SPOTIFY PREDICTIONS

Birgit Sõrmus
Kaspar Kadalipp

Business understanding

Identifying business goals

Our project is about predicting the popularity of songs based on their attributes given on Spotify. This topic was chosen because it is an interesting topic to research especially as music plays a huge part in a lot of people's lives, including ours.

The goal of our project is to analyze the data in different Spotify datasets we've gathered and determine what are the attributes given on Spotify like for songs that have been popular among listeners in the years 2017 and 2018 and compare whether the attributes were similar or different in those years. Based on the information we will get from the first step in our analysis, our next goal is to predict whether songs gotten from a different datasets are popular or not.

Assessing your situation

For this project we have two computer science students to do the project and various datasets, including two datasets that each have a top 100 most popular songs of the year (2017 and 2018) and datasets of a lot of different songs, all of which include information about the attributes related to each song given on Spotify.

The project must be done by December 16th along with a poster to present at the poster session. Each student must put in at least 30 hours worth of work.

There aren't many risks related to the project as the data is available to all on Kaggle and the code will be added to github regularly so even in case of a computer malfunction or something similar, not much of the work would be lost.

The same applies to terminology, meaning there isn't much to explain as the topic is easily understandable to most people. To clarify however, attributes of each song mean the characteristics of it, aka attributes like danceability, energy, acousticness etc. The audio features for each song were extracted using the Spotify Web API and the spotipy Python library. Credit goes to Spotify for calculating the audio feature values.

The data is provided for free on Kaggle and the software we will be using for this project is also free, so there will be no related costs. The benefits are mostly insights and knowledge received.

Defining your data-mining goals

The goals are to take the datasets we've gathered for this project, figure out the attributes to use and train a model that could predict the popularity of a song based on its attributes.

Data understanding

Gathering data

The data we will need for the project will have to include different attributes describing songs and some way to verify popular songs. The datasets we've gathered have that, as they all include lists of attributes describing each song. They also offer a way to see which songs are popular, as we have gathered two datasets that were the top most listened to songs of the year. Those two datasets will be the ones we'll use to train the model.

We also have another dataset with 176 774 different songs from 14 564 different artists that can be used to predict the popularity.

Describing data

Below is a table describing the attributes that the datasets include about each song.

song_title	categorical	Name of the song
artist	categorical	Artist(s) of the song
duration_ms	numeric	The duration of the track in milliseconds.
key	numeric / categorical	The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.

mode	numeric	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
time_signature	numeric	An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
acousticness	numeric	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
danceability	numeric	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
energy	numeric	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
instrumentalness	numeric	Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
liveness	numeric	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

loudness	numeric	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
speechiness	numeric	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
valence	numeric	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
tempo	numeric	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

Exploring data and verifying quality

Overall the datasets seem good to use for the project. The large dataset of songs to use later on to predict popularity had quite a lot of duplicates initially, but since the dataset was very large, we are still left with well over 100 000 songs to use which is well over the amount that we'll need. There seem to be no missing values in the datasets.

Planning your project

Project related tasks

Task 1. Finding relevant datasets and gaining initial understanding of the data that we'll use. Most of the data search was done on Kaggle which is where we ended up getting our datasets from as well. Initial data exploration was done using Excel before loading the data on Jupyter Notebook. (4 hours per person)

Task 2. Creating the slide for the project pitch and presenting it in class. (1 hour per person. This includes the discussion of different possible topics and coming up with the one that we are actually going to use)

Task 3. Doing homework 10 aka writing the report on business and data understanding and creating the project plan. (3 hours per person)

Task 4. Preprocessing the data, further improving our understanding of it and deciding on inital attributes to use for training. (2 hours per person)

Task 5. Testing out different models with different parameters and different data attributes. For this we will use different classifiers from sklearn and see which ones give the best results. (12 hours per person)

Task 6. Analyzing the results, meaning analyzing the predictions of popularity, comparing the attributes of popular songs from 2017 and 2018 and possibly creating graphs where possible. (8 hours per person)

Task 7. Creating the poster for the poster session. This will use the analysis from task 4. (3 hours per person)

All of the tasks altogether take at least 30 hours per person. More precise work divisions will develop during the project.