Q - 1. Describe your research scenario. This is a general description of the use case. Follow class examples, first describe the overall scenario and then specify a specific research question. - 200

A - 1. As a music listener, it’s intriguing to see how well Spotify understands the user’s music taste. Two key features of Spotify are:

1. **“Shuffle song”**: which though is a random operation, is critical to form a good user experience. This operation should be random enough, but confined to the user's taste.
2. **“Skip button”**: This article explains about the relationship between skip button and user experience. So, a skip button gives us information about the user’s music taste and indeed, skip rate depends on a lot of factors like the user's age, current hour of day, whether day is weekend or weekday.

So, the question that comes is **whether we can predict whether a user will skip a song shortly after play time** which is an indication of non-interest, when a playlist of songs are played sequentially based on historical play data of the user. This will give us an intuition about the user's music preference and help to design a “Shuffle Song” feature which is more focused on user experience.

Q - 2. Briefly, describe the data set, including each data field. If possible provide a Link to the main data set source. - 200

A - 2. There is a current challenge going on by Spotify: https://www.aicrowd.com/challenges/spotify-sequential-skip-prediction-challenge .
The dataset used for this model is derived from this challenge. Currently, we would be focusing on the sample dataset which has limited rows but will give us an intuitive idea about how our models are performing.

A dataset is divided into different sessions and every session has some sequence of tracks within a range of 0-20:

1. Session ID
2. Track ID

Every track has metadata which gives us details about music like acousticness, beat strength etc:

1. Duration - duration of the song
2. Release_year - release year
3. Us_popularity_estimate - US popularity
4. Acousticness
5. Beat_strength
6. Bounciness

7. Danceability
8. Dyn_range_mean
9. Energy
10. Flatness
11. Instrumentalness
12. Key
13. Liveness
14. Loudness
15. Mechanism
16. Mode
17. Organism
18. Speechiness
19. Tempo
20. Time_signature
21. Valence
22. Acoustic_vector_0
23. Acoustic_vector_1
24. Acoustic_vector_2
25. Acoustic_vector_3
26. Acoustic_vector_4
27. Acoustic_vector_5
28. Acoustic_vector_6
29. Acoustic_vector_7

Rest further, dataset contains details about user interaction and playlist details:

1. Session_position - Position in sequence in a single session
2. Session_length - Total session length
3. Skip_1 - Boolean indicating if the track was only played very briefly
4. Skip_2 - Boolean indicating if the track was only played briefly
5. Skip_3 - Boolean indicating if most of the track was played
6. Not_skipped - Boolean indicating that the track was played in its entirety
7. Context_switch - Boolean indicating if the user changed context between the previous row and the current row. This could for example occur if the user switched from one playlist to another.
8. No_pause_before_play - Boolean indicating if there was no pause between playback of the previous track and this track
9. Short_pause_before_play - Boolean indicating if there was a short pause between playback of the previous track and this track
10. Long_pause_before_play - Boolean indicating if there was a long pause between playback of the previous track and this track
11. Hist_user_behavior_n_seekfwd - Number of times the user did a seek forward within track
12. Hist_user_behavior_n_seekback - Number of times the user did a seek back within track

13. Hist_user_behavior_is_shuffle - Boolean indicating if the user encountered this track while shuffle mode was activated
14. Hour_of_day - {0-23} - The hour of day
15. Date - The date
16. Premium - Boolean indicating if the user was on premium or not. This has potential implications for skipping behavior.
17. Context_type - what type of context the playback occurred within
18. Hist_user_behavior_reason_start - - the user action which led to the current track being played
19. Hist_user_behavior_reason_end - the user action which led to the current track playback ending

Q - 3. Describe briefly in one or two sentences the main research question. - 100

A - 3. The main research question is whether we can predict the user interaction with the skip button after a song is played in some particular song sequence, given information about user and track. Moreover, can we analyze what kind of behaviour follows a strong pattern like the user listening to the whole music or not and what kind of interaction can be predicted more accurately like, will the user listen to whole music, or some part of it? Also, another question that arises is what quantity of historical data is sufficient to give us a better accurate prediction about skip behaviour. In this project, we would be answering these questions.

Q - 4. State the conclusion so that a none-data scientist can understand. - 100

A - 4. We can accurately predict the user skip behaviour interaction based on recent historical data of songs played by the user with more than 87% accuracy. Indeed, the best model predicts very accurately whether a user will listen to the full song with more than 99 % accuracy with knowledge of the last song played by the user. Also, we can conclude that the recent history of songs upto 15 songs play a very important role in providing very accurate predictions. We can use knowledge of skip behaviour in songs to provide user recommendations based on their music taste.