

Design Principles of Robust Multi-Armed Bandit Framework in Video Recommendations

BELHASSEN BAYAR* and PHANIDEEP GAMPA*, Amazon Prime Video

AINUR YESSENALINA and ZHEN WEN, Amazon Prime Video

Current multi-armed bandit approaches in recommender systems (RS) have focused more on devising effective exploration techniques, while not adequately addressing common exploitation challenges related to distributional changes and item cannibalization. Little work exists to guide the design of robust bandit frameworks that can address these frequent challenges in RS. In this paper, we propose a new design principles to (i) make bandit models robust to time-variant metadata signals, (ii) less prone to item cannibalization, and (iii) prevent their weights fluctuating due to data sparsity. Through a series of experiments, we systematically examine the influence of several important bandit design choices. We demonstrate the advantage of our proposed design principles at making bandit models robust to dynamic behavioral changes through in-depth analyses. Noticeably, we show improved relative gain compared to a baseline bandit model not incorporating our design choices of up to 11.88% and 44.85%, respectively in ROC-AUC and PR-AUC. Case studies about fairness in recommending specific popular and unpopular titles are presented, to demonstrate the robustness of our proposed design at addressing popularity biases.

CCS Concepts: • **Computing methodologies** → **Online learning settings**; **Regularization**.

Additional Key Words and Phrases: video recommendations, constrained optimization, temporal signals

ACM Reference Format:

Belhassen Bayar, Phanideep Gampa, Ainur Yessenalina, and Zhen Wen. 2023. Design Principles of Robust Multi-Armed Bandit Framework in Video Recommendations. 1, 1 (September 2023), 9 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Capturing users' interests that change over time is a key point in personalized video recommendations. This is because users' interests in recommender systems (RS) are typically time and catalog dependent, which can potentially make recommendations stale. This becomes very challenging when the promotional basket is organically updated with a set of entering and exiting titles; each associated with different degrees of popularity and release recency that evolve over time. In order to address recommendation staleness in RS with respect to both user's and title's behavioral changes, researchers developed numerous algorithms to solve exploration/exploitation in multi-armed bandit problems [4]. In video recommendation settings, the objective is to select for a user the arm (title) with the highest expected reward while exploring other arms to gain information about their expected rewards. Contextual bandit models are cheap to train, and have shown promising capabilities in a continual incremental training setting [4, 5, 13, 15].

*Both authors contributed equally to this research
Copyright held by the author(s).

Authors' addresses: Belhassen Bayar, bayarb@amazon.com; Phanideep Gampa, phanide@amazon.com, Amazon Prime Video; Ainur Yessenalina, yessenal@amazon.com; Zhen Wen, zhenwen@amazon.com, Amazon Prime Video.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.
Manuscript submitted to ACM

While most of the bandit research in RS has focused on developing exploration techniques to address several challenges associated with recency and distributional changes [6, 7, 18], there is little research work that specifically focused on addressing exploitation related bandit issues. However, exploitation issues are very common in RS in general, including bandit models, such as distributional and categorical biases induced by data sparsity or lack of granular representation of both title and user entities [1, 8, 10, 16]. Specifically, because of their small size and linear nature, linear bandit model parameters in continual incremental training setting can be associated with abrupt and non-smooth changes along with biases towards recommending more popular titles or categories defined through the input contextual signals. Additionally, many behavioral and time-variant metadata signals in RS can change dramatically during title promotion initial phase, creating distributional discrepancy in the user and/or title representation between training and inference phases which can drastically impact a bandit model performance.

In this paper, we present our design principles to tackle common exploitation challenges in building a multi-armed bandit production framework for video recommendations. We systematically investigate three common known issues in bandit models, and RS in general while proposing a solution for each of the challenges: (1) A novel data augmentation (DA) solution in RS inspired by computer vision research [12] that can help the bandit model to generalize to different distributional changes of time-variant metadata such as title recency. DA is also designed to address sparsity issues associated with this category of time-variant metadata. (2) Temporal title signals that can model an arm performance over consecutive time slots to overcome cannibalization issues due to lack of granular representation of an arm when multiple titles share similar contextual metadata information. This type of signal can also help the bandit model to quickly adapt to the very recent changes in the viewership trends. (3) L_2 smoothing of bandit’s parameter weights to attenuate their fluctuations and abrupt changes due to data sparsity in a continual incremental training setting. Our smoothing method incorporates an adaptive regularization techniques where the smoothing parameters are learned from the bandit historical weights, and is designed to prevent popularity biases towards a title or a category of titles.

Through a set of experimental results, we demonstrate the advantage of our proposed design principles to make a multi-armed bandit framework robust to common dynamic challenges encountered in RS; and particularly more frequent in the context of promotional catalogue with cold start titles. We thoroughly evaluate our proposed design choices through traditional AUC scores, along with a customized nDCG (normalized discounted cumulative gain) ranking metrics that can be leveraged in any bandit offline setting. The new designed framework typically outperforms a baseline bandit model used in production over 40 consecutive offline runs. Noticeably, the proposed model can achieve 11.88% and 44.85% relative gain compared to baseline, respectively in ROC-AUC and PR-AUC.

2 Related Work

The existing contextual bandit frameworks balance exploration/exploitation trade offs either by updating the confidence intervals in UCB based algorithms [7] or by updating the priors in Thompson sampling based algorithms [6, 18]. While UCB based algorithms use confidence intervals for exploring the arms, Thompson sampling based bandits use variance of the prior distributions to control the exploration. Also, collaborative filtering based algorithms use side information for exploiting and enriching the user and item information to deal with sparse data [1, 8, 16, 17]. Furthermore, using one hot encoding can result in group category based biases where positive rewards of one group within the same category can drive the value of the category weight [10]. Target encoding based techniques that use continuous values for encoding can help prevent categorical sparsity biases [10, 11]. Tree based models use heuristics like distribution based imputation for adding pseudo instances to deal with various categorical sparsity and biases [2]. Additionally, adapting machine learning based models to various invariances in the input data distributions through data augmentation

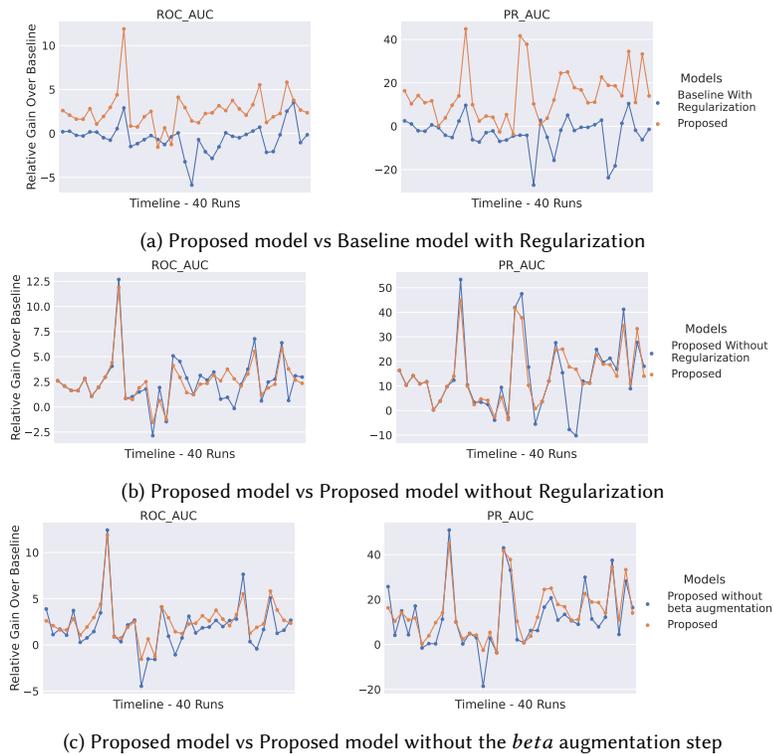


Fig. 1. The above figures contain the relative AUC scores gain of various model configurations over the baseline model. The relative gain is calculated for 40 consecutive training runs. We can observe that the proposed techniques help the model achieve better performance consistently.

techniques is a well studied problem in computer vision [12]. Also, data augmentation by corrupting the item sequences to deal with data sparsity helps improve the performance of sequential recommender systems [14].

3 Current Bandit Ranker

The current model is a logistic regression based Bayesian bandit with thompson sampling [6, 18] where each arm represents a title. We start with a random prior for arm feature weights that gets updated during training. The model takes input a set of features corresponding to the arms such as one hot encoded signals, e.g., Marketing Classification, content category, and Hours lapsed since the launch of the title. The model is trained to predict the user streaming action as a positive reward within a time-based attribution window, given a user-title pair. During the model training, we employ a binary cross entropy loss [3, 6], and the mean weights are shared across the arms as we try to rank the titles among themselves at any given time. The covariance of features is tracked separately for each title. During inference, given a customer and set of titles the models emits scores for each title which are used to rank the titles. Similar to all online models, this model is incrementally trained every T hours. The set of arms is updated every time a new title enters or exits the recommendation catalogue. Refer to the section 5 for more details.

3.1 Challenges

Because of their online and linear nature, bandit based recommendation systems commonly have to deal with 1) categorical data sparsity where the number of training examples for certain categories are very less; 2) categorical cannibalisation due to the lack of granular metadata; and 3) time based feature columns that dynamically update the input representations. These challenges are common in any ML models in RecSys and particularly in exploitation.

Hours lapsed since the launch of the title (recency) This feature helps to factor in the recency of the title launch and to tune the exploration/exploitation trade-offs accordingly. The feature value is set based on a time based binning of first hr_1 hours, hr_1-hr_2 hours, hr_2-hr_3 hours and greater than hr_3 hours respectively. While other features used by the model are fixed given a title-user pair, this feature can change over time. Because of the dynamic nature of this feature, the performance of a title moving from one bin to the next bin with time could be affected by the new bin weight. A high performing title in old bin can be down ranked when it moves to new bin if the weight of the new bin is less when compared with other time based bins, and vice versa with low performing title being up-ranked.

Categorical Feature Biases and Sparsity Apart from the dynamic time based feature, the model takes fixed one hot encoded categorical features as the input. Such encoding comes with a bias where a high performing title from one category will drive the performance of other titles from the same category by giving higher scores than expected, and vice versa. At the same time, one hot encoded features suffer with data sparsity in the context of online models where the set of titles and their categories can change. This data sparsity results in abrupt fluctuation of weights and can impact the next incremental training.

4 Approach

To build a robust multi armed bandit framework by efficiently dealing with dynamic changes in recency features and the categorical feature biases, we propose the following design principles.

Data Augmentation (DA) This approach is designed to fix fluctuations and feature mismatch sparsity issues of the hours since launch bin feature. As part of this approach, we perform a two stage data augmentation process. Specifically, 1) for each title, we create a copy of a small $\alpha\%$ ratio from its training examples where we transition the recency bin to the next bin. Titles that belong to the last bin will remain in the same bin in the new $\alpha\%$ sampled data; and 2) even after the first stage of title level sampling, there can be sparsity in some bins, we similarly upsample them with randomly selected training examples to guarantee at least $\beta\%$ shares with respect to the overall bin distribution. If there are zero examples from a particular hours since launch bin during training, the previous biases present in the weights would propagate to the next title that enters into that bin. The $\beta\%$ upsampling step helps attenuate this problem by providing non zero feedback from all the titles as an average. During training phase, we exclude the $\beta\%$ randomly upsampled data from updating the variance vectors of the title arms in order to prevent additional noise influencing the exploration. For our experiments, we decided to use $\alpha = 15\%$ and $\beta = 10\%$ based on empirical study of various configurations.

Temporal Title Signals (TS) This approach is aimed at tackling the categorical bin level cannibalization by providing granular features that can differentiate at title/arm level. At the same time, these signals incentivise exploitation by factoring in the relative popularity of the titles as input features. In this approach, we add three new columns containing title level Normalised Distinct Stream (NDS) values calculated in the last three p hour slots. Specifically, following the definition given in equation 1, we calculate $NDS_{title,(T,T-p)}$, $NDS_{title,(T-(p+1),T-2p)}$, $NDS_{title,(T-(2p+1),T-3p)}$ values for each title during the run hour T . The three new columns encode the popularity of each title relative to other titles in the three different time slots. A title would be up-ranked if it consistently has high NDS in all three slots and vice versa. For cold start titles that may not have data for all the three columns, we use average NDS observed during

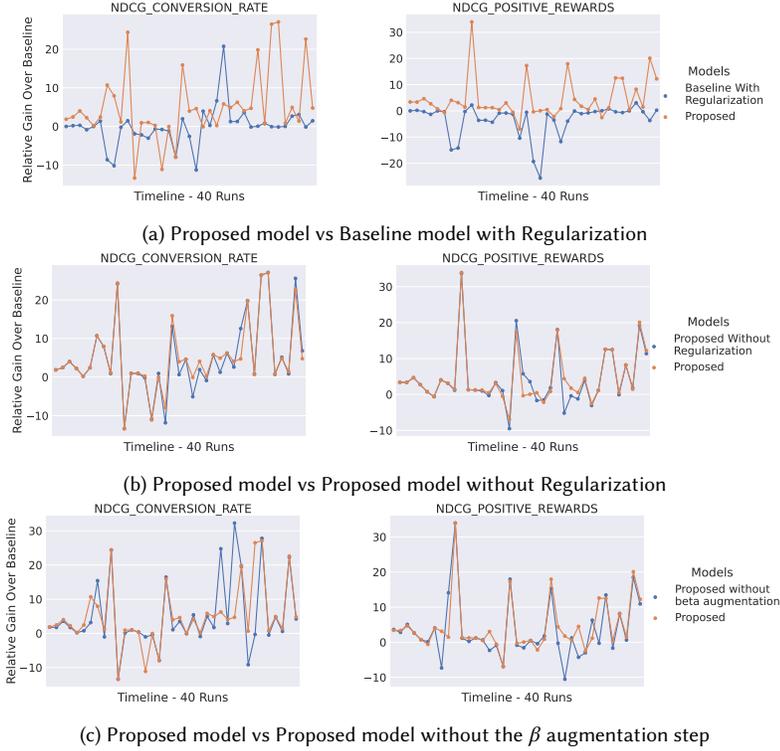


Fig. 2. The above figures contain the relative nDCG metrics gain of various model configurations over the baseline model. As defined in section 5.3, the aggregated nDCG metrics quantify the alignment between predicted title SOVs and production time title conversion rates and title positive rewards. The relative gain is calculated for 40 consecutive training runs.

training for the respective columns. For cold start titles with high priority marketing classification, we give train time maximum NDS value for all columns during their first hr_1 hours after launch. We observed that this strategy boosts share of voice (SOV) of such titles without impacting the performance of other titles. Share of voice of a title is defined as the percentage of users for which this title is present in top position when the titles are sorted by model scores.

$$NDS_{title,time-period} = \frac{distinct - streams(title, time - period)}{\sum_{t \in title-set} distinct - streams(t, time - period)} \quad (1)$$

L2 smoothing of weights To prevent the fluctuation of weights due to the data sparsity of various categorical features, we regularize the weights of the model by taking historical average of the weights as reference for smoothing. Specifically, we apply L2 smoothing for feature columns corresponding to marketing classification, content category, hours since launch bin, and the newly added NDS features by adding the L2 cost defined in equation 3 during training. As part of this approach, we use a smoothing vector as a reference to control the movements of weights and attenuate sudden spikes. As given in equation 2, we use the average weights of the previous q train runs as the smoothed weight vector for the regularization process. Given a regularization parameter λ , the L2 smoothing loss at training run hour T is defined as:

$$W_{T,smooth} = \frac{\sum_{t=1}^{t=q} W_{T-t}}{q} \quad (2)$$

$$L2 - smoothing - loss(W_T) = \lambda \|W_T - W_{T,smooth}\|_2 \quad (3)$$

5 Experiments

5.1 Data

For the experiments, the training and validation data is created by down sampling the production logs to maintain an X number of negatives for each positive user and title pair. This data is curated every T hours from the prod logs containing user request ids and the recorded streaming actions. For evaluation, we take a down sampled version of the prod logs gathered every T hours following the cadence for training. The production data consists of titles in the order of 50 and several dozens of millions of users. At any given time, the set of titles are updated based on a title’s promotional window.

5.2 Experiment Setup

Thompson sampling based Bayesian bandit model [6] consists of one linear layer with a sigmoid activation function, where Bayesian priors are maintained per feature and arm/title. As described in section 3, the mean weights are shared across the arms, but variance is kept separate. The proposed model takes the 3 additional temporal signal (TS) features compared to the baseline. For training, we use Adam [9] optimizer, and the models are trained end-to-end for 3 epochs using the binary cross entropy loss. We also apply a custom regularization term on the Bayesian priors. The model is incrementally trained on data gathered every T hours. For robust evaluation, we perform incremental training and evaluation for 40 consecutive runs on the data gathered from the production. For L2 regularization, we use a 0.25 regularization constant.

5.3 Results

We evaluate the models for stream action prediction given a user-title pair by calculating ROC-AUC and PR-AUC scores for each training run. We also report two aggregated nDCG (normalized discounted cumulative gain) based metrics that are useful for determining whether the title share of voice (SOV) re-adjustments are happening in the right direction when compared with production logs. Specifically, we report the nDCG metric calculated between the ranked lists of predicted top1 sov and the production time conversion rates (nDCG Conversion Rates), positive rewards (nDCG Positive Rewards) for all the titles. So, a higher nDCG metric and a drop in top1 SOV for a specific title implies that the re-adjustments are happening in the direction at an aggregate level. For all the metrics reported, we provide the relative gain of the proposed model and its configurations over the baseline model. We compare the metrics of the below model configurations against the baseline model and report the relative gain over baseline: 1) Proposed: This refers to the proposed model where we apply 3 methods described in section 4; 2) Baseline With Regularization: This is the baseline model but with regularization applied to categorical weights; 3) Proposed without Regularization: This is the proposed model without the regularization loss; 4) Proposed without *beta*: This is the proposed model, but without the second level of augmentation using the *beta* ratio.

As shown in the figures 1 and 2, we can observe that the proposed model is consistently doing better than the baseline with respect to both the AUC scores and nDCG metrics. We observe similar patterns in both the metrics across various configurations. This implies that the proposed techniques not only help in improving the title relevancy prediction performance, they also re adjust the title SOVs in the right direction. From figures 1a and 2a, we can understand

that using regularization for baseline doesn't necessarily help. Also from figures 1b and 2b, proposed model without regularization has slight drops in some of the runs signifying the positive impact of smoothing. From the figures 1c and 2c, it is evident that the second level of augmentation using β parameter is helping the model to better deal with the sparsity issues of the time based feature.

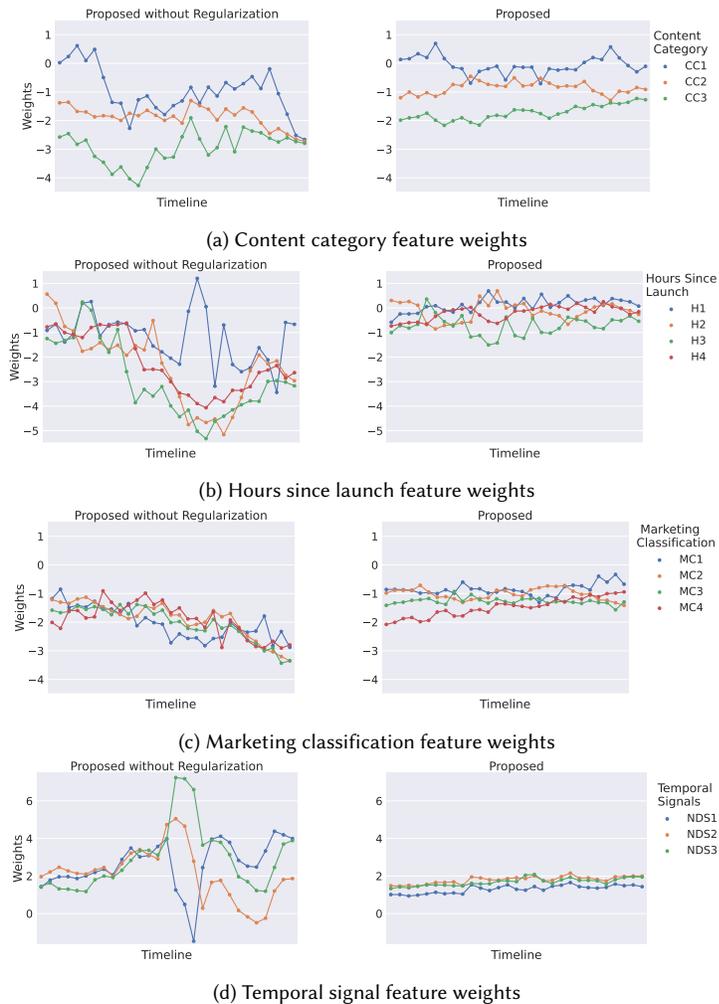


Fig. 3. Model feature weights before and after L2 regularization with a smoothing vector averaged over previous q runs.

5.4 Case Studies: SOV allocation for specific titles

The biases described in the section 3.1 can impact the SOV allocation of titles in an undesirable way. For example, an high performing title can get less SOV when it moves to the next recency bin that can be sparse at that point of time. At the same time, a low performing title can get more SOV than required when it belongs to a same category as another high performing title. As part of this case study, we analyse two such cases observed with prod model and how the proposed model can attenuate the biases by assigning correct SOV. As part of this case study, we report SOV of

a title in its first position and also average rank across all customers. We report these metrics as a relative gain over baseline for two titles, a popular title for reference and also the title under study.

Over Exposure As given in table 1, the baseline model is giving more SOV to an over exposed title at the expense of a popular high performing title. This scenario was when a popular title moved to its next recency bin that was sparse. Observe that during the same runs, the proposed model with the help of data augmentation and temporal signals not only maintains the SOV of over exposed title, it still gives high SOV to the popular title than the baseline.

Table 1. Average Rank and Top1 SOV relative gain of proposed model over the baseline for a popular title and overexposed title for three consecutive runs. For average rank, negative gain implies higher rank and for Top1 SOV, positive gain implies higher SOV than baseline.

Titles	Run1		Run2		Run3	
	Avg Rank	Top1 SOV	Avg Rank	Top1 SOV	Avg Rank	Top1 SOV
Popular Title	-43.41%	19.73%	-57.86%	149.37%	-22.56%	32.51%
Over Exposed Title	-18.34%	21.72%	-25.77%	41.88%	-13.80%	-3.82%

Under Exposure As given in table 2, the baseline model is giving significantly less SOV than expected to an under exposed title. This scenario is when the title belonged to a category with other low performing titles driving the performance. Observe that during the same runs, the proposed model with the help of temporal signals not only maintains the SOV of popular title, it gives significantly more SOV to the under exposed title than the baseline.

Table 2. Average Rank and Top1 SOV relative gain of proposed model over the baseline for a popular title and a underexposed title for three consecutive runs. For average rank, negative gain implies higher rank and for Top1 SOV, positive gain implies higher SOV than baseline.

Titles	Run1		Run2		Run3	
	Avg Rank	Top1 SOV	Avg Rank	Top1 SOV	Avg Rank	Top1 SOV
Popular Title	-13.32%	10.41%	-56.62%	200.43%	-21.80%	50.82%
Under Exposed Title	1.59%	-30.44%	-29.06%	710.05%	-63.32%	4871.34%

5.5 Regularization of weights

As given in the figure 3, regularization not only constrains the weights of the input features in a smaller region, it also reduces the fluctuations of weights over the training runs. While the smoothing of weights is clearly evident in non dynamic features like content category, marketing classification and temporal signals as given in figures 3a, 3c and 3d. For dynamic features that change over time like hours since launch, the fluctuations can still be seen as evident in figure 3b.

6 Conclusion

In this paper, we presented the primary steps to design a robust multi-armed bandit production framework in video recommendations with high-value marketing content. We showed that data augmentation method, bandit weights regularization, along with temporal granular signals can address most common biases and title cannibalization challenges in RS. Through our experiments and case studies, we demonstrated the effect of using each of our design choices. Experimental results from offline analyses showed that our proposed method can achieve up to 11.88% and 44.85% relative gain compared to a bandit baseline model used in production, respectively in ROC-AUC and PR-AUC.

References

- [1] Himan Abdollahpour, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*, 2019.
- [2] Timothy C Au. Random forests, decision trees, and categorical predictors: the "absent levels" problem. *The Journal of Machine Learning Research*, 19(1):1737–1766, 2018.
- [3] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- [4] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020.
- [5] Emanuele Cavenaghi, Gabriele Sottocornola, Fabio Stella, and Markus Zanker. Non stationary multi-armed bandit: Empirical evaluation of a new concept drift-aware algorithm. *Entropy*, 23(3):380, 2021.
- [6] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.
- [7] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.
- [8] Jiayu Han, Lei Zheng, Yuanbo Xu, Bangzuo Zhang, Fuzhen Zhuang, S Yu Philip, and Wanli Zuo. Adaptive deep modeling of users and items using side information for recommendation. *IEEE transactions on neural networks and learning systems*, 31(3):737–748, 2019.
- [9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [10] Carlos Mougán, Jose M Alvarez, Gourab K Patro, Salvatore Ruggieri, and Steffen Staab. Fairness implications of encoding protected categorical attributes. *arXiv preprint arXiv:2201.11358*, 2022.
- [11] Florian Pargent, Bernd Bischl, and Janek Thomas. *A benchmark experiment on how to encode categorical features in predictive modeling*. PhD thesis, Master Thesis in Statistics, Ludwig-Maximilians-Universität München ..., 2019.
- [12] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.
- [13] Nicollas Silva, Heitor Werneck, Thiago Silva, Adriano CM Pereira, and Leonardo Rocha. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. *Expert Systems with Applications*, 197:116669, 2022.
- [14] Joo-yeong Song and Bongwon Suh. Data augmentation strategies for improving sequential recommender systems. *arXiv preprint arXiv:2203.14037*, 2022.
- [15] Alane Suhr and Yoav Artzi. Continual learning for instruction following from realtime feedback. *arXiv preprint arXiv:2212.09710*, 2022.
- [16] Qinyong Wang, Hongzhi Yin, Hao Wang, Quoc Viet Hung Nguyen, Zi Huang, and Lizhen Cui. Enhancing collaborative filtering with generative augmentation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 548–556, 2019.
- [17] Lei Zheng, Vahid Noroozi, and Philip S Yu. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 425–434, 2017.
- [18] Li Zhou. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326*, 2015.