Classifying Songs in Spotify Playlists

Capstone Project for IBM Data Science Professional Certificate Specialization

by Tauno Tanilas - 2019

Introduction

Business problems:

- Understand the user’s needs and preferences in evaluating musical tracks
- Knowing these characteristics, provide better services

Main Goals:

- Determine characteristics that define the user's musical taste
- Using these characteristics compare the music user likes or dislikes
- Create a predictive model on whether user likes or dislikes a song
- Determine the main musical genres that are in user’s preferences
- Using Foursquare API, construct a map of accommodation options in one of the selected concert places.

Target audience:

- Audio streaming providers
- Any person willing to get insights about using classification modeling in solving machine learning problems

Data

- **Data Source** - Spotify platform
- 80 playlists in total - 45 liked, 25 disliked and 10 for evaluation
- **Data Acquiring** - Spotipy API as a lightweight Python library for the Spotify Web API
- 3620 songs in total after Data Cleaning - 1612 liked and 2008 disliked
- **10 audio features** that describe each track

Data

Audio features

- **Acousticness:** A measure from 0.0 to 1.0 of whether the track is acoustic.
- **Energy:** A measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
- **Danceability:** Describes how suitable a track is for dancing
- **Instrumentalness:** Predicts whether a track contains no vocals.
- **Liveness:** Detects the presence of an audience in the recording.
- **Loudness:** The overall loudness of a track in decibels.
- **Speechiness:** Detects the presence of spoken words in a track.
- **Tempo:** The overall estimated tempo of a track in beats per minute.
- **Valence:** A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.
- **Duration:** The duration of the track in milliseconds.

Exploratory Data Analysis

Most important audio features that determine my musical taste:

- **Acoustic:** I don’t like songs that are not acoustic at all.
- **Dance:** I prefer songs that are moderately danceable.
- **Energy, Loud, Tempo:** I prefer songs that are less energetic, fast and loud.
- **Valence:** I don't like songs that have little valence but the distribution of values in valence rate is quite equal.

Feature Selection

Independent data features:

- Acoustic
- Dance
- Duration
- Energy
- Instrumental
- Live
- Loud
- Speech
- Tempo
- Valence

Dependent data feature:

- Preference (’GOOD’, ’BAD’)

Data Splitting and Normalization

- 2896 rows in Train set (80%)
- 724 rows in Test set (20%)
- sklearn.preprocessing.StandardScaler package for Data Normalization

Prediction Metrics used in modeling

Three well known prediction metrics:

- **Accuracy** - the proportion of total number of predictions that were correct
- **Precision** - the proportion of positive cases that were correctly identified
- **Recall** - the proportion of actual positive cases that were correctly identified

Algorithms used in modeling

Three well known classification algorithms:

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machines

Model Optimization

Used Hyperparameters:

- 'C': (0.001, 0.01, 0.1, 1, 10)
- 'kernel': ('linear', 'poly', 'rbf', 'sigmoid')
- 'class_weight': ('balanced', None)
- 'gamma': ('scale', 'auto')
- 'shrinking': (True, False)

Prediction accuracy after GridSearch didn’t improve compared to default parameters.

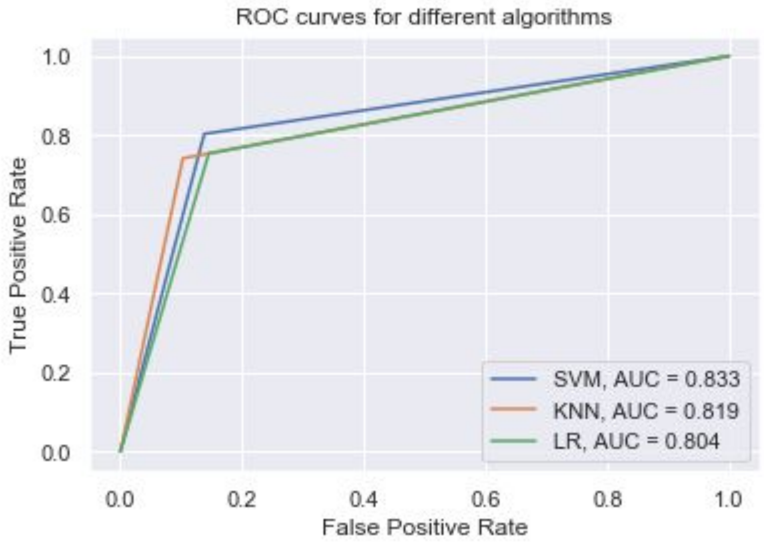
Results

Support Vector Machines as a best classifier:

	ACCURACY	PRECISION	RECALL
Support Vector Machines	0.835635	0.834543	0.832616
K-Nearest Neighbors	0.827348	0.832282	0.819391
Logistic Regression	0.809392	0.809279	0.804241

Results

Best AUC by SVM 83.3%, indicating a good level of prediction accuracy.



Results

- 10 playlists with different genres for **evaluation**
- One random song from each of them
- Accuracy prediction (**80,0%**) was proportional to the train/test part

Discussion

Possible suggestions for further developments:

- Increase the data set
- Try different proportions between liked and disliked playlists
- Try different normalization methods
- Try different modeling algorithms

Conclusion

- Most important features that determine my musical taste are Acousticness, Danceability, Energy and Valence.
- I prefer Country, Indie, Jazz, Pop, Rock, Soul and Afro music but Electronic, Hip-Hop and Metal are not so much in my favor.
- The best classifier was Support Vector Machines and the model optimization didn’t give any better results.
- Always normalize data before modeling. The difference in prediction accuracy with normalized and not normalized data was depending on the algorithm up to a twenty percentage point.
- Despite the quite small dataset the prediction model I created achieved a good level of prediction accuracy.