Why People Skip Music? On Predicting Music Skips using Deep Reinforcement Learning

Francesco Meggetto
NeuraSearch Laboratory, University of Strathclyde
Glasgow, UK
francesco.meggetto@strath.ac.uk

Crawford Revie
University of Strathclyde
Glasgow, UK
crawford.revie@strath.ac.uk

John Levine
University of Strathclyde
Glasgow, UK
john.levine@strath.ac.uk

Yashar Moshfeghi
NeuraSearch Laboratory, University of Strathclyde
Glasgow, UK
yashar.moshfeghi@strath.ac.uk

ABSTRACT

Music recommender systems are an integral part of our daily life. Recent research has seen a significant effort around black-box recommender based approaches such as Deep Reinforcement Learning (DRL). These advances have led, together with the increasing concerns around users' data collection and privacy, to a strong interest in building responsible recommender systems. A key element of a successful music recommender system is modelling how users interact with streamed content. By first understanding these interactions, insights can be drawn to enable the construction of more transparent and responsible systems. An example of these interactions is skipping behaviour, a signal that can measure users' satisfaction, dissatisfaction, or lack of interest. In this paper, we study the utility of users' historical data for the task of sequentially predicting users' skipping behaviour. To this end, we adapt DRL for this classification task, followed by a post-hoc explainability (SHAP) and ablation analysis of the input state representation. Experimental results from a real-world music streaming dataset (Spotify) demonstrate the effectiveness of our approach in this task by outperforming state-of-the-art models. A comprehensive analysis of our approach and of users' historical data reveals a temporal data leakage problem in the dataset. Our findings indicate that, overall, users' behaviour features are the most discriminative in how our proposed DRL model predicts music skips. Content and contextual features have a lesser effect. This suggests that a limited amount of user data should be collected and leveraged to predict skipping behaviour.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Spotify, Music, Skipping, User Behaviour, Prediction, Deep Reinforcement Learning

CHIIR '23, March 19–23, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0035-4/23/03.
https://doi.org/10.1145/3576840.3578312

ACM Reference Format:
Francesco Meggetto, Crawford Revie, John Levine, and Yashar Moshfeghi. 2023. Why People Skip Music? On Predicting Music Skips using Deep Reinforcement Learning. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3576840.3578312

1 INTRODUCTION

In recent years, online music streaming services (e.g., Spotify) have seen substantial growth. With the rise of digital music distribution, the related success of such streaming services, and the ubiquitous availability of music, a new listening paradigm has emerged. Users can access any song, at any time, and within a few clicks. As a result, there has been a significant change in users' behaviour and interaction with these systems [19]. Music recommender systems (MRS) aspire to tackle the problem of providing the users the support they need to access these large collections of music items and find songs that match their interests and needs. Recent research has seen a significant effort towards black-box based approaches such as Deep Reinforcement Learning (DRL) [3, 56]. This is motivated by the possible radical changes of behaviour from one song to another, or even within the same song, but at different points in time. Users' behaviour is influenced by external (trends) and internal (individual changes of personal interests) factors. The users' shifting interests and behaviour make it hard to learn a generalisable model to tailor the user's specific needs at any given time; it is a case where DRL is required due to continuous learning and adaption [9, 29, 61]. These advances, however, have inevitably led to rising concerns about how users' data is collected, stored, and used. This is leading to a strong research interest in building responsible systems and data collection procedures [45]. Constraints should be put in place when considering what data is collected and then presented to a model to measure user behaviours. This is due to the potential hazard of introducing errors and biases. Therefore, minimising and selecting high-quality data features is of important consideration.

With explicit rating data relatively scarce and rare in today's systems, modelling implicit feedback is becoming of acquired importance. For example, in a *lean-back* formulation, the case of automatic playlists or radio streaming, the user interaction is minimised. Users are presented with a single song at a time. The MRS needs to rely almost entirely on implicit feedback signals such as the skipping or scrubbing (i.e., seeking forward and backward by moving the

cursor [35]) to predict satisfaction and engagement [29, 61]. By understanding these interactions, insights can be drawn for the construction of more transparent and responsible systems. The skipping is a signal that can measure users' satisfaction, dissatisfaction or lack of interest, and engagement with the platform [44]. In a *lean-back* formulation, the MRSs are often designed to be more conservative, prioritising *exploitation* over *exploration* to minimise negative feedback (in this context, skips) [53]. Thus, one of their goals may be determined as recommending songs that yield the highest listening activity (i.e. no skip). However, understanding the users' skipping behaviour is still an under-explored domain [11, 34, 44]. It is a challenging problem due to its noisy nature: a skip may suggest a negative interaction, but a user may skip a song that they like because they recently heard it elsewhere. In this work, we aim to understand why people skip by comprehensively analysing the utility of users' historical data. In particular, we analyse the impact and effect of the users' behaviour (e.g., the user action that leads to the current playback to start), listening content (i.e., the listened song), and contextual (e.g., the hour of the day) features in the classification task of predicting the users' music skipping behaviour. We propose a novel approach that leverages and adapts DRL for this classification task. This is to most closely reflect how a DRL-based MRS could learn to detect music skips.

Prior works in analysing the skipping behaviour revealed an universal behaviour in skipping across songs, with geography, audio fluctuations or musical events, and contextual listening information affecting how people skip music [14, 44, 48, 49]. Recently, the effectiveness of deep learning models has also been explored for the task of predicting the users' sequential skipping behaviour in song listening sessions [1, 8, 12, 23, 31, 58, 68]. While they made a significant contribution towards this direction, their process is usually seen as an independent and static procedure. They may not account for the dynamic nature of the users' behaviour, and do not intuitively optimise for the long-term potential of user satisfaction and engagement [29, 40, 54, 61, 66, 67]. Overall, this motivates the investigation of the DRL's applicability in predicting music skips and a comprehensive investigation on the relation of the skipping signal with users' behaviour, listening context, and content. This paper aims to investigate the following two important research questions: *can DRL be applied to the users' music skipping behaviour prediction task, and if so, would it be more effective in the music skip prediction task than deep learning state-of-the-art models?* (**RQ1**); *what historical information is considered discriminative and serves as a high-quality indicator for the model to predict why people skip music?* (**RQ2**). To investigate our RQs, we have conducted an extensive study on a real-world music streaming dataset (Spotify). Our comprehensive analysis demonstrates the effectiveness of our approach and a temporal data leakage problem in the historical data. Overall, our findings indicate that the most discriminative features for our proposed DRL model to predict music skips are some users' behaviour features, with content and contextual features reporting a lesser effect. This suggests that a limited amount of user data can be leveraged to predict this behaviour, thereby offering implications in the building of novel user-centred MRSs and responsible data collection procedures. This is a necessary step in creating a holistic representation of the listeners' preferences, interests, and needs. The main contributions of this paper are:

- We demonstrate the applicability and effectiveness of DRL in predicting users' skipping behaviour from listening sessions. A framework is devised to extend the DRL's applicability to perform this classification and offline learning. This is the first time that DRL has been explored in this task. The effectiveness of our approach is empirically shown on a real-world music streaming dataset (Spotify). Our proposed approach outperforms state-of-the-art models in terms of Mean Average and First Prediction Accuracy metrics.
- We perform a comprehensive post-hoc (SHAP) and ablation analysis of our approach to study the utility of users' historical data in detecting music skips. We reveal a temporal data leakage problem in the historical data. Further, our results indicate that overall users' behaviour features are the most prominent and discriminative in how the proposed DRL model predicts music skips. The listening content and context features are reported to have a lesser effect.

2 RELATED WORK

A successful MRS needs to meet the users' various requirements at any given time [24, 55, 64]. Thus, user modelling is a key element. A line of research has tried to untangle the relationship between personality and the users' musical preferences [37, 51, 52]. Volokhin and Agichtein [60] introduced the concept of music listening intents and showed that intent is distinct from context (user's activity). A different, and arguably complementary, research direction is trying to understand and model how users interact with the underlying platform. This is a long-standing and under-researched problem of online streaming services [11]. An example of these interactions is the skips between songs. Its modelling and understanding during music listening sessions plays a crucial role in understanding users' behaviour [44]. The skips are often the only information available to the underlying MRS, and therefore they are used as a proxy to infer music preference [53].

The skipping signal has already been used in prior works, as a measure in heuristic-based playlist generation systems [10, 50], user satisfaction [24, 64], relevance [25], or as a counterfactual estimator [43]. Furthermore, given its universality and presence in other domains, recent research has also investigated its effect in ads on social media platforms [5–7]. Despite being abundant in quantity, it is a noisy implicit signal [53, 62]. A skipped track does not necessarily imply a negative preference. Multiple hypotheses can be formulated on why users skip songs, with recent research suggesting that people manifest an universal behaviour in skipping across songs, dictated by time, geography, and reaction to audio fluctuations or musical events [14, 48, 49]. Moreover, it has been shown in [57] that people who usually listen to songs in their entirety, show higher listening duration that those who do not. Most recently, Meggetto et al. [44] proposed a clustering-based approach that clearly identifies four user types with regards to their session-based skipping activity. These types, namely *listener*, *listen-then-skip*, *skip-then-listen*, and *skipper*, are influenced by the length of the listening session, time of the day, and playlist type. The main limitation of these prior works is that they explore the relation between listening context and content with the skipping behaviour. They do not explore how the user interactions with the

platform influence the detection of skips. This is a limitation that this works addresses.

In 2019, Spotify identified music skip prediction as an important challenge and organised the *Sequential Skip Prediction Challenge* [1] to explore approaches that could alleviate this problem. The challenge focused on predicting whether individual tracks encountered in a listening session will be skipped or not. To respond to this challenge, several deep-neural networks [1, 8, 12, 23, 31, 58, 68] and supervised learning [18] models were proposed. Afchar and Hennequin [2] proposed using interpretable deep neural networks for skip interpretation via feature attribution. Whilst neural networks, and in particular Recurrent Neural Networks (RNNs), have been shown to effectively model sequential data, they consider the procedure as a static process. They do not intuitively provide a mechanism for the long-term optimisation of user satisfaction and engagement, continuous learning, and the modelling of the dynamic nature of the user's behaviour [29, 54, 61, 66, 67]. Therefore, it is a case where DRL is required, an investigation and application of which has never been explored before. A research gap this work aims to address.

The *Sequential Skip Prediction Challenge* is a binary classification task. Despite receiving limited attention to date, DRL has been shown to be suitable and effective in classification tasks. It can assist classifiers in learning advantageous features [15, 27] and select high-quality instances from noisy data [17]. Wiering et al. [65] demonstrate that RL is indeed suitable for classification. Their model slightly outperforms existing classifiers, but training time and extra computational requirements are major drawbacks. With the recent advances in the field, a body of research is showing the superiority of DRL-based approaches for classification tasks [17, 27, 28, 39, 42]. In particular, the authors in [27, 28] show that a Vanilla Deep Q-Network (DQN) [47] approach is superior and more robust to state-of-the-art algorithms.

In this work, we explore, for the first time, the applicability of DRL in the task of sequentially predicting users' music skipping behaviour. This is motivated by the limitations of existing approaches and the advantages of DRL. By comprehensively analysing users' historical data, we study its utility and effect in our approach for this task. This work is the first step in understanding why people skip music.

3 APPROACH

In this section, we present our framework to facilitate the application of DRL to the problem of sequentially predicting users' skipping behaviour from listening sessions. To do so, we model this problem as a Markov Decision Process (MDP) and a mechanism is introduced in the RL problem formulation to correctly exploit logged interactions and thus perform offline learning. The details of this framework are as follows:

State: it is the record-level representation of a listening session at a discrete time step (i.e., position in the session). The state, i.e. a record in a listening session, includes various user's contextual information about the stream, their interaction history with the platform, and information about the track that the user listened to. An episode is the entire listening session, with sessions containing from 10 up to at most 20 records.

Actions: it is a discrete action space which is a binary indicator of whether the current track is going to be skipped or not by the corresponding user. Effectively, the problem formulation can also be thought of as a binary classification problem $A = \{0, 1\}$, where 0 represents a no skip operation and 1 represents a skip.

Reward: a positive reward of 1 is given for a correctly predicted skip classification, 0 reward (i.e., no penalty) otherwise.

Motivated by the discrete action space and off-policy requirements of the music skip prediction task, we leverage DQN[2]. These requirements preclude the use of algorithms such as Deep Deterministic Policy Gradient (continuous action space) and Proximal Policy Optimization (on-policy learning). Whilst the problem is formulated as an MDP, it is partially observable (POMDP) by definition. This is because only partial information about the listening context and of the user is available [16]. Hence, in our problem formulation, we consider MDP and POMDP to be equivalent. This means that we do not perform any further processing of the state representation (e.g., masking of some features).

This classification formulation can be seen as a guessing game, where a positive reward is given for a correct guess, and no penalty is given for an incorrect one. Long-term optimisation via discount factor γ can be thought of as a way to correctly guess as many records in an episode as possible. Since there is a sequential correlation among records within an episode (i.e., a music listening session), a high γ value should be used. This corresponds to optimisation on the total number of correct guesses in an episode (long-term) rather than optimisation on the immediate ones (short-term). By taking into account previous points in time and the past interactions with the environment, the DRL agent makes fully informed decisions.

3.1 Offline Mechanism

The DQN's standard training procedure is entirely online. Online learning is an iterative process where the agent collects new experiences by interacting with the environment, typically with its latest learned policy. That experience is then used to improve the agent's policy. However, exploiting logged data may be helpful and informative for the agent as a form of (pre)training. In offline learning (Batch RL [36]), the agent's task is instead defined as learning from a static dataset. Policies are learnt from logged data, and no interactions with the underlying environment are required. Whilst our prior formulation would work in an online learning setting, it presents a major problem when performing offline learning. A misclassification would cause a transition to a new state, which is, however, not part of the original trajectory and thus not represented in the dataset as well. The agent will generate and associate a (discounted) cumulative reward to a wrongly generated trajectory that is substantially different from the original. Thus, a pure offline algorithm has to exclusively rely on the transitions that are stored in the dataset provided in the input. From our initial formulation, we need to account for those out-of-distribution actions.

Within the definition of the reward function itself, the out-of-distribution, untruthful action is marked as invalid and, if sampled

[1] https://www.aicrowd.com/challenges/spotify-sequential-skip-prediction-challenge

[2] Due to space limitations, we refer the readers to [47] for the necessary background and overview of the algorithm.

by the agent throughout learning, it causes the current episode to be terminated. In other words, an incorrect guess (0 reward) leads to a terminal state. This simple constraint forces a minimisation of estimation errors and therefore it avoids the creation of potential estimation mismatches. As such, the untruthful action that causes the current episode to terminate avoids the future propagation of incorrect bootstrapped return estimations in the Temporal Difference target. This is to minimise the distributional shift issues due to differences between the agent's policy and the behaviour policy. More specifically, it explicitly ensures that regardless of the next sampled action, the current policy $\pi(a'|s')$ is as close as possible to the behaviour distribution $\pi_\beta(a'|s')$. The Q-function is queried as little as possible on out-of-distribution and unseen actions since this will eventually increase errors in the estimations.

This error, i.e. "extrapolation error" [22], is introduced when an unrealistic and erroneous estimation is given to state-action pairs. This is caused when action a' from estimate $Q(s, a)$ is selected, and the consequent state-action pair (s', a') is inconsistent with the dataset due to the pair being unavailable. It provides a source of noise that can induce a persistent overestimation bias and that cannot be corrected, in an off-policy setting, due to the inability to collect new data [21, 22]. Directly utilising DQN in an offline setting may result in poorer performance and a resemblance to overfitting [38]. Our proposed mechanism minimises these errors. It is important to note that the "correct" action is not forcefully fed to the agent as in Behaviour Cloning based approaches. We let the agent deterministically decide as if it were a live interaction with the environment, thus keeping the general workflow of the original algorithm intact. This provides a single interface to easily transition from offline to online learning and vice versa.

Finally, it is important to note that the aim of this work is to enhance our understanding of why people skip music and identify the high-quality features for its detection. To this end, we analyse the applicability of DRL in predicting this behaviour. We leave further tailoring of the approach to the music skip prediction task and an evaluation with recently proposed offline model-free algorithms [4, 13, 20, 33] for future work. Nevertheless, our proposed approach requires no architectural or algorithmic modifications. It offers the potential for a swift transitioning from online to offline learning and vice versa. It can be also be considered as a swift pre-training of an agent that can later be deployed online for continual learning.

4 EXPERIMENTAL SETTINGS

4.1 Dataset

We conduct our experiments on the real-world Music Streaming Sessions Dataset (MSSD) provided by Spotify [11]. The publicly available training set consists of approximately 150 million logged streaming sessions, collected over 66 days from July 15th and September 18th 2018. Each day comprises ten logs, where each log includes streaming listening sessions uniformly sampled at random throughout the entire day. Sessions contain from 10 up to at most 20 records and are defined as sequences of songs/tracks that a user has listened to (one record per song). Each record includes various user's contextual information about the stream (e.g., the playlist type) and interaction history with the platform (e.g., scrubbing, which is the number of seek forward/back within the track). Although the track titles are not available, descriptive audio features and metadata are provided for them (e.g., acousticness, valence, and year of release). It is important to note that there is no user identification, nor access to demographic or geographical information. Hence, by not knowing whether two sessions have been played by the same user or by two different users, this study revolves around the modelling and understanding of the users' skipping behaviour.

4.1.1 Temporal Correlation. There is no temporal correlation among listening sessions, i.e. the sessions are not presented in historical order, which is reflected in the chance of consecutive sessions having a considerably different hour of the day (e.g., morning and evening). Also, there is no order to the ten logs within a given day (i.e., the 1st log of the first day does not necessarily occur before the 2nd of the same day). This does not preclude the potential applicability of DRL for the skip prediction task since the hour of the day in which a song was played is provided. Thus, it allows for the modelling of skipping behaviour dependent on the hour of the day.

4.1.2 Creation of Training and Test Sets. In this work, we only leverage the training set since, in the test set, most of the metadata and the skipping attributes used as ground truth in our evaluation are not provided. By selecting logs from the original training set, statistics for our training and test datasets are presented in Table 1. As it can be seen from the statistics, the ratio of skip values for all sets is balanced between True and False values. This balanced distribution is an intrinsic property of the dataset and of any of the available logs. Due to the large amount of data, and therefore computational and execution time requirements, the first four logs of the first available day are used for training. Testing is performed on various logs in order to test the models' generalisability for different days. Except for T1, which is the 5th and next immediate consecutive log after the training set collection, all the other logs are of a random index, day and/or month. This random selection approach is justified by the fact that there is no temporal correlation among logs of the same day. This is to show the generalisation capabilities of our proposed approach and to allow for the comprehensive analysis of the importance of the users' historical data.

4.1.3 Data Preprocessing. All available features, with a full description available in [11], are included in the state representation, except for the skip features, session and song identifiers. Categorical features, such as the playlist type and the user's actions that lead to the current track being played or ended, are one-hot encoded. All the audio features are standardised to have a distribution with a mean value of 0 and a standard deviation of 1. Overall, this results in a state representation consisting of 70 features. For ease of discussion, they are grouped as follows:
User Behaviour (UB):

- **Reason End (RE)** is the cause of the current playback to end. This is a one-hot encoded feature that thus groups various encoded features such as *Trackdone*, *Backbtn*, *Fwdbtn*, and *Endplay*.
- **Reason Start (RS)**. Similar to *Reason End*, it is the type of actions that cause the current playback to start.

Table 1: Summary of datasets used for experiments after pre-processing. log(s) # indicate which log(s) are selected out of the available ten. skip (%) refers to the ratio between True and False values.

Dataset	Date	log(s) #	# of records	# of sessions	skip (%)
Training Set	15/07/2018	[0, 3]	11,927,861	711,838	51.20%
Test Set (T1)	15/07/2018	4	2,991,438	178,419	51.21%
Test Set (T2)	19/07/2018	8	3,395,883	204,145	50.53%
Test Set (T3)	27/07/2018	0	3,447,209	207,060	50.76%
Test Set (T4)	10/08/2018	6	3,407,685	205,267	50.42%
Test Set (T5)	09/09/2018	1	2,588,711	155,617	51.48%

- **Pauses (PA)** is the length of the pause in between playbacks. It consists of *No*, *Short*, and *Long Pause*.
- **Scrubbing (SC)** is the number of seeking forward or backward during playback. They correspond respectively to *Num Seekfwd* and *Num Seekback*.
- **Playlist Switch (PS)** indicates whether the user changed playlist for the current playback.

Context (CX):

- **Session Length (SL)** is the length of the listening session.
- **Session Position (SP)** is the position of the track within the session.
- **Hour of Day (HD)** is the hour of the day in which the playback occurred ([0..23]).
- **Playlist Type (PT)** is the type of the playlist that the playback occurred within. Examples are *User Collection*, *Personalized Playlist*, and *Radio*.
- **Premium (PR)** indicates whether the user was on premium or not.
- **Shuffle (SH)** indicates whether the track was played with shuffle mode activated.

Content (CN). This third and final category groups all the **Track (TR)** metadata and features, as they constitute the only content-based information in the MSSD. It includes 28 features such as *Beat Strength*, *Key*, *Duration*, and the eight *Acoustic Vectors* ([0..7]).

4.2 Evaluation Metrics

To perform an evaluation of our proposed approach, we adopt the evaluation metrics from the *Spotify Sequential Skip Prediction Challenge*. This is also to provide a fair comparison with the selected baselines, since they were proposed on this challenge and for the following task: *given a listening session, predict whether the individual tracks encountered in the second half of the session will be skipped by a particular user*. Therefore, every second half of a session in the selected test set is used for prediction. If a session has an odd number of records, the mid-value is rounded up. This is motivated by the fact that an accurate representation of the user's immediately preceding interactions can inform future recommendations generated by the music streaming service. Hence, it is important to infer whether the current track is going to be skipped as well as subsequent tracks in the session. First Prediction Accuracy and Mean Average Accuracy are adopted as metrics.

First Prediction Accuracy (FPA) is the accuracy at predicting the first interaction for the second half of each session.
Mean Average Accuracy (MAA) is defined as:

$$MAA = \frac{\sum_{i=1}^{T} A(i)L(i)}{T} \tag{1}$$

where T is the number of tracks to be predicted within the given session, $A(i)$ is the accuracy up to position i of the sequence, and $L(i)$ indicates whether the i^{th} prediction is correct or not. Intuitively, in these evaluation metrics higher importance is given to early predictions. In our setting, however, we do not exploit this specification in the problem formulation. Instead, the agent is instructed to optimise the total number of correct predictions in the session. This is to keep the system's specifications simple and easily adaptable to different metrics and/or tasks. In the dataset schema, prediction is based on the *skip_2* feature. It indicates a threshold on whether the user played the track only briefly (no precise threshold is provided) before skipping to the next song in their session.

4.3 Models

4.3.1 Baselines. To identify state-of-the-art baselines on the music skip prediction task, we performed an extensive search on prior works that utilise the MSSD dataset. We identified the following 4 of the top-5 ranked submissions to the *Spotify Sequential Skip Prediction Challenge* and presented at the WSDM Cup 2019 Workshop:

- **Multi-task RNN**: RNN-based approach that predicts multiple implicit feedbacks (multi-task) [68].
- **Multi-RNN**: Multi-RNN with two distinct stacked RNNs where the second makes the skip predictions based on the first, which acts as an encoder [23].
- **Temporal Meta-learning**: A sequence learning, meta-learning, approach consisting of dilated convolutional layers and highway-activations [12].
- **Weighted RNN**: RNN architecture with doubly stacked LSTM layers trained with a weighted loss function [31].

They respectively reported the 1st, 2nd, 3rd, and 5th best overall performance on the Spotify Challenge, with Multi-task RNN being the strongest and Weighted RNN being the weakest baselines. The exclusion of the 4th overall best model on the challenge in our evaluation is because no manuscript and code repository were found. For the selected baselines, we use the code accompanying the papers (GitHub links available in cited manuscripts). We then

reproduced their results locally by running their provided public code locally, to the best of our abilities and with an optimised set of parameters. However, despite our best efforts, we reported consistently worse results than the ones in the Spotify Challenge public leaderboard and/or accompanying papers. The test set used in challenge is not fully released. No ground truth is available, thereby not allowing for a local evaluation. However, given our procedure for the creation of the train and test sets (Section 4.1.2), i.e. the training is performed on the first available day and the evaluation is for different days/months, we make the strong assumption that the overall data distribution of our selected test sets and the one used in the public challenge are similar. For a fair comparison, we thus report the results from the public leaderboard since they are better than the ones from our local evaluation.

4.3.2 DQN Architecture. For this work, we explored nine state-of-the-art DQN architectures. By adhering to our proposed framework, they have been thoroughly investigated in the users' music skipping behaviour prediction task. They are the Vanilla [47], Double [59], Dueling [63], and their respective n-step learning variants [46]. Partially observable architectures have also been explored, with observations stacking [47] and Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) based architectures [26].

Due to space limitations, a comparison among all these architectural variants is not reported. We note, however, that Vanilla DQN achieves the best performance. This is given its comparable performance and the advantage of a significantly simpler architecture with lower complexity. Therefore, the reported results are only for the Vanilla DQN architecture (hereafter referred to as "DQN").

4.4 Experimental Procedure

We trained our DQN using the following set of parameters: experience replay memory is 10000, batch size and frequency of updates are set as 256, the learning rate is 0.001, and the discount factor is 0.9. The policy network consists of three fully-connected layers (of size 128) and a final action-value linear output layer of size 2. This final layer computes the Q-values for each action. Hyperparameters were selected by random and Tree-structured Parzen Estimator search, with the best set selected for evaluation on the test collections. The implementation of the DQN agent is provided by the Tensorforce [32] library. For complete reproducibility of our work, the code for this work is available at https://github.com/NeuraSearch/Spotify-XRL-Skipping-Prediction

To explore the potential instabilities and divergences during training, the proposed DQN approach is run five times per test set. The reported results represent the mean across all test sets. Lastly, during the training phase, learning is constrained with out-of-distribution actions, and therefore, some state-value pairs in the dataset are not experienced by the agent due to early termination. During the testing phase, all episodic records are sequentially retrieved, and the agent acts deterministically on the complete episodes for its evaluation.

4.4.1 Post-Hoc Analysis. In order to carry out an analysis on the importance and validity of the users' historical data in predicting music skipping behaviour, we first leverage the Shapley Additive Explanations framework (SHAP) [41]. It is a game-theory based approach that explains the predictions of machine learning models. In particular, we adopt the Kernel Explainer, which is a model agnostic method to estimate the SHAP values. This is because there exists no DRL specific explainers. However, since the Kernel Explainer makes no assumptions about the model to explain the predictions of, it is a highly expensive computational approach. This means that it is slower than the other model type specific algorithms. By considering these extensive computational requirements, for each test set, we estimate the feature importance values for the first 50 episodes (i.e., listening sessions) and with 200 perturbation samples per record. Given the high similarity across all test sets, we only report the results for T1.

4.4.2 Ablation Analysis. To validate the SHAP results, we perform an ablation analysis on the input state representation. We study the effect that the category (e.g., *UB*) and type (e.g. *RS*) features have on the DQN's performance. To this end, we train and evaluate (following the same above-mentioned experimental procedure) the proposed approach on a state representation that does not include the selected features' type. This iterative approach, whereby only a single type is removed for each evaluation, is repeated until all types that comprise the input state representation are evaluated.

4.4.3 Temporal Data Leakage. A closer investigation of the MSSD dataset, and validated by the post-hoc and ablation analysis, reveals a temporal data leakage of some features. These features have been left unnoticed and they have inadvertently affected the Spotify Challenge and thus the baselines. These features correspond to the length of session (*SL*) and the user actions that lead the current playback to end (*RE*). This is because they provide to the model information from the future that should not be available in a live predictive system. Although we recognise and acknowledge this to be a problem, the reported results on the comparison with the selected baselines are without the removal of such features. This is to provide a fair comparison with the selected baselines, since they include these features in their input representation. These features are removed from the state when we investigate why people skip music, and it is referred to as the "corrected" state.

5 RESULTS

First, the validity of our approach to predict users' music skipping behaviour is demonstrated against the state-of-the-art deep learning based models. Our analysis of the music skipping prediction task and of the MSSD dataset reveals a temporal data leakage problem (Section 4.4.3). With a "correction" of the state representation by removal of such features, we report the comprehensive investigation on how the skipping behaviour can be detected by analysing the importance of UB, CX, and CN.

5.1 Applicability of DRL to Music Skip Prediction (RQ1)

On our local evaluation, Multi-RNN and Temporal Meta-learning, despite outperforming Weighted RNN in the challenge submissions, perform consistently worse on our selected test sets. Multi-Task RNN, the best performing baseline on the public challenge, achieves

[3]Leaderboard results available on cited manuscripts and/or at https://www.aicrowd.com/challenges/spotify-sequential-skip-prediction-challenge

Table 2: MAA and FPA results for our proposed DQN approach and baselines. The reported results are the averages across all test sets for DQN (with 95% CI). For the baselines, we report the publicly available results from the Spotify Challenge[3]. This is to provide a fair comparison since they are better than those obtained from our local evaluation. No CIs are reported for the baselines due to their unavailability. The best performing model is highlighted in bold.

		MAA		*FPA*	
		Mean	95% CI	Mean	95% CI
	DQN	**0.820**	[0.818 - 0.822]	**0.881**	[0.880 - 0.882]
Public Leaderboard	Multi-task RNN	0.651	—	0.812	—
	Multi-RNN	0.641	—	0.807	—
	Temporal Meta-learning	0.637	—	0.804	—
	Weighted RNN	0.613	—	0.794	—

slightly inferior performance compared to Weighted RNN. Overall, we note that all the baselines perform consistently worse on our local evaluation than in the public challenge. We observe decreases in performance of 4.9, 16.2, 4.9, 0.8 (%) and 2.4, 8.2, 2.2, 0.4 (%) in MAA and FPA and for Multi-task RNN, Multi-RNN, Temporal Meta-learning, and Weighted RNN respectively. Therefore, in Table 2, we report results in terms of MAA and FPA metrics for our proposed DQN approach with the baselines' public results from the Spotify Challenge. This is because they are better than those that we obtained from our local evaluation and to provide an as fair as possible comparison. Our proposed approach exhibits significant improvements over all baselines on both MAA and FPA metrics. Our proposed DQN registers an increase of performance for both MAA and FPA of 17% and 7% respectively with regards to Multi-task RNN, the best performing baseline from the public challenge.

Overall, our results demonstrate the validity and applicability of DRL to predict users' music skipping behaviour. A Vanilla DQN architecture can outperform the more complex deep learning based state-of-the-art models. Furthermore, the results and a thorough analysis, omitted from this paper due to space limitations, also indicate that convergence is achieved using a significantly lower number of episodes, at around 2×10^5 ($\sim 1/4$ of the episodes in the training set). This suggests sample efficiency and swift convergence of our proposed approach. Thus, it also addresses the well-known problem of DRL, which is its computationally intensive and slow learning. Our approach converges swiftly and, in contrast to the selected baselines, it does not require GPU access. The low variability in performance across multiple runs and during the learning process also indicates stable and effective learning.

5.2 Identification of Temporal Data Leakage

In the previous section, we compared our proposed DQN against the selected baselines in order to demonstrate the validity of our proposed DQN. By performing an as fair as possible comparison, empirical results indicate the superiority of our approach. However, this benchmarking introduced errors into the model. This is because, as described in Section 4.4.3, we recognise that there are data leaking features in MSSD. The *SL* informs the model of how many songs a given user will listen to. This should not be made available because it is impossible to know how many songs a user will listen to in their current listening session. Further, the *RE* features provide information about how the current stream ends. This information should also not be exposed to the model. However, to provide a fair comparison with the baselines, since they are included in their input representation, these features were not removed despite our acknowledgement.

The temporal data leakage problem is validated by Figure 1, which reports the analysis of the average impact on model output (SHAP) of all features in the input state representation. It can be noted how the most discriminative feature to detect music skips is *RE Trackdone*, followed by *RS Trackdone*, *RS Fwdbtn*, and *Short PA*. *SL* is also found to have a relative impact (19th). It is clear that the proposed DQN considers these features to be of high quality and prominent importance for predicting the users' music skipping behaviour. However, they introduce a data leaking problem. By their removal from the input state representation, we observe a decrease in performance for our proposed DQN of 16% and 11% in MAA and FPA respectively. Further, we observe decreases in performance of 5.2, 26.2, 7.6, 0.6 (%) and 3.5, 28.4, 6.0, 1.3 (%) in MAA and FPA for Multi-task RNN, Multi-RNN, Temporal Meta-learning, and Weighted RNN respectively (differences calculated from the results obtained in our local evaluation after removal of the features with those reported in the public challenge). Overall, these results validate our initial intuition and demonstrate the data leakage problem. This finding provides a strong implication for a future outlook on creating attentive data collection procedures for transparent measurements of user behaviours. Offline benchmarks should be an as truthfully as possible reflection of real-world (online) tasks.

5.3 The Role of User Behaviour, Context, and Content in Detecting Music Skips (RQ2)

In this final section, we aim to address our main research question: why people skip music? To this end, we acknowledge and thus remove the leaking features from the state representation to enable for a correct modelling of the users' music skipping behaviour.

5.3.1 ***User Behaviour (UB).*** Figure 2 reports the SHAP features importance analysis of the proposed DQN on the "corrected" state representation. It can be observed that how the user interacted with the underlying platform to start the current playback (i.e., the *RS* type) is considered being the most discriminative feature to detect music skips. *Trackdone* and *Fwdbtn* are the highest negatively and positively correlated features in predicting a skip. They correspond

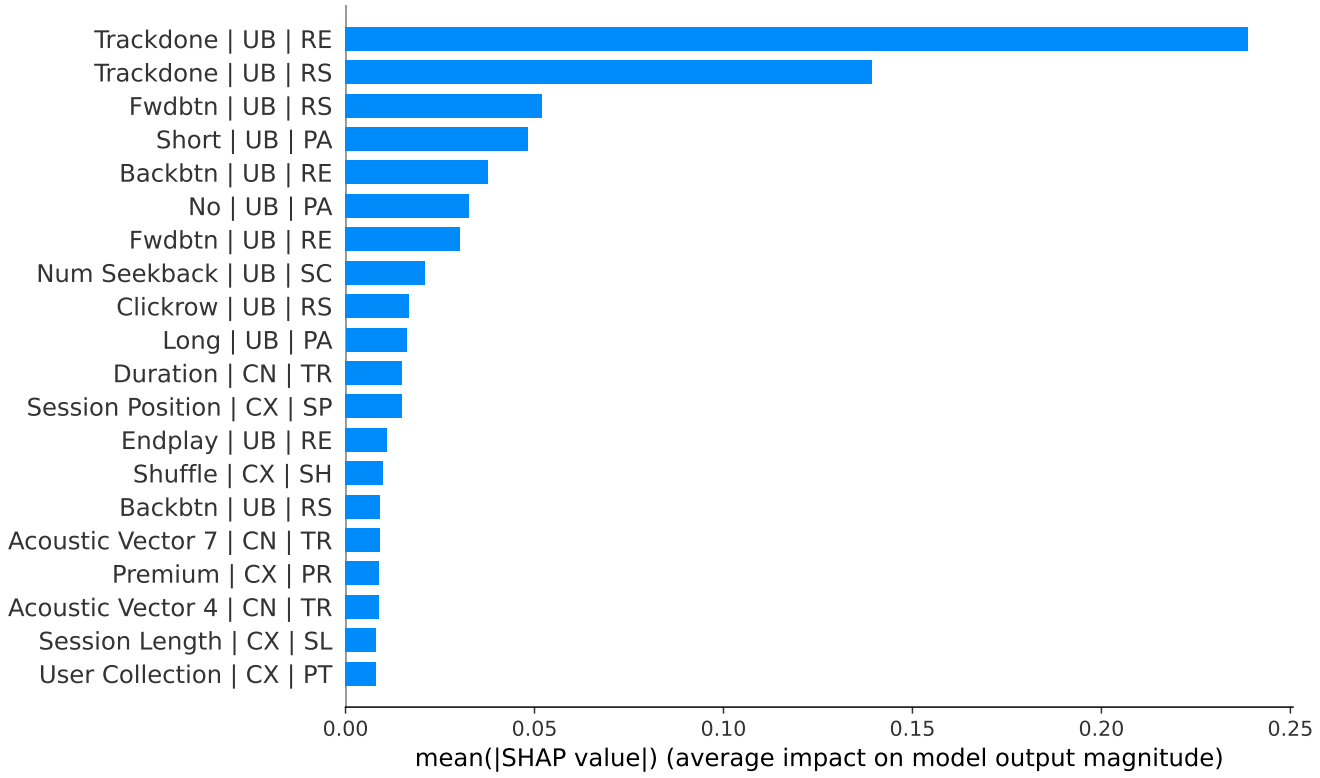

Figure 1: SHAP features importance analysis of the proposed DQN. The categorisation of the features and an explanation of the used acronyms is described in Section 4.1.3. Features are ranked in order of importance and they are reported as "[Name] | [Category] | [Type]".

to the user starting the current playback having listened in full or having pressed the forward button (i.e., skip) on the previous playback. These findings validate the recent observations by Meggetto et al. [44]. By considering their defined listener and skipper user types, we hypothesise that the user behaviour that can inform the membership of a user to one of these two types is a *RS Trackdone* or *Fwdbtn*. From our results, it is clear that how a person interacted with the previous song appears to greatly affect the DRL's ability to detect how they will interact next. Another UB that appears to have a prominent effect is the pause in between playbacks. A *Short PA* and a *No PA* are shown to highly and weakly suggest a music skip respectively. In the case of a *Long PA*, our results strongly indicate that the user will not skip their current song. This finding validates our initial hypothesis. It may correspond to a person searching the catalogue for a song they would like to listen, and hence a long pause. Therefore, it is intuitive that it may not be skipped. However, the effect of a short pause in detecting music skips is of surprising effect. This may be justified by a user's exploratory state where they browse the catalogue and briefly listen to multiple songs until they find a match for their needs.

5.3.2 ***Context (CX)***. We observe that users that listen in *Shuffle* mode and/or with a *Premium* account are associated with less skipping activity. Listening with a *User Collection* PT is associated with a higher skipping rate. It is also shown that listening under a *Personalised Playlist* or *Radio* is subject to more listening and thus less skipping activity. This finding could suggest that they have a higher users' engagement. However, this is not possible to quantify, and further evaluation is required in order to understand this phenomenon. This could be explained by the noisy nature of the skipping activity and the possibility, as in the example of radio listening, of passive (background) consumption of the music. Although the *PT* findings appear to partially validate prior work [44], in our ablation analysis we see that their removal from the state representation registers no significant effect on the DRL's ability to predict music skips.

5.3.3 ***Content (CN)***. The only content-based features in the MSSD are related to the track being listened by the user (*TR*). The correlation between skipping activity and the *TR* features is less obvious since they appear to be less discriminative and promiment in detecting music skips. *Beat Strength* and *Key*, although mostly centred around a zero impact, suggest that a high beat strength is associated with more listening, and a high-pitched song (*Key*) with higher chances of skipping. Further, longer songs (*Duration*) are usually associated with higher listening activity, although they may also correspond to skips. However, in our ablation analysis, we observe the no effect in the DQN's performance by the removal of all *TR* features. We find this to be of surprising effect, since it appears

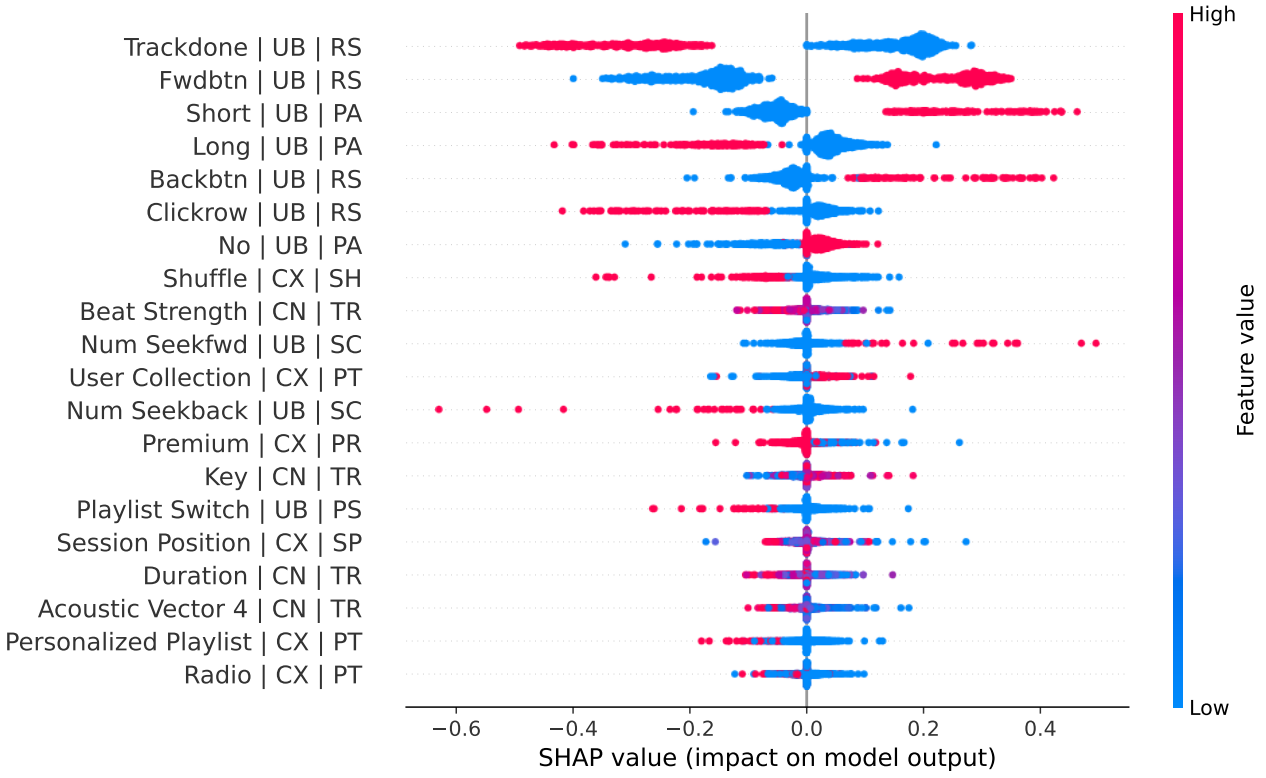

Figure 2: SHAP features importance analysis with positive (skip) and negative (no skip) impact values of the proposed DQN on a "corrected" state representation (i.e., after addressing temporal data leakage). The Feature Value axis refers to high or low observational values. For Boolean features (e.g., *RS Trackdone*), high/red is a True value, and low/blue is False. The categorisation of the features and an explanation of the used acronyms is described in Section 4.1.3. Features are ranked in order of importance and they are reported as "[Name] | [Category] | [Type]".

Table 3: MAA and FPA results for our ablation analysis on the proposed DQN on the corrected state representation. The reported results are the average across all test sets and the 95% CIs. (*) and (**) indicate that the selected type of features had a statistically significant effect in performance in the proposed DQN (on a "corrected state") on MAA or FPA. This is based on confidence levels ($p < .05$) and ($p < .001$) respectively.

		MAA		*FPA*	
		Mean	95% CI	Mean	95% CI
	Corrected State	0.664	[0.662 - 0.666]	0.773	[0.772 - 0.774]
UB	Reason Start (RS)	0.389 (**)	[0.378 - 0.400]	0.479 (**)	[0.464 - 0.494]
UB	Pauses (PA)	0.659 (*)	[0.657 - 0.661]	0.769 (*)	[0.768 - 0.770]
UB	Scrubbing (SC)	0.659	[0.655 - 0.663]	0.770 (*)	[0.768 - 0.772]
UB	Playlist Switch (PS)	0.662	[0.659 - 0.665]	0.773	[0.772 - 0.774]
CX	Hour of Day (HD)	0.663	[0.661 - 0.665]	0.773	[0.772 - 0.774]
CX	Playlist Type (PT)	0.663	[0.661 - 0.665]	0.772	[0.771 - 0.773]
CX	Premium (PR)	0.664	[0.662 - 0.666]	0.773	[0.772 - 0.774]
CX	Shuffle (SH)	0.663	[0.660 - 0.666]	0.774	[0.773 - 0.775]
CN	Track (TR)	0.664	[0.661 - 0.667]	0.773	[0.772 - 0.774]

to contradict prior research suggesting that audio characteristics influence how people skip music [14, 49].

5.3.4 Ablation Analysis. In order to validate our findings and to demonstrate the impact, whether statistically significant or not, that these features have on the DQN's performance, in Table 3 we report the results for the ablation analysis. We performed paired t-tests on the prediction accuracy of the proposed DQN (on the "corrected" input state representation) with each of the selected type of features (e.g., *RS*). We use (*) and (**) to denote the fact that the removal of the selected type of features had a statistically significant effect in performance in the proposed DQN on MAA and FPA. This is based on confidence levels ($p < .05$) and ($p < .001$) respectively. We note how the *RS* features type, as previously shown in Figure 2, is the highest quality estimator to detect music skips. Its removal registers a decrease in performance of 28% and 29% in MAA and FPA respectively. The *PA*s also register a significant impact. All the remaining features, including the CX and CN categories, do not appear to show a statistically significant effect on the DQN's performance. These results, therefore, suggest that a limited amount of users' data can be indeed leveraged to predict the users' music skipping behaviour, with only the *RS* and *PA* user behaviours showing a statistically significant effect.

6 DISCUSSION & CONCLUSIONS

In this work, we aim to understand why people skip music. To carry out such an analysis, we first proposed to leverage DRL to the task of sequentially predicting users' skipping behaviour in song listening sessions. By first understanding how a DRL model learns individual user behaviours, we can then help the process of explaining recommendations of a DRL-based MRS. To this end, we extended the DRL's applicability to this classification task. Results on a real-world music streaming dataset (Spotify) indicate the validity of our approach by outperforming state-of-the-art deep learning based models in terms of MAA and FPA metrics (**RQ1**). By empirically showing the effectiveness of our proposed approach, our main post-hoc and ablation analysis revolves around a comprehensive study of the utility and effect of users' historical data in how the proposed DRL detects music skips (addressing **RQ2**).

Our findings indicate that how users interact with the platform is the most discriminative indicator for an accurate detection of skips (i.e., *RS* and *PA*). Surprisingly, the listening CX and CN features explored in this work do not appear to have an effect on the DRL model for the prediction of music skips. Our analysis also reveals a temporal data leakage problem derived from some features in the dataset and used in the public challenge, since they provide information from the future that should not be made available to a live predictive system. Overall, this work shows that an accurate representation of the users' skipping behaviour can be achieved by leveraging a limited amount of user data. This offers strong implications for the design of novel user-centred MRSs with a minimisation and selection of high-quality data features to avoid introducing errors and biases. The results and a thorough analysis of our proposed approach indicate sample efficiency, swift convergence, and long-term stability of our proposed approach. With convergence reached using a significantly lower number of episodes, training time can be greatly reduced by early termination. With no GPU access required (in contrast to the state-of-the-art deep learning based models), our approach also clearly addresses the well-known limitation of DRL being a computationally extensive approach. These findings and the consistent performance with no signs of instability make this work of great interest for future research.

With the importance of modelling and understanding the users' skipping behaviour, we believe this work to be an important step towards improving user modelling techniques. An accurate representation of the skipping behaviour can provide an invaluable stream of information to the underlying recommendation process. For example, we expect our findings, e.g. the *RS* type, to be highly relevant in the downstream task of capturing, in real-time, a user's skipping type [44]. By extending our approach to predict and understand other users' behaviours, we can create a holistic representation of the listeners' preferences, interests, and needs. We also advocate for thoughtful considerations when collecting and then presenting data to a model for measuring user behaviours. With increasingly rising concerns around users' data collection and privacy, the need for minimal data collection is paramount. Our proposed approach can be extended in future works to predict *when* the song is likely to be skipped. This level of information could allow to predict moments in a song where skips are most likely to occur, which could be of great value for the underlying platform. Considering *how* user's emotions or current psychological state affect their skipping behaviour is also an interesting venue for further research. With access to richer behavioural data and non-anonymised listening sessions, another line of research can investigate the relation between skipping signal and the individual user's preferences (e.g., situation-aware MRS). Finally, although not the aim of this work, performance improvements are to be expected by further tailoring our approach to the music skip prediction task. Given the user-based exploratory nature of this work, we leave further experimentation and evaluations with emerging DRL model-free offline algorithms and architectures (e.g., extending our analysis to transformer-based DRL models [30]) for future investigation.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/R513349/1].

REFERENCES

[1] Sainath Adapa. 2019. Sequential modeling of Sessions using Recurrent Neural Networks for Skip Prediction. *arXiv preprint arXiv:1904.10273* (2019).
[2] Darius Afchar and Romain Hennequin. 2020. Making neural networks interpretable with attribution: application to implicit signals prediction. In *Fourteenth ACM Conference on Recommender Systems*. 220–229.
[3] M Mehdi Afsar, Trafford Crump, and Behrouz Far. 2021. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys (CSUR)* (2021).
[4] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. 2020. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*. PMLR, 104–114.
[5] Snehasish Banerjee and Anjan Pal. 2021. Skipping skippable ads on YouTube: How, when, why and why not?. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. IEEE, 1–5.
[6] Daniel Belanche, Carlos Flavián, and Alfredo Pérez-Rueda. 2017. User adaptation to interactive advertising formats: The effect of previous exposure, habit and time urgency on ad skipping behaviors. *Telematics and Informatics* 34, 7 (2017), 961–972.
[7] Daniel Belanche, Carlos Flavián, and Alfredo Pérez-Rueda. 2020. Brand recall of skippable vs non-skippable ads in YouTube: Readapting information and arousal to active audiences. *Online Information Review* 44, 3 (2020), 545–562.

[8] Ferenc Béres, Domokos Miklós Kelen, András Benczúr, et al. 2019. Sequential skip prediction using deep learning and ensembles. (2019).

[9] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. 2018. Latent cross: Making use of context in recurrent recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 46–54.

[10] Klaas Bosteels, Elias Pampalk, and Etienne E Kerre. 2009. Evaluating and Analysing Dynamic Playlist Generation Heuristics Using Radio Logs and Fuzzy Set Theory.. In *ISMIR*, Vol. 9. 351–356.

[11] Brian Brost, Rishabh Mehrotra, and Tristan Jehan. 2019. The music streaming sessions dataset. In *The World Wide Web Conference*. 2594–2600.

[12] Sungkyun Chang, Seungjin Lee, and Kyogu Lee. 2019. Sequential Skip Prediction with Few-shot in Streamed Music Contents. *arXiv preprint arXiv:1901.08203* (2019).

[13] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. 2018. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[14] Jonathan Donier. 2020. The universality of skipping behaviours on music streaming platforms. *arXiv preprint arXiv:2005.06987* (2020).

[15] Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, and Patrick Gallinari. 2011. Datum-wise classification: a sequential approach to sparsity. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 375–390.

[16] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. 2021. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning* (2021), 1–50.

[17] Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 32.

[18] Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. 2019. Skip prediction using boosting trees based on acoustic features of tracks in sessions. *arXiv preprint arXiv:1903.11833* (2019).

[19] Benjamin Fields et al. 2011. *Contextualize your listening: the playlist as recommendation engine*. Ph. D. Dissertation. Goldsmiths College (University of London).

[20] Scott Fujimoto, Edoardo Conti, Mohammad Ghavamzadeh, and Joelle Pineau. 2019. Benchmarking batch deep reinforcement learning algorithms. *arXiv preprint arXiv:1910.01708* (2019).

[21] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*. PMLR, 1587–1596.

[22] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*. PMLR, 2052–2062.

[23] Christian Hansen, Casper Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. 2019. Modelling sequential music track skips using a multi-rnn approach. *arXiv preprint arXiv:1903.08408* (2019).

[24] Casper Hansen, Christian Hansen, Lucas Maystre, Rishabh Mehrotra, Brian Brost, Federico Tomasi, and Mounia Lalmas. 2020. Contextual and sequential user embeddings for large-scale music recommendation. In *Fourteenth ACM Conference on Recommender Systems*. 53–62.

[25] Christian Hansen, Rishabh Mehrotra, Casper Hansen, Brian Brost, Lucas Maystre, and Mounia Lalmas. 2021. Shifting Consumption towards Diverse Content on Music Streaming Platforms. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 238–246.

[26] Matthew Hausknecht and Peter Stone. 2015. Deep recurrent q-learning for partially observable mdps. In *2015 aaai fall symposium series*.

[27] Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. 2019. Classification with costly features using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3959–3966.

[28] Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. 2020. Classification with costly features as a sequential decision-making problem. *Machine Learning* 109, 8 (2020), 1587–1615.

[29] Dietmar Jannach, Massimo Quadrana, and Paolo Cremonesi. 2022. Session-based recommender systems. In *Recommender Systems Handbook*. Springer, 301–334.

[30] Michael Janner, Qiyang Li, and Sergey Levine. 2021. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems* 34 (2021), 1273–1286.

[31] Olivier Jeunen and Bart Goethals. 2019. Predicting Sequential User Behaviour with Session-Based Recurrent Neural Networks. (2019).

[32] Alexander Kuhnle, Michael Schaarschmidt, and Kai Fricke. 2017. Tensorforce: a TensorFlow library for applied reinforcement learning. Web page. https://github.com/tensorforce/tensorforce

[33] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems* 32 (2019).

[34] Paul Lamere. 2014. *The Skip*. Retrieved Sept 22, 2022 from https://musicmachinery.com/2014/05/02/the-skip/

[35] Paul Lamere. 2015. *The Drop Machine*. Retrieved Sept 22, 2022 from https://musicmachinery.com/2015/06/16/the-drop-machine/

[36] Sascha Lange, Thomas Gabel, and Martin Riedmiller. 2012. Batch reinforcement learning. In *Reinforcement learning*. Springer, 45–73.

[37] Alexandra Langmeyer, Angelika Guglhör-Rudan, and Christian Tarnai. 2012. What do music preferences reveal about personality? A cross-cultural replication using self-ratings and ratings of music samples. *Journal of individual differences* 33, 2 (2012), 119.

[38] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).

[39] Enlu Lin, Qiong Chen, and Xiaoming Qi. 2020. Deep reinforcement learning for imbalanced classification. *Applied Intelligence* (2020), 1–15.

[40] Feng Liu, Ruiming Tang, Xutao Li, Weinan Zhang, Yunming Ye, Haokun Chen, Huifeng Guo, and Yuzhou Zhang. 2018. Deep reinforcement learning based recommendation with explicit user-item interactions modeling. *arXiv preprint arXiv:1810.12027* (2018).

[41] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[42] Coralie Martinez, Guillaume Perrin, Emmanuel Ramasso, and Michèle Rombaut. 2018. A deep reinforcement learning approach for early classification of time series. In *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2030–2034.

[43] James McInerney, Brian Brost, Praveen Chandar, Rishabh Mehrotra, and Benjamin Carterette. 2020. Counterfactual evaluation of slate recommendations with sequential reward interactions. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1779–1788.

[44] Francesco Meggetto, Crawford Revie, John Levine, and Yashar Moshfeghi. 2021. On skipping behaviour types in music streaming sessions. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3333–3337.

[45] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *Ai & Society* 35, 4 (2020), 957–967.

[46] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1928–1937.

[47] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[48] Nicola Montecchio, Pierre Roy, and François Pachet. 2020. The skipping behavior of users of music streaming services and its relation to musical structure. *Plos one* 15, 9 (2020), e0239418.

[49] Aaron Ng and Rishabh Mehrotra. 2020. Investigating the Impact of Audio States & Transitions for Track Sequencing in Music Streaming Sessions. In *Fourteenth ACM Conference on Recommender Systems*. 697–702.

[50] Elias Pampalk, Tim Pohle, and Gerhard Widmer. 2005. Dynamic Playlist Generation Based on Skipping Behavior.. In *ISMIR*, Vol. 5. 634–637.

[51] Peter J Rentfrow and Samuel D Gosling. 2003. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology* 84, 6 (2003), 1236.

[52] Peter J Rentfrow and Samuel D Gosling. 2006. Message in a ballad: The role of music preferences in interpersonal perception. *Psychological science* 17, 3 (2006), 236–242.

[53] Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2022. Music Recommendation Systems: Techniques, Use Cases, and Challenges. In *Recommender Systems Handbook*. Springer, 927–971.

[54] Guy Shani, David Heckerman, Ronen I Brafman, and Craig Boutilier. 2005. An MDP-based recommender system. *Journal of Machine Learning Research* 6, 9 (2005).

[55] Yading Song, Simon Dixon, and Marcus Pearce. 2012. A survey of music recommendation systems and future perspectives. In *9th international symposium on computer music modeling and retrieval*, Vol. 4. Citeseer, 395–410.

[56] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[57] John R Taylor and Roger T Dean. 2021. Influence of a continuous affect ratings task on listening time for unfamiliar art music. *Journal of New Music Research* 50, 3 (2021), 242–258.

[58] Charles Tremlett. 2019. Preliminary Investigation of Spotify Sequential Skip Prediction Challenge. (2019).

[59] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[60] Sergey Volokhin and Eugene Agichtein. 2018. Understanding music listening intents during daily activities with implications for contextual music recommendation. In *Proceedings of the 2018 conference on human information interaction & retrieval*. 313–316.

[61] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z Sheng, Mehmet A Orgun, and Defu Lian. 2021. A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)* 54, 7 (2021), 1–38.

[62] Wenjie Wang, Fuli Feng, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2021. Denoising implicit feedback for recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 373–381.

[63] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. 2016. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*. 1995–2003.

[64] Hongyi Wen, Longqi Yang, and Deborah Estrin. 2019. Leveraging post-click feedback for content recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 278–286.

[65] Marco A Wiering, Hado Van Hasselt, Auke-Dirk Pietersma, and Lambert Schomaker. 2011. Reinforcement learning algorithms for solving classification problems. In *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*. IEEE, 91–96.

[66] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. 2018. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 95–103.

[67] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*. 167–176.

[68] Lin Zhu and Yihong Chen. 2019. Session-based Sequential Skip Prediction via Recurrent Neural Networks. *arXiv preprint arXiv:1902.04743* (2019).