Predicting User Skip Interaction on Spotify

This project has been submitted by Achint Sharma, Student ID U23627091 and done under the guidance and supervision of Prof. Farshid Alizadeh-Shabdiz, Boston University as a part of the Term Project for the course CS677 - Introduction to Data Science with Python. I would also like to thank the Teaching Assistant for this course for sharing their valuable experiences and insights. I would also like to thank Spotify for making this data available through AI Crowd Challenge[1].

Abstract

Music forms an important part of human society and culture with historical roots emerging from prehistoric times[2]. As of 2019, 68 % of US adults report that they listen to music everyday[3]. The population of music listeners has grown substantially even after 2019. The field of data science has given engineers and researchers the ability to unleash powerful insights about music listener's psychology. As a consequence, we can now understand the user's taste better and create playlists and music recommendation systems focused on the user's interest. In this document, we would be focusing on one of the aspects, where we would be analyzing the user's interaction with a skip button and would try different data science models to analyze and predict the user's taste for music.

Introduction

A user's skip behaviour is defined as the time interval between the start of the song and the time at which the user skips the song. A user's skip song interaction forms a core aspect of user experience as it gives more control to the user and helps engineers and researchers to better understand the user's taste. This gives us information about the user's music interest which can help music applications to recommend and play songs which are more likely to be listened by the user. In this project, we are focusing on answering three main questions about the user and data:

1. Can we predict a user's skip song interaction based on historical data about the user's skip behaviour and music metadata and some other related user information?

2. What type of skip behaviour follows a more strict pattern and can be predicted more accurately from the data science models?

3. What amount of historical session data is sufficient to make a more accurate prediction? For example, can we make more accurate inference about the user's skip behaviour knowing about just the last 3 songs played by the user rather than the last 5 songs?

Dataset Description and Analysis

This dataset is provided as a part of the AI Crowd challenge by Spotify. Currently, we are using a sample dataset of the challenge which consists of 167880 rows. This dataset consists of 30 song related features and 21 other user interaction based features.

The dataset can be divided into 10000 sessions based on session_id and each track is analyzed in context of the session data. Each session consists of sequence of songs ranging from 10 to 20 as shown in the image below:

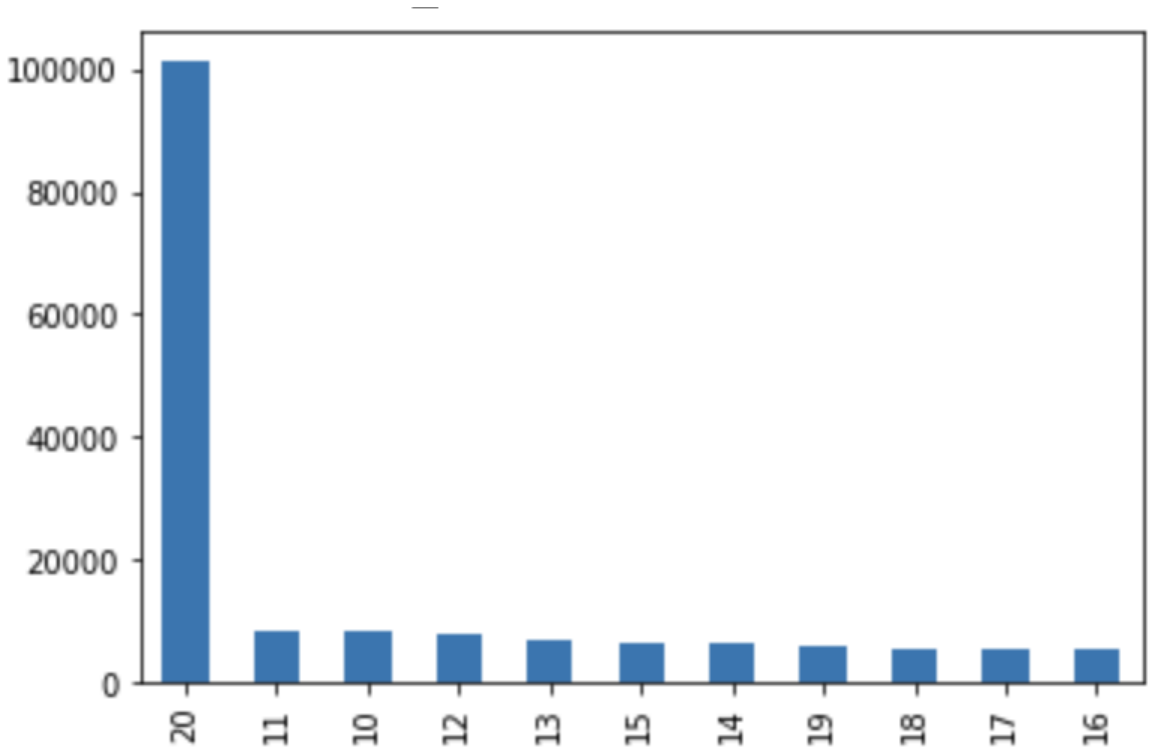

Fig 1: Count of total sessions vs Session length

In this project, we would be focusing on following 4 types of skip behaviours, as provided by dataset:

1. **Skip_1**: User skips the songs after playing very briefly
2. **Skip_2**: User skips the songs after playing briefly
3. **Skip_3**: User skips the songs after listening most of the song

4. **Not_skipped**: User listens the full song and doesn't skip.
5. **Otherwise:** Error data; to be ignored

The figure below shows the count of different skip behaviours:

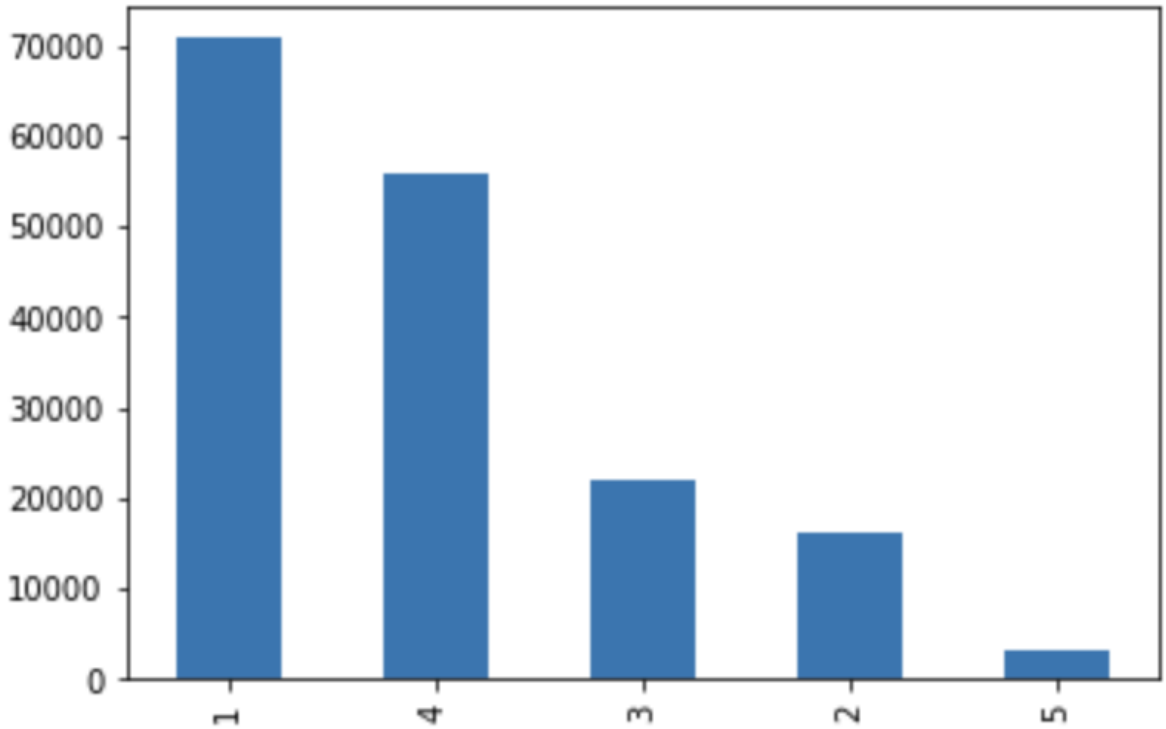

Fig 2: Count of Skip behaviour vs Skip Behaviour Type

As hour of the day is an important factor in predicting the skip behaviour type, we can see that the data is distributed well amongst all the day hours.

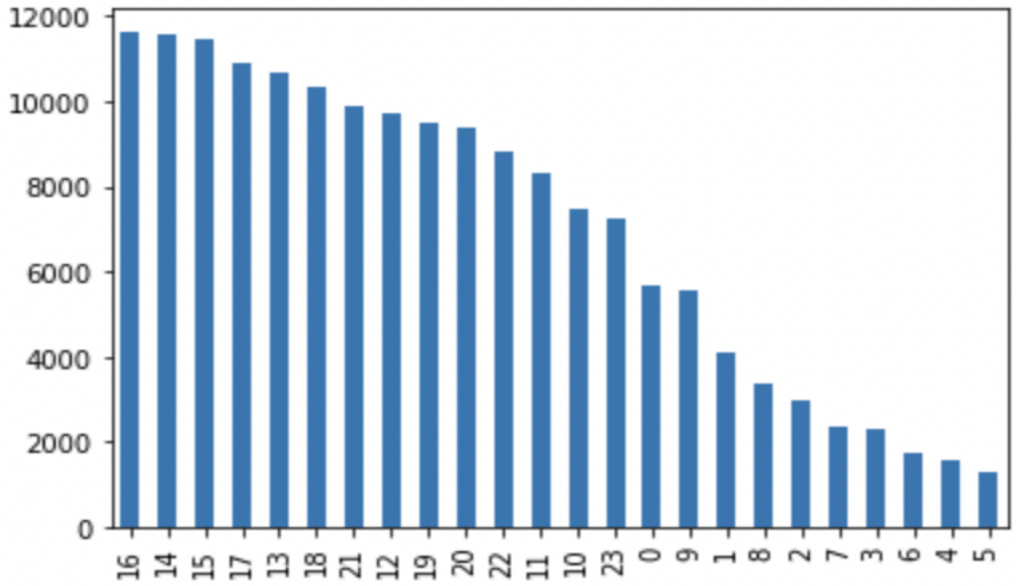

Fig 3: Distribution of data based on hour of day

Data Preprocessing

A very important aspect in data preprocessing that has been done related to this dataset is imposing features of previous historical sessional song data as individual features of the row. Consider n as a hyperparameter of the dataset, which is the concerned last n songs played. This means, all the data points of last n songs will be added to the concerned row as an individual row. If n = 5, all the data points of the last 5 songs will be added as a feature of the row. For example, for n = 0, total data points are 45 and for n = 5, we have 276 data points.

Further, to answer each of the questions given above, we would be reframing the problem of predicting skip behaviour to One vs Rest classification problem, where we would be predicting each skip behaviour type individually. This would help us in analyzing each question individually and with more granular inferences.

We would be scaling the dataset using standard scale because we would be trying distance-based classification techniques. We would be dividing the dataset into a 50-50 ratio for training and testing. As we can see above in Fig 2, the dataset is unbalanced in context of different skip behaviours. To counteract this, we would be using upscaling in the training data set to rebalance the data.

Analysis

We would be applying following classification techniques and compare results of each model for each question:

1. Logistic Classifier
2. Random Forest Classifier
3. Linear Discriminant Classifier
4. Quadratic Discriminant Classifier
5. Linear Regression
6. K-neighbours Classifier with p = 1, n = 3
7. K-neighbours Classifier with p = 2, n = 3
8. K-neighbours Classifier with p = 2, n = 5
9. Gaussian Naive Bayesian Classifier
10. Gradient Boosting Classifier
11. Adaptive Boosting Classifier

We would be training the model for a discrete value of n: [1, 3, 5, 10, 15, 19]. Further, as discussed above, we would be predicting each skip behaviour individually. So all our analysis will be based on n and skip behaviour. We would be using accuracy and F1 score as our measure to analyze the best model amongst all.

Results & Conclusion

Skip Type	**Best n**	**Best model for accuracy**	**Accuracy**	**F1 score**
1	10	**Adaptive Boosting Classifier**	0.882587	0.875035
2	15	**Gradient Boosting Classifier**	0.887757	0.900088
3	15	**Gradient Boosting Classifier**	0.986865	0.989939
4	1	**Random Forest Classifier**	0.990271	0.985426

As we can see, the ensemble classifiers provided us with best accuracy and f1 score amongst all other types of model. The neighbours classifiers didn't perform very well for all the types of

skip behaviour types. The accuracy reaches till 0.990271 with historical data of just the last one song for skip behaviour type 4, i.e., playing the whole song. For skip behaviour type 2 and 3, recent history of the last 15 songs provided us with the best accuracy and f1_score in each case. For skip type behaviour 1, best n choice is 10 which provides us an accuracy of 0.882587 and F1 score of 0.875035.

We can now conclude that indeed, we can predict each type of user skip behaviour with a very high accuracy for a sequence of songs. Here, the historical sequence of songs plays a very important role as we can see, the knowledge of the last 10-15 songs provides us with a better prediction about user behaviour.

References

1. https://www.aicrowd.com/challenges/spotify-sequential-skip-prediction-challenge
2. https://en.wikipedia.org/wiki/History_of_music
3. https://www.statista.com/statistics/749666/music-listening-habits-age-usa/