# Spotify Skip Prediction

Alex Thach – alcthach@gmail.com
BrainStation Data Science Capstone Project
April 4, 2022

## Introduction

The goal of this project is to predict whether a user will skip a given track on Spotify, a music streaming platform. Tracking skip is often overlooked as a research topic on music streaming platforms, often overshadowed by research related to recommendation systems. However, uncovering trends about track skipping may allow for new approaches to music curation for the listener. Such as ad-hoc or on-the-fly playlist curation. In essence, understanding the probability of a user skipping a certain track may also clue us in as to whether or not we want to include the song in the playlist for the user. Although Spotify markets unlimited skips as a feature for premium users, it appears to be a feature that is least important in the order of premium features. Which might suggest there is room to develop a feature that allows for the platform to consider transient listening moods in users.

## Data Acquisition/Description:

The data for this project was acquired from AIcrowd, a data science competition platform. The dataset belonged to the "Spotify Sequential Skip Prediction Challenge".

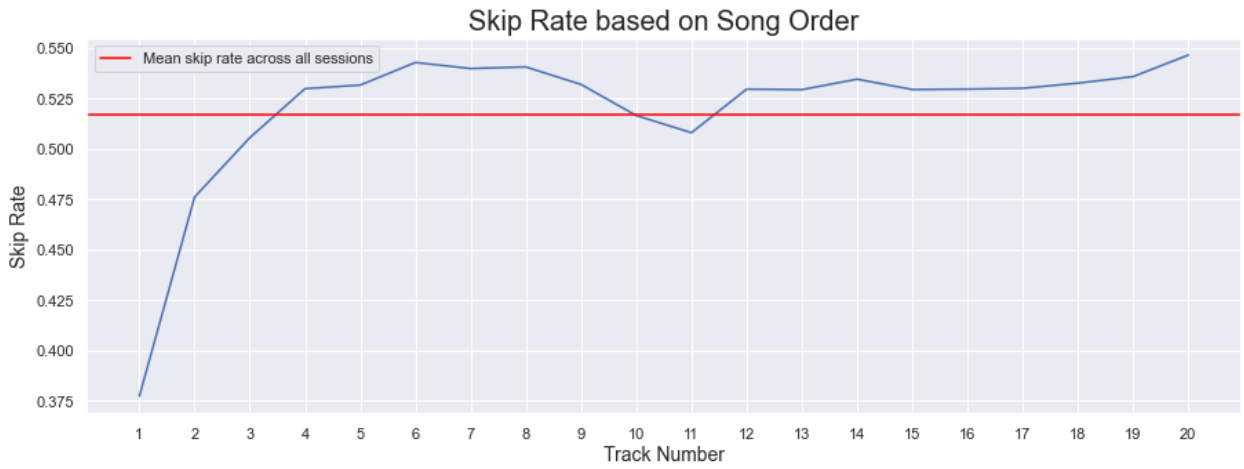
## Characteristics of the Data

The data was made available in a tabular format. Specifically in two comma-separated files. One file contained the user session logs, outlining the history of songs presented to the user and their interactions with each song. In addition, a table with the track features of the songs found in the session logs was also provided. In this table the song characteristics were provide for each song found in the session logs. Some track features include acousticness, valence, and speechiness. In the sessions logs were features like skip, pause, seek, and contextual features like type of playlist the track belonged to and if they were a premium user.

Together these two tables help into inform what the characteristics of the songs within a given user's listening session. And how the user interaction with each track during the session.

There was a total of 167,880 rows in the session logs dataset, which represented the total amount of songs presented to the user in 10,000 user listening sessions. The data types found in the dataset included: object, numeric (float, integer) and boolean.

## Summary of Findings from Exploratory Data Analysis




Based on the figure above, we can see that beyond the first few tracks during user listening sessions. Average track skip rate was between 50-55% across positions in the listening session. With an overall track skip rate around 52%.

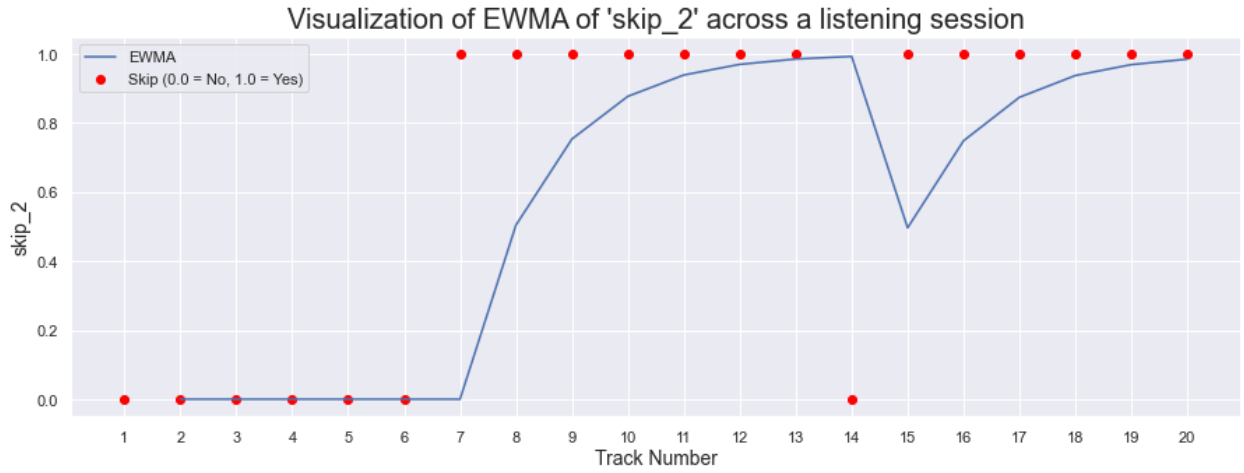
---

## Feature Engineering:

Methodologies employed include:

- Splitting of date column into day/month/year
- Computing cumulative average of session characteristics across the listening session
- Computer of exponentially-weighted moving average of session characteristics across the listening session

## Double-clicking on Exponential-weighted Moving Average




The figure above visualizes the relationship between actual skip outcome of a track and the exponentially-weighted moving average (EWMA). The red dots indicate the skip outcome of a track in the listening session, and the blue line indicates the exponentially-weighted moving

average throughout the course of the listening session. As you can see, the blue line plot appears to be staggered. This was done to mitigate data leakage. The EWMA includes the current observation in its computations, meaning that without the stagger, the model would indirectly understand if the track was skipped or not.

By employing this methodology, each observation now contains information about the aggregate characteristics of the session. For example, at track 8 in the example above, the EWMA jumps from 0 to about 0.5. The model would recognize that the user hasn't been in a skipping mood, but to the previous track that's changed, and as the session progresses, the EWMA climbs further as the user skips the following songs. And the EWMA reflects that the user has been in a 'skipping mood'. This methodology was extended to features such as danceability, year, and popularity estimates. Allowing the model to understand the general characteristics of the session at a given point in time. For example, the beginning of a session might be characterized by high skip rate and high loudness, then towards the middle skip rate is lower with lower loudness. This feature allows us to capture these characteristics.

---

## Modelling

*Key Findings:*

- A baseline logistic regression on the encoded categorical columns and exploded date column had a classification accuracy of 76%
- Classification accuracy was better when the model knew the history of user interactions
    - Strong predictor of track skip was if the forward button was used to arrive at the current song
    - Strong predictor of track non-skip was is the previous was played to the end
- Without this information the model performed slightly better than a naive model would (at 52% accuracy)
- My best performing model was a Random Forest classifier trained on the EWMA of user interactions and track features, with a classification accuracy of 78%

*Discussion:*

Although the classification accuracy of my best model outperformed the naive model by a considerable margin. There are some concerns. The model relies heavily on prior history of user interaction to drive its predictions. Without this information, it suffers a large decrease in predicting power. Because of this mechanic, its practicality in deployment is questionable. It's best served as an exploratory tool. As it's helped me to better understand the patterns in the data. But in terms of providing a solution to the problem of improving listening experience, it might serve as an intermediary step. It can interject during a session and classify a song as likely to be skipped, maybe signalling that the song should be removed. But it's quite static in that sense. The overall problem of improving listening experience through addressing track skipping likely requires a more sophisticated approach.

*Closing remarks:*

As I progressed through the project, I grew to appreciate the complex of the problem. The data analysis, feature engineering and modelling allowed me to build a deeper understanding of the problem space. I intend to iterate through the project again, attempting to uncover more patterns in the data and employ more sophisticated approaches to the problem. Which while likely require literature reviews of similar problems. And deep learning methodologies as they look well equipped to handle the complexity of this problem.