ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

Predicting Skipping Behavior in Music Streaming: The Impact of Recommendation Types on User Engagement

Bence Bilibók (663332)

Supervisor: Sean Brüggemann

Second assessor:

Date final version: 2024.07.18.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract:

Music streaming is a topic of importance in the twenty first century since it fundamentally changed the distribution of audio content. These services provide recommendations for the users in several ways on what to listen to, however there is a lack of research comparing the effects of different recommendation types on user engagement. The skipping behavior of the listeners is a good indicator of user engagement. The research was conducted on a real-world session level dataset provided by Spotify, where multiple potential skipping points during the song's duration were recorded, which all have importance regarding user engagement. Prior research only focused on the fact whether a track was skipped or not. The exact differences are not known between these points, hence the data needed to be treated as ordinal, which is a novel approach to this topic. Ordered Logistic Regression and Ordinal Forest were applied to make prediction on skipping behavior, and the results were evaluated, the tree-based method only performed slightly better. To understand the impact of the predictors, the SHAP values were utilized. This approach provided valuable insights on the effects of different recommendation types and other variables (pause before play, session position etc.) on the four different skip points. Users are generally most engaged in algorithmically created personalized recommendations and less with mood or genre based automated radio stations. The users own created playlists performed better than radios overall, and no real conclusion can be said about the song recommendations made by industry professionals.

Keywords: Music streaming, Spotify, Music recommendation, User Engagement, Ordinal Forest, SHAP

Table of Contents

Introduction

Digitalization and the rise of the internet changed almost every aspect of life in the twenty first century. One of these changes is the focus of this thesis: online music streaming. This technology provides a new way to distribute and consume music and has fundamentally changed the music industry over the last two decades. The changes affected the creators, the distributors, and the consumers as well, creating a vast number of effects and behavior changes to be studied in an academic context. One of the most active areas of research is on recommendation systems, which suggests (new) songs to the users of music streaming services. The idea behind recommendation systems is to aid users in selecting which songs to play by recommending tracks that match with their preferences. This way users do not have to find each one they want to listen to in a seemingly endless library, but rather rely on suggestions based on their listening history and taste.

There are three distinct categories of these recommendation types. The first is the user's own playlists, where they can create their own collection of songs they like to listen to and are already familiar with it. It can be argued whether this type can be considered recommendation. Having said that, in this case users gathered their preferred songs based on previous listening history, and they are recommending them for themselves later, which would definitely indicate that these are a form of recommendations as well. In contrast, there are playlists curated by the streaming platform's professionals, based on several criteria. The third type of playlists are also created by the platform; however these are all created algorithmically based on the users' preferences and listening history. These can be personalized for each listener, or contain songs based on chosen genres, artist seeds, or mood selections, these are the automated radio stations. The goal of this thesis is to explore the effectiveness of the different types of recommendations regarding user engagement.

My research will focus on the following research question:

How do users engage with different types of music recommendation (e.g., owned playlists, curated playlists, algorithmic playlists, and radio)?

Background

Streaming

The rapid advancement of digital technologies has transformed the way society consumes and interacts with multimedia content. One of the most interesting developments in this field is the emergence of streaming technology, which has revolutionized the delivery and consumption of digital media. Streaming was defined as the following (Küng, 2017, p. 34): “The customer requests content from a streaming service via smartphone, tablet, TV or PC. The main server of the streaming provider (…) then sends the content in data packets over wireless or cable to the consumer’s internet-connected device where packets are assembled (and deleted after consumption).” The expression streaming itself is not about the contents, but rather about the data transfer method. Streaming refers to the process of transmitting data, such as audio or video, in a continuous flow, allowing users to access and experience the content in real-time without the need to download the entire file. This way consumers never own the content, they only rent it from the provider. The definition also emphasizes that content is delivered to consumers in data packets, rather than in a single large file as with traditional downloads. This has the advantage that immediate access is provided to specific content without waiting for the entire file to be downloaded from the internet.

Streaming also has some further advantages compared to regular media, such as Radio or Television. In my opinion, the most significant aspect of streaming services is the user's freedom to choose the content they consume. These services typically offer a vast amount of content in various categories and quality levels, catering to every user's preference. Online streaming allows users to constantly interact with the platform in order to pause, continue, skip, or replay video or audio content at their convenience, giving them the freedom to manage their experience. In contrast, traditional media present a predetermined program at a specific time, limiting the user’s choice to selecting the channel or time slot. The contents’ pausing and playback options enables users to consume movies or lengthy podcasts in multiple installments, further enhancing their experience.

Another reason for streaming is the opportunity to explore content without actually owning any, hence no need to purchase it. This concept offers consumers the chance to browse through enormous libraries to find their optimal content with respect to their current preferences, without the need to buy any of it and rather just rent it.

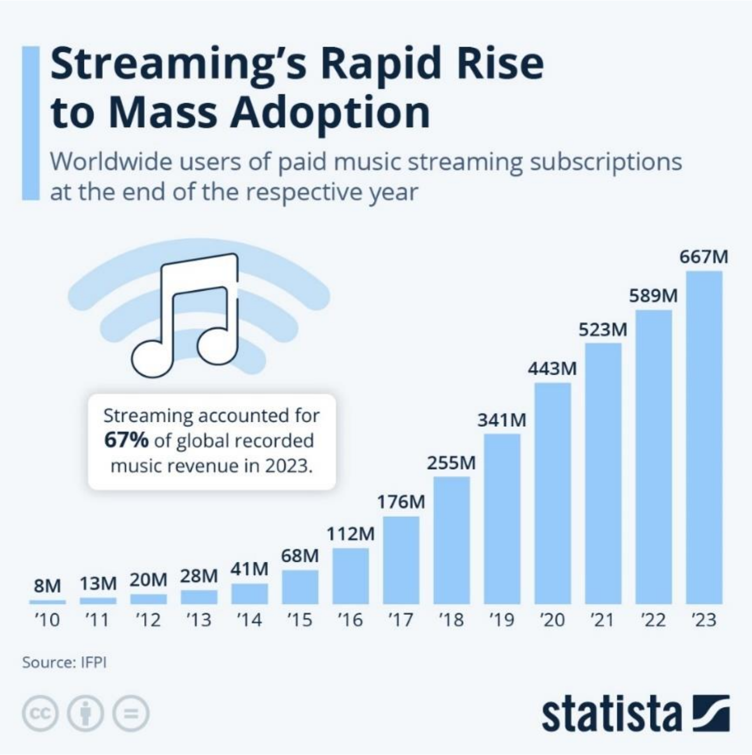
Naturally streaming has some downsides as well, as any service. Although the majority of the developed world has access to the internet, it is worth mentioning that this concept requires constant internet access, as well as a smart device, such as smart phone, smart tv or a laptop/PC. If the quality of the internet service is not optimal, the content can easily be lower in quality, or the constant playing can even be stopped, causing a disturbance for the consumer. Another aspect is the quality of the content. A headphone or an at home stereo speaker cannot be compared to a concert or a music hall in terms of quality and experience in the case of music streaming. Naturally, the purpose of these events is different than everyday use of streaming, but it is worth noting that.

By 2024, online content and services were well known, particularly after the pandemic, with many options available in online space. Just to mention a few important one: online banking services, consultations, news, and various forms of media content. Online media has a diverse range of formats, from live sports broadcasts to audio-only content like music and podcasts. Online education could also be mentioned, which has also undergone a remarkable transformation in recent years. A portion of course materials is now usually accessible via streaming, and online classes are conducted as live streams as well in some cases. The significance of this topic has been explored well before the widespread adoption of streaming services and the covid-19 lockdowns (Hartsell et al., 2006).

Music streaming

While video steaming is a constantly expanding industry as well, which is indicated by the popularity of several huge providers, such as Netflix, Amazon Prime or Disney Plus just to mention a few, the focus of this research will only be on music streaming. While it is mainly referred to as music streaming, the term stands for transmitting audio data online in real time without the actual need to download the data. This indicates that the technology can be used with other types of audios, such as podcasts and audio books. The rise of music streaming platforms fundamentally changed music consumption behavior and has provided users with instant and on-demand access to a huge library of songs from any web-connected device, overcoming the limitations of physical media and traditional radio.

The early rise of music sharing in the online space already began in the 2000's with the appearance of Napster, an illegal file sharing website, which greatly impacted the revenue loss of the music industry (Zehr, 2021). Zehr's article mentions how the digitization of music at the start of this century led increased online piracy, which caused revenue decline for the music industry. This opened the gate for the emergence of music streaming as an alternative. The paper suggests that while there is less industry profit from music ownership compared to the pre-streaming era, this model has helped to repel online music piracy, allowing the industry to return to its original levels of profit. With the rapid widespread of the internet and digital technology, consumer behavior has also changed, streaming services now provide access to a wider variety of music than ever before. This has led to an increased music discovery and consumption. Music streaming services became widely known and used around 2015-2017 according to Statista. Figure 2.1 shows the number of subscribers at the end of each year.



*F*igure 2.1: Music Streaming Service Subscribers Worldwide in millions between 2008-2023 (adapted source: Richter, 2024)

Music streaming services can operate based on two major business model types, premium or freemium. (Carroni & Paolini, 2017). Premium works as subscription-based service, for a monthly or annual fee the costumers gain unlimited access to a large amount of high-quality audio content and the means to listen to them however they see fit. However, freemium provides a service without costs, but with certain restrictions. These could include a limited number of listening or the use of certain features, such as searching or jumping between tracks. However, the main attributes of a freemium model are the advertisements, which the consumers on the freemium model need to listen to. In this version, the revenue from the services is generated through advertisements, based on either clicks or views. While these two models seem distant, the most common business model of the streaming platforms is the combination of the two, where both types of service are provided, which is called the two-tier freemium model (Thomes, 2013).

The third key players of the revenue streams are the artists, creators, who are receiving royalties after their content based on the number of streams their music receives. The structure of the payments can be quite complex, for example different rates for different models (freemium or premium), moreover it could also vary between different record label companies and artists. Per-stream payments are typically only a fraction of a cent, though it can be dependent on specific agreements as well. Another factor in the payment structure is the label companies, who are taking a share from the major artists, leaving them with a smaller percentage of their revenue compared to independent artists, who are operating without a label company.

Music streaming has a definite impact on today's music consumption, and it also fundamentally changed how musicians receive money after their content, which naturally created some controversy (Hesmondhalgh, 2020). From the perspective of Musicians, a couple of reasons can be mentioned against the Music Streaming Services. First of all, they argue that the royalty rates per stream are quite low, which makes it difficult for the artists to maintain a sustainable income from streaming alone, which is especially true for small and independent artists. Another concern of the lesser-known artists is the unequal distribution of the streaming revenue. Some argue that major artists and label companies receive a larger share compared to independent artists, making their situation even more difficult. Naturally, Streaming services had some positive effects on the industry, such as increased accessibility of audio content. Providing a platform where listeners can access a more audio content than ever made music much more available to anyone, which helped lesser-known artists to gain exposure. Before online streaming, the distribution of their content was much more difficult for these artists. A great problem for the music industry in the beginning of the 21st century was piracy (Zehr, 2021). However, by providing a legal alternative, and a more accessible platform, streaming services greatly reduced the revenue loss of the music industry, which was a historically significant issue in the early 2000s. On one hand, royalties may be low, on the other hand, streaming provides an additional revenue to the artists, completing their other sources of income, such as live concerts and partnerships. These arguments can provide a clearer picture of the industry and the streaming services' impact on it. However, complex issues continue to be present, and while these services greatly increased music consumption, they often fail to account for the diversity of the artists and genres, often to the detriment of smaller, lesser-known artists.

The competition of the rising market produced several relevant companies, whose platforms are being used by millions of people every day. As of 2023 Q3, this number reached 713 million subscribers worldwide (Mulligan, 2024). There are several prominent players on the market, such as Apple Music with a 12.6 % market share or Amazon Music with 11.1 percent. The previous platforms have the majority of their customer base in Westerns countries, while Tencent Music operates in China have a 14.4 percent market share considering the total global market. The highest number of subscribers (226 million) belong to Spotify, with 31.7 percent of the total music streaming subscribers. It cannot be said that they absolutely dominate the global market, however the platform has more than twice as many subscribers as the second and third competitor, which makes them the most relevant streaming platform today. Based on these facts, it is clear that a research would have the most relevance, if it is done on data about Spotify's music streaming platform.

Spotify

In order to thoroughly understand the background of the research, a few key details about the examined music streaming platform need to be discussed as well. The platform was launched in 2008 in Sweden, and since then became the world's leading music streaming platform. The platform operates on the two-tier freemium model, which was discussed earlier. According to the company's 2023 annual report, in 2023 Q4 Spotify had 602 million active users, of which 236 were premium subscribers (Spotify Technology S.A., 2024). While only around 40 percent of the users subscribe to the platform's premium plan, this segment provides 87.32 percent of the company's revenues, which by the end of 2023 reached 11.57 billion euros. This is a 12.88 percent increase compared to the previous year, proving that the company is still expanding. Compared to this, the ad-supported revenue was only 1.68 billion euros (12.68 %). Another interesting fact from the report is the geological distribution of the subscribers, almost 70 percent of their premium subscribers are from Europe and North America.

The compatibility of the application covers a wide range of devices: smartphones, laptops, tablets, Smart TVs (etc.). The platform is compatible with all the major operating systems as well, such as Windows, Android, and iOS. This broad accessibility has enabled Spotify to reach a great international audience, making it available in numerous countries and regions. The platform has some additional features that make it engaging, for example, the user-friendly interface and the ability to download content for offline listening. However, undoubtedly the most important feature of the platform is the recommendation system.

Multiple types of playlists are being utilized by Spotify to provide recommendations for their uses, among these are human created as well as algorithmically created. Spotify uses advanced deep-learning algorithms to curate personalized playlists and recommend new tracks to users. The algorithm analyzes a user's listening history, preferences, and behavior to generate recommendations suited to individual preferences. Numerous factors are considered such as the audio features of the songs (e.g., tempo, key, acousticness), the artists and genres the user listens to. The listening patterns and behaviors (e.g., skipping songs, pausing between songs, or listening to entire albums) are also being used in this process. By analyzing this data, the recommender system can identify similarities between songs and create playlists that align with the user's preferences (Freeman et al., 2022). Another key aspect of the algorithmically curated playlists is a technique called collaborative filtering, which analyzes the listener's preferences and finds users with similar preferences to make more valid recommendations. Furthermore, Spotify's algorithms are developed in a way to balance familiarity and novelty in their recommendations. While they aim to suggest music that aligns with a user's current preferences, they also introduce some novelty by recommending lesser-known artists or songs that are further from the user's typical choices. Algorithmic playlists are not limited only to personalized playlists. The platform also has so-called automatic radio stations, which can play thematic music, similar to traditional radio stations, however providing an improved experience through customization. Listeners can easily create such radio stations based on artists, albums, or songs. Spotify’s algorithm then creates a playlist that gathers similar tracks based on the above-mentioned factors, allowing a continuous listening experience. The algorithms and recommender systems are constantly evolving and adapting based on user feedback and interactions, such as skipping a song or listening to it in its entirety. These data-driven insights help the algorithm to be up to date on user preferences and evolve over time. Naturally, the exact working of the company's recommendation system is a commercial secret. These personalized playlists include "Daily Mixes" for individual day-to-day recommendations in multiple themes, "Release Radar" featuring newly debuted songs, or "Discover Weekly" with multiple individually recommended songs encouraging the adoption of new tracks each week.

Playlists are also curated manually by Spotify professionals who are constantly monitoring songs, analyzing streaming numbers and deciding what to include in certain thematic playlists. While it would not seem that straightforward, human-created playlists also heavily rely on data-driven insights (Freeman et al., 2022). Several statistics and insights can help a human professional what to include in certain editorial playlists. These could also be created based on mood, genre, or activities (study, meditation, or sports). Another version of editorial playlists is country specific, which provides localized recommendations based on local culture and music genres, but Spotify also partners with celebrities and influential artists to create playlists based on their own tastes. Through these diverse types of editorial playlists, the platform provides the opportunity to explore music beyond our individual tastes and preferences. This leads to the exploration of new types of songs and increasing music consumption overall. These playlists include among others Today's Top Hits" for the most streamed songs each day, "Rap Caviar" for genre specific music. Other recommendations provide tracks during sporting activities, such as the "Gym mix", and "Hot Hits Netherlands" represents country specific music suggestions for the Netherlands.

Among all these specific playlist types, the platform allows the freedom for its users to create their own playlists, consisting of songs they want to listen to in a continuous experience without the need to constantly search for tracks in the library. The unique feature of these playlists is that all the songs present are known to the user since they have chosen to put it on the lists. On one hand, these playlists lack the ability to offer new music for users, on the other hand they consist of songs that quite possibly are the most engaging for users that created them. Having said that, conducting different playlists for several situations and moods requires great effort and time, which not everyone is willing to do, and rather leave these decisions to be made by algorithmic suggestions and industry professionals.

The success of Spotify's music streaming platform is indicated by two major factors. The first is the number of customers that choose to subscribe for the premium version of the service. The second is the success of the recommendation, since it describes how users react to the music presented to them. Their goal is to be as engaging as possible to keep them on the platform, avoiding churning or switching for a concurrent platform such as Apple Music or YouTube Premium. The success of playlists and different recommendation types can be measured in the fact that whether the tracks were listened to or skipped. A low skip through rate is the goal of the platform because that way the users are most engaged with the recommended songs and the platform itself.

Skipping in another domain

Not finishing a product's consumption is not an entirely new phenomenon, and it is not specific to music streaming. Skipping a song can be considered as not finishing the product provided by a service. This kind of consumer behavior can be examined in other areas of business as well. Newspapers are particularly relevant here, since a lengthy article can easily be abandoned by readers, if it is not engaging enough, leaving it unfinished. Since the rise of online newspapers, this became easier to measure as well, compared to regular printed newspapers. Another interesting topic could be food consumption and plate abandonment in restaurants, where customers have not finished eating their food. This is easily measurable as well, and analyzing and understanding which types of dishes are mainly left on the plate can help restaurants improve them and help reduce food waste. The reasons behind not finishing a meal could depend on several factors, such as customers not communicating their taste, proportion preferences and allergies clearly to the stuff (Cerrah & Yigitoglu, 2022). On the other side restaurants have responsibilities as well, providing a clear menu with dishes that are generally well consumed, and accepting feedback from customers to improve their service and products, meals in this particular case.

Literature review

Music streaming has been the topic of several papers in the past. Multiple articles were published regarding different types of playlists, the prediction of skipping, and several works with different methods and datasets.

The core topic of this research is the playlists, and their differences. The work of Pachali and Datta (2023) focuses on Spotify editorial playlists, which are professionally curated. The paper reflects on how they are created and discusses two major effects. The findings include that it raises the number of daily followers of a playlist by 0.95% if it is featured in the applications search page and the positive effect on the followers is approximately 0.45 % if a major label artist's song is added to it. The paper focuses solely on editorial playlists, and their role, but not comparing them to other kinds of music recommendations. Playlists can be looked at from a different angle as well, Spotify's playlists have effects of their own, for instance on the user behavior during the playing of the songs and the discovery of new artists. These effects were analyzed on editorial playlists as well (Aguiar & Waldfogel, 2018), since these general playlists have the widest audience reach. The authors of the paper employ four empirical approaches to measure the impact of being included in a playlist on a song's performance. They have found that being added to Today's Top Hits, a list followed by 18.5 million people during the sample period (2018), raises streams by almost 20 million. If a song is included in the New Music Friday lists, the probability of song success greatly increases, including for new artists. The paper investigates the performances of songs, which are measured by the number of streams by users, however other engagement metrics have not been discussed.

Another type of playlist is the ones that are created by the users themselves, the work of Pichl, Zengerle and Specht (2017) focuses on that topic. In their paper they have examined Spotify data with PCA and Clustering techniques to create meaningful segments of playlists. They have found that playlists show distinct patterns based on factors like mood, activity, genre, and time of day. According to their results, user-curated playlists on Spotify reflect specific contextual factors and listening situations. For instance, playlists created early in the morning tend to contain a different mood and genre mix compared to playlists created in the evening. Playlists associated with activities such as working out or studying also show distinct musical characteristics. Based on these findings, they were able to predict the context of a playlist (e.g., mood, activity) only from musical characteristics with 80 % accuracy.

Modeling user behavior, especially track skipping, was the subject of various research. The work of Hansen (et al., 2020) examines sequences of Spotify listening sessions, although different from the one that will be used for this research. They have proposed a deep learning framework that incorporates both contextual information (e.g. time, location, mood, device) and sequential listening patterns to generate user embeddings. The model was able to capture complex relationships, when predicting whether a song was skipped or not. The findings included that recent listening history has a significant effect on skipping behavior, also their model's effectiveness greatly increases, if the type of the playlists (in their research stream source) is known. This validates the idea that playlists among other contextual factors have a significant effect on skipping predictions. However, in their research the roles of various kinds of playlists were not explored. The analysis of skipping behavior was covered by Meggetto (et al., 2021) as well. The paper investigates different behaviors during entire listening sessions analyzing the users' session-based skipping activity on music streaming platforms. K-means clustering was applied to group the customers into four distinct groups, additionally they have split the data based on session length into three parts (10-15-20). The researchers managed to identify four main types of session skipping behavior, providing insights into how users interact with musical content during their listening sessions. The four types were Listener, Listen then skip, Skip then listen, and Skipper based on their skipping behavior through their entire listening session. The analysis of short, medium, and long listening sessions reveals that these skipping types remain consistent across sessions of varying lengths, indicating stable patterns of user behavior regardless of session's duration. They have found how skipping behavior changes throughout the day, but also how it changes between different playlists. The Skipper behavior type was lower for Spotify generated personalized playlists and for catalog as well. The highest degree of skipper activity was found in the group's playlists of: User collections, radios and charts. These were all consistent on all session lengths. While these findings are interesting and valid basis for further research, the findings do not include when these behavior groups are skipping the songs and how the different playlists influence that. Therefore, it leaves room for further research in looking into the different kinds of skipping that can happen during the entire track.

These papers focused on editorial playlists and user-created playlists on their own, examining how curation and user behavior influence playlist creation and impact. These studies focused on one type of playlist and did not dive more into their comparison. Multiple models were built on Spotify session level data, predicting skipping behavior, however only the fact that a track was skipped or not was predicted. A previous study used the same data to identify skipping behavior groups, and the role of different playlists in these groups, however no additional information was found about the effects of playlists on the different skipping points. This research will focus on comparing different playlists, to see how they influence the prediction of skipping behavior. However, in these cases it will not be considered whether it occurred or not, but rather which point did it occurred.

Managerial relevance

The purpose of the recommendation system is to maximize user engagement (or minimize disengagement, such as skipping). However, user preferences can vary across several factors, such as play contexts, situations and/or mood. The context of listening in this case means the playlist type, which the songs are a part of. As I have mentioned above, there are three types of recommendations: The user's own playlists, professionally curated playlists, and algorithmic recommendations. Navigating between these types of recommendations can be challenging for a streaming service, each customer's reaction and action to these can differ.

Understanding the differences in playlist types can help music streaming services in the optimization of their content recommendations to better fulfill the users' preferences. By identifying the playlist types that are most likely to lead to a lower skipping rate, services can adjust their algorithms to prioritize content that is more likely to engage users. This can most of all lead to increased user satisfaction, which is also important in reducing churn rates and improving overall experience on the platform. The findings of this study could also inform the development of personalized playlists that are tailored to individual user preferences. By analyzing the recommendation types that are most well-received among users, services can create playlists that are more likely to increase user engagement.

The findings of the research could also encourage streaming services to review the available playlist options within the three main types and optimize skipping rates by creating new types of recommendations, which are the combination of different recommendations. For example, users could mix personalized playlists and owned playlists by creating playlists that recommend the artists, genres, and moods that they select for that exact playlist.

In addition to the benefits of music streaming services, the findings of this study can also have a significant findings for the content creators as well. For example, if a song is skipped within the first 30 seconds, artists and labels are not receiving any payment for their product. Their goal is also to extend the time while their song is listened to. Understanding the effects on different skipping types can create a difference in their work as well. By understanding the key elements that influence skipping behavior, musical and other content creators can better adjust their content to engage users and increase their chances of receiving some amount of payment after the rights. This can be particularly important for independent artists, who may rely heavily on streaming revenue to support their careers.

In conclusion, the topic of playlist types and their impact on skipping behavior is relevant for all three participants of the streaming market: The service, the customer, and the content creator. By understanding the differences in playlist types and skipping points, music streaming services can optimize their content recommendation systems, develop new playlist features, and reduce skipping activity. Additionally, the study's findings can help content creators as well, informing them of the differences, which can have a significant impact on their work as well.

Academic relevance

The topic has relevance in the academic prospect for several reasons. Firstly, the rise of music streaming has fundamentally transformed the way people consume and interact with music, reshaping the user's preferences, and moreover their behaviors. This study will contribute to a deeper understanding of these evolving behaviors, and the interactions with different recommendation systems, playlist types.

Secondly, the paper's emphasis on analyzing multiple types of skipping behavior, rather than simply examining whether it occurs or not, is an innovative approach that has not been extensively explored in previous research in the context of music streaming. Treating the skipping as an ordinal variable during a track can provide insights into when skipping occurs other than the fact that it happened. This aspect could also open the door to apply previously unused methods in the context of streaming, which are tailored for ordinal type of data. This approach to the topic could reveal insights that may have been overlooked in previous studies, advancing our understanding of skip behavior in music streaming contexts.

The role of playlists is highly relevant, as they have turned into a crucial aspect of music streaming platforms, influencing user consumption through different recommendation types. By investigating potential behavior patterns related to different playlist types, this research could make contributions to our understanding of how users interact with them. It allows for a deep examination of the role these playlists have in creating certain listening behaviors, such as skipping patterns, and how different curation techniques may influence user engagement and the consumption of music. Potential behavior patterns could reveal that for example listening to an algorithmic playlist decreases the possibility of skipping at the beginning of the song, since the user is not familiar with it, but increases the skip ratio later on during the track if the algorithm is not working fine and the user is not interested in the song.

To summarize, this paper could offer relevant findings in music streaming behavior, present a new method in the domain of skip predictions, and analyze differences in certain music recommendation systems. The results could encourage new theoretical discussions and empirical investigations in areas such as user experience design, exploring the effects of digital music platforms on user engagement on a subconscious level. The study could also have contributions to the field of human-machine interaction revealing how people interact with streaming applications and recommendation systems. Data scientists could also find value in the applied methodologies for analyzing complex user behavior patterns within streaming data.

Conceptual model

The dataset that provides the means for this research to be conducted was published by the Spotify music streaming service in 2019 (Brost et al., 2019) in the form of a competition with the task to predict track skipping. The variable of interest regarding this study will be the context variable, in which the track was listened to. The context has six possible values, which includes multiple types of playlists. These could be the previously described recommendation types, however the users have the chance to completely control their listening sessions and listen to music after a direct search. The context has six possible values, which includes multiple types of playlists among others, a summary of that is displayed in Table 4.1.

Table 4.1: Summary of listening contexts

Name of the variable	Description
User collection	Song was played from the user's personal library of saved tracks or albums.
Radio	Song was played from Spotify's automated radio stations, selected based on chosen genres, artist seeds, or mood selections
Personalized playlist	Songs was played from algorithmically generated playlists based on the listeners history, preferences, and similar artists or genres
Editorial playlist	Song was played from professionally curated playlists, focused on specific themes, genres, moods, or artists
Charts	Song was played from one of the top 100 chart where trending music is gathered based on different criteria like genre or release date
Catalog	Song was played directly from the platform's general music library, where users can search for individual tracks or albums

Charts and Catalog are not really considered recommendations in this study, since listening straight from the catalog is the means the users are searching directly for a certain song, and charts are just containing the currently popular songs without any exact recommendations. Having said that, listening straight from the catalog could serve as a baseline for the recommendation systems, since there the selection of the songs is in the users' hands. These context types could give some interesting insights into user interaction as well.

Skip types

As for the dependent variable, the skipping of a track, the data contains multiple Boolean variables that describe various kinds of skipping. Skip 1 indicates that a song was played very briefly, possibly skipped right away, before even listening into the track. This suggests that the user was possibly not interested either in the artist, genre, or the song itself, or was rather looking for a specific song in the playlist. Skip 2 means that the song was played only briefly. That could mean the track was not skipped right away but was skipped after listening into it. Compared to the previous variable, in this case the customer actually started to pay attention to the current song but decided that it was not in their current interest and skipped it. In this case, the user interaction was done based on the song itself as well, although different factors could play a part too. This was the dependent variable of Spotify's competition in 2019. Skip 3 is true at tracks, for which the majority of it was listened to, however the track was not entirely finished. Unfortunately, there is no exact information on what that means in seconds, or percentages. This variable could indicate that the recommendation was mainly successful, but in the end the user was not entirely engaged, and a skip was triggered. The Not skipped variable is true in case the song was not skipped at all; however it provides the least information on when the skip occurred. There were some tracks that had negative value in the Not skipped variable, however none of the skip variables were true. Based on this, there were some cases when the listening of song was almost complete, however a skip occurred before the end of the song. These cases only took 1.5 percent of the data, it is possible that there was a point between the Skip 3 and the end of the songs that were not recorded if skipped, however in this research these skips will be added as skipped at the third point for clarity.

The skip variables represent points during the songs until the song was played. If a skip did not occur, the song was listened to in its entirety. However, there is no information on which point of the songs these variables represent, how much difference is there between these points. The description of the variables was quite vague in this case. Based on these, the data has to be considered as ordinal data. The differences in the categories are clear, and they have a clear order as well, but no exact known distance between them: Skip 1, Skip 2, Skip 3, No Skipping.

Relationship of the skipping rates and playlist types

To start the exploration of the complex relationships between the context types and the different points of skipping, the preliminary analysis of skipping occurrence revealed the average skipping rates of all context types, which is shown in Table 4.2. For a more clarity, the results are also visualized on Figure 4.1.

Table 4.2: Average skip rates by context type in percentages

		Skip types			
		Skip 1	**Skip 2**	**Skip 3**	**Not skipped**
Name of the context	**User Collection**	46.32%	7.60%	11.55%	32.73%
	Radio	43.79%	10.12%	13.91%	31.09%
	Personalized playlist	30.33%	13.91%	21.45%	32.62%
	Editorial playlist	39.46%	10.82%	13.01%	35.01%
	Charts	40.42%	10.47%	11.11%	36.07%
	Catalog	32.89%	10.99%	15.40%	38.71%

Skipping right at the beginning before even listening to the song properly was the highest on the user's own collection. This makes sense since the listeners are familiar with the contents of their own songs and can perfectly decide if they want to skip to the track before listening to it. The lowest Skip 1 skipping rate was produced by personalized playlists, which is a bit lower than catalog. This suggests that this type keeps the users generally the most engaged, since they are not constantly skipping through song, but rather start to listen to them at least until a little bit.

Overall statistics suggest that if a track is not skipped right at the beginning, it will be listened to completely, or at least for the majority of it. The skipping rates were the lowest at Skip 2. At that point, user collection has the lowest skipping rates, suggesting that if somebody starts to listen to a track, their intent would not change after briefly listening to the song. Personalized playlists have the highest rate here, which could suggest the algorithm tends to play songs that the listener is already familiar with, while their intent is to find new tracks to listen to. However, it could also mean that the suggestion was not good enough to keep them engaged. The other types have somewhat similar rates at this skipping point. The skip rates at Skip 3 are averagely 3.75 percentage points higher than at the previous point. This indicates tracks are a bit more likely to be skipped after most of it was listened to rather than just briefly played. Personalized playlists have an exceptionally high average skip rate here compared to the other types. A probable reason behind this could mean that users are more likely to actively search for new songs with personalized playlists, and not wait for every song to end, but they listen to the most of it before skipping.

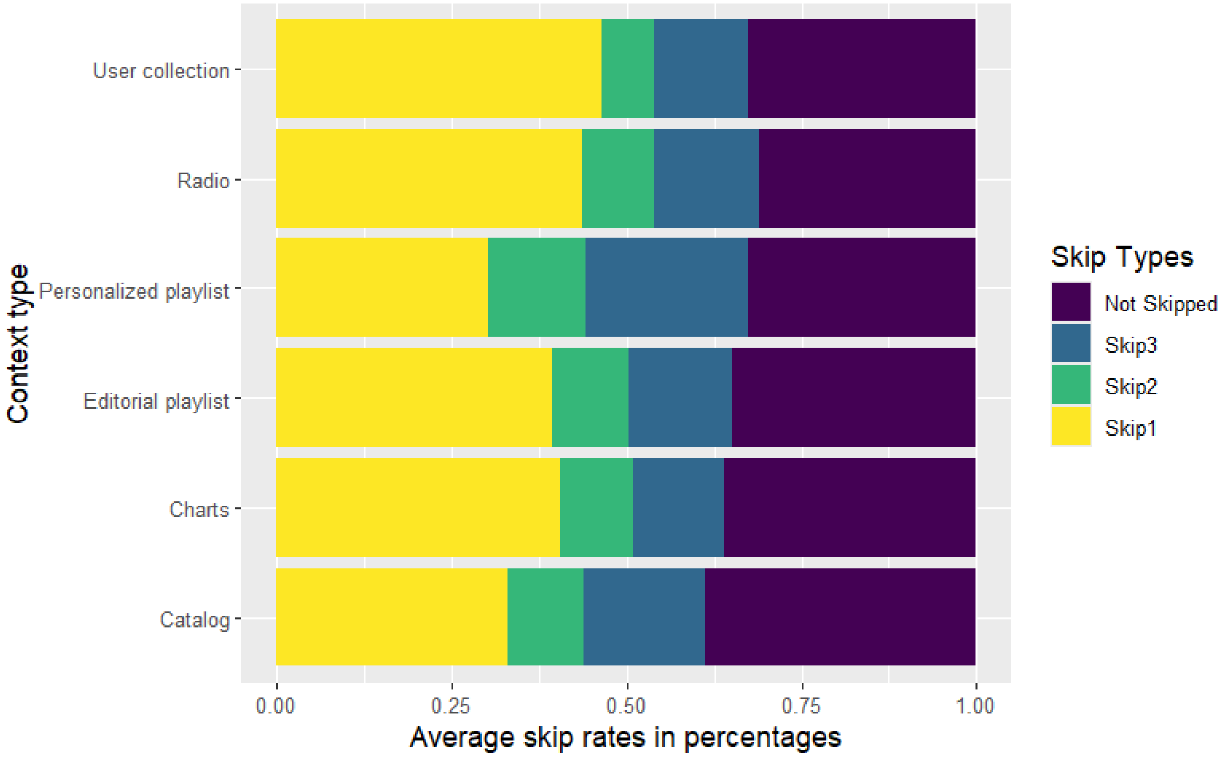


Figure 4.1: Average skipping rates by context type

Looking at the overall not skipping rates, Catalog is higher than the rest. If a user searches for a song in the Spotify library, it is most likely they will not skip the song. Radio, personalized playlists, and user collection have somewhat similar overall (not) skipping rate, which is a great validation for Spotify's algorithmically created playlists, as they engage the users similarly as their own song collection. Based on these ratios, playlists created by Spotify's professionals have quite an average value at all skipping points, and overall as well. Hopefully, further analysis will reveal more insights on this type of playlist.

These informations provide a good picture of the relationship between the most important variables in the framework of the research, and the ordinal nature of dependent variables were also explained. However, analyzing average skip rates cannot provide a clear picture of how these playlist types really affecting the success of a song, and also other variables are possibly playing a part in that.

Models

The primary aim of the research is to predict skipping behavior as a function of play context and assess the relationship between user behavior and session engagement (I.e., skipping). Reviewing the competition from which the data is available, in most cases a form of deep or reinforced learning was applied, in forms of recurrent neural networks (Hansen et al., 2019; Meggetto et al., 2023) to predict skipping. Gradient Boosting was used as well (Béres et al., 2019; Ferraro et al., 2019), however only in fewer cases and with a different approach. In the competition, the given task was to predict only the Skip 2 variable. The other types were not examined thoroughly, hence neither of the models accounted for the Skipping variables' ordinal nature. Given this nature, most of the above-mentioned approaches are not the best way to predict skipping, when it is considered ordinal, and more than one point of skip occurrence is of interest.

The nature of ordinal data lies in the fact that it is somewhere between the numerical and categorical types. There are clearly separate groups, and the order between them is also given compared to regular categorical data, however the class withs are unknown, such as in ratings (e.g., 1 to 5 stars). Most regular models can manage the data type, although they are treating it mostly as categorical or numerical data not accounting for the ordinal nature.

Generalized Ordered Logistic Regression

One of the most common ways to predict ordinal data is Ordered Logistic Regression (McCullagh, 1980). It is a type of regression analysis that uses a logistic function to model the probability of an observation falling into a particular category of the ordinal variable. However, this version of logistic regression assumes that the relationship between the independent variables and the log odds of being in a higher category versus a lower category is constant across all categories of the ordinal dependent variable (Williams, 2006). This is otherwise called as the proportional odds assumption. Unfortunately, it is highly unlikely that this assumption holds in a real-world dataset, however this can be tested with the Brant test (Brant, 1990). A solution to this problem could be the use of the Generalized Ordered Logit model. Unlike the traditional Ordered Logit model, the gologit model relaxes this assumption and allows the effect of predictors to vary across different categories by making multiple binary logistic regressions at the same time. With this approach, the model estimates a set of coefficients and intercepts for each cut-point (between skipping intervals), which can make the interpretations quite complex. A halfway solution exists between the two absolutes, such as the Partial Proportional Odds model, which only relaxes the assumptions for the variables that do not meet it. This model was applied in research context before (Agga & Scott, 2015), and it also accounts for the ordered nature of the data, unlike multinomial logistic regressions, however estimating this many coefficients can be computationally expensive.

Ordinal Forest

An alternative and more complex method, which is rather Machine Learning than statistical, could also be considered for handling ordinal data, namely Ordinal Forest. The basis of it are Decision trees, and Random Forests, which originally were not developed for ordinal data in comparison to Ordered Logit, however recently a method was proposed how it could be altered for predicting ordinal data (Hornung, 2020). The model assumes an underlying latent, continuous variable behind the ordered categories. The Ordinal Forest process starts by assigning initial score sets to the ordinal categories, then building decision trees and optimizing the score sets. This is done by minimizing the MSE of the out-of-bag predictions of each tree (which are approximately one-third of the data) measured by a performance function, and a n number of scores will determine the optimal score set for the latent variable. This is where the model differs from traditional random forests. After the model conducted the optimal score set, it works like a traditional regression forest, using the optimized score sets instead of the classes for the splits. Each tree is built on a different bootstrap sample of the data, and at each split only a subset of the variables is selected. The possible splits are evaluated on the remaining data (out-of-bag data), and the one with the largest differences in average scores is chosen for the split. The splits are keep continuing until a given stopping criterion, which could be a minimum node size or a maximum depth of the tree. The model conducts a predetermined number of these trees, hence the name forest. All the trees in the forest then calculate a score, and the average score of the forest will be mapped back to the ordinal class whose optimized score is closest to it. In most cases, predictions are made by the described majority voting of the trees, however when the performance measuring function of the model is based on class probabilities, the highest average probability across all the trees will determine the outcome.

Hornung (2020) presented in his paper, how his method is performing against a multi class random forest, and a naïve ordinal forest, which is essentially a regression forest, where the class widths were 1,…J of the ordinal response used as score values. The ordinal forest model outperformed the other two models in most cases of a study conducted on real data and also on a simulated one. As this is a relatively new method, it has not been widely used so far, for these reasons it would be a good contribution of this thesis as well. The model works well for low- and high-dimensional data as well.

Evaluation Metrics

In the case of ordinal data, measuring the performance of the model is not as straightforward as in the case of numerical or categorical trees. As for the evaluation metric, Hornung proposed the Weighted Kappa to measure the quality of the predictions (Hornung, 2020). This is the case because in the presence of ordinal data the predictions that are close to a given class also need to be considered, not just the classes that measure pure accuracy. Kappa is an evaluation metric that is widely used for assessing the agreement between two raters, or the predictions and the real classes. It also accounts for the agreement that might just occur by chance, as presented by Cohen (1960). His original version was suited for only nominal data, however a modified version, the weighted Kappa could also be used, which can account for the level of disagreement between raters, embracing the ordinal nature of the data. The two most popular weights are linear and quadratic weights. Quadratic weights put a larger emphasis on predictions that are far from the original class, penalizing them more. Linearly weighted kappa is a balanced route between Cohen’s kappa, where no benefit is attributed to classes other than the true class, and the quadratic weighted kappa. Since the differences between the classes are far from each other, and the difference between skipping right at the beginning or not skipping at all should be more heavily penalized, the quadratic weighted Kappa should be a more suitable choice for this type of data. The interval of Kappa goes from -1 to 1, where -1 is complete disagreement between the raters and 1 is the perfect. 0 stands for the occurrences are solely can be accounted for chance. This method considers the nature of the ordinal data and can give a clearer understanding of the model's predictions. It cannot be as straightforward as accuracy, but for comparing two models run on the same data should be suitable. Having said that, looking at accuracy and individual class Precision rates could also provide some useful information about the models.

The benefits of the Ordinal Forest model include the capturing of more complex patterns in the data, that simpler methods might not recognize, although there is of course no guarantee of a better performance. The point of this study is to find the difference between the different playlist types at different points of the songs, not to train perfect predictor or analyze the class distances. For this, the predictions of these methods need to be analyzed instead. The Generalized Logistic regression (and the Partial Proportional Odds model) gives coefficients that are a bit complex to interpret, but it is still an option to see behind the model. However, the Ordinal Forest is considered a Black Box machine learning model, which means that the interpretation of factors behind the predictions is not that straightforward. For this purpose, Black Box interpretation models could be utilized to understand the contribution of certain variables, such as playlist. The SHAP Values are commonly used in these cases and can provide great insight into the model's workings.

SHapley Additive exPlanations (SHAP)

The idea behind this method was first described in 1953 (Shapley, 1953), as a contribution the game theory. It is a way to calculate the gains and the losses among players based on their contributions to the outcome of the game. This idea was used to create SHapley Additive exPlanations (SHAP), which adapts the method to the field of machine learning interpretations (Lundberg & Lee, 2017). In this context, the features are considered the players, and the predictions are the outcomes of the games. The Shapley values measure how much each player/feature value contributes to the prediction compared to the average of all predictions. The process of calculating the marginal contribution first considers all possible subsets of the features, and then computes the difference with and without the given feature. Naturally, this process takes up a large amount of time, and grows exponentially with the number of features. A possible solution for this is to use Monte Carlo samples for calculating approximate values for the model, which can greatly speed up the process (Štrumbelj & Kononenko, 2014). This way at each simulation the method only considers a different subset of permuted features, and it averages the conditional expectations of the subsets to estimate the Shapley values. This way the process is sped up, but it can also produce less accurate values so the choosing of the optimal number of Monte Carlo simulations is important. Individual SHAP values can provide insights on individual predictions, however aggregating them could also provide results on the global scale, and on the influence on the whole model. A huge advantage of the model is its model-agnostic nature, so it can be used with any machine learning model, this includes the ordinal forest model, and it can provide rich visualizations as well. The downsides include the computational costs, which are still significant with the above-mentioned method. Comparing to the regressions output, this method does not produce exact numbers on how the variables influence the log odds of falling into one category, however for the comparison of different features it is perfectly suitable as interpretable numerical values on the contributions.

To summarize the framework, first the two proposed model will be trained on the same training data, then be evaluated on another proportion of the data with Cohen's Kappa and Kappa with quadratic weights, perhaps with Precision on certain categories as well. Then insights on the predictions of the better model needs to be evaluated in order to understand the effects of different recommendation types on the predictions of the model. For that, if necessary, the Shapley Values of the Ordinal Forest model's predictions will be conducted as well.

Results

Data

The dataset was released in 2019 to provide grounds for the competition (Brost et al., 2019). It consists of more than 160 million streaming sessions with user interactions. Each listening session contains 10-20 tracks, and the information of how long the songs were listened to. In order to work with the data with restricted amount of computational resources, a subset of the given data was created, with 3000 full length sessions, which approximately consists of 50 000 songs. The data's observations happened on a session levels, and multiple variables are containing information about the sessions themselves, therefore only full sessions are going to be examined during the research.

Among these 50 000 tracks, 41.12 percent belonged to the Skip 1 type, which means approximately 20 000 tracks were skipped right at the beginning. However, the amount of Skip 2 was significantly smaller, only 9.5 percent and Skip 3 accounts for 15 percent of the songs. There were overall 17 thousand songs (34 percent), which were not skipped at all. It seems quite clear that users usually skip a song without listening to it, or they do not skip at all. However, between the two ends of the scope, users rather listen to most of the songs rather than skip it after a brief period.

The distribution of the context types is also not very equal, as it can be seen on Figure 6.1. Track listening from User generated playlists are more than 42 percent, Spotify's editorial playlists are 17 percent and personalized playlists are only responsible for 2.7 percent of the listenings. Charts have an even lower cut, however algorithmically created radio stations provide 12 %, and listening straight from the general catalog created 24 percent of the examined songs. It is quite clear that users prefer their own playlists most times, and for new song discovery they tend to listen to editorial playlists and automated radio-s rather than personalized playlists.

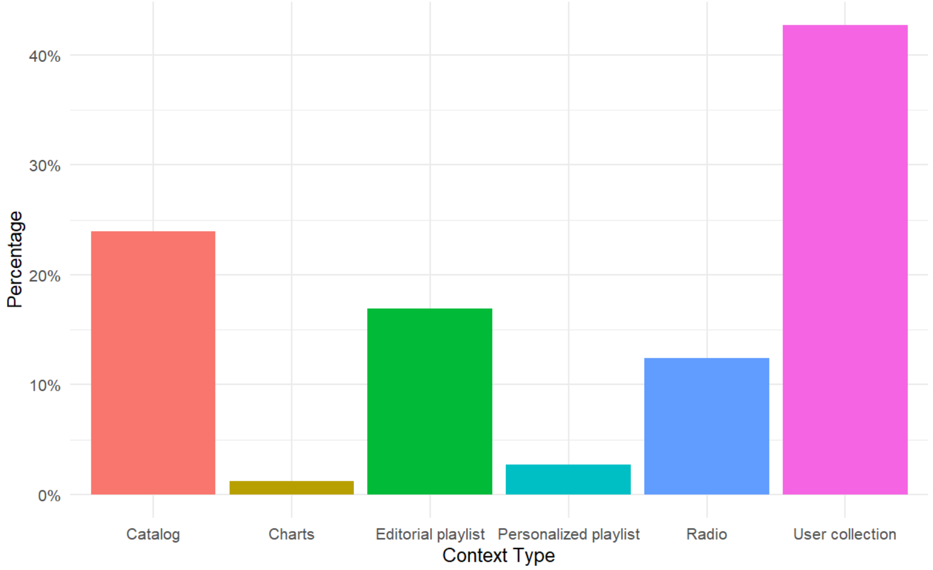


Figure 6.1: Proportions of the Context types

The data has several more variables that could potentially aid the predictions, describing several aspects of the listening sessions. Session length describes how many songs are in one session in total. Sessions from 10 to 20 in length were recorded, among these almost half of the sessions (47 %) were 20 in length. Session length 10 were the second largest with 7.97 percent and after that the pattern seems to descend until 19, which is only 3.07 percent. Each songs positions in these sessions were also captured by the session position variable, however since different sessions have different lengths, session position as percentage of the total session length could also have meaning, for that reason the position percentage variable was created.

The Hour of day provides information about the time when the tracks were played. The listening hours are almost normally distributed, rising from 6 o'clock, peaking during 16-17 and reaching the lowest around 4 o'clock in the morning. There only 0.72 percent of the tracks are listened to, which is less than a tenth of the peak hours' numbers. The premium type records whether the users had a premium a premium subscription or not, 84 percent tracks were listened to with premium subscription.

During listening, Spotify's application provides several interaction possibilities for the user, skipping the track, either go forward or backward in the recommendations, pausing the track, seeking forward or backward in a song, shuffle the songs on the playlist or changing context. All these user behaviors are recorded as well. Two variables record the fact how a track was started and ended, in both cases pressing the forward button and the track being done were the most common. The following were also recorded: whether or not the user did a pause and was it long or short before playing the given track. Only around 20 percent of the tracks had a pause before play. Two variables recorded the number seek forwards and seek backwards during each song, in general users tend to seek forward more often than backwards. Spotify also offers a function to randomly play songs from a given playlist, which is called shuffle. Thirty-four percent of the songs were played from the playlists in a random order rather than the given/intended order of the tracks.

The users are also able to change between the given listening context, however this does not necessarily mean the changing of the type of the recommendations, since listeners can change between their own curated playlist as well. The context switch Boolean variable captures this kind of behavior. Twenty percent of the 3000 listening sessions had at least one context switch. Based on the context switch, two additional variables were also created. One measures the number of contexts in the listening sessions and the other displays that in which one did each song occurred. One would assume that changing the listening context could mean that the user is not satisfied with the recommendation, which would also be seen in the skipping behavior. The variable in the dataset only has TRUE as a value, if the context was changed before a track, although it perhaps has more significant information for user engagement if the context was changed after a song. In order to measure that, another variable was created. If the proportion of the context switches in the different context types are examined relatively to the total number of tracks played in that context, the followings can be found in both cases (switched to and switched from). In both cases, the users' own playlists were involved relatively the least in context switching, in both cases approximately 1.4 percent. Personalized playlists, editorial playlists and charts show around 1.5-3 percent in both cases. While users tend to switch to radio stations and to catalog around the same proportion, 4.6-

4.8 percent, the difference is a bit bigger when examining the switched from contexts, where catalog is more than 5.5 percent, while users switch from radio stations only 3.7 percent of the time. Looking at the three main types of recommendations, it can be said that users tend to switch to rather than from editorial playlists and user collection, however in personalized playlists the tendency is the opposite. It could be also interesting to see how users change between these contexts and recommendation types.

There are 2035 switches in the data. As changing context in catalog is true even if the user is looking for a new song, switching from catalog to catalog takes up 30 percent, however this is not especially relevant here. The second highest number belongs to switching from catalog to radio, which could suggest a behavior when listeners are starting a radio station from a song they found in the catalog, this happened 12 percent of the time. Catalog is part of every context switch that has the proportion above 6 percent. Looking into context switches that did not include catalog and changing the type of the context as well, changing between radio and user collection has the highest percentage, around 3 percent in both ways. Users also tend to switch between professionally curated playlists and their own collection around the same around the same time, both around 1.8 percent. The rest is all below 1 percent of the switches and other significant differences between the different playlist types cannot be seen here.

The relationship between dependent variable(s), the skip types and the rest of the independent variable are worth taking note of as well. The relative positions of the tracks with respect to skipping behavior show quite a normal distribution in each kind, as displayed on Figure 6.2. However, the average position in the tracks where Skip 1 was present tends to be a bit later than Skip 2 and Skip 3. These two also have larger interquartile distance, indicating a less dense distribution. The average position of the tracks that were not skipped is at 54 percent of the session’s length, similarly, to Skip 1. Overall, these numbers suggest that listeners are more engaged with the songs in the beginning of the sessions, since they tend to skip later during the tracks, and later on the session skipping the tracks sooner, or not skip at all, depending on how engaging they are.

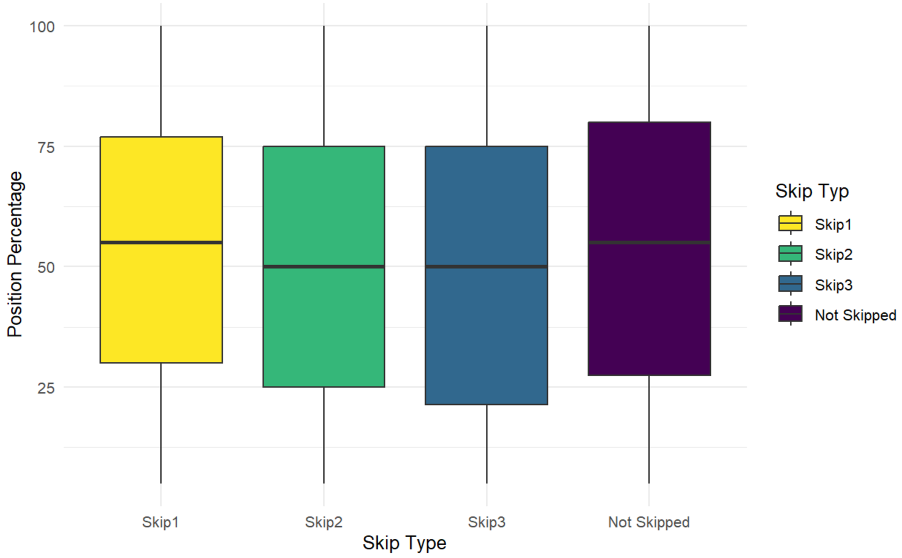


Figure 6.2: Position percentage by Skip type

Context switches also usually happen after skipping a track during the third interval, more than 47 percent of the switches happen there, 30 percent after Skip 2 and only 12 percent happens after Skip 1 and also if the track was not skipped. This is interesting, on one hand listening to most of the song would indicate that the recommendations were successful, but on the other hand the context switch suggests the context was not engaging enough for the listeners. Following this line of thought, the relationship between the number of contexts in a session and the skipping behavior was also examined. If there is only one context (the same context, not the same type of context multiple times) in a session, then 45 percent was skipped at the first point and 34 percent not skipped at all. However, when a session contains more context, Skip 2 but especially Skip 3 becomes a more popular user interaction. This could indicate a behavior pattern, when users are exploring multiple contexts, listening through most of the tracks in order to find the best one for their taste at the moment. Truth be told, this pattern is hard to observe, since only 20 percent of the sessions contain a switch in context. Seventeen percent has two contexts in them, after that with every number the percentages are halved.

The main goal is to examine the differences between different playlist types and a few interesting details can also be noticed among the relationships between the playlist types and the other independent variables. A good example for this was the differences in context switches discussed above. Looking into the premium and freemium subscribers, it can be found that premium users tend to play more songs in personalized playlists and automated radio stations, while freemium users are relatively more focused on editorial playlists and their own collection. This suggests that premium users are usually more open for new music discovery, while freemium users with restricted skips and additional advertisement are more focused on their own songs and editorial playlist, which offers them a safer way to explore music without unwanted songs. The connections with the hour of the day and session length were also examined, however no relevant differences were found.

During the process of examining the data, and creating the described additional variables, the dataset was of course checked for missing values and duplicated instances, fortunately none were found in the selected random three thousand sessions. For the models to handle the context variable appropriately, it needed to be one-hot encoded, this way each playlist types have its own dummy variable. In the final dataset for the analysis, twenty dependent variables were present and the ordinal independent variable with four levels, as described previously. Before running the models, the dataset was split into two parts, the models were trained on 70 percent of the data, and they were assessed on the remaining, previously unseen 30 percent of the data.

Generalized Ordered Logistic Regression: Performance evaluation

As mentioned before, the generalized ordered logistic regression probably could not be used here, since the proportional odds assumption would not hold on a real-world dataset. Naturally, this needed to be assessed before turning to the generalized model. For this, the Brant test was utilized, which has the null hypothesis that the parallel regression assumption holds. The test was done on each of the variables, interestingly for the hour of the day variables the H0 could not be rejected on any significance level. For the rest of the variables H0 was rejected with a p value lower than 0.01. This way the Model technically will be a Partial Proportional Odds model, since for this one variable only, the relationship is the same across all dependent variables, hence only one logistic regression needs to be calculated for it across all skipping types. After training the Model, predictions were made on the test set to prevent overfitting. The confusion matrix for this can be seen in Table 6.1.

Table 6.1: Partial Proportional Odds Model confusion matrix

		Reference			
		Skip1	**Skip2**	**Skip3**	**Not Skipped**
Predictions	**Skip1**	4928	622	907	3020
	Skip2	5	41	34	4
	Skip3	405	469	784	352
	Not Skipped	1025	270	501	1687

The Model's overall accuracy was 49.42 percent, which means it has predicted almost half of the skips correctly, which is not that bad, given that there were four outcomes. Cohen's Kappa is 0.2045, and the weighted Kappa was 0.1816, which means that if at the evaluation the data's ordinal nature is accounted for, the model's performance is slightly worse. According to Cohen's framework, this result stands for none to slight agreement between the two ratings (McHugh, 2012). If we zoom in on the four levels, it can be seen that the model performed best in the Skip1 category, 77.45 percent of skips were identified correctly there. Unfortunately, this is not the case in the second category, where only 3 percent were correct, which means basically that the Model cannot predict if a song is skipped after only a brief period of time. The case is slightly better at the last two categories, with 35 and 33 percent precision. It is clear from the Model, that it is poor at predicting Skip 2, and rather not predicts it in most cases. This is possibly due to the data imbalance, since only 9.3 percent of the data contains songs that were skipped at this point. It makes perfect sense, that the participants trained overly complex deep neural networks the capture patterns in this variable.

Ordinal Forest: Performance evaluation

The training of the second model required a bit more training and fine-tuning. The tuning of an ordinal forest model basically can be divided into two parts. Firstly, the training to get the optimal score sets for the class values, and secondly the training of the actual regression forest. The number of score sets tried had the biggest computational impact on the model, so it was started with 50 sets and increased after (default values is 1000). Between 100 and 1000 the results did not change, so 100 was the final number of score sets tried for the optimization to speed up the process. In order to get the class probabilities for each class, probability was chosen for the performance function. This proved more successful than the default function, and it also provided class probabilities crucial for the SHAP values. In this particular case, the data had a large number of observations and low dimensionality. It was pointed out by Hornung (2020) that in these cases going with the default parameters could take a considerable amount of time. However, a small value can be chosen for the trees of the score optimization, since the trees in the forest are more precise because of the considerable number of observations, so this value was set to 10 according to the paper. The 10 best score sets were used to calculate the optimized score sets. In the second stage of the training, the bigger forest was trained as well. During the first stage of the training, in most cases the forest was greatly overfitted on the training data. In order to handle that problem, the minimum node size for the splits were raised to 40 and number of variables to consider during each split was lowered to 3 from 4 (the rounded down square root of the number of variables). These measures did help to prevent overfitting to some extent. Finally, the ideal number of trees in the final forest was set to 3000, since after that the performance did not improve.

After training the Model, the predictions were calculated on the left-out test set. Looking at the confusion matrix in Table 6.2, it can be said that the evaluation numbers of the predictions only slightly improved compared to the Partial Proportional Odds Model.

Table 6.2: Ordinal Forest Confusion Matrix

		Reference			
		Skip1	**Skip2**	**Skip3**	**Not Skipped**
Predictions	**Skip1**	4870	699	960	2787
	Skip2	10	21	17	2
	Skip3	265	297	539	198
	Not Skipped	1218	385	710	2076

The overall accuracy of the predictions are 49.86 percent, which is basically the same, the Model only did 66 more correct predictions. Cohen’s Kappa was 0.2007 and weighted Kappa was 0.2124. When accounting for the random chance of choosing correctly, the Model performs the same as the previous one, however when the other categories have a weight as well, the Model performs slightly better. This weighted Kappa value now belongs to the interval which stands for fair agreement, although is quite far from an optimal performance. Looking into the induvial skipping points, Skip 1 and Skip 2 have a bit worse Recall, but only around 1 percent. The significant difference between the two models can be explored in the Skip 3 and Not Skipped categories. While the gologit Model correctly predicted 35 percent of the third skipping variable, this one only guessed correctly 24 percent of the time. In contrast, the ordinal forest performed significantly better in the not skipped category, with 8 percentage points better (41%). This is the main difference between the two Models, which also explains why the ordinal forest performed better when looking at the weighted Kappa. Ordinal Forest performs better in a larger class, which is at the end of the spectrum as well. When giving more accurate predictions for listening to the song in its entirety, less wrong predictions are penalized for being too far from the first and second points compared to the previous Model, hence the weighted Kappa is higher here. This is exactly how a metric can measure the performance of an ordinal model.

Overall, it can be said, that these two Model performed similarly, however Ordinal Forest did slightly better after tuning. As written previously, there is no guarantee a more complex model could present more accurate predictions in every case. Viable solutions for improving the predictions could be the solving the data imbalance problem with stratified sampling of the skip types. Having said that, it is only can be done if the data is sampled on track level, breaking up the sessions, and losing some session level variables, potentially some contextual features such a session position and the number of different playlists in a session. Since the goal of this research is to evaluate the role of different recommendations in these predictions, it is more important to see behind the models than having the perfect model, and breaking up a real-world dataset which was collected on session level might be counter effective for that purpose.

Interpretation of the results with SHAP values

After training the two models, the ordinal forest's performance turned out to be slightly better. To understand the effect of the variables on the predictions of this Model, the SHAP values needed to be computed. Additional reasons to interpret the results of this technique compared to the proportional odds model is the more clear and visualizable results, and I also believe that applying the SHAP values to an ordinal Forest model could be an interesting combination and a great contribution of the research.

To understand how predictions were made on previously unseen data and provide a more realistic assessment of the features' effects in a real-life scenario, the SHAP values were conducted on the test set, on which the Model was evaluated. As described in the theoretical model, the computational limitations only allow the calculate the approximation of the values. The computation started first with 10 Monte Carlo samples, and step by step increased by 10 samples, until the values seemed relatively stable with different random seeds considering the computational time it required. The final values were calculated with 50 Monte Carlo samples for each level skipping.

First take a look at the Mean Absolute SHAP values across all skip types to understand the general idea about which variables are the most influential in the Model. Not pausing the track before playing it had the largest impact on the predictions, especially in the Skip 1 and Skip 3 categories. Taking a short pause before the track and the length of the listening session also had a huge magnitude in the predictions, however the distribution looks a bit different. They both contribute around an average of 6 percent to the songs' probability of being skipped at different points.

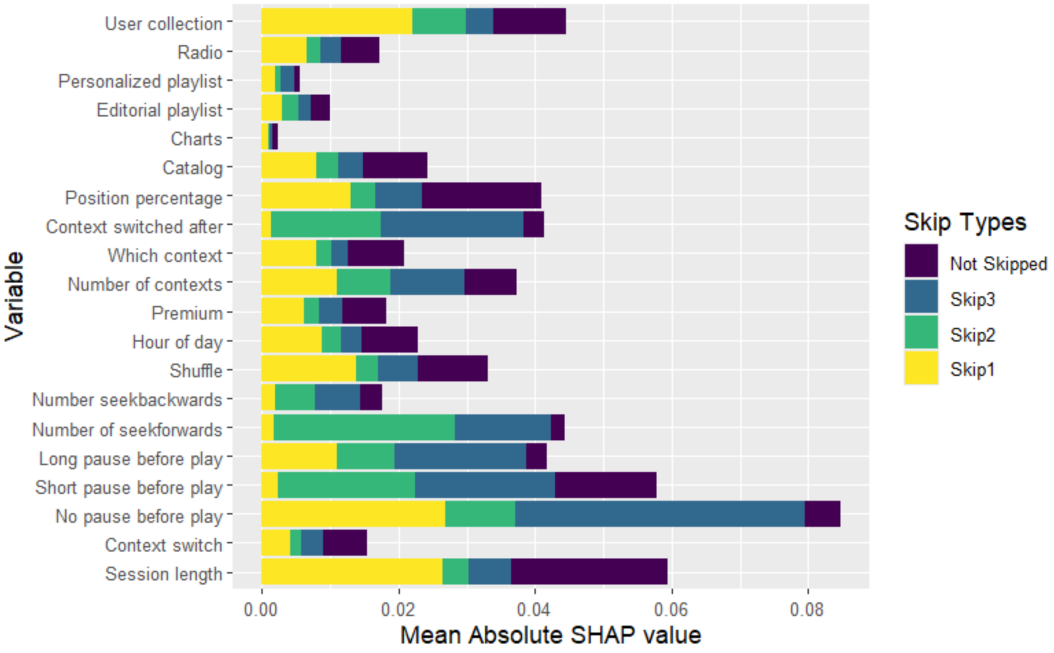


Figure 6.3: Average impact on model output magnitude by all variables

As it can also be seen, the track's positions within the session and the number of contexts had profound impact on the predictions, validating the approach to work with the data on a session level. While these are interesting results as well, the main topic of this research is to find to differences between different playlist types, Figure 6.4 highlights only the dummy variables that are representing these listening contexts.

It is clear that being in the users own playlist has the greatest impact on model output, it is almost twice as strong as the effects of the catalog variable, which stands for listening for a track after directly searching for it. The former has the biggest impact on the Skip 1 point then the Not skipped, the latter has a bit bigger impact on the Not skipped prediction. Personalized playlist and charts have less than 1 percent average absolute contribution to the skip probabilities, indicating a minor magnitude overall in the Modell. All in all, it can be said that these total magnitudes correlate with the occurrence of the playlist types in the dataset. This could be due to several factors, such as the Model is more likely to learn from more frequently occurring features, which are reflecting on the average impact in the SHAP values.

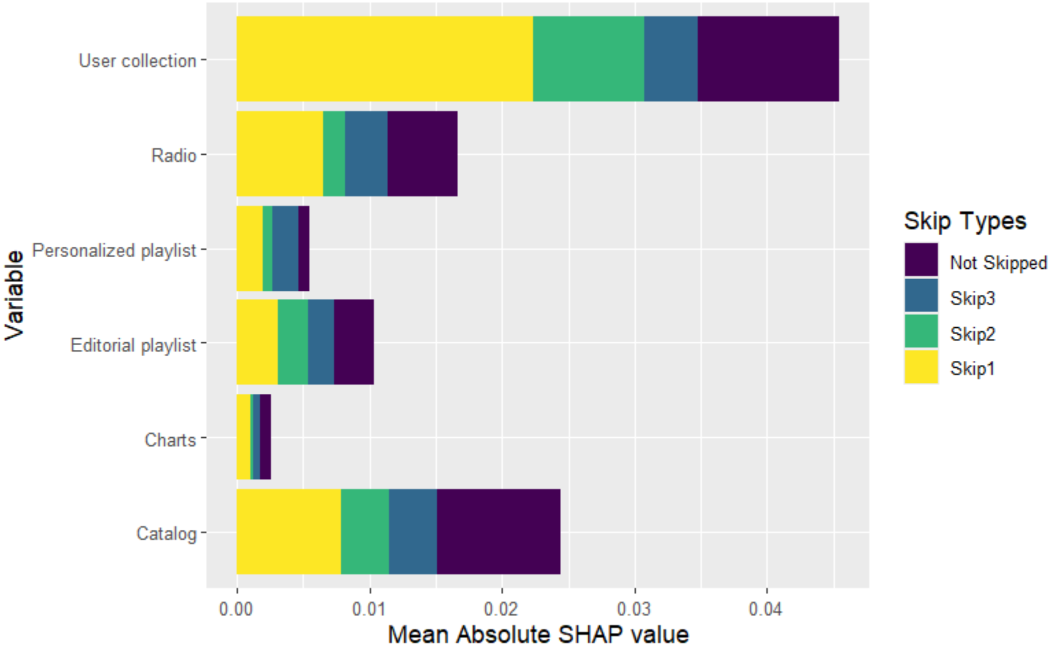


Figure 6.4: Average absolute impact on model output magnitude by playlist types

There cannot be much meaning deducted from these informations on how the values of the variables influence the SHAP values themselves, only the average (absolute) magnitude. For this purpose, the bee swarm plots allow us to zoom in on the individual skip types, and the variables contribution to them, highlighting each specific predictions as a point, bee if referred to the plots type. The SHAP values were calculated for each of the skip types, the following informations can be deducted from the values influencing the predictions of skipping at the very beginning. In total, no pause before play and session length are the two most influential variables (see in Appendix

Fig. 1). If no skip occurred before the playing of the song, that had a positive effect on the probability of skip 1, however if this was not true, that had a bigger impact on the predictions in the negative directions. To put this in more clear terms, holding a break before the playing of a song can actually decrease the chance of skipping a song, while listening consecutively could lead to skipping at this point, although the impact is smaller. As for the session length, a session of 20 songs leads to a higher skipping probability in each song, and shorter session are indicating lower skipping probability at the very beginning, and the latter with a twice stronger impact.

Zooming in on the variables of interest, Figure 6.5 shows the bee swarm plot with the six context types and the SHAP values. Each point represents a non-overlapping data point, and the color corresponds to its raw feature value. Since all variables here are dummy variables, yellow indicates that the variable belongs to that category, and purple that it is not. The positions of the points indicate whether they have a positive or a negative impact on the predictions. User collection have the biggest impact, which also have symmetric magnitude. If a track is listened to from the user's own playlists, then it contributes to a higher skipping probability at the first point.

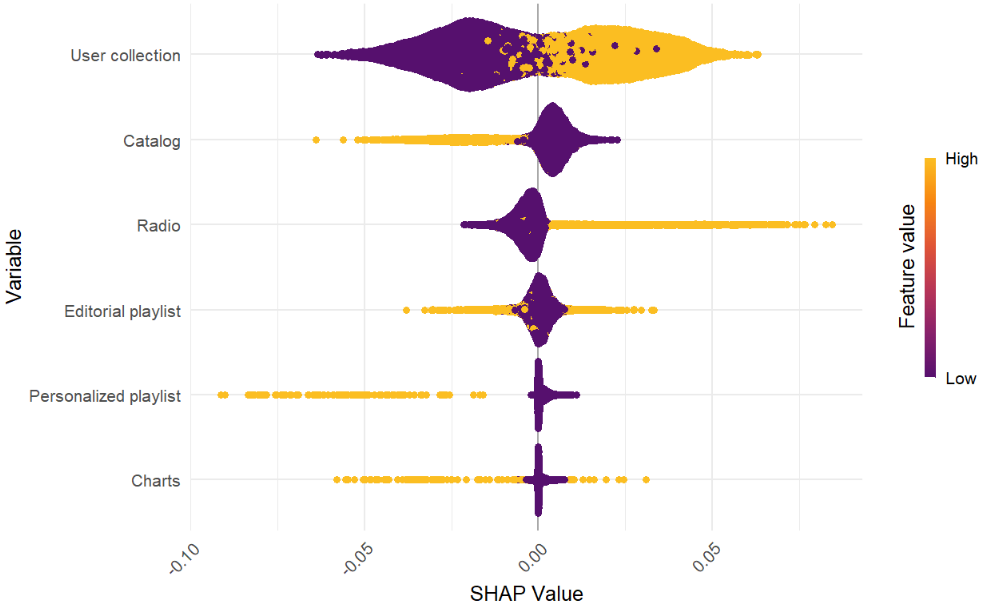


Figure 6.5: Beeswarm Plot of SHAP Values: Influence of Playlist Types on Skip1 Probability

Automated radio stations have a strong positive impact on the predictions as well in case the song was listened to from this context. On the other hand, not being in this context has a less impactful meaning, which is suggested by the density of the low value points around the baseline (average skipping rate at Skip1). Editorial playlists have a somewhat neutral impact on this type of skipping, given that the belonging to this context have the same negative and positive effect as well, while not being in that type has virtually no impact to the Model. Personalized playlists have clearly the biggest negative effect on skipping probability at this point. It is an interesting insight, that while both radio and personalized playlists are automatically generated, they have opposite effect at this point. Listening straight from the catalog have also a strong negative effect on skipping probability, however charts have basically no real impact on the Model's predictions. The last two types are not recommendation types, hence they are less in focus of this research, although their effect is worth noting since it is a great counterpart to the playlists.

The next point represents, that the song was listened to, but only for a brief amount of time, not enough to engage the listeners. The three most important variables seem very interesting here (see in Appendix Fig. 2). The number of seek forwards within the song have the highest impact. If none were made, that have a small negative effect, but on the other hand if some seeking forward (even one) was made during the listening, that had a high magnitude positive impact on the probability of being skipped after a brief period. The same is true if a short pause was held before the listening of the song, or if the listening context was changed right after the song. Both have a positive impact, the context switch is higher in that case, although in absolute value the short pause has a higher impact on the predictions.

Figure 6.6 displays the effects of the context types for the second skipping point. User collection's contributions seem reversed and more asymmetric here. Listening to song from these playlists indicate a lower skipping probability, while having a zero value for this variable might lead to skip here. The dense cluster of the low values close to the baseline value also suggests that the positive effect is weaker than the negative, which have a more elongated shape. Professionally generated playlists have a clearer SHAP values in this case, values of 1 on this variable suggest a higher likelihood of skipping. Radio paint an opposite picture, where high values indicating a negative input on the Model's output. Personalized playlists here have a semi-neutral role, although some high values of the variables shown to have rather positive than negative effect.

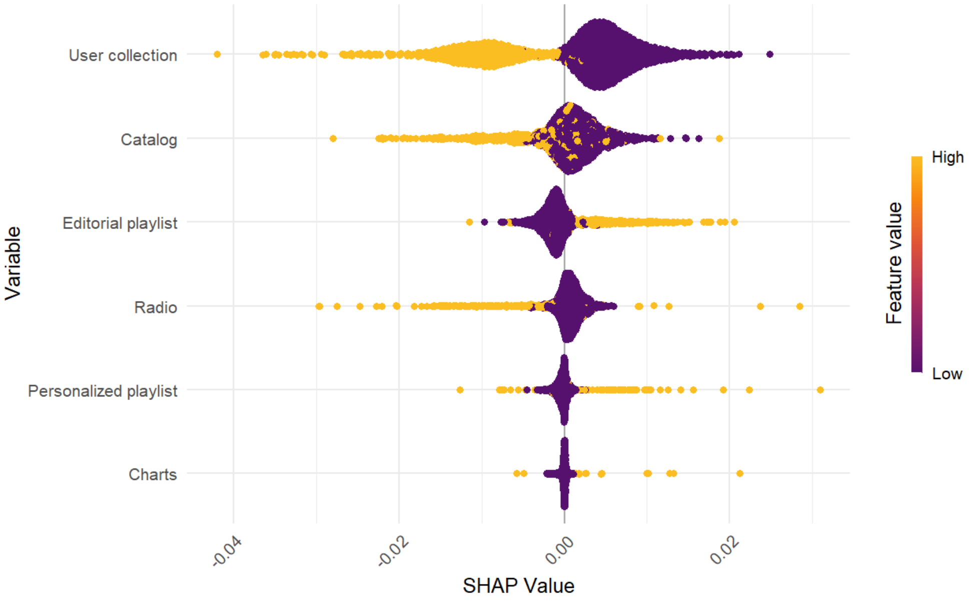


Figure 6.6: Beeswarm Plot of SHAP Values: Influence of Playlist Types on Skip2 Probability

As of the case of catalog, it is still clearly having a stronger negative effect on the skipping probability, however there are some values with positive effect as well. Charts does not have any significant magnitude here neither. While these SHAP values were stable, and the picture seems clear on how the variables influence the predictions made by the Model, it is needed to keep in mind that it performed very badly in this category, so the real-world implications of these values could be misleading. It needs be evaluated, if the effects of the playlists at this point fit into a potential pattern with the other skipping points, otherwise the implications need to be treated with caution.

The biggest impact in the Model can be found at Skip3, not pausing the track before the play has the largest average absolute magnitude across all four skip points, it contributes an average 4.24 percent to each instances probability of being skipped (see in Appendix Fig. 3). Interestingly at this point, a high value has negative effect on the predictions while low value has the same positive effect, in contrast to skip 1. Switching context after a song and holding a small break suggest a higher skipping probability. These findings suggest that a continuous listening behavior means lower chance of skipping a song after listening to the majority of it. When user interactions with the application happen, in other words the user is less engaged, such as in the case of switching context after the track or holding a short/long pause before it, the probability of being skipped at this point increases.

User collection’s impact direction is not as clear here as in sooner skip points, however it can still safely be said that being in this type still decreases the probability, and the magnitude of the effects seems similar in both directions. Radio has a truly clear effect here, belonging to this type of playlists have a great negative impact on the probability, somewhat greater than at the second skip point. The effects of personalized playlists are the opposite, meaning it has a positive effect on the chance of the songs being skipped. In both cases, the low values have small-to no effect. Editorial playlists show the signs of rather decreasing then increasing the probability, but the effect could also be interpreted as neutral, since high values are present on both sides of the spectrum. It is the same case with being search directly through the catalog, while charts have no real effect here as usual.

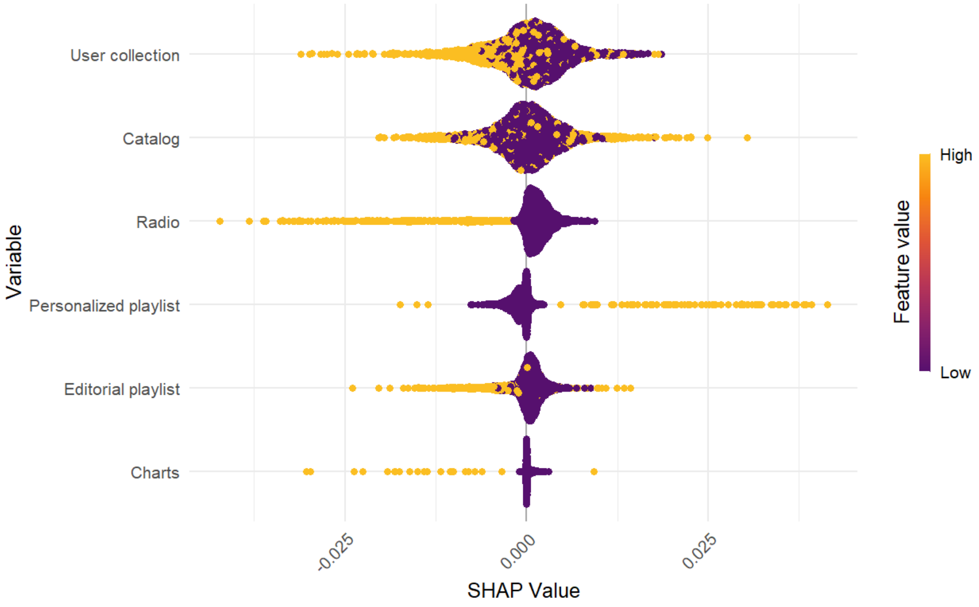


Figure 6.7: Beeswarm Plot of SHAP Values: Influence of Playlist Types on Skip3 Probability

The last variable in the order is only reached when the users do not skip the songs at all. The most meaningful factors here are session level variables, such as the length of the session and the position of the song divided by the session’s length (see in Appendix Fig. 4). Full length sessions tend to have negative impact, while shorter ones indicate a higher chance of being not skipped. The percentage of the positions paints a more interesting picture. It appears as having a high value there suggest higher probability of not getting skipped, having low have a small magnitude effect toward the same directions. The values around the middle of the sessions have however a negative effect on the probability here. These informations could indicate that a track played around the middle of a 20-song long session have the least chance of not getting skipped (more chance of being skipped at some point), and shorter session last few songs have a higher probability of getting listened to in total. Short pause before play also has a larger negative effect here, which aligns with the SHAP values of the previous variables.

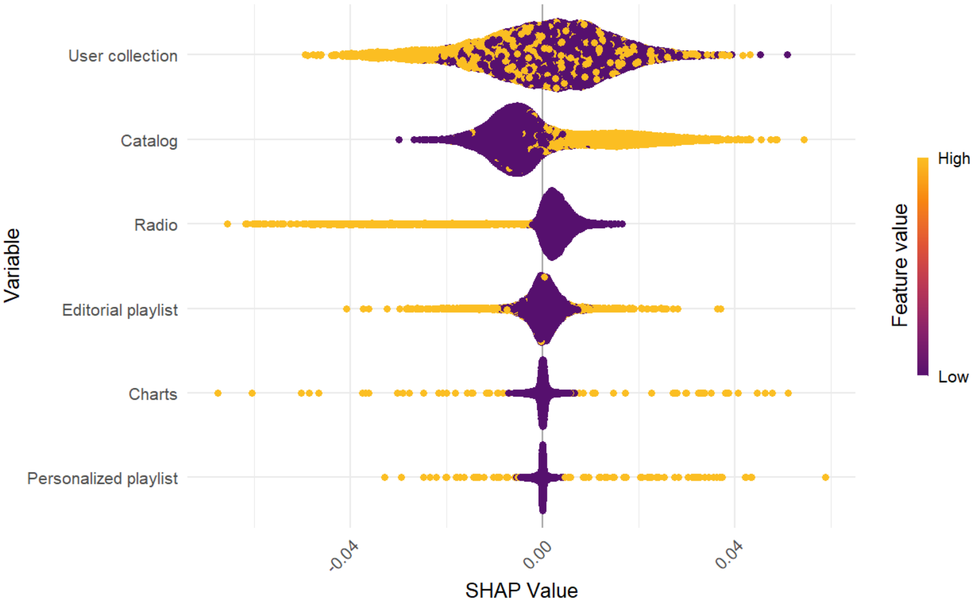


Figure 6.8: Beeswarm Plot of SHAP Values: Influence of Playlist Types on Not skipped Probability

Zooming in again on the context variables, User collection have an even more mixed effect on the probabilities. On the negative side still only high values are present, however the large cluster around the baseline seem the be evenly distributed between 0 and 1 raw variable values. The type has still a large impact on the Model, and only the high values have strong negative effects, however the contribution of being in this type to the predictions is kind of balanced around the baseline. This is suggested by the overlapping instances, which are on the two ends of the feature value's spectrum. Overall, being in the users own playlists might decrease the chance a bit of a song to be listened to in total.

The Spotify generated radio stations kept their effect to the last point, the higher values negative impact is even bigger than previous cases, when the low values are centered close to the baseline, suggesting a minimal positive impact if a song is not in this category. Editorial playlist have an average absolute impact on the Model, however looking at this plot a clear conclusion cannot be drawn on which way does it contributes to the probability of not skipping a song. To my surprise, personalized playlist basically does not have an effect at this point, neither does charts. Seemingly listening to a track straight from catalog is the only thing, which could positively impact the probability of the predictions at the end of the spectrum, meaning a directly search song is more likely to get listened to in its entirety.

Discussion

The bee swarm plots summarized really well, how each variable impacts the predictions at each skipping points. To acknowledge the effects across all model outputs, the findings of the SHAP values need to be looked at together. First discuss the effects of the generally most important variables. It has been clear from the plots that holding or not holding a long/short break before playing the track has a serious impact. At the very beginning, not holding a break has very negative effects on skip probability, at the next point have a small positive effect, while after listening to most of the song not holding a break before have the most positive effect for the skipping occurrence. Holding a rather short break have a huge negative impact of a song not getting skipped at all. In summary a pause before the song is most likely to cause a skipping after most of the song is played, however not the whole song. Looking at the session length, it seems clear that songs in long sessions have the most positive impact at skip1, from skip 2 they already have a negative effect, which is again the case of skip 3. At Not skipped, they have the largest negative effect on the probability. At shorter sessions, the tendency is the opposite, however the effects are consistently stronger on the model output. Being listened to in a shorter session overall have a more positive effect on the song's success.

The main idea behind the research was to find a pattern on how users interact with different recommendation types, and for that skipping behavior proved a great metric. The different context variables' impact on the probability of skipping at different points in the song's duration can provide an answer for that. Visualizing the average SHAP values of the positive feature values in each playlist type's dummy variable (if a song is being played in that playlist, the value is 1, in other cases 0) can aid to understand this. Being in the users own collection means that they are already familiar with the tracks and gathered them together to fit their musical taste. These playlists are not providing anything new to the users. The impact started positively in the beginning, then turned into negative for the second point. From there, the impact of being in this type was still negative, however the strength of it first decreasing then increased again a bit at the last point, being or not being a song in a user own collection had an slight negative effect on the predictions after the first point. This suggest, that listening to a familiar song either cause it to be skipped before even listening to it, or if started properly, a negative impact with different strength can be seen. It needs to be kept in mind, that the second point here might not be accurate, since it does not fit clearly into a pattern. It might be possible that the features have an overall decreasing impact on the probability of the skips in the order.

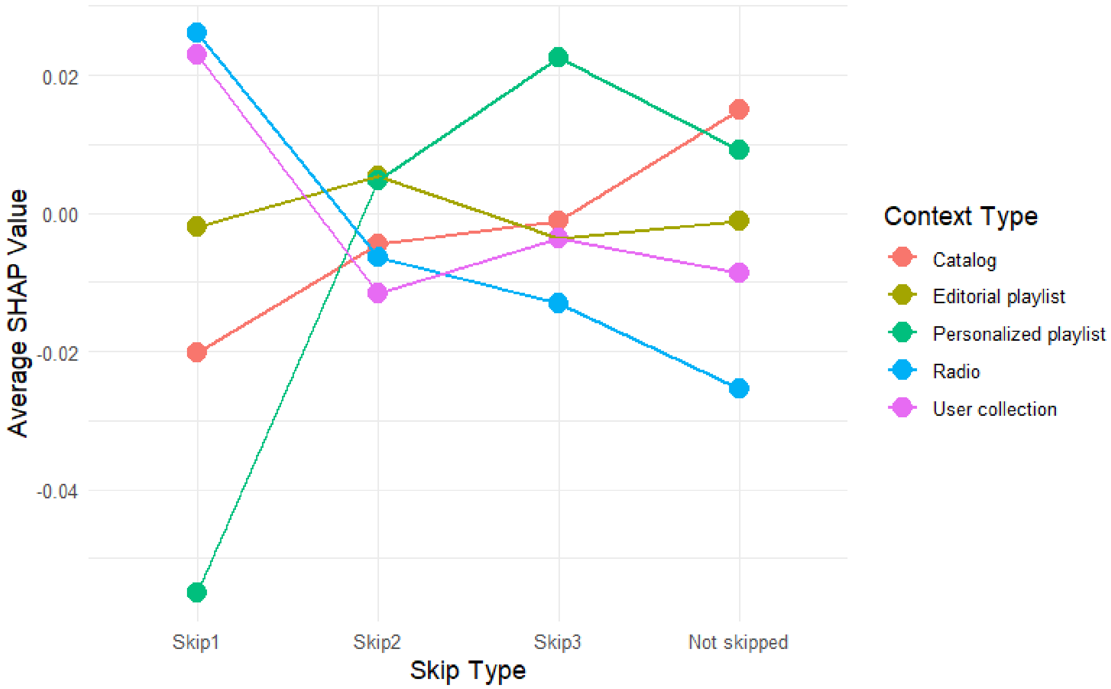


Figure 6.9: Average SHAP values of high feature values by Skip type and Context type

The platforms automated radio stations have a really large positive effect towards skipping at the first point, and after that the probabilities turn into negative effects, which are keep increasing until the not skipped points. This suggests that these radio stations are overall less successful than the user’s own collection. Radio stations can be created to provide similar songs to certain genres, artists, playlists, but even to a single song. The findings suggest that these songs are not that successful in terms of skipping as the users own familiar songs.

The other algorithmically generated playlists are tailored for each individual user. Personalized playlists create a balance between familiar and unknown song, based on the interest and listening history of the user. This approach seems more successful, as the variable influences negatively the Skip 1 probabilities (radio and user collection at this point had a strong positive effect). The direction of the Skip 2 influence is not clear, it might be a bit positive, but considering the model performance this might be misleading. Having said that, the chanced have a big positive influence from this type at third skipping points, suggesting an increasing pattern in the direction of the influence. Unfortunately, no real conclusion can be drawn from the not skipped SHAP values for this variable, however it can be said that a song being played in the personalized playlist context usually have a higher possibility of being played longer, than other recommendation types.

At three out of four points, being part of an editorial playlist had positive and negative effect as well, around the same magnitude. At the second skipping point the effects were rather positive, however since the model performance and the inconsistency with the other skipping variables I would rather not interpret those SHAP values further. Based on these findings, in case of listening to playlists generated by Spotify's professionals no real conclusion can be drawn for the skipping behavior. Most likely in these cases the fact that a user is listening to an editorial playlist is not influencing their average skipping behavior, other factors are more important in these cases, perhaps the sessions length or pausing behavior.

Listening from the platforms general catalog can serve as a baseline for the recommendation types. The SHAP values of this variable shows a definite increasing pattern, starting from a strong negative impact on skip 1 and finishing with a similarly strong positive impact on the probability of the song not getting skipped. These indicates that if a user is listening to a certain song he/she is particularly interested in, the track will more likely be listened to until a later point on the song, meaning playlists are still have room to adapt to fully understand the users' needs. As a last mention, charts do not really have much effect on the predictions, the logit model also suggested that it is the only not a significant variable in that model.

To summarize these findings, many interesting information can be explored in the SHAP values of the ordinal forest models predictions, however analyzing all of them were not in the scope of this research. The most important one suggests the importance of a break before the songs and the length of these examined sessions. After evaluating the different recommendation types, it seems clear that users are the most engaged with the sessions when listening to their individually tailored playlists. Comparing with the other algorithmically created playlist type, radio is less successful in the terms of influencing the skipping behavior in a positive way. The users own created playlists performed better than radios overall, however they do not show any unfamiliar songs to the listeners, as the algorithmically created ones do. Unfortunately, no real conclusion can be said about the editorial type. Spotify's services of providing access freely to a vast amount of music proves successful, as the directly searched tracks have a very positive impact, meaning the listening in those cases is more likely to end when the song is finished.

Conclusion

In this research, the topic of music streaming has been explored, especially focused on Spotify's platform, which is currently leading the market. The users can choose from multiple types of music recommendations while using the application instead of searching for every track by themselves. The general workings of these types have been described, in order to understand the background of the research. Several articles explored the topic of different playlists and their individual effect; however their differences were not really well analyzed before. One of the most suitable metrics for measuring the success of a song on a streaming platform is whether it is being listened to or rather being skipped. The topic of skipping behavior has been of importance before. The company held a competition for freelancers to develop machine learning solutions to predict the skipping behavior of their customers, where several complex solutions were developed and described in papers. The data that was released contained multiple types of skipping point during a song's listening interval, however the purpose of the competition was only focused on one of those. The problem with these different skip points was that the description did not quantify them, only a brief description was attached to them, from which only the order of them were clear. Considering different skipping points as ordinal data has not been considered when building these deep or reinforced learning solutions, allowing grounds for this thesis topic. Most methods could possibly handle ordinal data, although most of these are treating it categorical, not acknowledging its ordinal nature or treating it is as numerical, giving the same differences between the categories, which is a wrong approach as well. To handle this data type properly, correct methods needed to be found. One was a more general statistical method, generalized ordered logistic regression (more precisely partial proportional odds model). Another option was the Ordinal Forest model, which is a more complex machine learning solution based on random forest, tailored for ordinal data. These two models were trained and evaluated on a feasible subset of the dataset. Interestingly, the more complex machine learning model only performed slightly better. To get insights into to effects of different recommendation types, the SHAP values were constructed for the predictions of the

Ordinal Forest Model, and the insight were evaluated. The most influential factors behind the predictions were identified, as well as the effects of different playlists on the probability of skipping the tracks at different points.

The research has provided new findings on the differences in user engagement between the music recommendation types. For that purpose, a relatively new machine learning approach was utilized to account for the ordinality of the skipping at different points during the song, which was not in center of previous research. To get insights into the workings of the model, SHAP values provided insights into the different skipping points. The usage of the ordinal forest with combination of a Black Box interpretation method proved a great academic contribution, as well as the understanding of the differences between the recommendation type. As it has been shown, Spotify's personalized playlist types perform the best among the recommendation types. This can encourage the platform to advertise the successfulness of their algorithm, since currently only a surprisingly small portion of tracks are listened to in these contexts. It has also been shown that the automated radio stations are on one hand listened to more often, however their success in terms of skipping is less optimal. This fact could be useful for competitors such as Apple Music and Youtube Premium, since leveraging this fact could encourage them to improve their similar features to outperform Spotify, and close in on them in terms of market share. It has also been found that editorial playlists do not really affect skipping behavior, this fact have to be evaluated and the company needs to find ways to improve the success rate of these type of recommendations. The above-mentioned findings could also be used for content creators. Being featured on an editorial playlist on Spotify can be a momentous success for an independent artist, however it is not guarantee, that the song will be successful in terms of user engagement, and the artists might not even get paid in the end. Composing a song that is very similar to the average of a certain type of music or mood might put the track on the automated radio stations, however that can even decrease the chance of being played longer, so it is not a solution for them. A better option for the creators is to find an audience, which will put their music in their own playlists, which they are listening to more often, and skipping the song generally less.

The limitations of the study also need to be mentioned. With limited computational resources, only a minor fragment of the total data could be used, since handling large amount of data, training machine learning models and especially computing SHAP values are requiring huge amount of computing power. This would indicate that the true impacts of the playlists might look a bit different on the whole dataset. Secondly, the chosen methods did not perform particularly well. This is most true in case of the variable, for which the said competition was organized. To put this in context, the largest streaming platform released 360 gigabytes of sensible company data, on which groups of experts trained overly complex deep learning solutions to leverage the patterns and predict this kind of skipping, which is makes perfect sense now. While the focus of this research was not to perfectly predict the skipping themselves and the different intervals between them, a more better performing model's insights might possibly be safer to interpret. The ordinal nature of the skipping variables was also a sort of limitation, since knowing the exact differences between these points would have possibly given more clear predictions on them, hence more interpretable differences on the recommendation types.

In spite of these limitations, this research has valuable findings, and could potentially spark future research. This could be the usage of new methods, that might can leverage the ordinal nature better, hence providing better performance. Other Black Box interpretation methods could also be used, to provide a different approach on getting relatable insights on the data. This would especially be interesting in the case of editorial playlists, where no real results were found. These results could also be good ground for further research in the differences on the distinct recommendation types, since that is still a relatively unobserved area of music streaming.

References

Agga, G. E., & Scott, H. M. (2015). Use of generalized ordered logistic regression for the analysis of multidrug resistance data. *Preventive Veterinary Medicine*, *121*(3–4), 374–379. https://doi.org/10.1016/j.prevetmed.2015.08.015

Aguiar, L., & Waldfogel, J. (2018). *Platforms, Promotion, and Product Discovery: Evidence from Spotify Playlists*. (Working Paper No. 24713) http://www.nber.org/papers/w24713

Béres, F., Kelen, D. M., & Benczúr, A. (2019). *Sequential skip prediction using deep learning and ensembles* (pp. 1–4). ACM Press. https://people.eng.unimelb.edu.au/jianzhongq/wsdm19-cup-reports/reports/report16.pdf

Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics*, *46*(4), 1171–1178. https://doi.org/10.2307/2532457

Brost, B., Mehrotra, R., & Jehan, T. (2019). The Music Streaming Sessions Dataset. *The World Wide Web Conference*, 2594–2600. https://doi.org/10.1145/3308558.3313641

Carroni, E., & Paolini, D. (2017). *Content acquisition by streaming platforms: Premium vs. freemium*.

Cerrah, S., & Yigitoglu, V. (2022). Determining the effective factors on restaurant customers' plate waste. *International Journal of Gastronomy and Food Science*, *27*, 100469. https://doi.org/10.1016/j.ijgfs.2022.100469

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Ferraro, A., Bogdanov, D., & Serra, X. (2019). *Skip prediction using boosting trees based on acoustic features of tracks in sessions*. https://doi.org/10.48550/arXiv.1903.11833

Freeman, S., Gibbs, M., & Nansen, B. (2022). 'Don't mess with my algorithm': Exploring the relationship between listeners and automated curation and recommendation on music streaming services. *First Monday*. https://doi.org/10.5210/fm.v27i1.11783

Hansen, C., Hansen, C., Alstrup, S., Simonsen, J. G., & Lioma, C. (2019). *Modelling Sequential Music Track Skips using a Multi-RNN Approach*. https://doi.org/10.48550/arXiv.1903.08408

Hansen, C., Hansen, C., Maystre, L., Mehrotra, R., Brost, B., Tomasi, F., & Lalmas, M. (2020). Contextual and Sequential User Embeddings for Large-Scale Music Recommendation. *Fourteenth ACM Conference on Recommender Systems*, 53–62. https://doi.org/10.1145/3383313.3412248

Hartsell, T., Yuen, S., & Yuen, Y. (2006). Video streaming in online learning. *AACE Journal*, *14*, 31–43.

Hesmondhalgh, D. (2020). Is Music Streaming Bad for Musicians? Problems of Evidence and Argument. *New Media and Society*, *23*. https://doi.org/10.1177/1461444820953541

Hornung, R. (2020). Ordinal Forests. *Journal of Classification*, *37*(1), 4–17. https://doi.org/10.1007/s00357-018-9302-x

Küng, L. (2017). *Strategic Management in the Media: From Theory to Practice* (p. 248). Sage. https://doi.org/10.4135/9781446280003

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Neural Information Processing Systems.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109–127. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 276–282. https://doi.org/10.11613/BM.2012.031

Meggetto, F., Revie, C., Levine, J., & Moshfeghi, Y. (2021). On Skipping Behaviour Types in Music Streaming Sessions. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3333–3337. https://doi.org/10.1145/3459637.3482123

Meggetto, F., Revie, C., Levine, J., & Moshfeghi, Y. (2023). Why People Skip Music? On Predicting Music Skips using Deep Reinforcement Learning. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, 95–106. https://doi.org/10.1145/3576840.3578312

Mulligan, M. (2024, January 30). *Music subscriber market shares Q3 2023 New momentum*. MIDiA Research. https://www.midiaresearch.com/reports/music-subscriber-market-shares-q3-2023-new-momentum

Pachali, M. J., & Datta, H. (2023). What drives demand for playlists on Spotify? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4079693

Pichl, M., Zangerle, E., & Specht, G. (2017). Understanding User-Curated Playlists on Spotify: A Machine Learning Approach. *International Journal of Multimedia Data Engineering and Management*, *8*(4), 44–59. https://doi.org/10.4018/IJMDEM.2017100103

Shapley, L. S. (1953). A Value for n-Person Games. In *17. A Value for n-Person Games* (pp. 307–318). Princeton University Press. https://doi.org/10.1515/9781400881970-018

Spotify Technology S.A. (2024). *Spotify Annual report 2023. https://d18rn0p25nwr6d.cloudfront.net/CIK-0001639920/26aaaf29-7cd9-4a5d-ab1f-b06277f5f2a5.pdf*

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Thomes, T. P. (2013). An economic analysis of online streaming music services. *Information Economics and Policy*, *25*(2), 81–91. https://doi.org/10.1016/j.infoecopol.2013.04.001

Williams, R. (2006). Generalized Ordered Logit/Partial Proportional Odds Models for Ordinal Dependent Variables. *The Stata Journal*, *6*(1), 58–82. https://doi.org/10.1177/1536867X0600600104

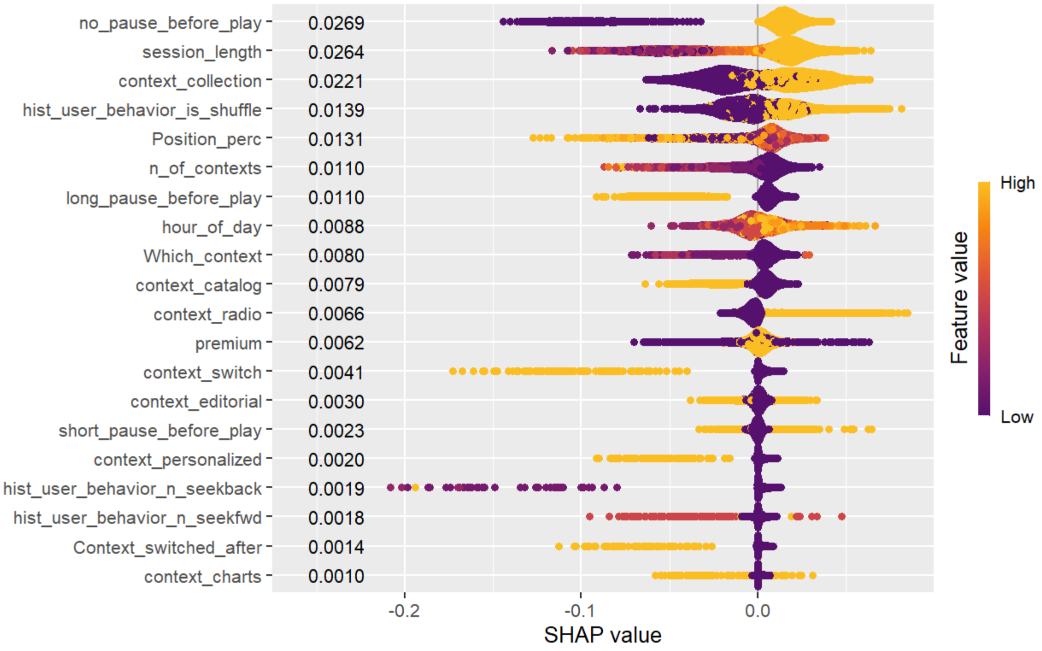
Zehr, H. (2021). An Economic Analysis of the Effects of Streaming on the Music Industry in Response to Criticism from Taylor Swift. *Major Themes in Economics*, *23*, 51–63.

Graphs:

Richter, F. (March 22, 2024). Streaming Rapid Rise to Mass Adoption [Digital image]. Retrieved May 28, 2024, from https://www.statista.com/chart/24506/users-of-paid-music-streaming-subscriptions/

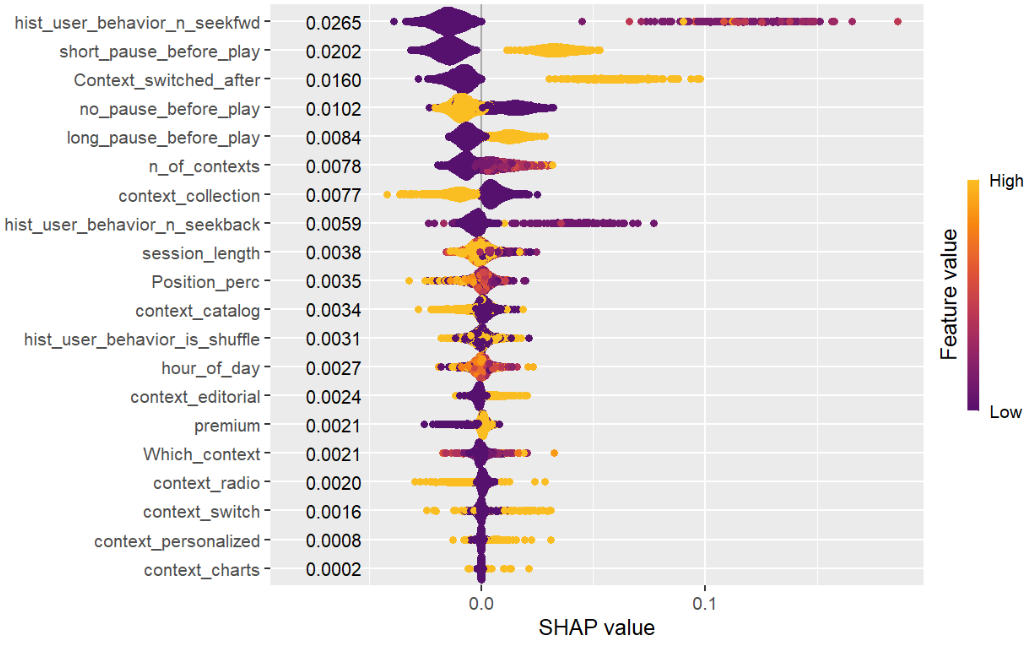
Appendix

Figure 1:



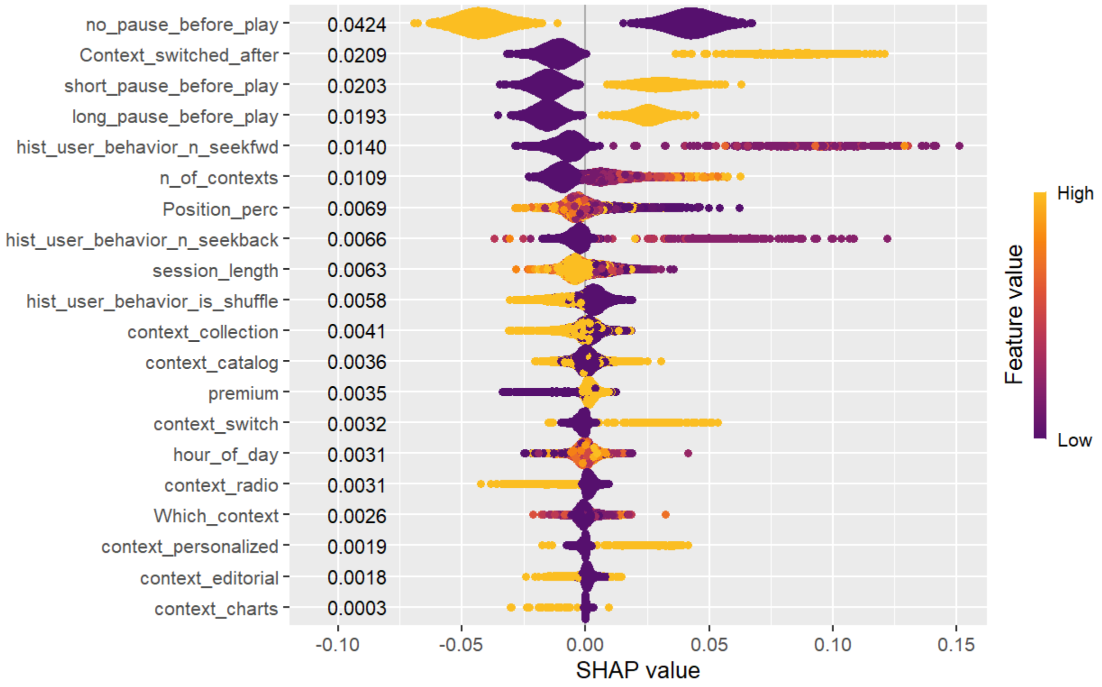
Beeswarm Plot of SHAP Values: Influence of all the variables on Skip 1 Probability

Figure 2:



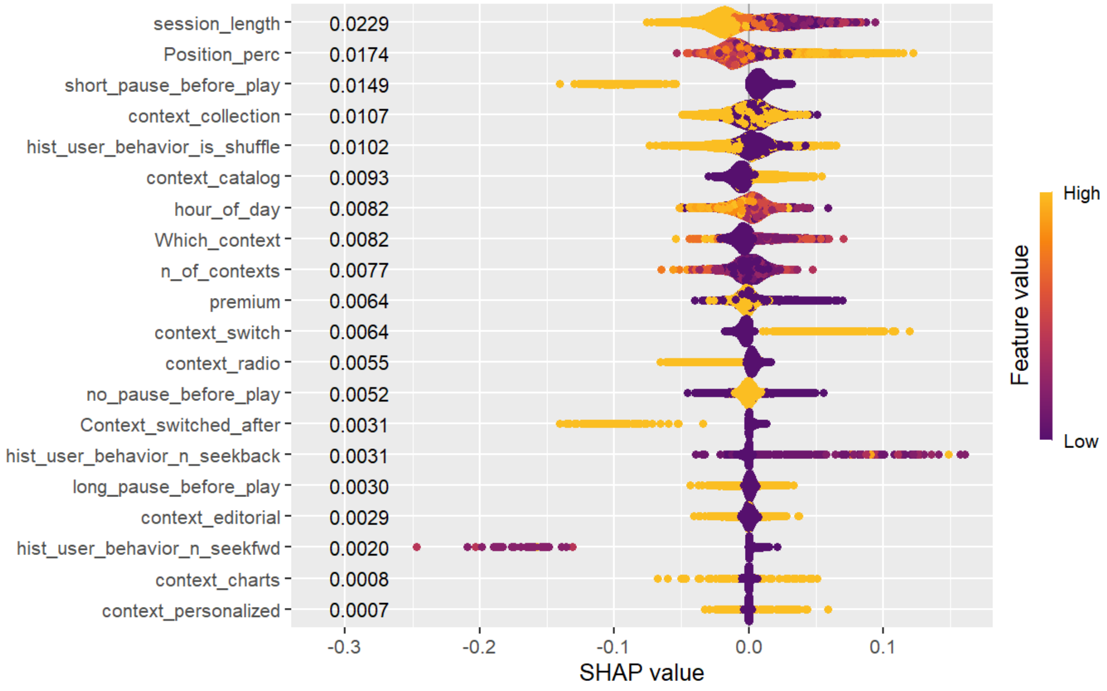
Beeswarm Plot of SHAP Values: Influence of all the variables on Skip 2 Probability

Figure 3:



Beeswarm Plot of SHAP Values: Influence of all the variables on Skip 3 Probability

Figure 4:



Beeswarm Plot of SHAP Values: Influence of all the variables on Skip 4 Probability