and optimizing the outcome. A higher total reward indicates better performance and success in maximizing the rewards in the given context.

The results of the experimentation showed in Figure 3 that the ϵ-greedy algorithm achieved the highest total reward among the tested algorithms. This indicates that the ϵ-greedy algorithm was successful in selecting arms that led to the highest cumulative rewards. Thompson Sampling, UCBRL, and UCB algorithm followed with progressively lower total rewards. Although they achieved lower rewards compared to ϵ-greedy, they still outperformed random selection in terms of making more effective decisions and generating more favorable outcomes.

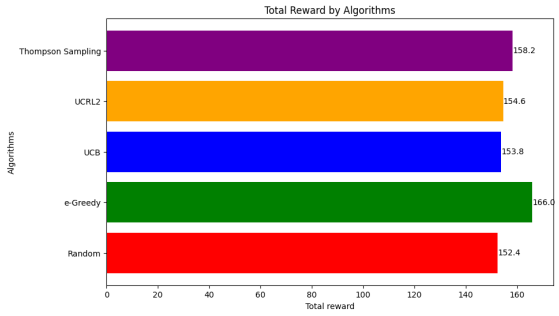

Fig. 3: Total Reward by Algorithms

V. Discussion

In the study, it was found that the ϵ-greedy algorithm performed the best among the different algorithms tested, followed by Thompson Sampling, Uncertainty-Based Confidence-Bound Reinforcement Learning (UCBRL), and Upper Confidence Bound (UCB) algorithm. This ranking raises interesting points for discussion, and several possible reasons can be explored:

1) Exploration and exploitation trade-off: The ϵ-greedy algorithm balances exploration and exploitation by choosing between the action with the highest estimated reward (exploitation) and a random action (exploration) with a certain probability. This balanced approach may have allowed the algorithm to discover and exploit the most rewarding arms more effectively.

2) Optimistic bias: UCB algorithm, which ranked last in the study, tends to have an optimistic bias by overestimating the potential rewards of each arm. This bias may have led to suboptimal decisions in certain scenarios, resulting in lower performance compared to other algorithms.

3) Robustness to uncertainty: Thompson Sampling, which ranked second, is known for its robustness to uncertainty. It samples arms based on their probability of being the best and updates these probabilities based on the observed rewards. This adaptive nature may have allowed Thompson Sampling to adapt well to the changing environment and make effective decisions.

VI. Conclusion

According to the experimental results, the ϵ-greedy algorithm demonstrates superior optimization capabilities compared to other Agent's decision functions. achieving a significantly higher reward value of 166.0 compared to other models. The ϵ-greedy algorithm performs well compared to other Agent's decision functions because it strikes a balance between exploration and exploitation. The algorithm chooses the action with the highest estimated value most of the time (exploitation), but also explores other actions with a small probability ϵ (exploration). This allows the algorithm to exploit the best-known actions while continuously exploring potentially better actions.

VII. Acknowledgements

I would like to express our gratitude to the University of California, Irvine (UCI) for providing the necessary resources and support for this project. The availability of the UCI dataset has been instrumental in conducting our research and analyzing the results.

I would like to express my sincere gratitude to the SCC462 course at Lancaster University for offering a remarkably flexible environment, enabling profound exploration of the applications and far-reaching extensions of artificial intelligence. The invaluable opportunity provided by this course has significantly enriched my understanding and appreciation of this field.

References

[1] R. C. Gatti, "A multi-armed bandit algorithm speeds up the evolution of cooperation," *Ecological Modelling*, vol. 439, p. 109348, 2021.

[2] N. Silva, H. Werneck, T. Silva, A. C. Pereira, and L. Rocha, "Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions," *Expert Systems with Applications*, vol. 197, p. 116669, 2022.

[3] D. I. Mattos, J. Bosch, and H. H. Olsson, "Multi-armed bandits in the wild: Pitfalls and strategies in online experiments," *Information and Software Technology*, vol. 113, pp. 68–81, 2019.

[4] C. Yan, H. Han, Y. Zhang, D. Zhu, and Y. Wan, "Dynamic clustering based contextual combinatorial multi-armed bandit for online recommendation," *Knowledge-Based Systems*, vol. 257, p. 109927, 2022.

[5] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks," *Neural Computing and Applications*, vol. 31, pp. 6893–6908, 2019.

[6] N. Radović and M. Erceg, "Hardware implementation of the upper confidence-bound algorithm for reinforcement learning," *Computers & Electrical Engineering*, vol. 96, p. 107537, 2021.

[7] M. Kamiura and K. Sano, "Optimism in the face of uncertainty supported by a statistically-designed multi-armed bandit algorithm," *Biosystems*, vol. 160, pp. 25–32, 2017.

[8] A. Dzhoha and I. Rozora, "Multi-armed bandit problem with online clustering as side information," *Journal of Computational and Applied Mathematics*, vol. 427, p. 115132, 2023.