Do Popular Songs Endure?

Progress Report

Prasoon Rai , Sushant Ojal
110921754 , 110944445
{prrai,sojal}@cs.stonybrook.edu

1 Introduction

The long time endurance of a song is a challenging problem to solve. In the project proposal, we pitched to utilize data from several areas to come up with diverse features that can potentially affect the popularity of a song over the decades that rolled by. We aimed to build a model that could predict the current popularity of the song, represented by an endurance score. We successfully obtained initial data sources and collected some data along with considering a preliminary feature set.

During the midterm report phase, we have come a long way in molding the rudimentary ideas to a concrete working model. We now have a rich data set comprising of more than 25,000 songs with over 25 features that characterize them. Our baseline model along with an improved version is complete and evaluation metrics for the same have been explored.

The following sections describe the data sources, methods employed towards its accumulation along with details on the machine learning models and features. At the end, we provide the upcoming enhancements towards features and proposal towards an advanced model to further improve our predictions. The current progress of our work can be found at: `https://github.com/prrai/Data_Science_Do_Popular_Songs_Endure`.

2 Data Scraping

In continuance with the data we collected for the initial progress report, we have scraped websites for features which add more information for each of the songs. We are enumerating the sources which we used for this purpose:

1. **Universal Music Database(UMD):** We had used this website to get weekly Billboard top 100 songs along with the number of weeks that song stayed in top 100, its peak position, entry into the Billboard and more. Using this data, we prepared a dataset with roughly 25918 distinct songs having above mentioned features which have made it to the Billboard Top 100 from 1959 to present date.

2. **Spotify:** We used *Spotipy*, a lightweight python API for the Spotify's web API, which provides access to the Spotify's music database. Using this we were able to obtain a substantial number of features for every song, including its audio features, the duration of the song and ts popularity. The details of these features are provided in the subsequent sections. It is worth mentioning that the the popularity of songs

on Spotify serves as an estimate of the present day popularity of the song. We also used Spotify to query the popularity of the artist of a song. We were able to gather data for approximately 22k songs out of the dataset of 26k distinct songs.

3. **Youtube:** We used YouTube's python API to query a song based on the name of the track and the name of the artist. The sum of the view counts of the first two videos (sorted by relevance), is taken to be another estimate of the popularity of a song.

4. **Billboard.com:** We have used this website to obtain the top 200 albums and artists per week, besides giving information on the number of songs per artist that made to Billboard top 100, 10 and peak (number 1).

5. **What-song.com:** We scraped data of more than 25000 unique artists to get the exact counts of every song that was featured in a movie or a TV show. A scraper and an extractor were created to appropriately obtain the feature counts.

6. **Awardsdatabase.oscars.org:** We used this site to get the list of every song that has ever received an Oscar award in the best song category.

7. **Grammy.com:** We were able to obtain every artist that has received a Grammy lifetime achievement award. We also scraped out the number of nominations and wins across all categories for every artist that features in over song database.

8. **AZlyrics.com:** We have opted not to opt for text analysis for the lyrics, since songs on a variety of topics can be popular. It is the overall eloquence and the meaning of the lyrics which has a greater impact on the popularity of a song. Analysis of such factors is significantly difficult.

9. **Wikipedia:** We also opted not to scrape Wiki pages of the songs since we were able to gather most of the relevant information using the above mentioned sources.

3 Features

This section highlights the features we have used to build out training set. We will also mention features we still plan to gather using more data collection. We had the following set of features during our project proposal stage:

1. **Title (Source: UMD)** The name of the song.

2. **Artist (Source: UMD)** The band or vocalist who created the song.

3. **Entry Date (Source: UMD)** The date the song made it to the Billboard Top 100 for the first time.

4. **Peak Position (Source: UMD)** The highest rank the track made in the charts.

5. **Total Weeks (Source: UMD)** The number of weeks the song stayed in the top 100 Billboard Charts.

6. **TopSongArtist (Source: Billboard.com)** The number of songs of that artist that have topped the charts.

7. **TopSongArtist10 (Source: Billboard.com)** The number of songs of that artist to have made it to the top 10 of the charts.

8. **TopSongArtist100 (Source: Billboard.com)** The number of songs of that artist to have made it to the top 100 of the charts.

In addition to these features, we acquired the following features:

1. **Duration ms (Source: Spotify)** The duration of the song in milliseconds.

2. **Danceability (Source: Spotify)** 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.

3. **Energy (Source: Spotify)** Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.

4. **Acousticness (Source: Spotify)** A confidence measure from 0.0 to 1.0 of whether the track is acoustic.

5. **Key (Source: Spotify)** The key the track is in.

6. **Liveness (Source: Spotify)** Detects the presence of an audience in the recording.

7. **Loudness (Source: Spotify)** The overall loudness of a track in decibels (dB).

8. **Speechiness (Source: Spotify)**Speechiness detects the presence of spoken words in a track.

9. **Tempo (Source: Spotify)** The overall estimated tempo of a track in beats per minute (BPM).

10. **Valence (Source: Spotify)**A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.

11. **Mode (Source: Spotify)** Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived.

12. **Movies_TV_feature_count (Source: what-song.com)** This feature indicates the number of movies and TV shows a particular song featured on.

13. **Oscars_won (Source: Awardsdatabase.oscars.org)** The number of Oscars won by the current song.

14. **Artist_lifetime_grammy_achievement (Source: Grammy.com)** The feature holds a value of 1 if the artist for the current song has received the lifetime achievement Grammy award, else 0.

15. **Artist_grammy_wins (Source: Grammy.com)** The number of Grammy awards won by the artist across all categories in their music career.

16. **Artist_grammy_nominations (Source: Grammy.com)** The number of Grammy awards nominations received by the artist across all categories in their music career.

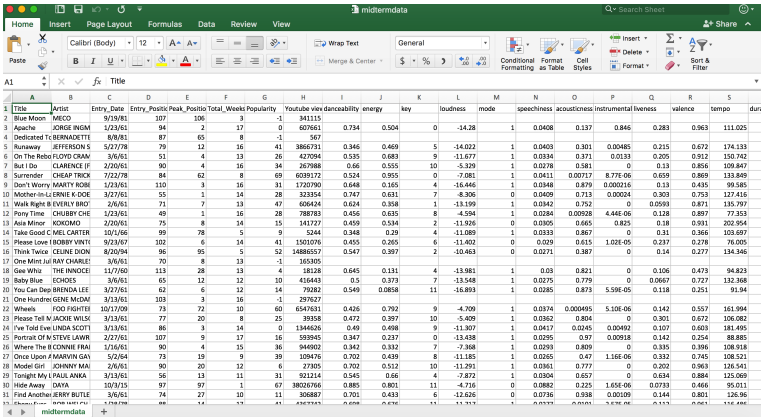


Figure 1: View of the data set showing features and few entries.

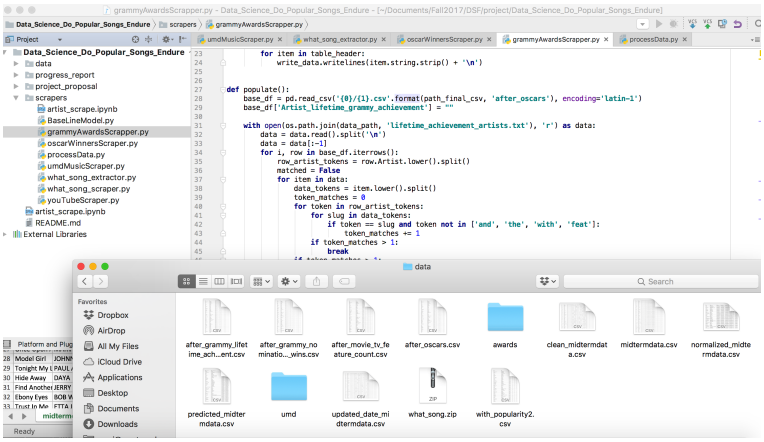


Figure 2: The figure shows various scrapers coded and the corresponding data collected as a result.

3.1 Estimation of Endurance Score

To estimate the present day popularity of any given song, we calculate an *Endurance Score* of a song. For this, we have taken into consideration, the following two features:

1. **Popularity on Spotify**: The popularity of the track. The value will be between 0 and 100, with 100 being the most popular. This value is determined by Spotify, and is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are.

2. **YouTube Viewcount**: The number of times a song has been viewed on YouTube. This value is the average of the first two most viewed videos of the song.

The sum of the normalized values of the popularity values of both YouTube and Spotify serve as the Y values for out training data. As is evident from the shown scatterplot, there is a positive correlation between both the popularity measures.

Another important fact worth mentioning is that we have removed the songs which were released after 2013, since there is an increase in both the YouTube view-count as well as number of streams on Spotify simply by the virtue of them being recent. After this, we were left with 20228 distinct songs in our database.

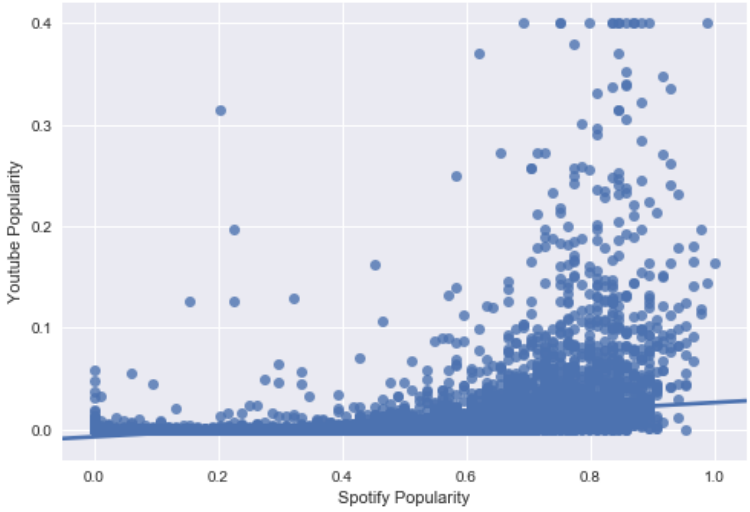


Figure 3: A scatterplot of the popularity of all the songs in the dataset on Spotify and YouTube. The values have been normalized on a scale of 0 to 1 on both axes.

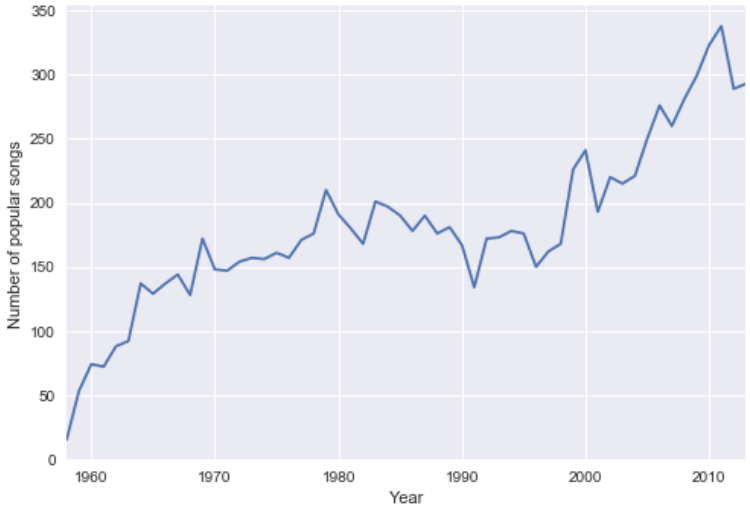


Figure 4: Trend of popular songs over the years. The count of popular songs of a year is calculated by the number of distinct songs in that year which exceed the median popularity of all the songs over all the years. The rapid increase in recent years is most probably because of the online presence of a song immediately on its release.

4 Baseline Model

4.1 Data Processing:

Before applying the model, we cleaned up the data. This comprised of removing rows that have missing several entries, filling empty column entries with median values, nor-

malization of data and filtering out songs after the year of 2013. The endurance score is calculated and added as a new column to the dataset.

4.2 Linear Regression model:

Once we finalized the data, we implemented a wrapper on top of sklearn that allows us to quickly apply models to our data. As a baseline, we implemented linear regression model. We split the data into 80% and 20% training and test data, followed by the prediction.

- Linear regression works by solving a linear equation where coefficients are estimated from the training data with intention to minimize residual sum of squares between difference in the training and test data. In scikit-learn, the coef_ contains a list of all the coefficients, whose number would be equal to the total features considered during modeling.
- Like any regression model, linear regression seeks to fit the data along a curve, in this case a straight line. The performance is gaged by observing the root mean square error (lower is better) and the coefficient of determination (closer to 1, the better).

4.3 Ridge Regression model:

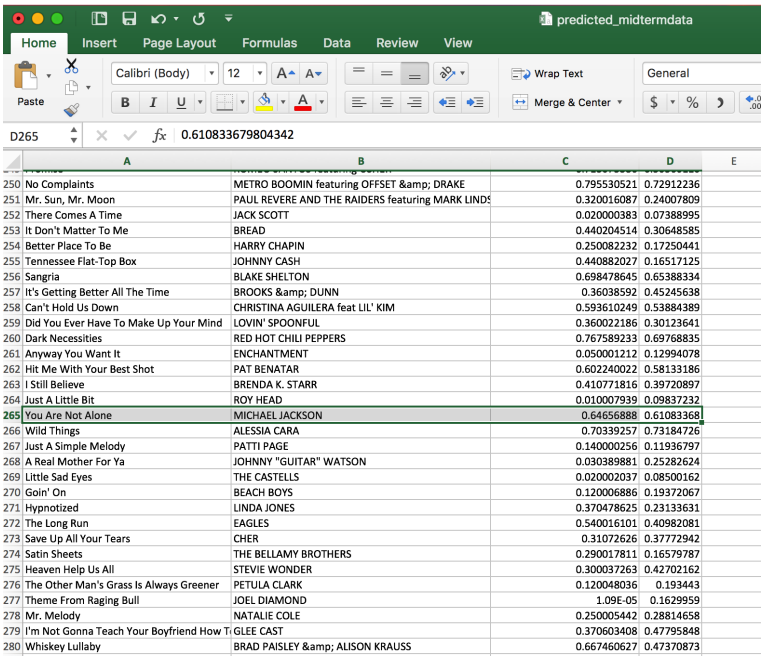
As an extension to the linear regression, we use ridge regression to further improve the accuracy of the model.

- Ridge regression introduces a trade off by allowing a bias in the coefficients but decreasing the variance of the parameters. So although, we know that the predictions are off by a little amount, the variance is less and overall the prediction is more reliable as compared to the OLS. To be more specific, assume the problem at hand is A.x = b, where A is the known matrix and b is the know vector. Ordinary least squares aims to minimize ($||A.x()b||^2$), which represents the L-2 norm. Ridge regression adds a penalizing term ($||T.x||^2$) to this L-2 norm, where the matrix T controls the extent to which the coefficients represented by the matrix x are allowed to grow.
- In the given dataset, there are a number of properties which are highly correlated to each other. It makes sense in such a case to do L2 regularization to prevent the coefficients of these properties from blowing up, so that there is less variance.

4.4 Model evaluation

We observed that our models perform reasonably well at the baseline. We use r2-score and root mean square error as metrics for estimating the accuracy. The results are as below. The ridge regression slightly improves the model's accuracy.

Model	**R2 score**	**RMSE**
Linear Regression	0.52473124364508	0.15421120189311405
Ridge Regression	0.5250676278818431	0.1541566186716756



predicted_midtermdata

Home | Insert | Page Layout | Formulas | Data | Review | View

Paste | Calibri (Body) | 12 | Wrap Text | General | Merge & Center

D265 | 0.610833679804342

	A	B	C	D	E
250	No Complaints	METRO BOOMIN featuring OFFSET & DRAKE	0.795530521	0.72912236	
251	Mr. Sun, Mr. Moon	PAUL REVERE AND THE RAIDERS featuring MARK LINDS	0.320016087	0.24007809	
252	There Comes A Time	JACK SCOTT	0.020000383	0.07388995	
253	It Don't Matter To Me	BREAD	0.440204514	0.30648585	
254	Better Place To Be	HARRY CHAPIN	0.250082232	0.17250441	
255	Tennessee Flat-Top Box	JOHNNY CASH	0.440882027	0.16517125	
256	Sangria	BLAKE SHELTON	0.698478645	0.65388334	
257	It's Getting Better All The Time	BROOKS & DUNN	0.36038592	0.45245638	
258	Can't Hold Us Down	CHRISTINA AGUILERA feat LIL' KIM	0.593610249	0.53884389	
259	Did You Ever Have To Make Up Your Mind	LOVIN' SPOONFUL	0.360022186	0.30123641	
260	Dark Necessities	RED HOT CHILI PEPPERS	0.767589233	0.69768835	
261	Anyway You Want It	ENCHANTMENT	0.050001212	0.12994078	
262	Hit Me With Your Best Shot	PAT BENATAR	0.602240022	0.58133186	
263	I Still Believe	BRENDA K. STARR	0.410771816	0.39720897	
264	Just A Little Bit	ROY HEAD	0.010007939	0.09837232	
265	You Are Not Alone	MICHAEL JACKSON	0.64656888	0.61083368	
266	Wild Things	ALESSIA CARA	0.70339257	0.73184726	
267	Just A Simple Melody	PATTI PAGE	0.140000256	0.11936797	
268	A Real Mother For Ya	JOHNNY "GUITAR" WATSON	0.030389881	0.25282624	
269	Little Sad Eyes	THE CASTELLS	0.020002037	0.08500162	
270	Goin' On	BEACH BOYS	0.120006886	0.19372067	
271	Hypnotized	LINDA JONES	0.370478625	0.23133631	
272	The Long Run	EAGLES	0.540016101	0.40982081	
273	Save Up All Your Tears	CHER	0.31072626	0.37772942	
274	Satin Sheets	THE BELLAMY BROTHERS	0.290017811	0.16579787	
275	Heaven Help Us All	STEVIE WONDER	0.300037263	0.42702162	
276	The Other Man's Grass Is Always Greener	PETULA CLARK	0.120048036	0.193443	
277	Theme From Raging Bull	JOEL DIAMOND	1.09E-05	0.1629959	
278	Mr. Melody	NATALIE COLE	0.250005442	0.28814658	
279	I'm Not Gonna Teach Your Boyfriend How T	GLEE CAST	0.370603408	0.47795848	
280	Whiskey Lullaby	BRAD PAISLEY & ALISON KRAUSS	0.667460627	0.47370873	

Figure 5: The figure shows the expected and the predicted values by the baseline model for the test data built through train-test split of the dataset

5 Future Work

The part of the project which is still left for us to do can be broadly divided into two parts:

1. More features: There are several important features which we are still missing and which we believe will impact the evaluation metrics on our models. They are the following:

   (a) Genre: We plan to add this categorical field, because this may have a direct impact on the popularity of a song.

   (b) Release date: We currently have the date on which the song made it to the top of the billboard charts. A large difference between the release date and the date on which it made it to the charts might be indicative of the fact that the song is there to stay.

   (c) Demographics: We are planning to include this external dataset, which will indicate the percentage of people who are in their teens and early adulthood over the years. People have a tendency to latch on to the music of their youth, and a higher percentage value of young adults during a certain era might indicate that the song will endure.

2. Better models: We still have to experiment with different machine learning models their parameters to get the best possible predictions we can possibly get.