KKBox's **Music Recommendation** Analysis

Ana Villalba

November, 2018

The problem

KKBox is leading the music streaming industry in Asia. Its wide Asian-Pop music library contains more than 30 million tracks.

The service includes a **recommendation system** that needs 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.

Improving the recommendation system can help **improving retention and increasing monetization**. KKBox offers subscriptions with trials, so giving users a really good first experience will lead to a high conversion after the trial.

The problem

In order to improve the recommendation system, we will focus on answering the following question:

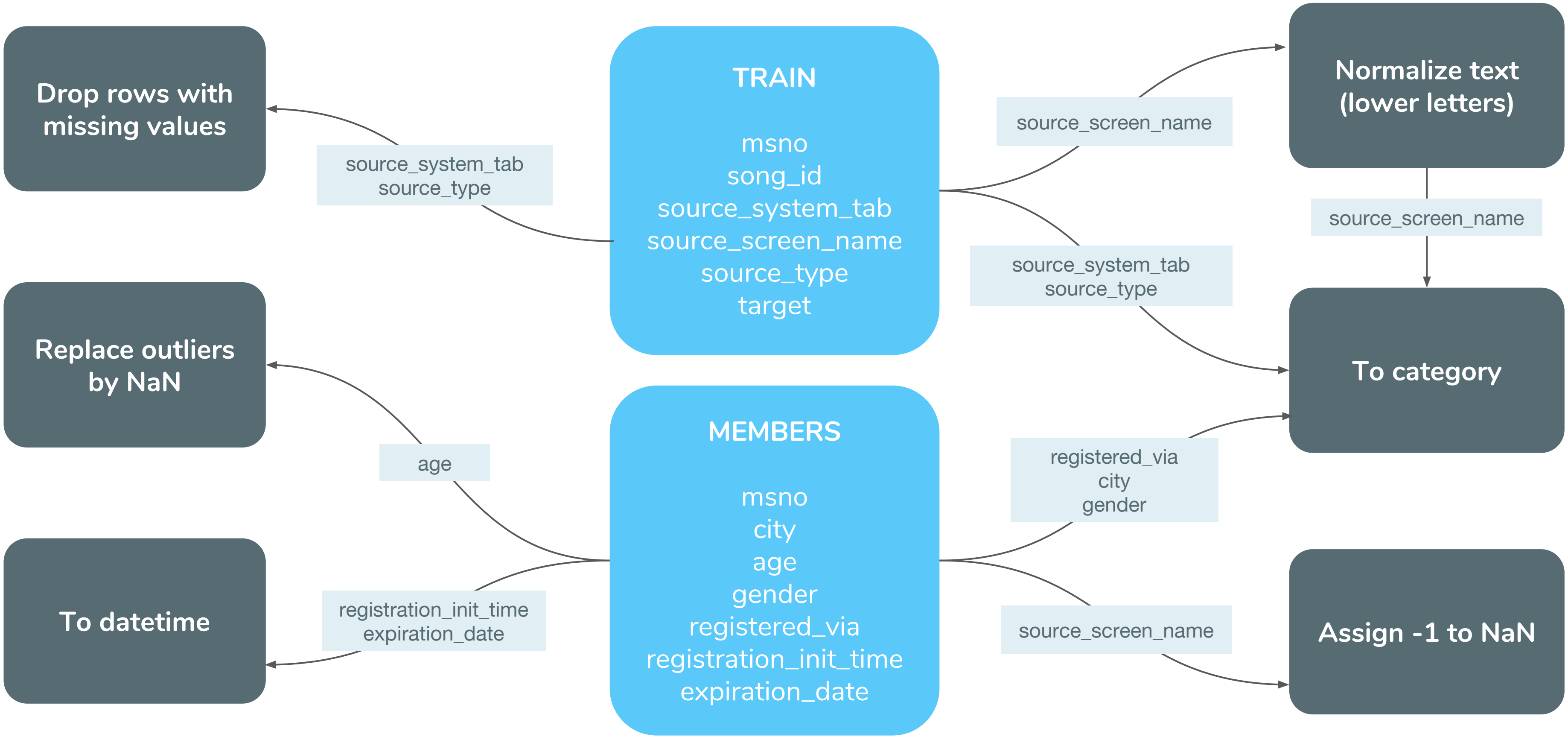
Will a user listen to a song again in less than a month after the first time listening to it?

The dataset

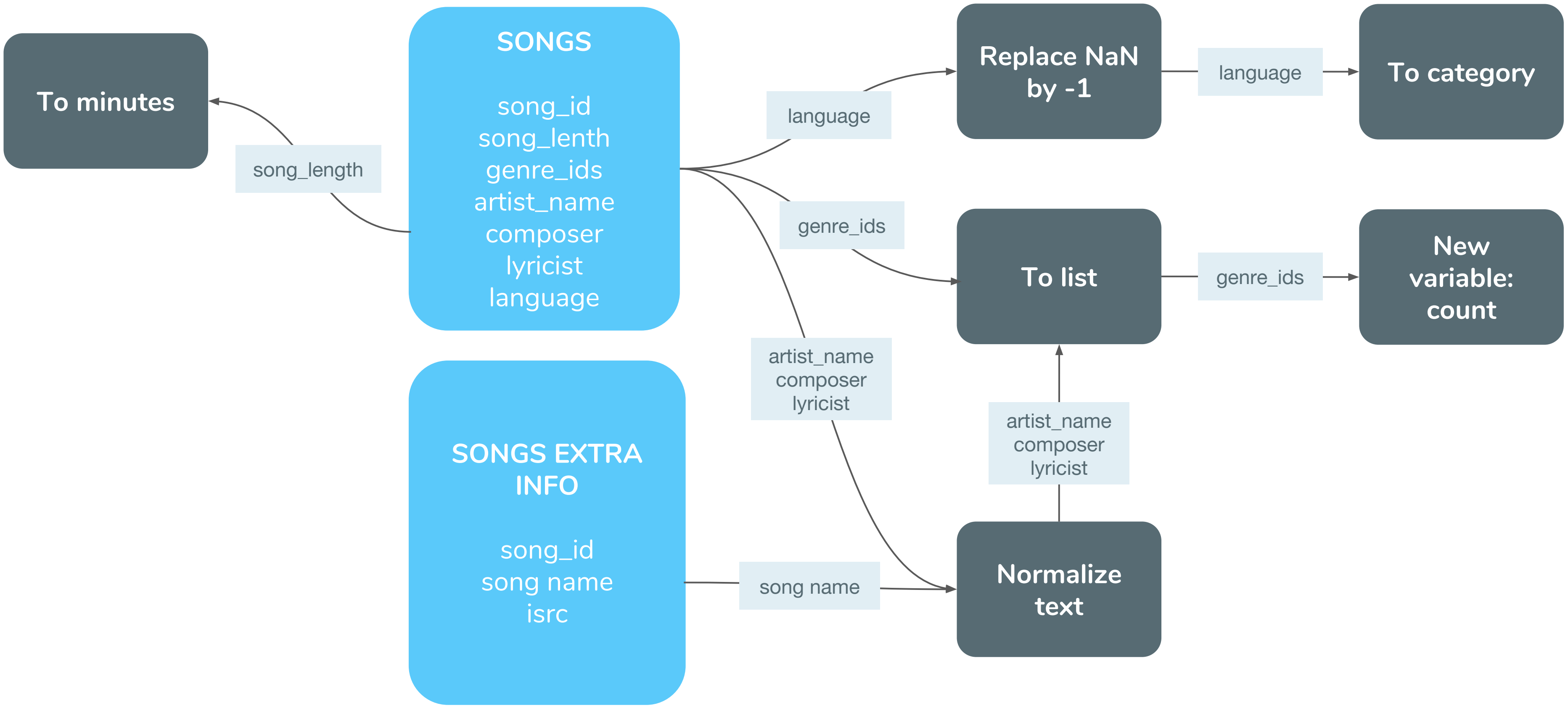
Data from the WSDM - KKBox's Music Recommendation Challenge was used.

- Over **7 million listening events** (each event is the first one performed for a user-song pair)
- More than **30,000 users**
- More than **2 millions songs**
- **4 data tables** (train, members, songs, song extra info)

Data wrangling steps

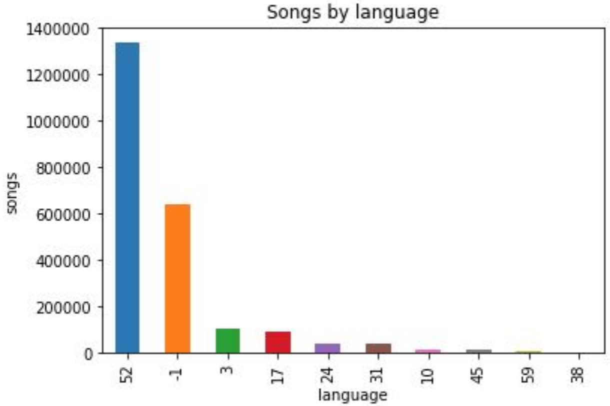


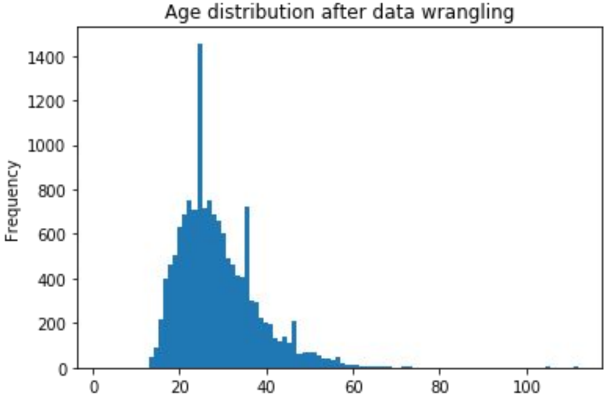
Data wrangling steps

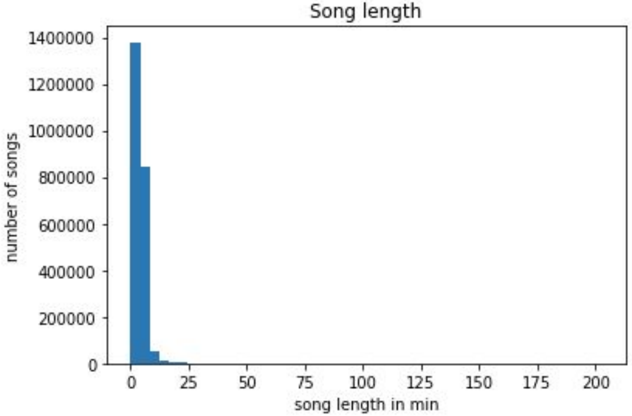


Data wrangling steps

Some features after transformation.



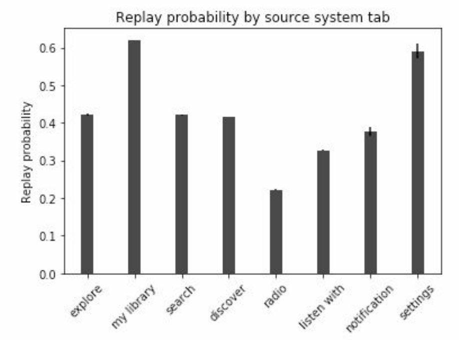


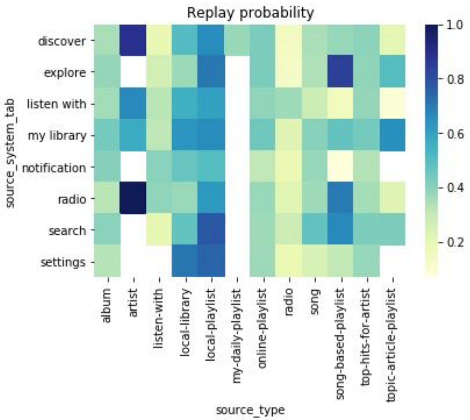


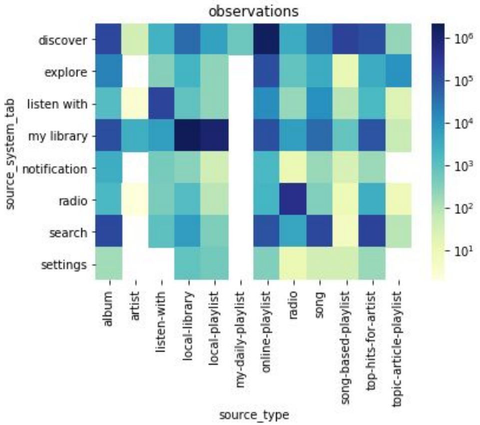
Exploratory analysis

Usability:

- Songs listened from source system tab **'my library'** and **'settings'** have the highest replay probability. Settings tab is not very relevant since there are few observations.
- Songs launched from tab **'radio'** have the lowest replay probability.
- Songs listened from **'local-playlist'** and **'local-library'** source types have higher probabilities to be replayed than other types.



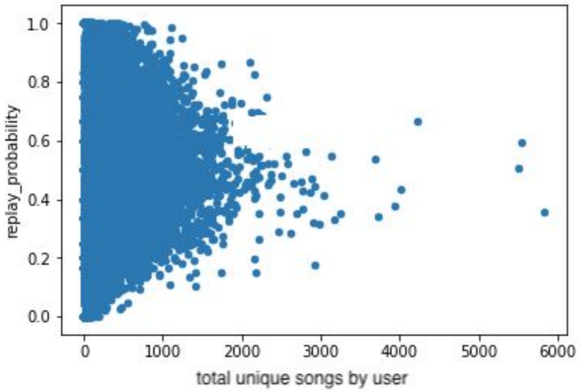


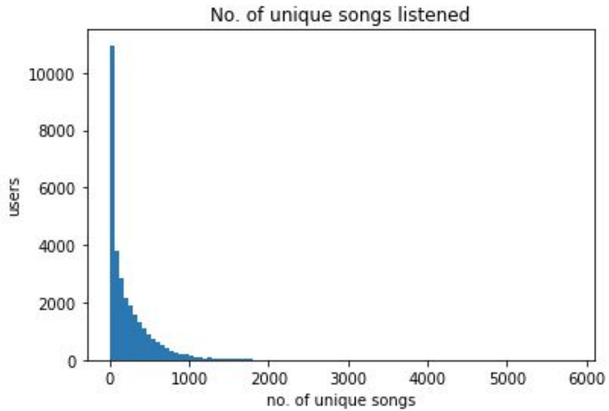


Exploratory analysis

Usability:

- No visual correlation between user **engagement** and replay probability.
- Engagement means that a user **listens more unique songs**.
- There is a big amount of users who have listened few unique songs.

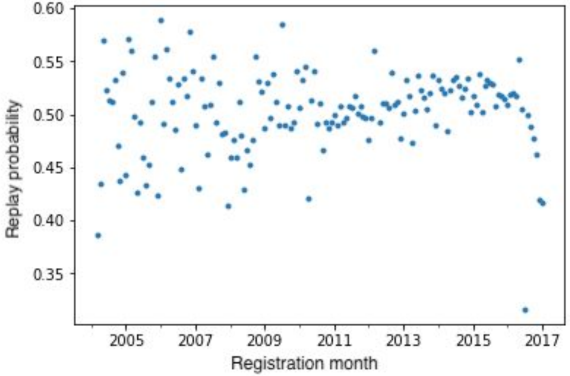


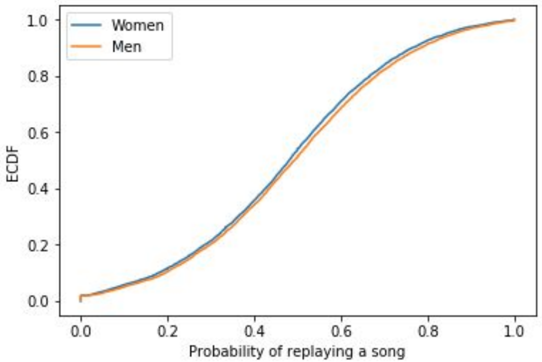


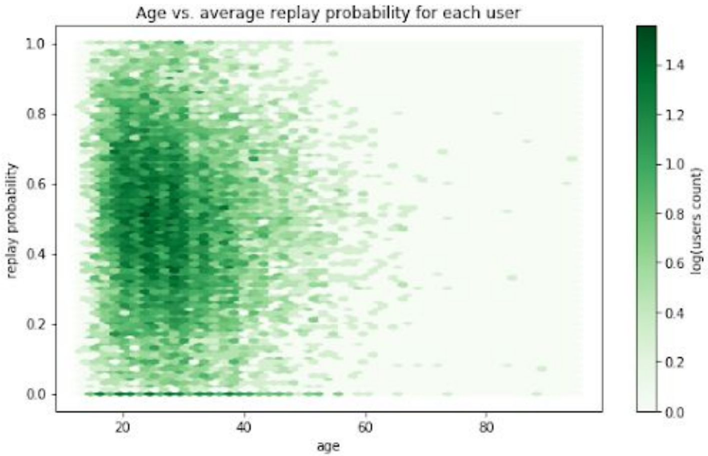
Exploratory analysis

User properties:

- No visual correlation between user **registration date** and user probability to replay songs.
- Slightly higher replay probability in **men** (48.9%) than in **woman** (47.7%).
- Small negative correlation between replay probability and user **age** (Pearson correlation: -0.09).



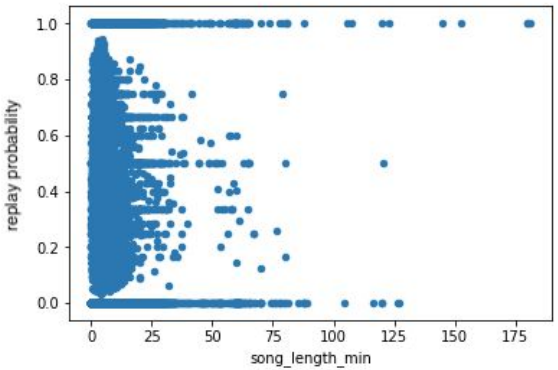


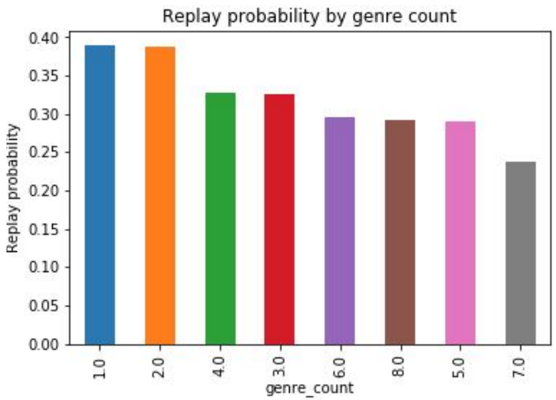


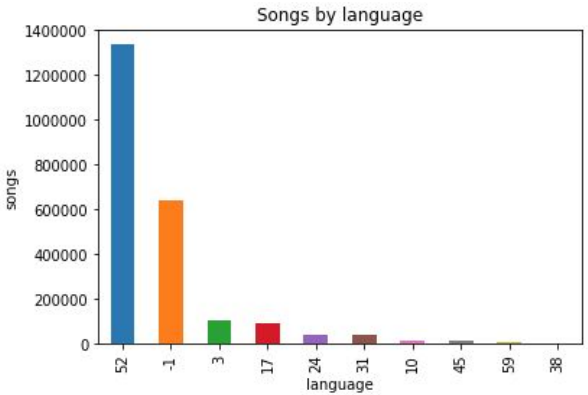
Exploratory analysis

Song properties:

- No visual correlation between **song duration** and replay probability.
- Small negative correlation between replay probability and **number of genres** a song belongs to (Pearson correlation: -0.02).
- Most of the **songs** belong to **language** 52. No interesting patterns were found regarding replay probability.







Machine learning: Models

1 Logistic Regression

The output is a function of the different features which are weighted by **coefficients.** This method is suitable for large datasets.

2 Linear SVC

Support vector machine that uses linear kernel. The **linear kernel** usually scales better for large number of samples than other kernels.

3 Decision Tree

Determines which features have the most **importance** to the decision. It performs really good in large datasets and the results are **interpretable**.

4 Random Forest

Creates **several** fully grown **decision trees** selecting a number of features and returns the best performing one. It usually outperforms decision tree.

Machine learning: Approach

TRAIN	TRAIN + MEMBERS	TRAIN + MEMBERS + SONGS	DIMENSION REDUCTION (PCA)
36 features	67 features	164 features	93 features
Only use **usability features** included in the train dataset, describing the tab or screen the song was listened from.	**Member features** such as age, city or registration year are added to the usability features.	On top of the previous features, also **song properties** such as duration or genres it belongs to are included.	Using usability, member and song features we apply dimension reduction using PCA.

Note: The analysis was performed in a subset of the data to reduce computational complexity. Events corresponding to users who have listened between 20 and 30 unique songs were selected. This increases the chances of users appearing both in the training and test data.

Machine learning: Results

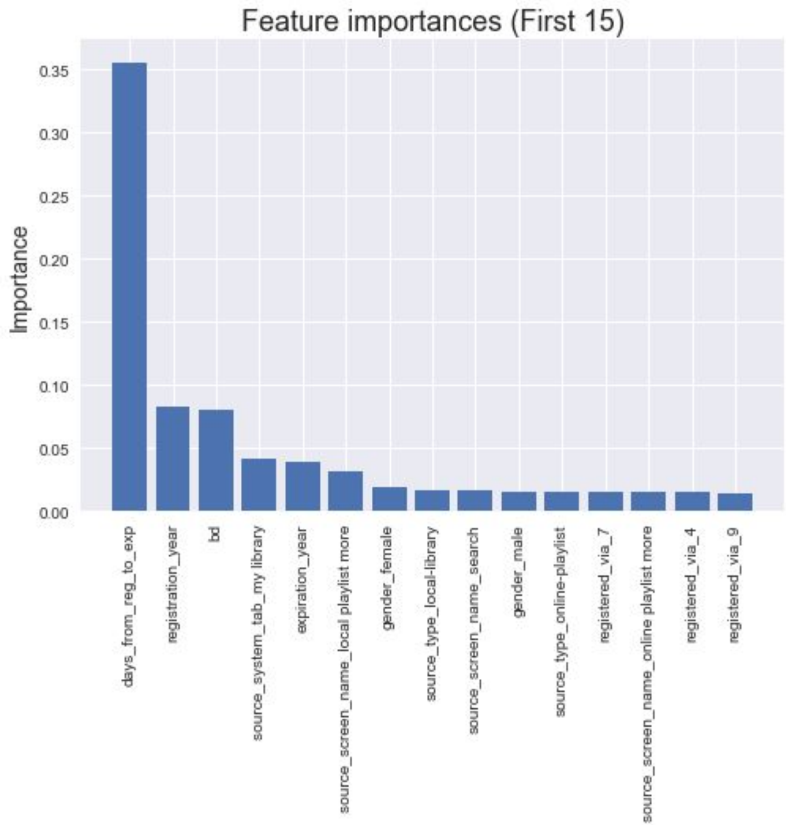
	Logistic Regression	Linear SVC	Decision Tree	Random Forest
Train	0.6644 0.6448	0.6645 0.6456	0.6674 0.6541	0.6670 0.6543
Train + members	0.6712 0.6595	0.6684 0.6589	0.7605 0.7880	**0.7618** **0.8021**
Train + members + songs	-	-	0.7238 0.7072	0.7430 0.7707
Dimension reduction (PCA)	-	-	-	0.7305 0.7588

Accuracy score   ROC-AUC score

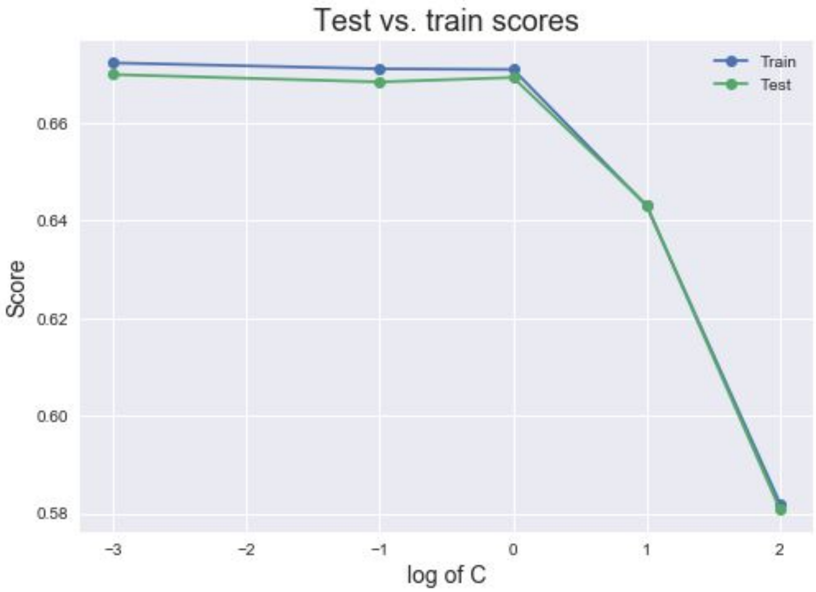
Machine learning: Results

Most important features

- User properties such as **days from registration to expiration**, registration year, age, expiration year or gender.
- Usability features related to **my library** and **playlists** such as source system tab my library, source screen local playlist more or online playlist more or source type local library or online playlist



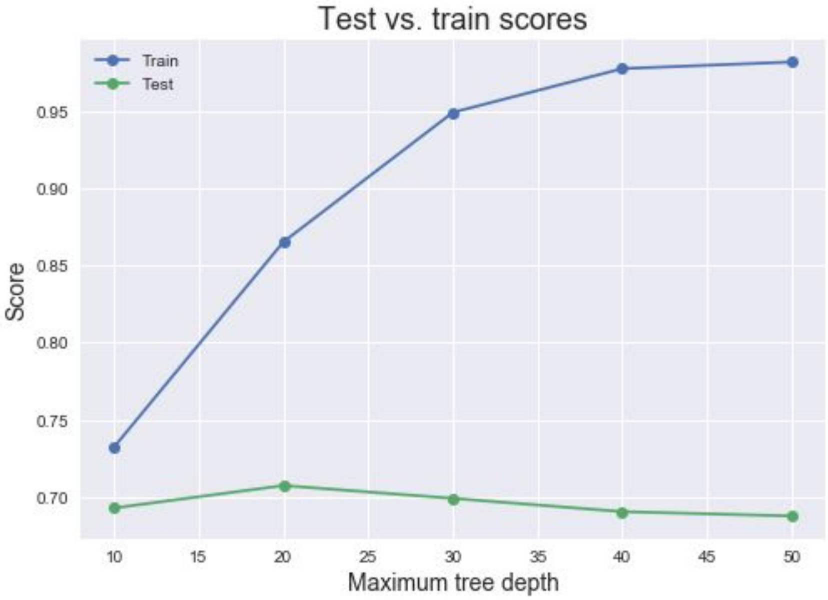
Machine learning: Hyperparameter tuning



Linear SVC. Train + Members

67 features

When we only select usability and member features there are no overfitting problems, since train and test scores are very similar. In this case **lower Cs performs better.**



Decision Tree. Train + Members + Songs

164 features

Introducing song features causes **overfitting** problems. In this tree decision model the best maximum depth is 20, after that train score increases until almost 100%.

Conclusions

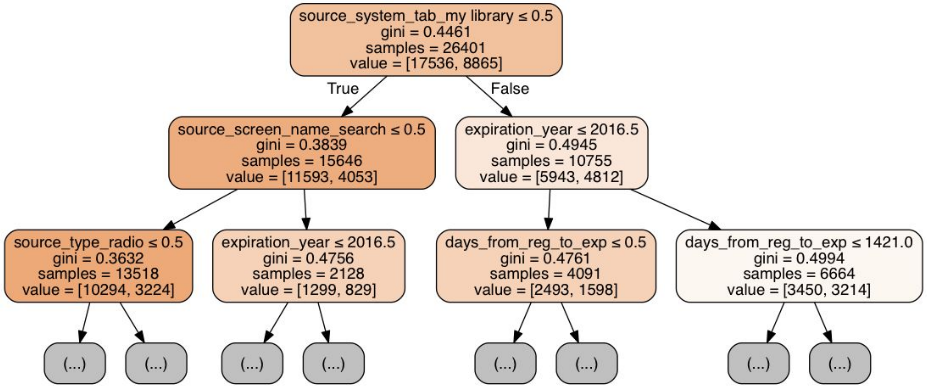
Consistent results

- Source system tab **'my library'** and source types **'local-playlist'** and **'local-library'** showed the highest replay probabilities in the **exploratory analysis.** Source system tab **'my radio'**, the lowest.

- It was proven with a **hypothesis test** that the source system tab 'my library' has higher replay probability than other tabs.

- In the best **decision tree**, source system tab 'my library' was the first decision feature. Source type 'radio' is among the most relevant features.

- Tab 'my library', source types 'local-playlist' and 'local-library' appear among the most relevant features in the **random forest** classifier.



Client Recommendations

1. ## Improve member data collection.

   Only around 42% of users have **gender** and **age** values. Member features are the most relevant for the model performance. In the best classifier (random forest including member and usability features) the age was the 3rd most important feature and gender_female and gender_male, the 7th and 10th.

2. ## Teach users how to add songs to ‘My Library’ and use playlists

   First user experience should make a big effort on teaching users how to add songs to ‘My library’ and use playlists. Some options are adding a tutorial in the onboarding, prompt users to add first listened songs to their playlists or make more visible icons to add songs.

3. ## Make less prominent the ‘Radio’ tab

   Radio tab can be included inside another tab, since it is not that popular and also it is less likely that users will listen again these songs. E.g. Spotify mobile app only has 3 tabs (‘home’, ‘search’ and ‘my library’) and radio option is inside ‘my library’.

Next steps

- Scale to bigger training datasets the analysis
  Does the performance of the model improves?
- Consider song popularity
  Include every listening event for each song and not only the first one, or the total number of times a song was listened by each user. This way song popularity can be measured which would help improve predictions.
- Try more ensemble algorithms
  Do boosting algorithms such as XGboost or LightGMB or voting ensemble perform better?