KAGGLE-SPOTIFY DATA

Jan Markus Rokka, Urmi Tari, Reena Seeba

INTRODUCTION

For the purposes of this project we are using the dataset “🎹Spotify Tracks Dataset” from Kaggle. This dataset consists of a large list of songs from the popular streaming platform Spotify, along with 20 parameters describing each track. Our goals were:

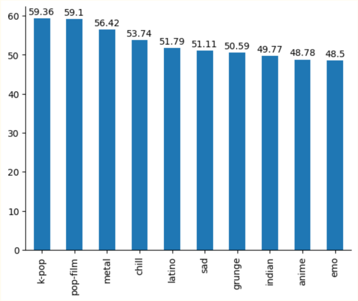
- Build a model to predict danceability based on the song’s other parameters.
- Find the most and least popular genres based on averages.
- Find the parameters which affect the track’s popularity the most.

METHODOLOGY

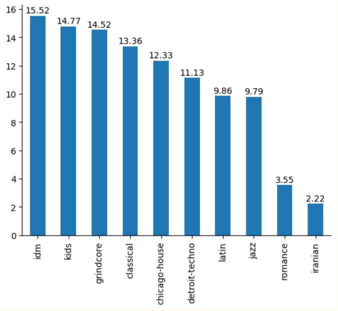
- To predict the danceability of a track we trained a random forest regression model to predict a track's danceability based on it's other parameters.
- To find the most and least popular genres we calculated the mean popularity values of all the different genres and sorted them. (Each of the tracks was already given a popularity score between 0-100)
- To find out which parameters affect a track's popularity the most we found the correlation values between a track's popularity and each of it's other parameters. The results were then sorted. We also repeated the process after having modified the data to only contain tracks from a particular genre.

Finding the most and least popular genres

Most popilar genres based on popularity

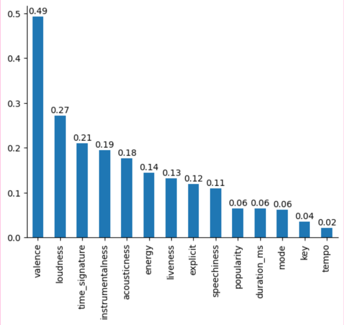


Least popular genres based on popularity

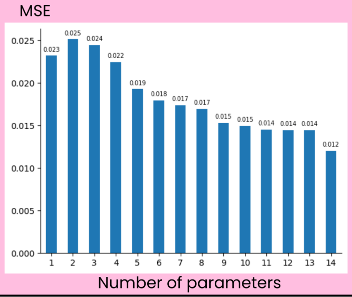


Predicting danceability

All other parameters ranked in terms their correlation to danceability.

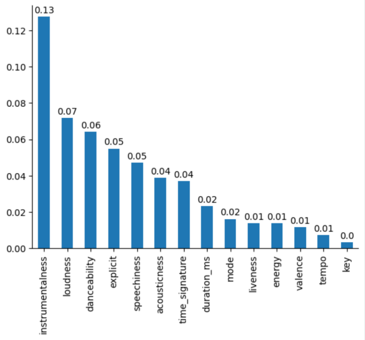


The model's mean squared error (high = bad) based on how many parameters it was given. The parameters were given in order of the most correlated.

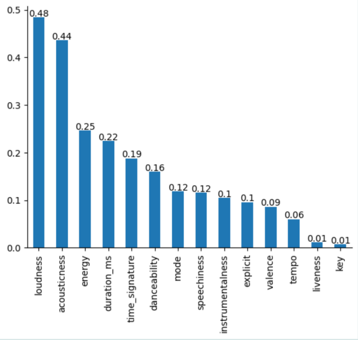


The secret of popularity

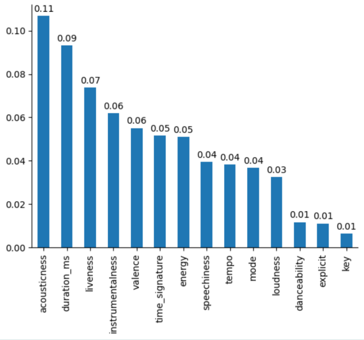
Correlation values between popularity and every other (potentially relevant) attribute.



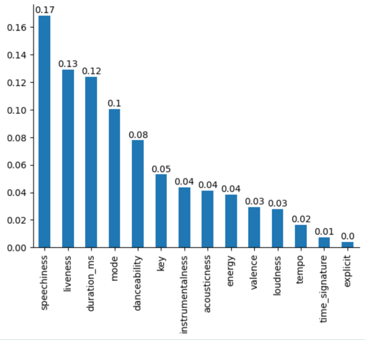
Parameters most correlated with the popularity of k-pop.



Parameters most correlated with the popularity of kids music.



Parameters most correlated with the popularity of sad music.



Overall, we didn't find strong correlations between popularity and the other parameters. However, when comparing the results between different genres we discovered that the correlations with popularity vary quite a lot. These results could also have a more immediate practical use.