Uncertainty-aware contextual multi-armed bandits for recommendations in e-commerce

Anantharaman Subramani, Niteesh Kumar, Arpan Dutta Chowdhury, Ramgopal Prajapat
Department of Data Science and Analytics, Tata Unistore Limited, Mumbai, India

Article Info

Article history:

Received Aug 27, 2024
Revised Feb 5, 2025
Accepted Mar 15, 2025

Keywords:

E-commerce
Multi-arm bandits
Personalization
Recommendations
Reinforcement learning

ABSTRACT

The growing e-commerce landscape has seen a shift towards personalized product recommendations, which play a critical role in influencing consumer behavior and driving revenue. This study explores the efficacy of contextual multi-armed bandits (CMAB) in optimizing personalized recommendations by intelligently balancing exploration and exploitation. Recognizing the inherent uncertainty in user behaviors, we propose an enhanced CMAB policy that incorporates item correlation matrix as an additional component of uncertainty to the conventional binary exploration and exploitation setup of bandit policies. Our approach aims to increase the overall relevance of recommendations through the 'triadic framework' of CMAB, that seamlessly integrates with existing bandit policies, enabling adaptive recommendations based on diverse user attributes. By outperforming traditional models, this uncertainty-aware method demonstrates its potential in refining recommendation accuracy, thus maximizing revenue in a competitive e-commerce environment. Future research will explore dynamic uncertainty modeling and cross-domain applications to further advance the field.

Corresponding Author:

Anantharaman Subramani
Department of Data Science and Analytics, Tata Unistore Limited
Mumbai, Maharashtra, India
Email: sar.anantharaman11@gmail.com

1. INTRODUCTION

Recommendation systems have become integral to enhancing user engagement and driving revenue across various industries, more specifically for e-commerce domain. For instance, Amazon attributes approximately 35% of its e-commerce revenue to product recommendations, while Netflix reports that 75% [1] of viewer activity is driven by personalized suggestions. Within the dynamic landscape of recommendation systems, a fascinating journey unfolds, marked by a relentless pursuit of precision and personalization.

Traditional paradigms, exemplified by collaborative filtering (CF) models, have seen triumphs but not without facing intricate challenges. Strategies like "people who bought X also bought Y" or "people like you liked X" have been the bedrock of recommendations, shaping the way users discover products and content [2]. Yet, the journey has been far from linear. The Achilles' heel of CF lies in the accuracy-diversity dilemma and the sparsity of interaction data [3], [4]. To grasp the magnitude of this challenge, consider the vast expanse of online platforms where users engage with myriad products and services. Amazon, for instance, boasts millions of products, and Netflix hosts an extensive library of movies and TV shows. Navigating through this sea of options requires not just accuracy but also diversity to cater to the eclectic tastes of users [5], [6]. To address these challenges, recommender systems have embraced auxiliary information. For instance, incorporating item attributes such as genre, price, or style has significantly enhanced recommendation quality [7], [8]. This diversification strategy mitigates the data sparsity issue inherent in CF by enriching the recommendation process with additional facets [9].

Drifting away from traditional paradigms, the enhancements in recommendation channels include the infusion of contextual information to the system inputs, along with the adoption of auxiliary information in form of product attributes and user contexts where the latter act as a key differentiator [10]. Context-aware recommendations, leveraging user-specific details like location, device, or time of day, have demonstrated remarkable efficacy [11]. Knowledge graph-based recommender systems represent the foundational chapter in this narrative. Companies like Alibaba have pioneered the integration of knowledge graphs to enhance the understanding of item relationships. By mapping entities and their relations, knowledge graphs offer a structured framework [12], [13] and this approach is supported by studies on the impact of knowledge graphs [14], [15], and it has exhibited promise in reinforcing the knowledge representation of users and items. As we navigate through the dynamic realm of online interactions, the limitations of static recommendation models become apparent and also the construct of knowledge-graphs lends to computationally intensive operations. The crux lies in their inability to capture or extend itself to the evolving nature of user preferences [16], [17]. Research on dynamic user preferences [18], [19] reveals that static CF models often fall short in accurately predicting users' changing tastes.

Enter the realm of exploration-exploitation methods, prominently illustrated by the multi-armed bandits (MAB) of reinforcement learning (RL) frameworks warrants a conducive pipeline to include contextual information of users to the recommendation ecosystem. The application of bandit algorithms in dynamic domains like news recommendations has been transformative [20], [21] and companies like Spotify have embraced bandit-based approaches to fine-tune their recommendation engines [22]. The essence of MAB lies in dynamically balancing exploration and exploitation through the estimation of ‘Reward’ for the generated recommendations at time (t), thereby, adapting to the ever-shifting landscape of user preferences which holds as a bottleneck in traditional CF techniques and in the construction of knowledge-graph based techniques. Contextual multi-armed bandits (CMAB), an extension of MAB, amplifies the sophistication by incorporating contextual information. Studies comparing traditional MAB with CMAB have found that CMAB outperforms its counterpart by considering the dynamic aspects of user context [23], [24] in the bandit policies.

In this evolutionary journey, a notable milestone is the knowledge-enhanced linear upper confidence bound (LinUCB) [25], here referred to as similarity-LinUCB. Imagine a scenario where the dynamism of CMAB converges with the knowledge-rich environment of a graph and this amalgamation [26] unlocks new dimensions in recommendation systems. Similarity-LinUCB captures intricate high-order connectivity in form of action-response mapping (ARM) as proposed in this research, transcends the limitations of static models [27]. In essence, the introduction of similarity-LinUCB signifies a paradigm shift a departure from static models to dynamic, uncertainty-aware recommendation strategies through the similarity component added to conventional LinUCB policies of CMAB. The forthcoming sections unravel the layers of the proposed approach, providing a deep dive into its background, the motivation fuelling its development, and the experimental substantiation of its prowess. Through meticulous experimentation, the effectiveness of similarity-LinUCB is showcased, setting a benchmark in the dynamic landscape of recommender systems. The narrative unfolds not just as a chronicle of evolution but as a testament to the resilience, adaptability, and user-centricity inherent in the realm of recommendations.

2. METHODOLOGY

The objective of this research paper is to innovate within the realm of RL by introducing an intermediate stage in bandit policies, transitioning from traditional exploitation-exploration strategies to a more nuanced exploitation-informed exploration-exploration framework, a ‘triadic framework’. This three-stage concept aims to enhance recommendation systems, particularly in less popular context scenarios, where conventional approaches may fall short. By leveraging the strengths of each stage, the proposed framework optimizes recommendations across a spectrum of contexts. In popular contexts, the exploitation component maintains relevance, while the exploration component accommodates new or unseen contexts. For less popular contexts, the introduction of an informed exploration component, driven by action similarity scores or ARM, facilitates informed decision-making, thereby enhancing recommendation accuracy. Through empirical evaluation and analysis, the research seeks to demonstrate the efficacy of this novel framework in optimizing recommendation systems and improving user satisfaction across diverse contexts and scenarios. Through the development and evaluation of an uncertainty-aware CMAB approach, the research endeavors to optimize customer engagement metrics, ultimately maximizing revenue and sustaining a competitive edge in the burgeoning e-commerce market. Addressing the scalability challenges inherent in traditional recommendation systems, particularly with non-stationary data and seasonal fluctuations, remains a central focus. The introduction of the 'informed exploration' component implicitly aids in overcoming these challenges, offering improved scalability compared to the conventional LinUCB setup in CMAB.

The key goals of the proposed research are as follows. i) introduce an approach that strategically incorporates a product similarity matrix into the CMAB framework to add an additional layer of uncertainty and enhance the ability to capture diverse user cohorts. ii) propose a method that addresses scalability challenges witnessed in previous bandit-based recommendation studies and is suitable for real-world e-commerce applications dealing with large product catalogues and non-stationary user behavior data. iii) evaluate the proposed uncertainty-aware CMAB approach and demonstrate through experimentation that it outperforms baseline method, i.e., conventional LinUCB setup in optimizing key metrics like customer engagement, satisfaction, and business outcomes like revenue for e-commerce platforms; and iv) validate the effectiveness of leveraging product similarity information within a bandit-based recommendation model for capturing the nuances of individual consumer choices in a personalized yet uncertain recommendation environment.

2.1. Conventional CMAB setup

Traditional recommendation systems struggle to effectively promote the discovery of new products due to their reliance on past user interactions, which often leads to reinforcing similar suggestions, i.e., over exploitation on historical performances and overlooking novel choices, i.e., indifferent to exploration of new actions. Random exploration, while a potential solution, can result in poor user experience. Contextual bandit algorithms offer a solution by intelligently balancing exploration and exploitation using varied bandit policies (π). These policies present users with multiple options, "arms" (a), and learn from feedback through reward (R) estimations, for generated recommendation in each trial (t), adapting recommendations over time to prioritize options likely to yield higher rewards while exploring new possibilities. Figure 1 depicts the conventional CMAB setup in production environment where customer/user interactions are captured at regular intervals (such as hourly, daily, weekly or monthly intervals) and persisted in fast-access databases (such as Hive) for speedy retrieval. The conventional CMAB setup is formulated in (1).

$$\pi^*(x) = \arg\max_a E[R|X = x, A = a] \quad (1)$$

Where $\pi^*(x)$ is the policy chosen, $arg\ max_a$ denotes the action a that maximizes the expected reward, and E[R|X=x, A=a] represents the expected reward R given context X=x and action A=a.

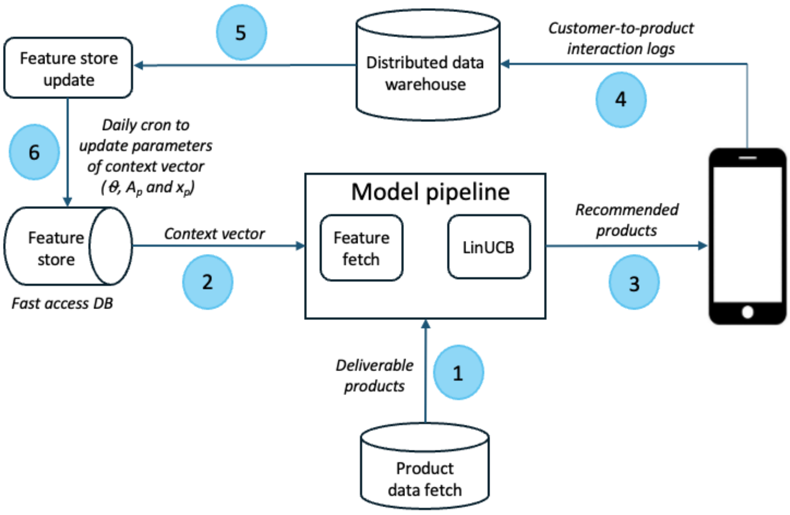

Figure 1. Conventional contextual-multi arm bandit setup

2.2. LinUCB based CMAB

In this study, LinUCB algorithm is employed as the primary policy for the CMAB setup. Since LinUCB is a linear contextual bandit algorithm, the expected reward is modelled as a linear function of the context vectors and uses upper confidence bounds to balance exploration and exploitation component.

Any action (a) represents in-stock and non-delisted products to be recommended, and our objective is to maximize cumulative reward and compare regret between the conventional LinUCB approach and our novel approach incorporating informed exploration component.

For each trial, t, LinUCB policy scores every action and chooses the action that maximizes the score. There are 2 components to the LinUCB policy, both formulated in (2) and (3). In (2) denotes the mean reward estimate of each arm (a), and in (3) denotes the corresponding standard deviations attributing the confidence bounds for the quantum of exploration.

$$\widehat{r_{t,a}} = x_{t,a}^{T}\theta_{a} \tag{2}$$

$$\widehat{c_{t,a}} = \alpha\sqrt{x_{t,a}^{T}A_{a}^{-1}x_{t,a}} \tag{3}$$

Where θ and A are model parameters equivalent to weights vector and context-action covariance matrix representations, x is the context vector, α is the hyperparameter, higher the value gives more emphasis on exploration component.

The model parameters θ and A are updated in accordance with the extracted rewards after each time-step, t and the update cycle of LinUCB is formulated in (4) and (6).

$$A_{a_t} = A_{a_t} + x_{t,a_t}.x_{t,a_t}^{T} \tag{4}$$

$$b_{a_t} = b_{a_t} + r_t.x_{t,a_t} \tag{5}$$

$$\hat{\theta}_a = A_a^{-1}b_a \tag{6}$$

Final, empirical estimate of LinUCB is mathematically represented in (7) with reward estimate and confidence bound parameters are summed together to derive the final score for each action, (a) for the given time-step, (t).

$$a_t \stackrel{\text{def}}{=} arg \max_{a\epsilon A_t}(\widehat{r_{t,a}} + \widehat{c_{t,a}}) \tag{7}$$

2.3. Uncertainty-aware CMAB setup

The proposed approach of this study incorporates similar products identified as an additional component in LinUCB policy. By combining LinUCB policy with our similarity matrix, we aim to provide more personalized and diverse recommendations attributed to impact the different user contexts, such as new/popular/less-popular segments. We quantify contextual information with features such as age, gender, user preferences and location derivatives to enhance recommendation accuracy and all the contextual features are chosen in such a way that they are applicable across category spectrum in the product catalog.

To enhance exploration in contextual bandits, especially in contexts with less historical data, we integrate our product similarity matrix. This matrix is derived by identifying similar products using bidirectional encoder representations from transformers-based embeddings, i.e., BERT, generated from product attributes, as denoted in (8) and top k similar products are selected, product (p) here translates to actions (a) in CMAB setup. This enables the identification of products similar to those previously purchased by users, allowing for intelligent exploration of new recommendations while considering user preferences and behavior.

$$S(i,j) = similarity(BERT(p_i), BERT(p_j)) \tag{8}$$

Where S(i, j) denotes similarity between products p_i and p_j and BERT(p) represents the BERT embedding of product p.

For less popular contexts, the similarity matrix is utilized to assign higher score to the set of actions in similarity matrix that are mapped to current action a_t. This approach ensures that we make an informed exploration of actions that are likely to yield higher rewards based on product similarity. For very popular contexts, the standard exploitation component of LinUCB is used, leveraging historical data and in completely unseen contexts, the exploration component of LinUCB allows the model to recommend new actions, course correct basis the rewards generated as feedback for recommendations and thereby improve the recommendation quality in future. The transformed LinUCB policy pertained to this study is formulated in (9).

$$a_t \stackrel{\text{def}}{=} arg \max_{a \epsilon A_t} (\widehat{r_{t,a}} + \widehat{c_{t,a}} + \beta \sum_{i=0}^{k} S(a_t, a_i')) \tag{9}$$

Where β is the hyperparameter for similarity term and $\sum_{i=0}^{k} S\left(a_{t,} a_i'\right)$ represents the summation of similarity scores between action a_t and k-most similar actions a_i'.

This enhanced LinUCB policy as formulated above combines both contextual information and product similarity to make more informed recommendations leading to potentially higher user satisfaction and engagement. Thereby, the proposed approach addresses all possible contextual states, such as i) new contexts; ii) less-popular contexts; and iii) popular contexts, leading to three novel states of CMAB: exploitation, exploration, and informed exploration. The similarity component introduced in the updated LinUCB equation adds the new informed exploration state to the existing exploitation and exploration states and the proposed architecture setup of CMAB is depicted in Figure 2.

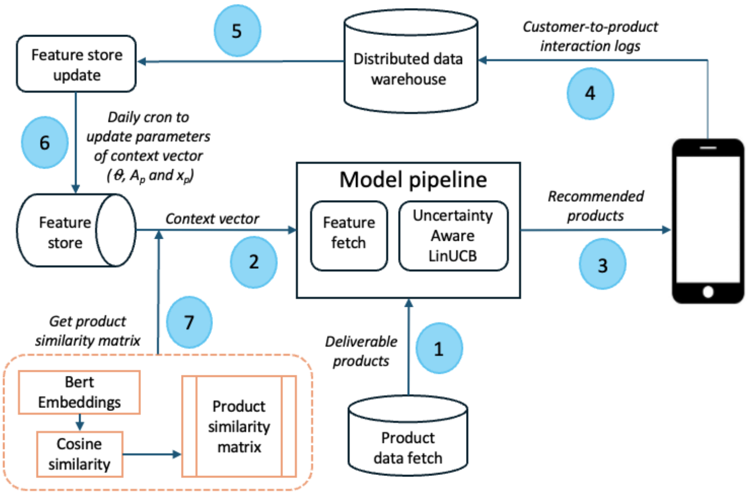

Figure 2. Uncertainty-aware LinUCB setup (proposed framework)

2.4. Experimentation setup

To evaluate the effectiveness of our proposed approach, the experimentation process focused on quantifying contextual information, defining reward metrics, finalizing actions and comparing regret between the conventional LinUCB setup and our novel approach of uncertainty-aware LinUCB setup. We utilized various user attributes, including age, gender, user profile attributes such as issues faced and preferences, along with location derivatives, to quantify contextual information. After multiple iterations and consolidating feedback from corresponding business verticals, we were left with 21 distinct user contexts that ensures comprehensive coverage of user demographics and preferences within 3 shortlisted categories. In addition to contextual feature engineering, we also employed tuning of hyperparameters pertained to RL techniques. The reward metric selected for this research was the 'purchase rate.' Each time-step, t, was equated to a daily interval, during which the user's purchases of the generated recommendations were recorded, and the associated rewards were stored in a fast-access database (here, Hive). Instead of considering the quantity of purchases, we focused on whether the recommended product was purchased or not, assigning a value of 1 if purchased and 0 otherwise. Actions as stated in previous section denotes in-stock and non-delisted products across three best-selling categories in focus for experimentation and analysis: i) apparel: kurta and kurtis; ii) beauty: perfumes; and iii) footwear: running shoes. The primary objective of our experimentation was to minimize cumulative regret while also quantifying the gain achieved through the introduction of the 'informed exploration' component to the conventional LinUCB policy. The three categories i.e. apparel, footwear and beauty were chosen for following reasons:

- Relevance to e-commerce: apparel, footwear, and beauty products are popular categories in 'TATA UNISTORE' platform both in terms of volume of products being sold as well as revenue generated.

- Diversity of products: these categories encompass a wide range of products with varying attributes, styles, and preferences. This diversity allows for testing the effectiveness of recommendation algorithms in catering to different user tastes and preferences.
- Varied consumer behavior: consumers exhibit diverse behavior when shopping for apparel, footwear, and beauty products. Some may prefer well-known brands, while others may prioritize specific features or price points. This diversity in consumer behavior provides a rich dataset for evaluating recommendation algorithms.

Cumulative regret as quantified in (10) serves as a metric to quantify the effectiveness of the recommendations by measuring the opportunity loss incurred due to suboptimal recommendations. This metric was then utilized to determine the purchase rate, reflecting the proportion of users who made a purchase based on the recommendations provided. This experimentation process was repeated for both the conventional LinUCB method and our proposed novel approach of LinUCB, enabling a comparative analysis of recommendation performance across different scenarios and user contexts shortlisted.

$$\mathrm{R_k} = \sum_{\mathrm{i=1}}^{\mathrm{k}} (\text{Optimal Reward} - \mathrm{r_i}) \tag{10}$$

Where R_k is the cumulative regret for k recommendations, optimal reward is the reward that would have been obtained with best possible recommendation, and r_i is the reward obtained for i-th recommendation.

3. RESULTS AND DISCUSSION

The experimental results demonstrate the efficacy of the proposed uncertainty-aware LinUCB approach in improving recommendation performance across different product categories. The time-step, t, is set to 10 iterations, highlighting the evaluation of recommendation outputs with 10 days of customer interactions spanning weekends and weekday sales, allowing to consider diverse customer purchase behavior patterns. The configured value of 10 is set to enable faster experimentation.

3.1. Key findings

Table 1 shows the experimental findings that underscore the effectiveness of our proposed uncertainty-aware LinUCB approach in enhancing recommendation performance within the same cohort across different product categories. Lower the cumulative regret implies better recommendation efficiency and thereby promoting the overall purchase rate for 3 chosen categories. By leveraging informed exploration and contextual information, our methodology provides a robust solution to enhance user experience and drive sales on the TATA UNISTORE platform. The reduction in cumulative regret observed for the proposed approach can be attributed entirely to the informed exploration component of the uncertainty-aware LinUCB policy, as it is the only distinguishing parameter compared to the baseline LinUCB policy.

Table 1. Evaluation metrics for baseline (LinUCB) vs proposed (uncertainty aware LinUCB) approach

Category	Approach	Cumulative regret @10	Purchase rate (%)
Footwear-running	LinUCB (baseline)	209	1.98
	Uncertainty-aware LinUCB	197	6.19
Beauty-perfumes	LinUCB (baseline)	202	2.37
	Uncertainty-aware LinUCB	176	16.19
Apparel-kurta and kurtis	LinUCB (baseline)	210	0.68
	Uncertainty-aware LinUCB	203	3.33

3.2. Category level result interpretation

In the footwear-running shoes category, the uncertainty-aware LinUCB approach demonstrated a substantial reduction in cumulative regret from 209 to 197, compared to the conventional LinUCB method and purchased rate increased notably from 1.98 to 6.19% indicating an efficient exploration-exploitation trade-off and driving better user engagement and conversion. The beauty-perfumes category showed the most remarkable performance, with significant decrease in cumulative regret from 202 to 176, reflecting a more optimized recommendation strategy. Moreover, the purchase rate increased from 2.37 to 16.19%, indicating a substantial improvement in user purchases. The informed exploration component proved effective in recommending products to less-popular contexts, leveraging limited purchase data to drive conversions.

The proposed approach demonstrated marginal improvements for apparel-kurta and kurtis, with cumulative regret decreasing from 210 to 203, and purchase rate too saw a modest increase from

0.68 to 3.33%. These results highlight the method's ability to improve conversions, albeit less pronounced compared to other categories. Comparatively, smaller improvement in this category underscores a need for further evaluation of recommendation strategies specific to apparel. Refining the product similarity matrix and/or devising category-specific strategies to generate contextual information for categories like Apparel could lead to greater engagement and improved conversion metrics. Across all 3 categories, the proposed approach consistently minimized cumulative regret and improved purchase rate, reassures the impact of 'Informed exploration component' and its ability to enhance user engagement and drive conversions. However, targeted refinements in categories with marginal gains could further elevate performance and ensure consistent improvements across all spectrums.

4. CONCLUSION

Based on the comprehensive research presented in this paper, it is evident that the e-commerce landscape is undergoing a significant transformation driven by personalized product recommendations. Leveraging advanced algorithmic frameworks such as CMAB is crucial for optimizing recommendation systems in a dynamic setting. The introduction of the uncertainty-aware CMAB approach, which incorporates a product similarity matrix to enhance recommendation efficiency by improving the purchase rate, marks a significant gain for our platform's revenue. As e-commerce sales continue to surge, the importance of uncertainty-aware CMAB in personalized recommendations cannot be overstated. Furthermore, this research implicitly addressed the scalability bottlenecks faced with knowledge graphs since the proposed approach functions at context level and not at user level where the latter is more granular leading to exponential increase in computational complexities. However, for use-cases requiring granular contextual information may still encounter scalability bottlenecks. Future work could focus on tactical refinements at the category level, specifically in generating more tailored contextual information that derives nuanced and distinct user contextual states. Additionally, leveraging the large language model (LLM)-based text embeddings to construct a more accurate product similarity matrix presents a promising avenue for enhancing the system's overall performance and efficiency. These enhancements have the potential to further improve user engagement and conversion metrics across diverse categories.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Anantharaman Subramani	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		
Niteesh Kumar		✓	✓	✓	✓			✓	✓					
Arpan Dutta Chowdhury				✓		✓			✓	✓	✓			
Ramgopal Prajapat				✓	✓	✓	✓			✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Derived data supporting the findings of this study are available from the corresponding author, AS, upon reasonable request. However, restrictions apply to the availability of these data, which were used under license from our organization and are not publicly available.

REFERENCES

[1] M. Hejazinia, K. Eastman, S. Ye, A. Amirabadi, and R. Divvela, “Accelerated learning from recommender systems using multi-armed bandit,” *arXiv-Computer Science*, pp. 1-8, Aug. 2019.

[2] L. Zhou, “A survey on contextual multi-armed bandits,” *arXiv-Computer Science*, Aug. 2015.

[3] A. Pilani, K. Mathur, H. Agrawald, D. Chandola, V. A. Tikkiwal, and A. Kumar, “Contextual bandit approach-based recommendation system for personalized web-based services,” *Applied Artificial Intelligence*, vol. 35, no. 7, pp. 489–504, Jun. 2021, doi: 10.1080/08839514.2021.1883855.

[4] Q. Shi, F. Xiao, D. Pickard, I. Chen, and L. Chen, “Deep neural network with linucb: a contextual bandit approach for personalized recommendation,” in *Companion Proceedings of the ACM Web Conference 2023*, New York, USA: ACM, Apr. 2023, pp. 778–782. doi: 10.1145/3543873.3587684.

[5] L. Li, W. Chu, J. Langford, and R. E. Schapire, “A contextual-bandit approach to personalized news article recommendation,” in *Proceedings of the 19th international conference on World wide web*, New York, USA: ACM, Apr. 2010, pp. 661–670. doi: 10.1145/1772690.1772758.

[6] N. Aramayo, M. Schiappacasse, and M. Goic, “A multi-armed bandit approach for house ads recommendations,” *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4107976.

[7] H. Singh, S. Yadav, A. K. Banyal, and S. N. Deshpande, “Recommendations engine with multi-objective contextual bandits (using reinforcement learning) for e-commerce,” *International Research Journal of Engineering and Technology*, vol. 7, no. 4, 2020.

[8] M. Gan and O.-C. Kwon, “A knowledge-enhanced contextual bandit approach for personalized recommendation in dynamic domains,” *Knowledge-Based Systems*, vol. 251, p. 109158, Sep. 2022, doi: 10.1016/j.knosys.2022.109158.

[9] I. Manickam, A. S. Lan, and R. G. Baraniuk, “Contextual multi-armed bandit algorithms for personalized learning action selection,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Mar. 2017, pp. 6344–6348. doi: 10.1109/ICASSP.2017.7953377.

[10] A. Slivkins, “Contextual bandits with similarity information,” *arXiv-Computer Science*, Jul. 2009.

[11] J. Abernethy, E. Hazan, and A. Rakhlin, “Competing in the dark: an efficient algorithm for bandit linear optimization,” in *21st Annual Conference on Learning Theory - COLT 2008*, Omnipress, 2008, pp. 263–274.

[12] R. Agrawal, “The continuum-armed bandit problem,” *SIAM Journal on Control and Optimization*, vol. 33, no. 6, pp. 1926–1951, Nov. 1995, doi: 10.1137/S0363012992237273.

[13] P. Auer, “Using confidence bounds for exploitation-exploration trade-offs,” *Journal of Machine Learning Research*, vol. 3, pp. 397–422, 2002.

[14] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine Learning*, vol. 47, pp. 235–256, 2002, doi: 10.1023/A:1013689704352.

[15] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, “The nonstochastic multiarmed bandit problem,” *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, Jan. 2002, doi: 10.1137/S0097539701398375.

[16] P. Auer, R. Ortner, and C. Szepesvári, “Improved rates for the stochastic continuum-armed bandit problem,” in *Learning Theory*, Berlin, Heidelberg: Springer, 2007, pp. 454–468. doi: 10.1007/978-3-540-72927-3_33.

[17] B. Awerbuch and R. Kleinberg, “Online linear optimization and adaptive routing,” *Journal of Computer and System Sciences*, vol. 74, no. 1, pp. 97–114, Feb. 2008, doi: 10.1016/j.jcss.2007.04.016.

[18] J. S. Banks and R. K. Sundaram, “Denumerable-armed bandits,” *Econometrica*, vol. 60, no. 5, Sep. 1992, doi: 10.2307/2951539.

[19] D. A. Berry, R. W. Chen, A. Zame, D. C. Heath, and L. A. Shepp, “Bandit problems with infinitely many arms,” *The Annals of Statistics*, vol. 25, no. 5, Oct. 1997, doi: 10.1214/aos/1069362389.

[20] S. Bubeck, “Regret analysis of stochastic and nonstochastic multi-armed bandit problems,” *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012, doi: 10.1561/2200000024.

[21] S. Bubeck and R. Munos, “Open loop optimistic planning,” in *COLT 2010 - The 23rd Conference on Learning Theory*, Omnipress, 2010, pp. 477–489.

[22] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari, “Online optimization in x-armed bandits,” in *Twenty-Second Annual Conference on Neural Information Processing Systems*, HAL open science, 2008, pp. 201–208.

[23] S. Bubeck, G. Stoltz, and J. Y. Yu, “Lipschitz bandits without the lipschitz constant,” in *Algorithmic Learning Theory*, Springer, Berlin, Heidelberg, 2011, pp. 144–158. doi: 10.1007/978-3-642-24412-4_14.

[24] S. Bubeck, N. Cesa-Bianchi, and S. M. Kakade, “Towards minimax policies for online linear optimization with bandit feedback,” *arXiv-Computer Science*, pp. 1-15, Feb. 2012.

[25] E. Turğay, D. Öner, and C. Tekin, “Multi-objective contextual bandit problem with similarity information,” *arXiv-Statistics*, pp. 1-12, Mar. 2018.

[26] S. Lipovetsky, “Prediction, learning, and games,” *Technometrics*, vol. 49, no. 2, pp. 225–225, May 2007, doi: 10.1198/tech.2007.s482.

[27] W. Chu, L. Li, L. Reyzin, and R. E. Schapire, “Contextual bandits with linear payoff functions,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, PMLR 15, 2011, pp. 208–21.

BIOGRAPHIES OF AUTHORS

Anantharaman Subramani holds a Bachelor of Engineering (B.E.) in electronics from Madras Institute of Technology, Chennai, India in 2017, PG Diploma (data science) from International Institute of Technology, Bangalore, India in 2019 and Master of Science in data science from Liverpool John Moores University, England in 2020 with the Dissertation “Hybrid approach of sentiment analysis using affective state feature embeddings” respectively. He is currently a Senior Data Scientist at Data Science and Analytics Department in TATA Technologies, Mumbai, India. He is involved with the design and optimization of ranking and recommendations workflows for the web and mobile platforms of the E-commerce application. His research interests are pertained to machine learning, embedding optimizations, reinforcement learning, and deep learning techniques. He can be contacted at email: asubramani@tataunistore.com or sar.anantharaman11@gmail.com.

Niteesh Kumar holds a Bachelor of Technology in Instrumentation and Control Engineering with a minor in Computer Science from the National Institute of Technology, Tiruchirappalli, class of 2022. He is currently a Data Scientist I at Tata Technologies, Mumbai, India. In his role, he is actively engaged in the development and enhancement of ranking algorithms and recommendation systems for both web and mobile platforms in the E-commerce domain. His work focuses on improving user experience and optimizing performance through advanced machine learning techniques. His research interests span across various fields, including machine learning, embedding optimization, reinforcement learning, and deep learning methodologies. He can be contacted at email: nsoundrapandian@tataunistore.com or niteeshkumar2001@gmail.com.

Arpan Dutta Chowdhury holds a Bachelor of Engineering (B.E.) in mechanical engineering from the Jadavpur University, Kolkata, class of 2022 where we were involved in development of models used for self-driving cars and models for finding industry standard equipment from AutoCAD drafts. He is currently a Data Scientist I at Tata Technologies, Mumbai, India. In his role, he is actively engaged in the development and enhancement of ranking algorithms, recommendation systems and visual search for both web and mobile platforms in the E-commerce domain. His work focuses on improving user experience and optimizing performance through advanced machine learning techniques. His research interests span across various fields, including machine learning, embedding optimization, reinforcement learning, and deep learning methodologies. He can be contacted at email: arpandc14@gmail.com or achowdhury@tataunistore.com.

Ramgopal Prajapat is M.Tech. in industrial and management engineering from Indian Institute of Technology, Kanpur, India and currently working as Sr Vice President for Allcargo Group. He has over 20 years of experience in AI, consulting, and entrepreneurship, and brings a wealth of expertise in harnessing data to drive innovation. His background includes managing diverse portfolios, leading teams, and implementing AI solutions across industries such as banking, finance, e-commerce, and logistics. As a co-founder of a startup, Ram thrives in fast-paced environments. He has also been actively engaged in academia, delivering AI and ML sessions at MICA Ahmedabad and IIT Delhi for over five years. He can be contacted at email: ramg_iitk@yahoo.co.in.