



CASE STUDY OF SPOTIFY SKIP OPERATION ON LIVING STREAMING DATA

ZAMMATH TV



INTRODUCTION

- Spotify — is a streaming service whose business model is centered around delivering song recommendations to users in order to keep them using their app.
- Spotify has over 190 million active users interacting with over 40 million tracks
- The goal of the challenge is to predict the likelihood of a user skipping any given song during a listening session.



DATA SETS


Spotify supplies two main sets of information. One table has information about user listening sessions. For example

- session_id
- session_position
- session_length
- track_id_clean
- skip_1
- skip_2
- skip_3
- not_skipped
- context_switch
- no_pause_before_play
- short_pause_before_play

Long_pause_before_play

- hist_user_behavior_n_seekfwd
- hist_user_behavior_n_seekback
- hist_user_behavior_is_shuffle
- hour_of_day
- date
- premium
- context_type
- hist_user_behavior_reason_start
- hist_user_behavior_reason_end

Note that the user sessions were all between 10 and 20 tracks long and didn't include any identifiable information about the users



The second table has metadata about the tracks (corresponding to the trackid features from session table) such as:

- track_id
- duration
- release_year
- us_popularity_estimate
- acousticness
- beat_strength
- bounciness
- danceability
- dyn_range_mean
- energy
- flatness
- instrumentalness
- key
- liveness
- loudness
- mechanism
- mode
- organism
- speechiness
- tempo
- time_signature
- valence
- acoustic_vector_0
- acoustic_vector_1
- acoustic_vector_2
- acoustic_vector_3
- acoustic_vector_4
- acoustic_vector_5
- acoustic_vector_6
- acoustic_vector_7



FEATURE ENGINEERING

- This data is inherently Sequential, The model skip prediction for any given song will be based on the songs that were skipped earlier in the user listening session.
- Generate features to account for a user's listening history.
- Added previous track features, including if that track was skipped.



MODEL IMPLEMENTATION

- Target Metric Accuracy Per competition guidelines.
- Logistic Regression

We can implement this model by assuming logistic regression ,relationship between variables

- Random Forest

Tree_based model which would automatically handle feature interaction

- XGBoost

Boosting is an ensemble method for improving the model prediction of any given learning algorithm



Results

ACCURACY:-

- Best test accuracy .73 (with lightGBM)
- Logistic regression -pretty poor performance(0.54)
- Random forest-Much better performance(0.70)
- Xgboost-Much more performance(0.72)



CONCLUSION AND FUTURE SCOPE

The Model Performance is reasonable, with an accuracy of .73 compared to completely a naive model that would have an accuracy of .51, but still with plenty of room for improvement.

SCOPE:-

Test with Other types of Algorithms

- Unsupervised learning to cluster song
- Supplement the dataset with more data from the Spotify API
- RNN to predict based on the sequence of tracks

Create DJANGO App to visualise predictions using D3



THANK YOU