**Million Song Project**

Group 3, Final report

Project statement

The project aim is to predict the song hotness - or ‘song hotttnesss’ [sic] - based on its features (tempo, loudness, segment data etc), and metadata (artist information, year of release etc).

Project background and literature review

This project is based on the Million Song Dataset (MSD) provided by labROSA at Columbia University.

The majority of record company profits are made from smash hits, and a great deal of money and effort is spent on seeking out the next success. Thus, predicting song hotness would be a useful way to test new songs and predict their commercial potential. To our knowledge, this aspect has not been extensively explored in the literature– previous analysis using the MSD has concentrated on predicting year, genre and providing recommendations to users.

Previous attempts to predict hit songs have had mixed success, possibly due to the unpredictability of cultural markets[[1]](#footnote-1). Although there are industry rules of thumb on how to write hit songs, there is comparatively little in the machine learning space to automatically detect hits. Early attempts used acoustic and lyric-based features to build support vector machines have not resulted in reproduceable successful models, which is explained by the fact that these features are not informative enough[[2]](#footnote-2).

More advanced features that have been extracted in the MSD may provide a better basis for song hotness prediction and have been used in a small number of studies to predict song popularity. One of the most relevant is Pham et al[[3]](#footnote-3), which used song characteristics consisting of both acoustic features and metadata from MSD. The authors applied a number of machine learning algorithms (SVMs, neural networks, logistic regression, Gaussian discriminant analysis, and linear regression) to predict whether a song is popular or not. When designed as a classification task, the best models achieved around 80% test set accuracy compared to a 75% baseline (labelling only the top 25% of songs as being popular). The best regression model achieved average error of 0.134 compared to a baseline of 0.159.

Preliminary EDA

The full dataset of a million songs is around 270 GB. There is also a subset of 10,000 songs (around 2.5 GB) that we will use for the purposes of developing the prediction approach, as indicated in the project guidelines. Both datasets are provided as compressed HDF5 files.

In the main dataset are audio analysis features for each track, including the target variable of song hotness. This has values ranging from 0-1, and indicates the popularity of a song. It is assigned algorithmically based on various metrics, including news/blog mentions, play counts, music reviews, radio airtime and Billboard rankings.

There are also an additional 53 variables that include:

* Several **artist and track IDs** that can be used for linking with other datasets: The Echo Nest[[4]](#footnote-4) ID, Music Brainz, 7digital and playme.com.
* **Echo Nest tags**, which include 'artist terms', ‘similar artists’.
* **Music Brainz tags** include the fields 'year', 'artist mbtags' and 'artist mbtags count'. These tags are applied by humans and there are fewer of them compared to the Echo Nest ones, but where they do exist they tend to be cleaner.
* **Artist data,** such as location, ‘hotttnesss’ and familiarity
* **Song data,** such as tempo, key, duration, title
* **Track analysis data,** containing arrays of features that correspond to different segments or sections of the track. The main acoustic features are pitches, timbre and loudness, as defined by the Echo Nest Analyze API. The API provides these for every “segment”, which are generally delimited by note onsets, or other discontinuities in the signal. The API also estimates the tatums, beats, bars (usually groups of 3 or 4 beats) and sections.

In terms of the data quality, we identified the following:

* `Song\_hotttnesss` is missing for 43% of the rows in the subset. If these are removed, around 25% of the remaining rows have hotness values of 0
* Several columns are non-informative and can be dropped:
  + `Analysis sample rate:` is the same value for all tracks
  + `Energy` and `Danceability`: contain only zeros
  + `Artist latitude` and `longitude`: have more than 60% missing data
* `Year` has around 25% of rows with zero, which can be considered missing data. These are left as they are for now.

The data is accompanied by additional files that provide the following information:

* List of track Echo Nest IDs
* List of all artist IDs
* List of all unique artist terms
* List of all unique musicbrainz tags
* List of the tracks for which there is year information
* List of artist latitude and longitude, where available
* Summary file of the whole dataset containing metadata but no audio analysis
* SQLite database containing track metadata
* SQLite database containing links from artist ID to tags
* SQLite database containing similarity among artists

The distribution of song hotness is skewed with around 25% of the tracks having zero values (of the 5648 that do not have NaN values). Average song hotness per year seems to have a slight upward trend. We must also consider that musical tastes change over time.

|  |  |
| --- | --- |
| image1.tif | /var/folders/t6/q8snfhlx5b1gbwddfx3vt16h0000gn/T/com.microsoft.Word/Content.MSO/55BBE8A2.tmp |

Exploration of correlations with the single-valued numerical columns shows that artist familiarity and artist hotness are strongly correlated with the target variable. As previously identified, there seems to be a positive correlation between song hotness and year, as well as with loudness.

image3.tif

|  |  |
| --- | --- |
| image4.tif | image5.tif |

Several of the columns also contain numerical arrays of data for distinct parts of each song:

1. **Bars**: A bar (or measure) is a segment of time defined as a given number of beats.
2. **Beats:** A beat is the basic time unit of a piece of music; for example, each tick of a metronome. Beats are typically multiples of tatums.
3. **Sections:** Sections are defined by large variations in rhythm or timbre, e.g. chorus, verse, bridge, guitar solo, etc.
4. **Segments:** A set of sound entities (typically under a second) each relatively uniform in timbre and harmony.
5. **Tatums:** Tatums represent the lowest regular pulse train that a listener intuitively infers from the timing of perceived musical events (segments).

All of these parts have two columns associated with them:

* Start: the markers explained above, in seconds.
* Confidence: indicates the reliability of its corresponding attribute. Elements carrying a small confidence value should be considered speculative.

It is already clear from the descriptions that the arrays will be of different lengths since the length of the tracks and frequency of the measures will vary. Below are histograms showing the frequency of array lengths for the songs in the subset.

image6.tif

The most detailed information is for segments, which contain several other features:

* Loudness\_max: peak loudness value within the segment
* Loudness\_max\_time: offset within the segment of the point of maximum loudness
* Loudness\_start: loudness level at the start of the segment
* Pitches: is given by a “chroma” vector, corresponding to the 12 pitch classes C, C#, D to B, with values ranging from 0 to 1 that describe the relative dominance of every pitch in the chromatic scale.
* Timbre: is the quality ('colour') of a musical note. It is what makes a particular musical sound different from another, even when they have the same pitch and loudness. The dataset represents timbre as a 12 dimensional vector of 'MFCC-like' features, each of which are unbounded values roughly centered around 0.

The number of segments per track is variable and each segment can itself be of variable length - typically they seem to be around 0.2 - 0.4 seconds but can be as long as 10 seconds or more.

Information on loudness at the start of each segment is shown below for three example songs. The outputs for loudness\_max are similar.

image7.tif

The 12 dimensional arrays contained in the pitch and timbre columns are shown below for the same three example songs.

image8.tif

image9.tif

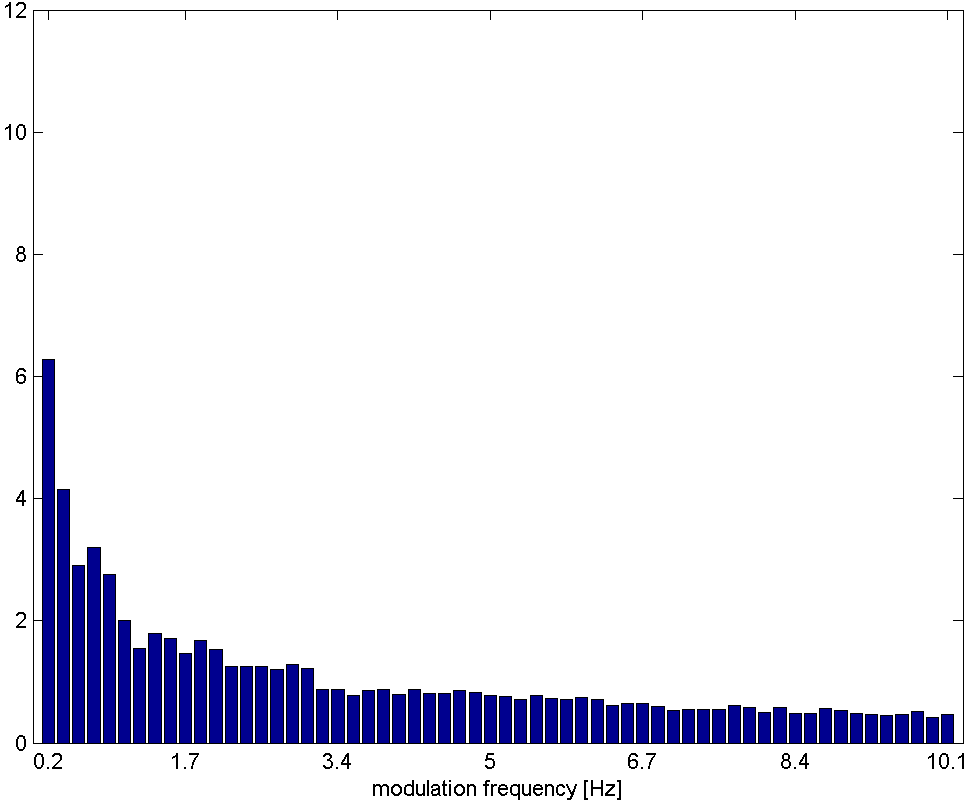
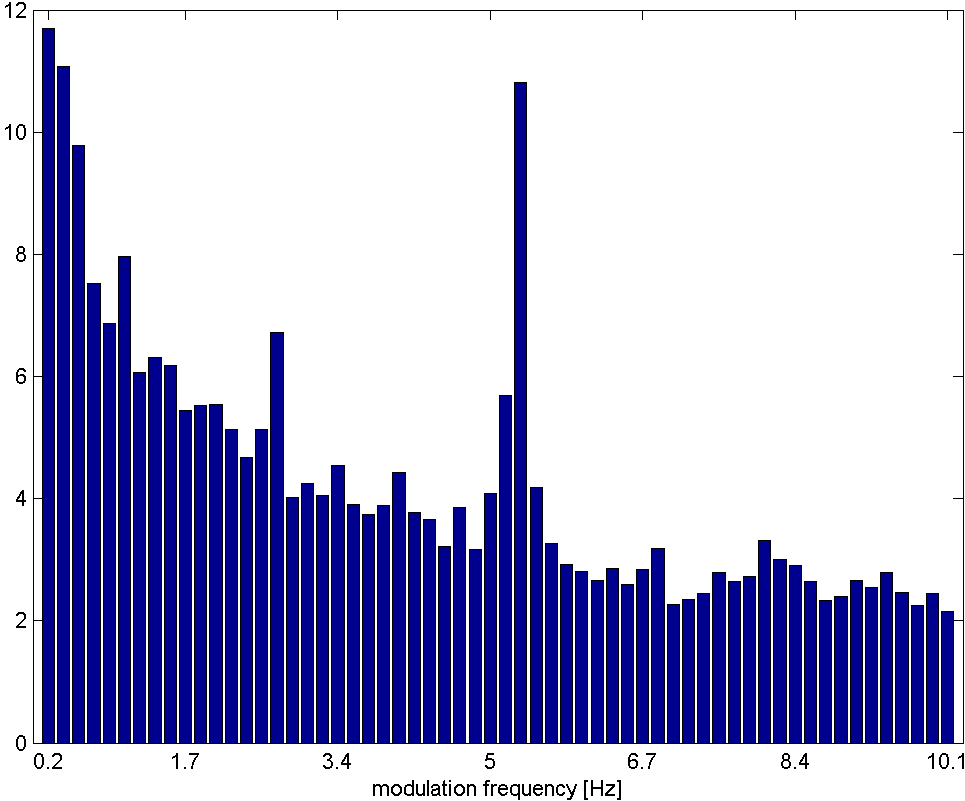
Genre and style information

We started to explore whether the MSD can be augmented with information from other sources. We tried to merge the data with genre and style information for Vienna University[[5]](#footnote-5) by joining on the track IDs. Unfortunately, the results were incomplete, with more than half of records having no genre information and around three-quarters having no style information.

Audio information

<http://www.ifs.tuwien.ac.at/mir/msd/download.html#groundtruth> provides data on audio-extracted information in ARFF format (<https://www.cs.waikato.ac.nz/~ml/weka/arff.html> ). We started to look at rhythm histograms (<http://www.ifs.tuwien.ac.at/mir/audiofeatureextraction.html#RH> ):

**The Rhythm Histogram** features are a descriptor for general rhythmics in an audio document. The magnitudes of each modulation frequency bin of all critical bands are summed up, to form a histogram of "rhythmic energy" per modulation frequency. The histogram contains 60 bins which reflect modulation frequency between 0 and 10 Hz. For a given piece of audio, the Rhythm Histogram feature set is calculated by taking the median of the histograms of every 6 second segment processed. E.g. classical vs. rock music:

ARFF format is difficult to read directly (version incompatibility) hence a direct CSV read is done. Match is done against track id. There are 60 measurements per track. The idea is to use these measurements as a feature to identify “hot” songs.

**Statistical Spectrum Descriptor**

The Sonogram is calculated as in the first part of the Rhythm Patterns calculation. According to the occurrence of beats or other rhythmic variation of energy on a specific critical band, statistical measures are able to describe the audio content. Our goal is to describe the rhythmic content of a piece of audio by computing the following statistical moments on the Sonogram values of each of the critical bands:

mean, median, variance, skewness, kurtosis, min- and max-value

Modelling

We created two sets of models, treating song hotness as either a binary variable for classification or a continuous variable for regression. The classification problems defined a song as being “hot” if they were in the top 25% of the song hotness scores. The reason for choosing this threshold is that it seemed a reasonable trade-off between having enough positive examples and the fact that not many songs are truly hot in practice.

Baseline

Rahul

Audio features in the dataset arrays

As found in the EDA, the core dataset contained various audio features in numerical arrays. The most informative of these was the segment data, which included features as well as the time stamps.

The data were processed into a 26 dimensional vector for each segment, consisting of concatenated data in the following order:

* loudness at start of segment
* maximum loudness in each segment
* 12 timbre features
* 12 chroma features

The arrays were cropped by taking the first 350 segments and excluding tracks that don't have at least this number of segments. This should be sufficient to capture the main parts of each song.

Two models were trained on this dataset:

**Model 1: CRNN network**

The CRNN first passes the data through a series of convolutions before feeding to a Gated Recurrent Unit (GRU). Given the imbalance in the training data, the classes are balanced in the model fitting.

* It uses **1D convolution layers** that perform convolution operations just across the time dimension. This is to extract features from each time slice (rather than 2D convolution more frequently seen in image processing). Dropout and max pooling are used to prevent overfitting.
* The output of the CNN step is fed into a **GRU, which should find the short and long term structure of the audio features**. Recurrent neural networks (RNN) are used for understanding sequential data. They make a hidden state t dependent on the hidden state in the previous time step.

**Model 2: Parallel CNN RNN network**

The CRNN model above uses RNNs as a temporal summariser, but the GRU uses the output of the CNNs. In a parallel CNN-RNN model, the architecture has the following structure instead:

* The signals are passed through 2D convolutions followed by max pooling and the final output is flattened. This is similar to image processing where the signals are now processed in 2D.
* Separately, the signals are passed through RNN
* The outputs of both steps are concatenated and passed to a dense network.

This approach allows us to preserve the temporal relationships of the original signals and use CNN and RNN in parallel.

Unfortunately, neither the classification nor regression version of these models yielded any improvement over the baseline.

Additional audio features (AudioModel.ipynb)

<http://www.ifs.tuwien.ac.at/mir/audiofeatureextraction.html> contains links to several datasets which were created from MSD audio features. Specifically, we looked at:

* Rhythm histograms: general rhythmics in audio document, stored as 60 bins which reflect modulation frequency from 0 to 10 Hz. Feature set is calculated by taking the median of the histograms of every 6 second segment processed (60 values per track)
* Statistical spectrum descriptors: according to the occurrence of beats or other rhythmic variation of energy on a specific critical band, audio content is described by statistical moments of critical bands (mean, median, variance, skewness, kurtosis, min- max-value) (168 vector of values per track)

These audio datasets were combined (following overall “collage” idea from <https://www.pyimagesearch.com/2019/01/28/keras-regression-and-cnns/>), using the preprocessing dataset (e.g. removal of song-track mismatches <http://millionsongdataset.com/blog/12-2-12-fixing-matching-errors/> , restricting the track set to match the given subset of MSD).

CNN was created to attempt to find meaningful relationship between audio features and 'song\_hotttnesss'. In this way, audio features represent a ‘picture’ on which CNN is trained.

* Input is reshaped to look like a ‘picture’, i.e. 228 per-track data values are reshaped to 12 x 19 ‘rectangle’
* Several sets of 2D convolution, activation (RELU) and max pooling filters (16, 32, 64)
* Single ‘channel’ is used as if it’s a monochrome ‘image’
* For regression, Linear activation is used in the last layer while for classification, Sigmoid is used

Unfortunately, regression version of this model hasn’t yielded any improvement over the baseline. Classification version of this model has done better than regression (~75%) but still worse than baseline.

Text to Topics to Hits (TextTopicHits.ipynb)

For some songs, there are lyrics available (https://labrosa.ee.columbia.edu/millionsong/musixmatch). To explore relationship of song words and hit topics, the following model was done:

* Combined dataset was created off Musixmatch
* Dataset was parsed to go from words and their counts (which was apparently done for copyright reasons) to “sentences” where each word appears as many times as in the original text
* Corpus and vocabulary mapping created
* Latent Dirichlet Allocation was executed on the full corpus to extract 4 topics, 10 words each
* Latent Dirichlet Allocation was executed on only hit songs (defined as top 75% quantile for ‘song\_hotttnesss’ field) corpus to extract 4 topics, 10 words each for hit songs
* Words which appear only in hits and not in all songs were detected (‘hit words’)
* Prediction of a hit song was made based on presence of the ‘hit words’

Results fluctuate a bit from run to run but resulting accuracy is still low (~40%), i.e. this model did not yield any improvement over the baseline.

Lyrics data

Rahul

Artist terms

The data set contains artist terms for each song. In all, there is 3100 artist terms in use in the subset. These tags may be terms like «new wave», «indie rock», «americana» or similar. Each song is tagged with a number of such terms. The song with most terms linked to it, had 82 terms. Some songs had 0.

We cleaned up the terms, and counted the number of terms each song was tagged with as a variable. Then we filtered out the terms that were used for less than 20 songs. This left us with 824 terms that we one hot encoded for the models, in addition to the parameter stating the count of terms used for the song (before the removal of the sparsely used terms).

To ensure that the artist terms are as common as possible across the different songs, we did also implement stemming.

We trained a number of models on this data; both classifications and regressions.

Classifications

We tried a number of different classification models:

* K-Nearest Neighbor
* Linear support vector machine
* Non-linear support vector machine
* Gaussian Process Classifier
* Decision Tree Classifier
* Random Forrest Classifier
* Multilayer Perceptron Classifier
* AdaBoost Classifier
* Gaussian Naive Bayes Classifier
* Quadratic Discriminant Analysis
* Neural networks

K-Nearest Neighbor and Non-linear support vector machine achieved the best results on our initial run, and these were ran again with cross validation to ensure a more valid result.

QDA and Random Forrest did not improve the result compared a baseline of declaring all songs as not hot. Naive Bayes did substantially worse. The rest achieved better results.

The two models selected for cross-validation achieved results up towards 90 %.

We tried a number of combinations of neural networks as well, and achieved similar results to QDA and the Random Forrest. All these came close to 90 %, which was the best results we achieved within this project. This is a substantial improvement above the baseline model, but unfortunately not a robust method to rely on if ones goal is to be a multimillionaire songwriter.

Regression

We also attempted to run regression models on the artist terms. We ran a few different models here as well:

* 2 different Neural nets with K-Fold Cross Validation
* Gradient Boosting

The results did not prove very successful.

When we compare the true hotness compared to the predicted hotness from the gradient boosting model; it looks as such:

The predicted song hotness increase slightly as the true song hotness increase. However, this effect is very weak. The model does not predict any songs above approximately 0.65, while we in the classification section only classified songs with a score above 0.75 as hot. Hence, the regression model have not classified any songs as hot according to our earlier definition.

We can see that there is a correlation between the songs hotness and the predicted hotness, but this correlation is too weak to be useful.

Conclusions

1. Salganik et al (2006) Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market <https://www.princeton.edu/~mjs3/salganik_dodds_watts06_full.pdf> [↑](#footnote-ref-1)
2. Herremans et al (2014) Dance Hit Song Prediction <http://antor.uantwerpen.be/wordpress/wp-content/papercite-data/pdf/herremans2014dance.pdf> [↑](#footnote-ref-2)
3. Pham et al (2015) Predicting Song Popularity <http://cs229.stanford.edu/proj2015/140_report.pdf> [↑](#footnote-ref-3)
4. Now owned by Spotify [↑](#footnote-ref-4)
5. Available at <http://www.ifs.tuwien.ac.at/mir/msd/download.html> [↑](#footnote-ref-5)