Université Jean Monnet

Data Mining and Knowledge Discovery

Project Report

A deep insight into the different song features.

Author:
Rohil Gupta

Supervisor:
FabriceMuhlenbach

March 17, 2018

Contents

List of Figures

1 Introduction

This project dives deep into the relation between the various latent attributes of the music answering the psychological factors behind peoples' different music preferences. Music is heard by people daily in many parts of the world, and affects people in various ways from emotion regulation to cognitive development, along with providing a means for self-expression. Music training has been shown to help improve intellectual development and ability.

There was a need at start of this project to find a dataset giving out various audio features of particular track. After the lot of research over Internet to find the reliable source, it was found that Spotify which is a music, podcast, and video streaming service that was officially launched on 7 October 2008. It was developed by Spotify AB in Stockholm, Sweden. It provides DRM–protected content from record labels and media companies. Spotify is a fermium service; basic features are free with advertisements or limitations, while additional features, such as improved streaming quality and music downloads, are offered via paid subscriptions.

Spotify Web Api provides various audio features of each track which is provides in the hyper link. Observing the diversity of audio features provided by the Spotify a huge potential was seen to analyze further to find the complex relation between them.

The other challenge was to accumulate a large diverse dataset by bulk calling the Spotify API to provide the audio features for this data mining project. The R package **Spotifr** which is spotifyr is a quick and easy wrapper for pulling track audio features from Spotify's Web API in bulk. By automatically batching API requests, it allows to enter an artist's name and retrieve their entire discography in seconds, along with Spotify's audio features and track/album popularity metrics.

Many interesting results were found in during this project which is explained further in the report with all the results obtained.

The Github repository can be found through this link.

2 Problem Understanding

The Spotify provides various audio features; among all of them the audio feature Valence is the positivity meter of a particular song, i.e a highly positive song will have a value close to 1. The other feature which seemed very interesting was Track Popularity which tells the measure of how popular a song is on meter scale defined from 0 to 90.

After analyzing all the features with a Musicology student, I came up with two very eventually interesting questions for the project which could provide us compelling new relations.

1. What features contribute to the high or low value of the valence? What is the importance order of those features and how they can be effectively used to predict the valence for a particular song?

2. What features contribute to the the high or low value of the popularity of a song? What is the importance order of those features and how they can be effectively used to predict the popularity of a song?

The basic motive behind is to find interesting relations which will facilitate us to make our recommendation systems for music more robust, providing more coherent suggestions to the users preferences.

3 Data Accumulation

The R-code provided on Github uses the spotifyr package which is a quick and easy wrapper for pulling track audio features from Spotify's Web API in bulk. By automatically batching API requests, it allows you to enter an artist's name and retrieve their entire discography in seconds, along with Spotify's audio features and track/album popularity metrics.

It was given a special attention that the diverse songs were included in the dataset so that there is no monotonicity in the accumulated dataset. Artist discography ranging from famous artist Eminem, AC/DC etc to not famous artist like Miles Davis were covered. The dataset comprises of 10,130 songs.

4 Data Understanding and Preparation

It was found that there are many features in the dataset which are not useful and can be removed to avoid the curse of dimensionality. The features giving the information about the artist album name, release date etc, doesn't seem to be of any importance in this project so they are removed.

The audio feature of **Mode** is converted to the binary values, i.e for Major it is '1' and Minor it is '0'. The value of **Valence** is also converted as a binary value with any value more than 0.50 is '1' and less than that is '0'. The value of **Track Popularity** is also converted to a binary value of '1' for anything greater than 45 and '0' for less than 45.

This is done to normalize the data and making use of classification task on these attributes. This makes the learning process fast and the analysis more robust to outliers and redundancy.

5 Modeling

The following steps were adopted to modeling the dataset after Data Understanding and Preparation:

1. This step involves removing the highly correlated features which increases the redundancy in our dataset, hence are not required for our task. It was found that for the target variable *Valence*, Track Popularity feature was redundant feature.When the target variable is *Track popularity* has a redundant feature of energy which is removed from the dataset for the further analysis.

2. This step involves using ROC analysis for each attribute to find the feature importance order . The **ROC Analysis** is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. It is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. It constructs an Learning Vector Quantization (LVQ) model. The varImp is then used to estimate

the variable importance.

3. This step involves Automatic feature selection methods can be used to build many models with different subsets of a dataset and identify those attributes that are and are not required to build an accurate model. A popular automatic method for feature selection provided by the caret R package is called Recursive Feature Elimination or RFE. A Random Forest algorithm is used on each iteration to evaluate the model. The algorithm is configured to explore all possible subsets of the attributes.

These steps were performed on the both the set of target variables and the results obtained are shown in Evaluation section.

6 Evaluation

6.1 Correlation Matrices

These are the following matrices that are obtained by calculating the correlation between the attributes to remove the redundant features.

1. The highly redundant feature when the target variable is *Valence* is Track Popularity and it shown in the figure below.

	Loudness	Energy	Tempo	Liveness	Danceability	Speechiness	Instrumentalness	Accousticness	Track_Popularity	Album Popularity	Mode
Loudness	1.00000000	0.69374040	0.089921372	0.051951779	0.07297257	-0.02404944	-0.223579556	-0.40470709	0.169052387	0.15683847	-0.057924858
Energy	0.69374040	1.00000000	0.157725875	0.221821265	0.06044426	0.12883417	-0.017869024	-0.67456807	0.221806688	0.20408866	-0.086703608
Tempo	0.08992137	0.15772587	1.000000000	-0.016389587	-0.08750832	-0.01321906	0.041023526	-0.11128283	0.036930514	0.02448412	0.007150207
Liveness	0.05195178	0.22182126	-0.016389587	1.000000000	-0.27352915	0.16856191	-0.008046307	-0.05579872	0.017016713	0.01864228	0.048991494
Danceability	0.07297257	0.06044426	-0.087508322	-0.273529150	1.00000000	0.16054488	-0.058583757	-0.13729888	-0.025099023	-0.01602631	-0.031660163
Speechiness	-0.02404944	0.12883417	-0.013219064	0.168561908	0.16054488	1.00000000	-0.057327455	-0.04040628	0.014631946	0.04168107	-0.073108664
Instrumentalness	-0.22357956	-0.01786902	0.041023526	-0.008046307	-0.05858376	-0.05732746	1.000000000	-0.03355118	0.049410418	0.03263083	-0.051395664
Accousticness	-0.40470709	-0.67456807	-0.111282832	-0.055798724	-0.13729888	-0.04040628	-0.033551177	1.00000000	-0.326993628	-0.32051276	0.069292869
Track_Popularity	0.16905239	0.22180669	0.036930514	0.017016713	-0.02509902	0.01463195	0.049410418	-0.32699363	1.000000000	0.90571268	-0.006915707
Album Popularity	0.15683847	0.20408866	0.024484116	0.018642276	-0.01602631	0.04168107	0.032630826	-0.32051276	0.905712681	1.00000000	0.022036249
Mode	-0.05792486	-0.08670361	0.007150207	0.048991494	-0.03166016	-0.07310866	-0.051395664	0.06929287	-0.006915707	0.02203625	1.000000000

Figure 1: Correlation Matrix for Target variable Valence

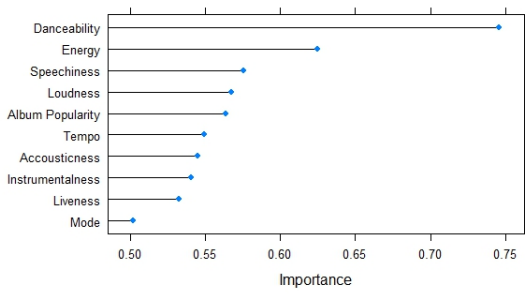
2. The highly redundant feature when the target variable is *Track Popularity* is Energy and it shown in the figure below.

	Loudness	Energy	Tempo	Liveness	Danceability	Speechiness	Instrumentalness	Accousticness	Album Popularity	Mode	Valence
Loudness	1.00000000	0.69374040	0.089921372	0.051951779	0.07297257	-0.02404944	-0.223579556	-0.40470709	0.15683847	-0.0579248576	0.1640400397
Energy	0.69374040	1.00000000	0.157725875	0.221821265	0.06044426	0.12883417	-0.017869024	-0.67456807	0.20408866	-0.0867036076	0.2491013697
Tempo	0.08992137	0.15772587	1.000000000	-0.016389587	-0.08750832	-0.01321906	0.041023526	-0.11128283	0.02448412	0.0071502066	0.1150705141
Liveness	0.05195178	0.22182126	-0.016389587	1.000000000	-0.27352915	0.16856191	-0.008046307	-0.05579872	0.01864228	0.0489914939	-0.0861066356
Danceability	0.07297257	0.06044426	-0.087508322	-0.273529150	1.00000000	0.16054488	-0.058583757	-0.13729888	-0.01602631	-0.0316601632	0.5147346454
Speechiness	-0.02404944	0.12883417	-0.013219064	0.168561908	0.16054488	1.00000000	-0.057327455	-0.04040628	0.04168107	-0.0731086637	0.0820600716
Instrumentalness	-0.22357956	-0.01786902	0.041023526	-0.008046307	-0.05858376	-0.05732746	1.000000000	-0.03355118	0.03263083	-0.0513956637	-0.1049869261
Accousticness	-0.40470709	-0.67456807	-0.111282832	-0.055798724	-0.13729888	-0.04040628	-0.033551177	1.00000000	-0.32051276	0.0692928689	-0.0909197739
Album Popularity	0.15683847	0.20408866	0.024484116	0.018642276	-0.01602631	0.04168107	0.032630826	-0.32051276	1.00000000	0.0220362491	-0.1460289998
Mode	-0.05792486	-0.08670361	0.007150207	0.048991494	-0.03166016	-0.07310866	-0.051395664	0.06929287	0.02203625	1.0000000000	-0.0005772688
Valence	0.16404004	0.24910137	0.115070514	-0.086106636	0.51473465	0.08206007	-0.104986926	-0.09091977	-0.14602900	-0.0005772688	1.0000000000

Figure 2: Correlation Matrix for Target variable Track Popularity

6.2 ROC Analysis Results

1. The results for the ROC analysis for target variable Valence are shown in figure below.



(a) Boundary of the calibration grid

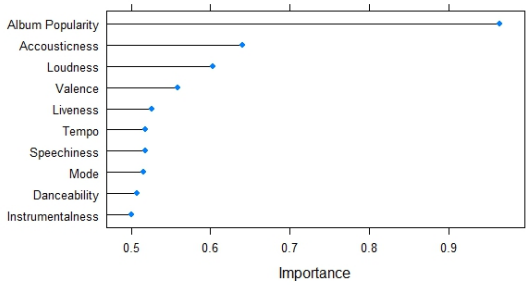
```
ROC curve variable importance

                  Importance
Danceability          0.7458
Energy                0.6247
Speechiness           0.5754
Loudness              0.5673
Album Popularity      0.5638
Tempo                 0.5494
Accousticness         0.5450
Instrumentalness      0.5408
Liveness              0.5328
Mode                  0.5021
```

(b) Predicted grid corners in absence of distortion

Figure 3: ROC analysis to find importance of features in relation to Valence as target variable

2. The Results of ROC analysis for target variable Track Popularity are shown in figure below.



(a) Boundary of the calibration grid

ROC curve variable importance

	Importance
Album Popularity	0.9642
Accousticness	0.6403
Loudness	0.6033
Valence	0.5590
Liveness	0.5258
Tempo	0.5179
Speechiness	0.5176
Mode	0.5161
Danceability	0.5074
Instrumentalness	0.5010

(b) Predicted grid corners in absence of distortion

Figure 4: ROC analysis to find importance of features in relation to Track Popularity as target variable

6.3 Recursive Feature Elimination Results

1. The Results of RFE for target variable Valence are shown in figure below.

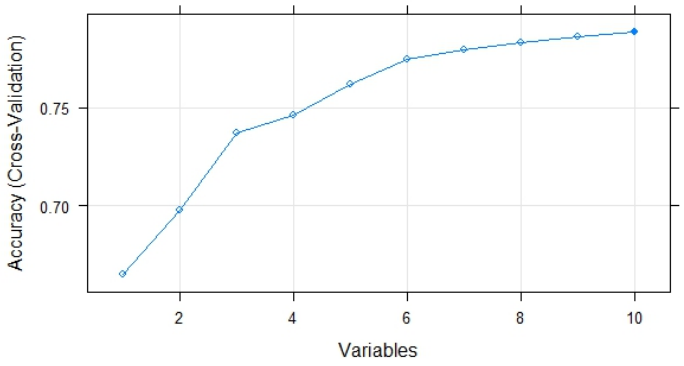


Figure 5: RFE analysis for target variable Valence plot

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         1   0.6649 0.3253   0.013699 0.02764
         2   0.6977 0.3913   0.013083 0.02630
         3   0.7373 0.4704   0.010891 0.02158
         4   0.7460 0.4882   0.008472 0.01704
         5   0.7618 0.5203   0.015306 0.03071
         6   0.7743 0.5456   0.015130 0.03031
         7   0.7792 0.5553   0.016375 0.03296
         8   0.7829 0.5632   0.013845 0.02783
         9   0.7861 0.5694   0.014502 0.02917
        10   0.7887 0.5749   0.009932 0.02004        *

The top 5 variables (out of 10):
   Danceability, Energy, Speechiness, Album Popularity, Loudness
```

Figure 6: RFE analysis for target variable Valence

2. The Results of RFE for target variable Track Popularity are shown in figure below.

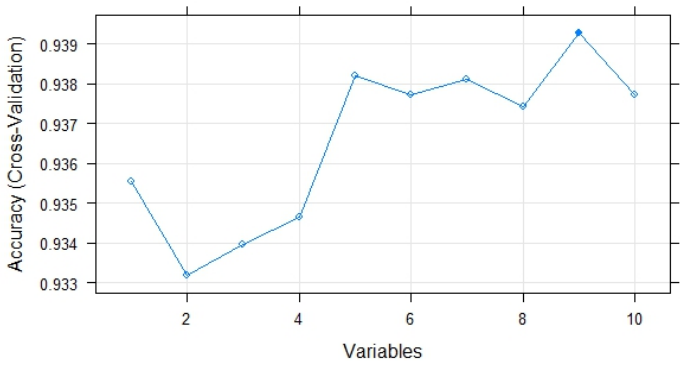


Figure 7: RFE analysis for target variable Track Popularity plot

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD  KappaSD Selected
         1   0.9355 0.6732   0.003983 0.020761
         2   0.9332 0.6712   0.004429 0.019883
         3   0.9340 0.6837   0.004673 0.022490
         4   0.9346 0.6900   0.005083 0.019886
         5   0.9382 0.7039   0.005076 0.019231
         6   0.9377 0.7013   0.003356 0.015190
         7   0.9381 0.7015   0.003053 0.015659
         8   0.9374 0.6973   0.003790 0.020634
         9   0.9393 0.7075   0.002426 0.009959        *
        10   0.9377 0.6998   0.004306 0.021660

The top 5 variables (out of 9):
   Album Popularity, Accousticness, Loudness, Instrumentalness, Speechiness
```

Figure 8: RFE analysis for target variable Track Popularity

7 Deployment and Conclusion

From the above analysis done, it can be inferred the following things:

- The most important features which can be used or gets associated with the Valence attribute the most according to ROC analysis are Danceability, Energy, Speechiness and Loudness in decreasing order. The decreasing order for target variable Track Popularity is Album Popularity, Accousticness, Loudness and Valence.

- The most important features which can be used or gets associated with the Valence attribute the most according to RFE Analysis are Danceability, Energy, Speechiness and Album popularity in decreasing order.The decreasing order for target variable Track Popularity is Album Popularity, Accousticness, Loudness, Instrumentalness.

- For the deployment of the found knowledge, **Spotify** can associate these attributes to Valence and Track Popularity to help target their customers who are looking forward to listen to a highly positive song or a popular song and vice versa, to make their recommender system more robust and increase it's accuracy.