KKBox's Music Recommendation Analysis

Introduction

Music is an important part of the daily life for lots of the people. Currently music is a billion-dollar global industry due to the high demand of the public for music consumption.

In 2017, 38 percent of the music industry revenues worldwide was attributed to streaming sector. Big players as Google with Youtube Music, Apple with Apple Music, Spotify or KKBox are trying to get their piece of the cake. In this highly competitive environment, being able to satisfy users' needs by implementing and optimizing music recommendation systems will be crucial.

By improving recommendation systems, business metrics could be as well improved. For example, offering songs adjusted to the user preferences can improve retention which is highly related to monetization. Many companies in the music industry have subscription business models where increasing retention is key to improve the revenue.

Client and problem definition

KKBOX was founded in 2004, and it is leading the music streaming industry in Asia. Its wide Asia-Pop music library contains more than 30 million tracks. The service allows you to listen to your music playlists from all your devices and uses a recommendation system combining matrix factorization and word embedding methods.

Nowadays, user music preference could vary from classical music to the latest hit. Recommendation systems need to predict whether a person will enjoy a new artist or a new song. This is especially challenging when the listener recently joined the service, since there is not enough historical data. Predicting user behaviour is always a difficult task. This project will focus on answering the following question:

- Will a user listen to a song again in less than a month after the first time that it was listened?

This problem was proposed by the International Conference on Web Search and Data Mining (WSDM 2018). One of the challenges about this problem is the fact that there are several datasets provided, so in order to combine those pieces of information makes the analysis harder. The dataset it very recent, from 2017, so there are not many references about its analysis.

Data description

The dataset contains a sample of over 7 million listening events performed by more than 30,000 users. Each event corresponds to the first one performed for a particular user-song pair in the selected time period. Data also contains details about the user, such as registration time or age, and song metadata.

Data is available on Kaggle:
https://www.kaggle.com/c/kkbox-music-recommendation-challenge/data

The full description of each of the files can be found below:

train.csv (7.4m rows, 30.8k users)

- msno: user id
- song_id: song id
- source_system_tab: the name of the tab where the event was triggered. System tabs are used to categorize KKBOX mobile apps functions. For example, tab `my library` contains functions to manipulate the local storage, and tab `search` contains functions relating to search.
- source_screen_name: name of the layout a user sees.
- source_type: an entry point a user first plays music on mobile apps. An entry point could be `album`, `online-playlist`, `song` .. etc.
- target: this is the target variable. `target=1` means there are recurring listening event(s) triggered within a month after the user's very first observable listening event, `target=0` otherwise .

test.csv (2.6m rows)

- id: row id (will be used for submission)
- msno: user id
- song_id: song id
- source_system_tab: the name of the tab where the event was triggered. System tabs are used to categorize KKBOX mobile apps functions. For example, tab `my library` contains functions to manipulate the local storage, and tab `search` contains functions relating to search.
- source_screen_name: name of the layout a user sees.
- source_type: an entry point a user first plays music on mobile apps. An entry point could be `album`, `online-playlist`, `song` .. etc.

songs.csv (2.3m rows)

The songs. Note that data is in unicode.

- song_id
- song_length: in ms

- genre_ids: genre category. Some songs have multiple genres and they are separated by `|`
- artist_name
- composer
- lyricist
- language

members.csv (34.4k rows)

User information.

- msno
- city
- bd: age. Note: this column has outlier values, please use your judgement.
- gender
- registered_via: registration method
- registration_init_time: format `%Y%m%d`
- expiration_date: format `%Y%m%d`

song_extra_info.csv (2.3m rows)

- song_id
- song name - the name of the song.
- isrc - International Standard Recording Code, theoretically can be used as an identity of a song. However, what worth to note is, ISRCs generated from providers have not been officially verified; therefore the information in ISRC, such as country code and reference year, can be misleading/incorrect. Multiple songs could share one ISRC since a single recording could be re-published several times.

Data wrangling and cleaning steps

In the **train.csv** document we performed the following steps:

- Identify ‘Unknown’ values as NaN.
- Transform source_screen_name values to lower letters.
- Transform source_system_tab, source_screen_name and source_type columns into categorical variables to reduce memory usage.
- Drop rows containing missing source_system_tab and source_type values, since they represent less than 5% of the data.

Here is the list of actions we did in the **members.csv** document:

- Transform registration_init_time and expiration_date columns into datetime variables.

- Transform city, gender and registered_via columns into categorical variables to reduce memory usage.
- Remove rows where expiration_date is before 2004, since KKbox was launched in 2004.
- Age values less than 0 and greater than 120 are converted into missing values (NaN).

In the **songs.csv** file we performed the steps described below:

- Replace missing language values for -1.
- Format language values removing decimals and convert language into categorical variable. This reduces the memory usage.
- Express song_length in minutes instead of milliseconds.
- Capitalize first letter of each word and remove unnecessary spaces (leading, ending and multiple spaces) in the artist_name, composer and lyricist columns.
- Transform genre_ids into a list. In the future probably it will be needed to create a categorical column for each of the genre to say whether a song belongs to a particular genre or not.
- Transform composer and lyricist into a list.
- Add column genre_count displaying number of distinct genres a song belongs to.

Some variables were not modified:

- Language values equal -1 sometimes refer to songs with only melody and other times to songs were the language has not been categorized. Therefore, no further steps have been taking to transform those values.
- Song_length outliers were kept since they correspond to playlists

Finally, in the **song_extra_info.csv** file we only did one change:

- Capitalize first letter of each word and remove unnecessary spaces in the name column.

Additional data sources

In order to improve the quality of the data we already have, we could try to find missing values related to the song information. In particular, there are many missing values in the composer, lyricist and language fields. We could try to find a website where the language of the song is accessible by looking at the combination of name and isrc code, since in most of the cases is available.

It would be interesting as well to have the total number of times a particular song has been listened by the user. Currently we only have data for the first time the user had listened to a

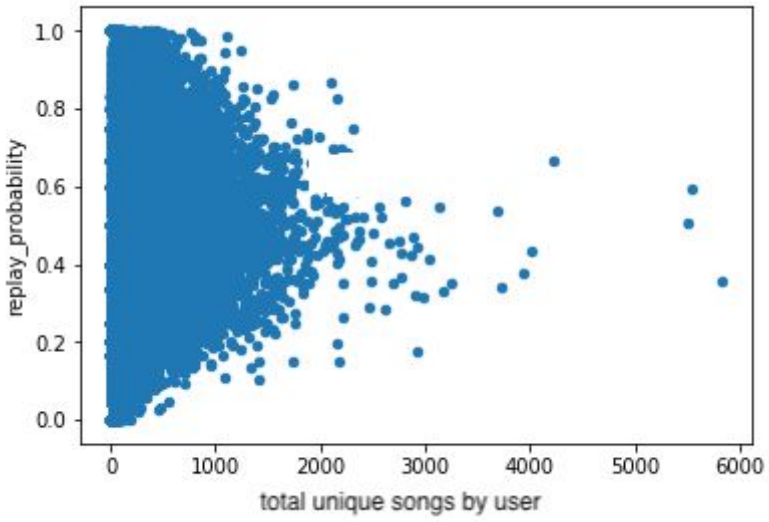
particular song and whether that user has listened that song again in a period of time or not. Because we are missing the total number of times that song was replayed, we can't identify popularity among the songs which definitely would improve our predictions.

Initial findings

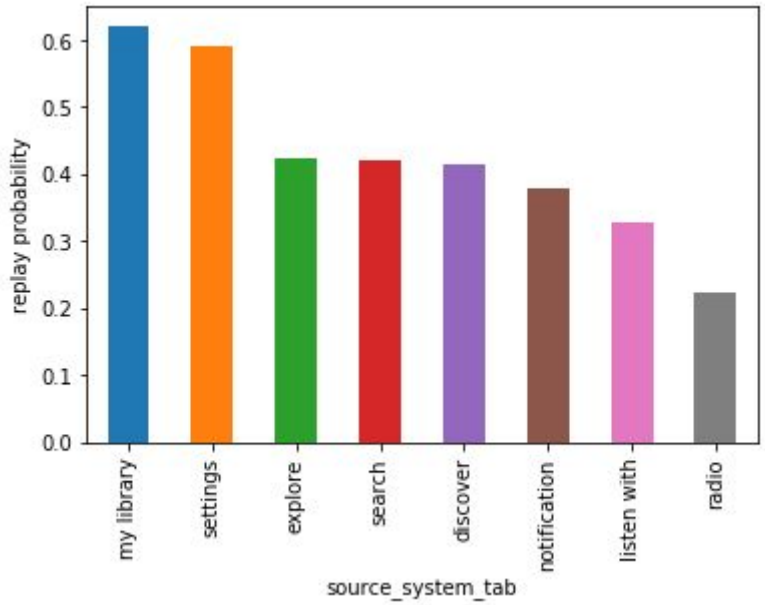
Our initial analysis was focused on trying to identify factors that affect the probability of a song to be played again within a month after the user's very first observable listening event. Many times we will refer to the 'replay probability' for convenience. We can divide the factors that we analysed in three categories: usability, user properties and song properties.

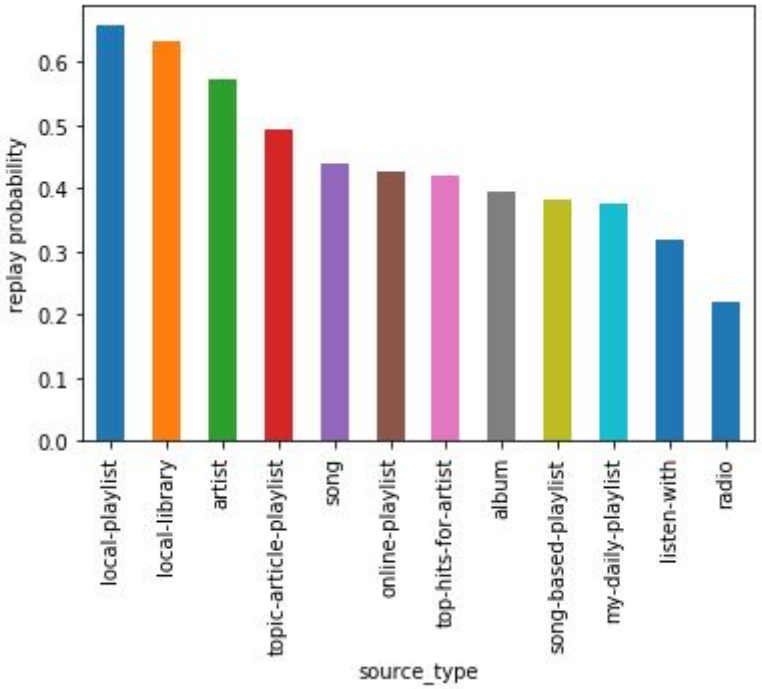
Usability

You could expect that highly engaged users are more likely to replay a song. In order to see if there was any correlation in this case, we have first considered that a user is more engaged if he has listened to more unique songs (in the dataset we don't have how many times each song was listened). From the plot below, we could not found visual evidence of any correlation between the engagement and the replay probability.



The more interesting findings were related to the screen/tab from which the song is accessed from. Songs listened from 'My library' tab, which happens to be the most popular one, and source types 'local-library' and 'local-playlist' seems to have the highest chances to be replayed. On the other hand, songs listened from radio tab have the lowest chances to be replayed.





After running a hypothesis test, we demonstrated that ‘My library’ tab has higher probability than the rest of the tabs. We obtained a p-value of 1. Since the p-value is really high, we rejected the null hypothesis that assumes the probabilities for ‘My library’ tab and the rest are the same. The minimum difference observed in the test between 'My library' tab probability and the others is around 23.2%.

We conclude that the probability of replaying a song when it is listened from 'My library' is higher than when it is listened from other tabs.

We wanted to analyzed all tabs individually to see if we could spot any other interesting differences.

For that purpose, we computed the probability of a song to be listened again depending on the tab the song was listened from. We calculated 95% confidence intervals of the probabilities. Here are the results:

- Explore tab: [0.420, 0.425]
- My library tab: [0.619, 0.620]
- Search tab: [0.420, 0.423]
- Discover tab: [0.415, 0.416]
- Radio tab: [0.221, 0.224]
- Listen with tab: [0.325, 0.329]
- Notification tab: [0.366, 0.390]
- Setting tab: [0.570, 0.611]

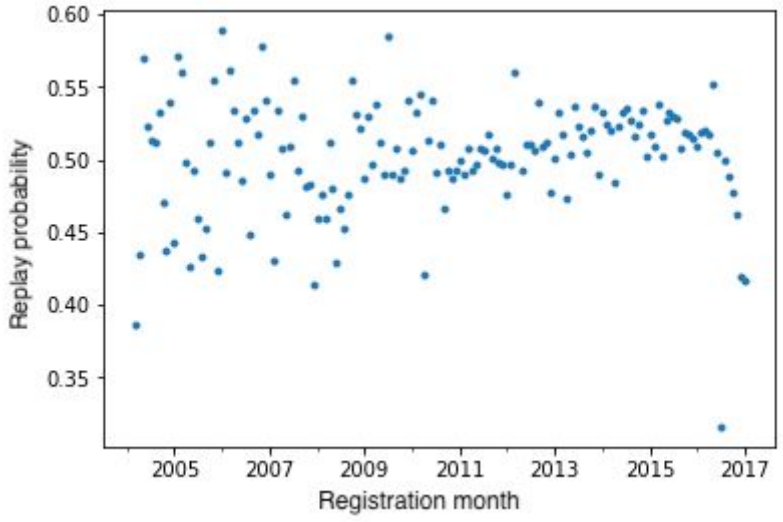
From the previous results, we can conclude that there are clear differences in probability depending where the song was launched from. The best tabs are ‘My library’ and ‘Settings’ and the worse one is ‘Radio’.

It makes sense that users visit frequently playlists and songs that are saved in their library. The company definitely should focus on teaching users how to build their playlist or save the songs in the library.

It is also logical that songs from the radio tab are not listened that frequent, because it is not the user who chooses what song is going to be played.

User properties

It could be expected that users who have been using the service more, will have higher chances to replay songs. We did not find any visual correlation between the registration date of the users and their probability to replay songs.



We were as well interested to check if there were differences between males and females. We did a hypothesis test where the null hypothesis assumes the mean replay probability in men and women is the same.

The observed probability in men is 48.94% and in women is 47.66%.

We obtained a p-value of 0.9998. At a confidence level of 95%, provided that this is a two tailed test, the null hypothesis would be rejected. The p-value is higher than 97.5%. We conclude that the difference is statistically significant and men have a higher probability of replaying songs.

Finally, we wanted to analyze if younger users were replaying songs more often than older users. We run another hypothesis test to check it. In this case, the null hypothesis states that there is no correlation between the user's age and the replay probability.

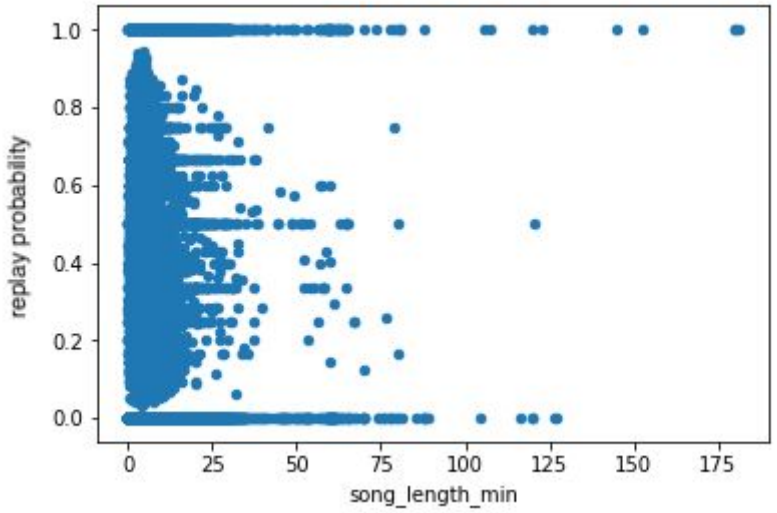
The observed Pearson correlation between the variables is -0.09.

We obtained a p-value of 1. Since the p_value is really high, we rejected the null hypothesis in favor of the alternative one. There is a week negative correlation between the users' age and the mean probability of a user to replay songs. Overall the probability or replaying songs decreases as users get older.

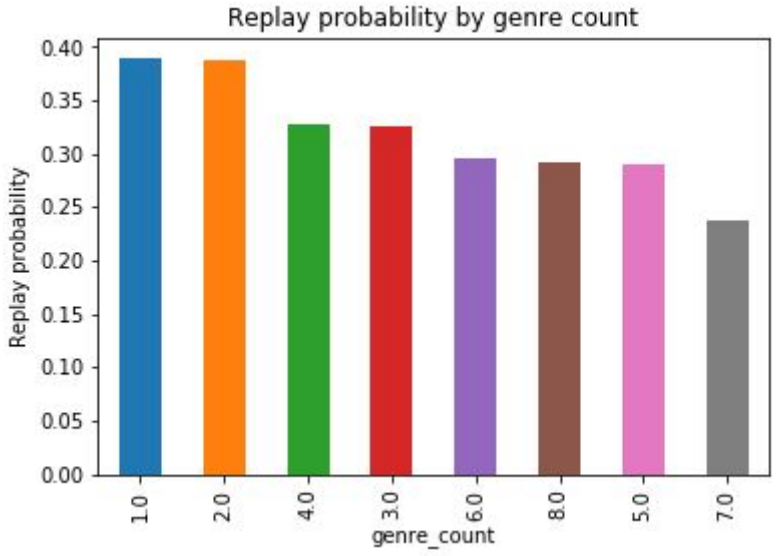
Song properties

We have analysed as well some characteristics of the songs such as language, number of different genres that belong to or duration.

We have not found any visual correlation between duration and replay probability.



In the plot below, it is represented the mean replayed probability depending on the number of genres a song belongs to.



Although songs that are belonging to 1 or 2 genres seem to have higher probability, there is not a clear negative correlation between the number of genres and the probability.

In order to confirm if there is any correlation, we did a hypothesis test. Our null hypothesis assumed that there is no correlation between the number of genres a song belongs to and the replay probability. We obtained a p-value of 1 and rejected the null hypothesis in favor of the alternative one. There is a negative correlation between the number of genres a song belongs to and the probability of that song to be replayed.

Nonetheless the observed correlation (-0.021) is very small and therefore is not useful for practical purposes.

We did not observe any interesting patterns regarding song language. Most of the language belongs to language with tag 52 or the language is unknown.