#### Springboard Capstone Project

#### **Benson yang**

#### **Bensonyang0326@gmail.com**

## Springboard Capstone 1

## Classifying spotify’s songs type

Benson yang

Introduction

Founded in 2006, Spotify has quickly risen to become the top music streaming service in the world, with an estimated [489 million monthly active users](https://www.businessofapps.com/data/spotify-statistics/) at the end of 2022 and 11.72 billion euros of annual revenue in 2022. In addition, Spotify leads global music streaming service providers with a 32% market share.

We can say Spotify is a leader in the music streaming industry. Therefore, knowing what types of music are popular in a specified country based on Spotify data is important to businesses.

The Data

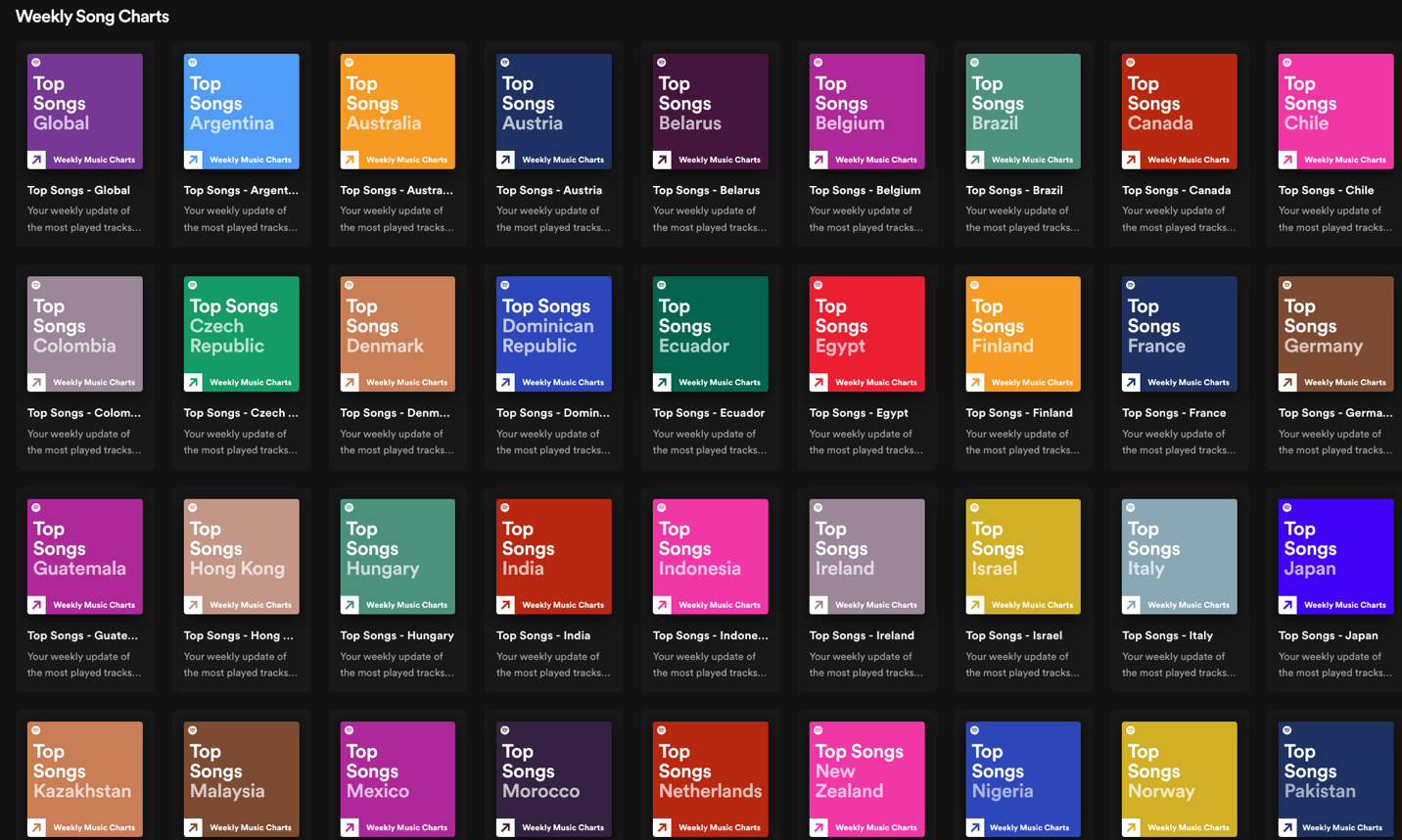
The dataset is collected from 5 popular types of music tracks and several countries’ weekly top 50 popular songs by the Spotify API and stored in the Google platform database.

This dataset features pre-analysis data from Spotify. The data set is a set of 18 features: 12 numerical and 6 categorical.

(information is based on Spotify Website.)

|  |  |
| --- | --- |
| acousticness | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
| analysis\_url | A URL to access the full audio analysis of this track. An access token is required to access this data. |
| danceability | 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. |
| duration\_ms | The duration of the track in milliseconds. |
| energy | 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. |
| id | The Spotify ID for the track. |
| instrumentalness | 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. |
| key | The key the track is in. Integers map to pitches using standard [Pitch Class notation](https://en.wikipedia.org/wiki/Pitch_class). E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1. |
| liveness | 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 | 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 typically range between -60 and 0 db. |
| mode | 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. |
| speechiness | 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. |
| tempo | 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. |
| time\_signature | An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4". |
| track\_href | A link to the Web API endpoint providing full details of the track. |
| type | The object type. |
| uri | The Spotify URI for the track. |
| valence | 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). |

The music type datasets has 27,894 entries, split into 3,788 entries, 8,512 entries, 3796 entries, 4350 entries, 7448 entries, which related on music types, and the testing datasets, which is top-50 popular song track, has 58 different dataset that related on the country.



Exploratory Data Analysis & Feature Cleaning

From the exploratory data analysis, we initially searched for any data-related issues.

First, we have to deal with the data returned from API requests.

Text

Description automatically generated

The spotify api returns only the song list information, not the attributes of each song, so we need to make the next request for each song in the list.

Text

Description automatically generated

Import to a pandas data frame, we got following column, and data.

Column:

### Text Description automatically generated

### **Observation:**

### 

### **Feature Cleaning :**

### we removed the **categorical** columns, which are

### **id,**

### **type,**

### **uri,**

### **track\_href,**

### **analysis\_url,**

### **time\_signature.**

### We leave this column out because those column are not quite useful for classifying, debugging and analysis.

### Text Description automatically generated

### In order to store data neatly from different types of music, we used MySQL database on **Google Cloud Platform** to store the dataset .

### **Table :**

### **Text Description automatically generated**

### **data:**

### 

### Before analyzing the dataset, we pulled back the data from the database and create a temp table, then removed the same observation because the same values would negatively affect the classifier.

### Additionally, add the music type related to which table it is from and removed the **classification** and **song\_name** columns.

### A picture containing text, scoreboard Description automatically generated

### Then clean up the missing values.

### 

### A picture containing table Description automatically generated

### **The visually difference of different music type:**

### PoP vs Chill : Chart, histogram Description automatically generatedPoP vs EDM : Chart, histogram Description automatically generated

### Pop vs Hip-Hop :

### Histogram Description automatically generatedPoP vs R&B : Chart, histogram Description automatically generated

### Chill vs EDM:

### **Chart, histogram Description automatically generated**

### Chill vs Hip-Hop:

### 

### Chill vs R&B:

### R&B vs EDM

### Chart, histogram Description automatically generated

### R&B vs Hip-Hop :Chart, histogram Description automatically generated

### EDM vs Hip-Hop :

### Chart, histogram Description automatically generated

### According to chart, those are the difference between each music type,

### In **pop** music, **acousticness** is more distributed in 0~0.2, **loudness** is more distributed in -10,

### In **Chill** music, **acousticness** is more distributed in0.8~1.0, **energy** is more distributed in 0~0.6 ,**valence** is more distributed in 0~0.4

### In **R&B** music, **energy** is more distributed in 0.4~0.6, **tempo** is more distributed in 80~120.

### In **EDM** music, **tempo** is more distributed in 100~150

### In **Hip-Hop** music, **speechness** is more distributed in 0.2~0.4

Chart, bar chart

Description automatically generated

Text

Description automatically generated

Taking a look at our feature importances, it is clear that the top 3 impactive contributors to the model accuracy is energy, instrumentalness and speechiness. Given these feature importances, we intend to reduce the number of features in our model to simplify our classification by removing the less important features. Using our feature importance chart. we will take off the lowest impactful feature: Key. In addition, after creating the simplififed model, we want to measure the effectiveness of our model by using cross validation.

Creating a Simple Model :

### In order for us to create a simplified model, we can use our feature importances, and only choose the top 9 most impactful features. These features are **acousticness, danceability, energy, instrumentalnessloudness, speechiness, temp, valence, duration\_ms**. In addition, in order to make sure that our model can be more accurate, we use Bagging (Bootstrap aggregating) to do Integrated Learning and make accuracy higher and measure the accuracy.

### We used F1-score to measure the accuracy. It combines Precision and Recall information to evaluate the model's performance more comprehensively. When the difference between Precision and Recall is significant, F1-score will tend to be smaller; when the Precision and Recall are equal, F1-score will reach the maximum value of 1.

### If just using normal decision tree, we only can have 55% accuracy.

Calendar

Description automatically generated

On the other hand, after using Decision Tree Bagging, bring every new data into each Decision Tree model for prediction, and average or vote the predictions from multiple models. The accuracy and precision increased significantly.

### 

Conclusion

Our model seems to be much more accurate than guessing. From the results of the decision tree analysis, we are able to achieve approximately 64% accuracy. If we put new music data into this model, it can predict the music type with 64% accuracy. Then we can use this model to measure each country's recently popular music types by importing the popular music data into this model.

Result :

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

We can said that the majority music type in America, Australia and Asia is Pop

music, the majority music type in Europe is EDM music and the majority music type in

Africa is R&B music.

Future Works

To further improve the model, we would need to diverge from the reliance on the Spotify track features. At this point, a majority of our model accuracy comes from this number. In addition, our dataset is imperfect, and despite covering the most popular type of music and training it on music track features, it would be in our best interest to bolster our dataset with examples that analyze the music track's pitches and timbre.

This model is a minimum viable product. In order to use the model, we have to specify a data file to read and make predictions. In most popular music type of country analysis situations, you would use this model to observe what type of music would be more appropriate to enter a specific country market in order to lift the conversion rate. By quantifying the likelihood of conversion, music brokerage companies or indie bands can optimize the success of their marketing Strategies.