Do popular songs endure?

Final Report

Introduction

Determining if a popular song has endured (will endure) over time is a tricky question to answer as the change in popularity of a song since it's time of release depends on a variety of factors, the most important of which is the shift in music taste across generations. We started out by defining a simple baseline model which followed an exponential decay and predicted the current popularity as a function of original popularity and the year of release. The sniff tests on this model correctly proved that song popularity cannot be defined by these factors alone(the model failed to capture correct popularity of older songs of a particular genre/songs from famous artists that had endured). In order to accurately predict current popularity(which is a measure of how well a song has endured) we need to take into consideration factors such as peak position of the song, duration for which it charted at the peak position, artist popularity, genre etc.

With the new models, we have tried to rectify the shortcomings of our baseline model by introducing new features. Before building the new models, we performed an analysis on trends for current popularity of songs that charted at a particular peak position for a given year. We also created multiple models and split our data based on year as well as peak position to correctly identify the specific cases for which our model fails. This report builds up on the work done until the progress report and details the new features, models and exploratory analysis focused around peak position.

Data Scraping

We have extracted data from the following sites :

- Billboard.com[1]: Scraped 100 most popular songs for each year from 1958 to 2017[1]. We use the weekly rank data to get the artist name, song name, entry date and peak position achieved.
- Universal music data (UMD)[4]: To get the number of weeks a particular song stayed on the billboard for.
- Spotify API (spotipy[2]): Spotipy is a lightweight python API for Spotify's web API. For every song, we extracted the song popularity(a metric that can be used as current popularity) and artist popularity.
- YouTube[3]: YouTube's python API was used to get the view count of a particular song based on the song and artist name. Sorted by most relevant song, view count of the first video is taken to be the estimate of popularity of a song.
- Grammy.com: We were able to obtain every artist that has received a Grammy lifetime achievement award. We also scraped the number of nominations and wins across all categories for every artist that features in over song database.
- Wikipedia: We wrote a Wikipedia scraper to obtain the release date for each of the songs in our database.

Estimation of current popularity

Issues with using only Spotify :

- Spotify, in some cases, returns a popularity score of 0 for obviously popular songs. One such example we found in our dataset was the song 'Are you lonesome tonight?' by Elvis Presley.
- Spotify heavily normalizes the popularity scores for songs irrespective of their date of release. This does not allow for any observations about the decay in popularity with passage of time.

In order to rectify this, we added YouTube view count as a contemporary measure of popularity. Our current popularity score is a product of Log of Spotify Popularity and Log of YouTube view count.

We found that adding YouTube count, which is an absolute score till date helps in observing how songs that have released a few decades ago wane in popularity.

Why product of the two ?

After evaluating the values of our popularity measures, we found that the log of YouTube view count was in the range 1-20 and the log of Spotify score was in the range 1-5. Taking a product of the two seemed like a valid assumption considering it brings the current popularity in a range of 1-100 (2.96 - 98.03 in reality). In short, it works for our dataset.

Data preprocessing and cleaning

- Remove rows that have missing column values (i.e no spotify rank or view count or peak position etc).
- Use Label encoders for categorical columns such as artist name.
- Remove rows having original popularity value as 0. The number of rows affected in the dataset was minimal and hence this isn't a major concern

Features Used

For our basic linear regression model, we had used the following features :

- SongId, performing artist: Name of the song, artist
- Song Release Date: Date of release
- Total Weeks: Number of weeks the song stayed on the billboard charts
- Artist Popularity: Value between 0 and 100(most popular). Calculated from the popularity of all the artist's tracks.
- Release Date: Date on which the song debuted. Obtained this information from Wikipedia.
- Original Popularity: Popularity at the time of release. Normalized between a value of 0 - 100(most popular).
- Current popularity: Product of logarithm of Spotify popularity score and YouTube view count.

As mentioned above, we add a few new features(which combine to convey the genre information) for our advanced model to better predict the current popularity.
Genre is a combination of all of these quantitative values rather than just one categorical field.

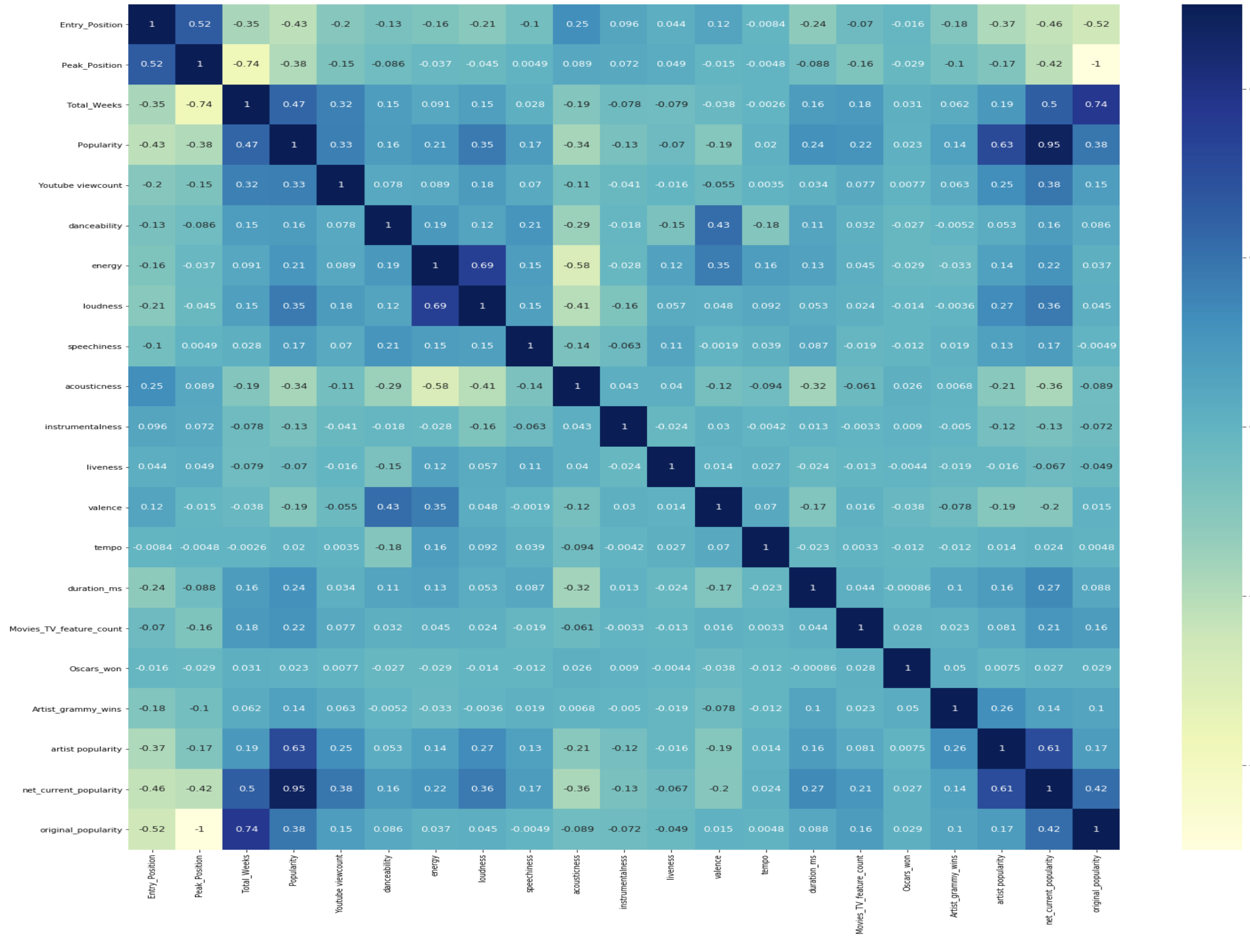
- Valence: Conveys how positive a song can make you feel. Range : 0 to 1 with 1 conveying happiness, euphoria etc and 0 conveying sadness, depression, anger, etc.
- Loudness: Conveys how loud a song is (metal songs will be louder as compared to country songs).
- Danceability: How suitable a track is 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.
- Speechiness: Number of words spoken in a particular track.
- Acousticness: How acoustic a particular song is. Songs with high acousticness will consist mostly of natural acoustic sounds (think acoustic guitar, piano, orchestra, the unprocessed human voice).
- Liveness: Probability that the song was recorded with a live audience.
- Energy: Represents a perceptual measure of intensity and activity. Typically, energetic tracks are fast, loud, and noisy.
- Instrumentalness: Represents the amount of vocals in the song. The closer it is to 1.0, the more instrumental the song is.
- Movies_TV_feature_count: Number of time a song has been featured in movies and tv songs.

Artist Features :

- Artist Popularity: The score of how popular an artist is . A number between 1 and 100.
- Artist Grammy Wins: The number of grammy wins of an artist
- Top Songs Artist 100: The number of songs of an artist in the top 100 charts.

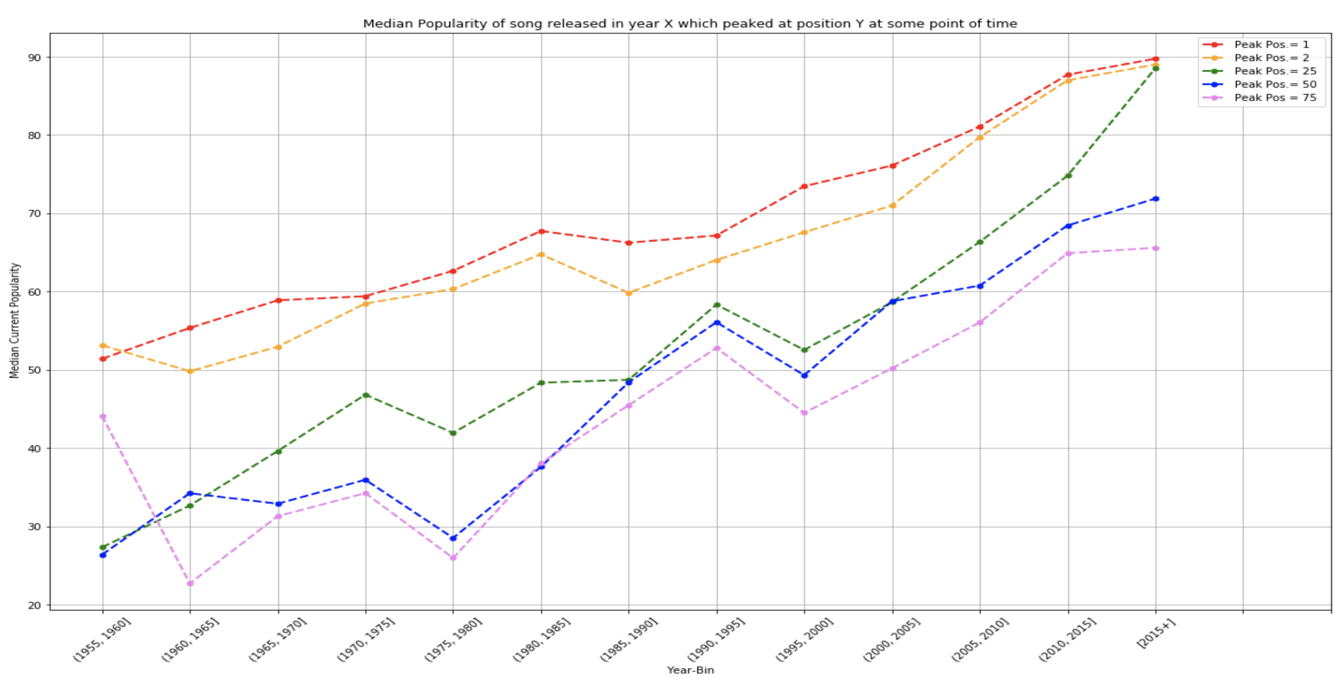
Correlation between numerical features

- Energy and Loudness are highly correlated. Loud songs are more energetic and vice versa.
- While we use peak position as an initial metric of popularity, the number of weeks for which a song remained on the charts(Total_Weeks) highly correlates with the initial popularity(it goes to reason that a song that charted for more duration would be more popular). Below we have checked with our model if this indeed is an important feature.
- We see a good deal of positive and negative correlation between different factors such as speechiness, acousticness, loudness etc. and current Spotify and Youtube popularity. We believe these features should play an important role in determining song endurance.



Song Endurance Modelling

We explore the trends in current popularity of songs that peaked at a particular position across the years. Intuition: the current popularity of songs from two decade ago would be less than songs that released a decade ago. But songs that charted at different positions would wane in popularity differently (a song that had peaked at 1 is much more popular than song that peaked at 5). Hence, a better intuition would be to see how songs that peaked at a particular position over the years have waned in popularity. For this, ideally, as peak position value increases (i.e song popularity is relatively less, say peak 50), the relative trend line for this should be below the trend line for songs that have a lower peak position (i.e more popularity, say peak position 1).

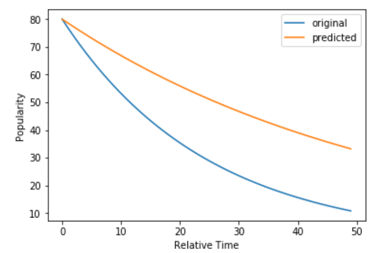


The graph has been plotted as follows : For every 5 year bin, we plot the mean current popularity of all the songs which released at year Y and peaked at a particular position (we have shown peak positions 1,2,25,50,75 to keep the plot clean) on the billboard chart at some point in the future. For lack of a better metric, we can call the set of songs, that peaked at number 1 as the most popular ones for that year bin. As you can see, the trend line for a higher peak position(closer to 1) is almost always above that for a lower peak position(closer to 100).

Brief Description of Models Used

Baseline Model

We started our analysis with a very simple baseline model involving Linear Regression. The initial thought process was that the current song popularity is a power law involving it's initial popularity(correlated with Peak Position) and year of release.



Sniff Test for Baseline Model

1960-1980:

Overperforming **:**

Bustin Surfboards appearance in Pulp Fiction (best movie ever?) explains why this relatively unknown song of the 60's is very popular today. The entire movie soundtrack is amazing and often played on Spotify and Youtube.
**A potential issue: Days of Pearly Spencer was not a very popular song back then but seems to be an over performer here. Wikipedia shows that a cover of this song actually became famous in the 90's. This could imply that there are still some issues with the youtube/spotify popularity API despite all the care we have taken.

Don't stop me now is a perfect example of this. This is a song which did not peak that high in the US when it initially released but through subsequent radio plays , use in advertisements and commercial promotions, the song has gone on to become one of Queen's most popular soundtracks

Underperforming **:**

Bits and Pieces is also a good example of a popular song that did not endure. Let's just say they couldn't capture the magic of another band comprising of 5 gentlemen. We heard this one along with Court Of Love and quite frankly, the rhythm and the beats did not appeal to us. This could be why these popular songs have underperformed and haven't endured.

Genre could be one of the reasons why Look What You've Done To Me by AL Green hasn't remained very popular today. Wikipedia describes the genre of the song as 'Soul'. Soul music dominated the U.S. R&B chart in the 1960s but since then , has seen a drop in popularity.
Smoke From a Distant Fire - a one hit wonder that has probably faded into oblivion possibly due to lack of success of the band as a whole. (Artist's popularity has some weightage on a song's popularity and endurance after all)

1980-2000

Overperforming :

Ozzy's Crazy Train is a very good example of an over performing song as well. Recent usage in movies like Ghost Rider and Megamind could explain the high current popularity of this heavy metal song

Last Night A D.J. Saved My Life is a song which is termed as a **one hit wonder.** It's considered as one of the greatest songs written about being a girl which can explain it's relatively high popularity even today. Mariah Carey's cover of this could be a possible reason as well of why the original song's popularity remains high.
Our very basic model goes a little haywire here. November Rain shouldn't be termed an overperforming song. It's definitely a highly enduring song but we need to evaluate more features to better predict the overperforming songs.

Pearl Jam's Jeremy[5] is an interesting pick here. Grunge in the 90's (read Nirvana) was the real deal but Jeremy is more than a song. It was a song surrounded in controversy and it's music video still remains chilly to watch. However it's underlying theme of depression and guns in school makes it a song that still sees high replays on the radio. Few grunge songs would be over performing and highly enduring but Jeremy isn't a surprise.

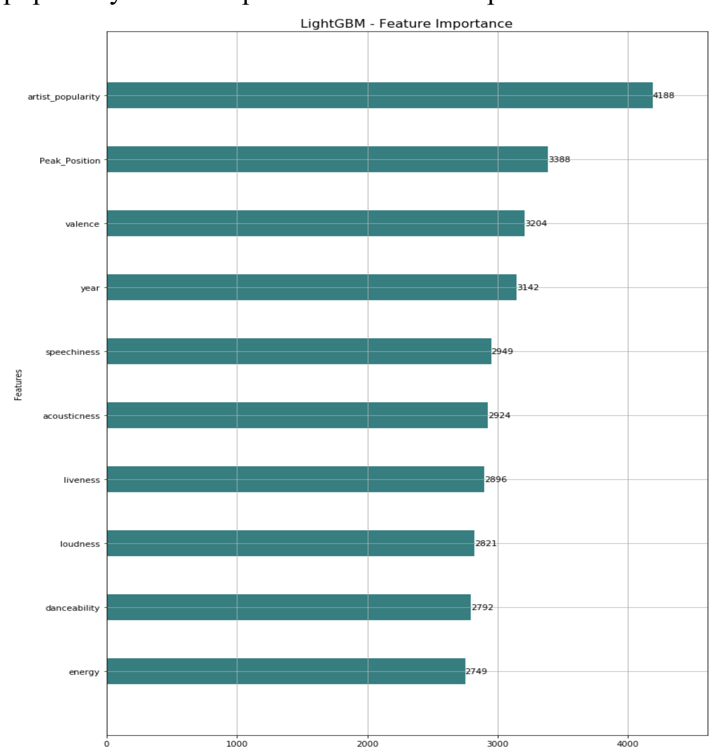
Underperforming :

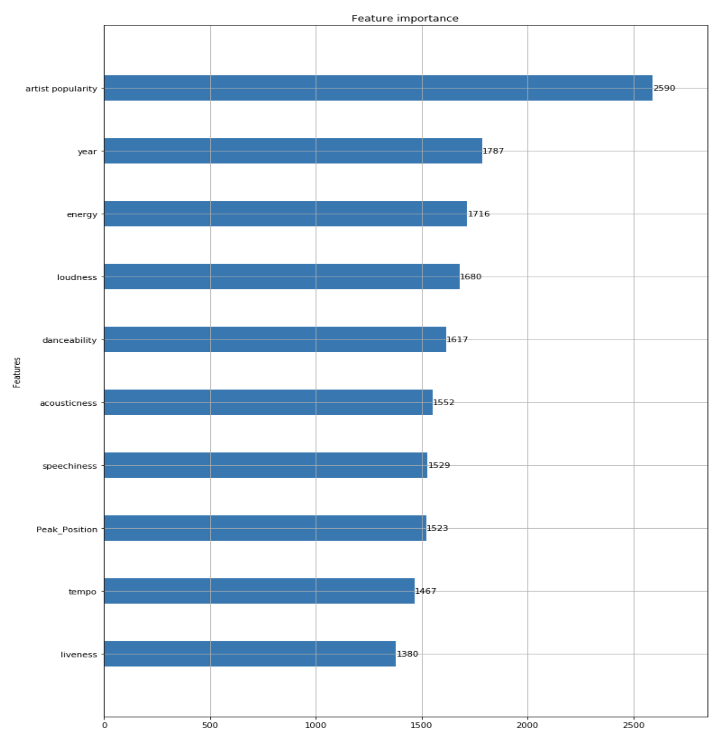
The late 90's probably laid the foundation for the formation of R&B. But one thing we know about this genre is that a lot of songs have a very short shelf life. 50 cent ,Jay-Z and Akon might have captured the space in the later years but the initial years saw a lot of new artists achieve short lived fame. I miss the homies is a good example of this. A song that enjoyed a good run in the year it was released but has since then dropped significantly in popularity.

The biggest concern here was that a simple power law was not generic enough to model the popularity trend for any given song. Hence, the new approach involved measuring the over performing and underperforming songs more closely to the peak position it achieved at any point of time and of course, it's year of release. This is a fair comparison since if the song's current popularity is say 10 or 15 more than the median popularity of all the songs of that year which also peaked at the same position, it can be deemed "over performing" and vice versa for "underperforming". We also collated a lot of different song and artist features mentioned above to more accurately model a song's endurance. These features are passed onto an advanced ML model to gather feature weights and predict current popularity/endurance of songs.

Advanced Model

We have used two different boosting algorithms - *Light GBM* and *XGBoost* to model our features and check just how important different features are. While we do know that a song's current popularity should depend on it's peak position, year of release and the number of weeks it stayed on the charts, we think the other features also play a significant role in shaping its current popularity. The comparison of feature importance for these models is as follows:





Light GBM uses a histogram based algorithm which results in faster training speed and higher efficiency. Since it uses a leaf wise split rather than a level-wise approach, it has better accuracy than any other boosting algorithm. Light GBM has known to have issues with overfitting and as a result we verified our feature important results with another Boosting algorithm - XGBoost.

Analysis: We can clearly see that the current song popularity is heavily dependent on the peak position and the it's year of release. Different song characteristics such as valence , loudness etc are also playing an important role. There are 2 interesting observations here - Artist Popularity plays a major role in determining the song's current popularity. This also raises the question of one hit wonders which we explore in the latter section of the report. On the other hand, width i.e. the number of weeks a song remained on charts is not playing a very important role in determining its popularity according to our model. Here, we don't assume width to be the difference between the date of first and last appearance of the song on the chart. It's the exact total of the number of weeks a song remained on the charts.

Feature Analysis

Below are the plots for the median over and under-performing songs with respect to the general trend lines of certain features for songs that peaked at position 1.

Valence vs Year

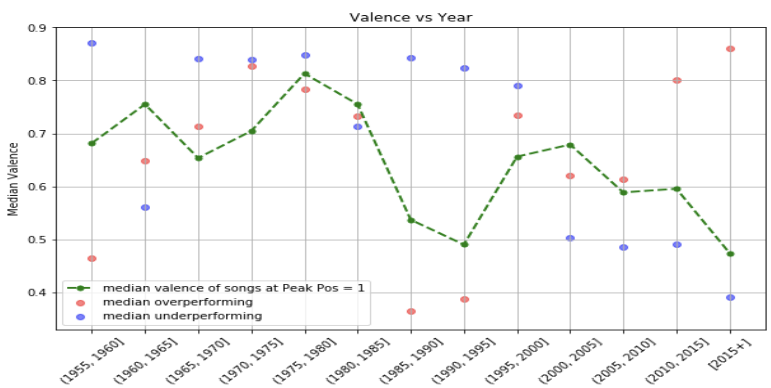


Fig 1

Loudness vs Year

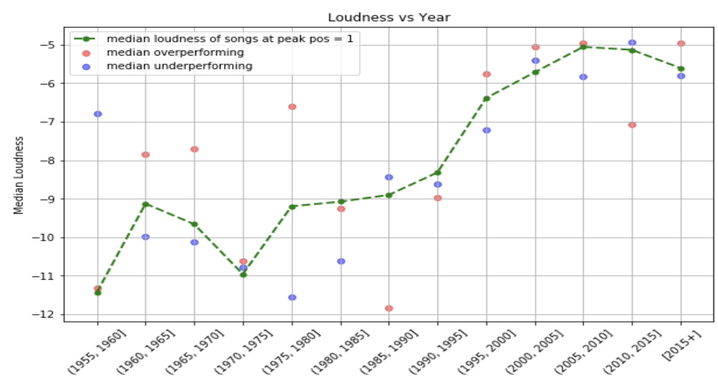


Fig 2

We see that the valence for songs has decreased over the years in Fig 1. There is also a reversal in trend for the over performing and underperforming songs in the last decade. Most underperforming songs have high valence. According to Fig 2, median loudness of songs has increased across the years. Thus we can observe that most over performing songs have high loudness quotient. Also another observation is that under-performing songs have loudness value around and below the median values.

Danceability vs Year

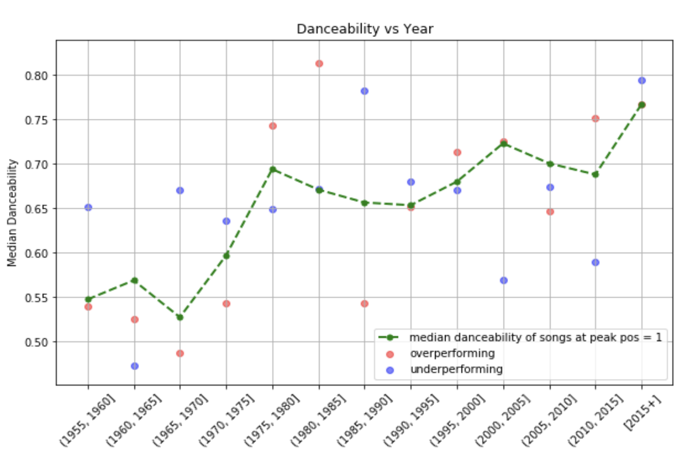


Fig 3

Acousticness vs Year

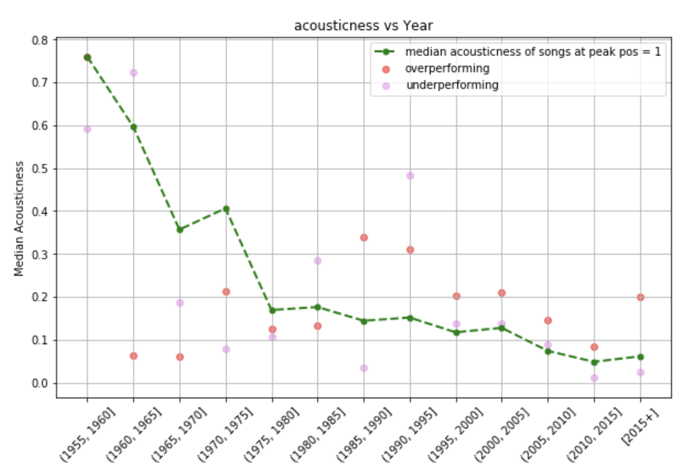


Fig 4

Fig 3 captures the median danceability of the songs across years from 1955-2017. One trend observed is after 1995, all underperforming songs have low danceability factor to it while over performing songs have no trend per se. In *Fig 4,* we can see that median acousticness decreases with time. One acute observation is that all the overperforming songs after 1985 are highly acoustic in nature. There is no exact trend observed for underperforming songs w.r.t acousticness. While most of the them are around the median value, very few of them are above the median acousticness value. After further analysis, we found that there is no single combination of all features pertaining to a song's success. Gloria Gaynor's iconic "*I Will Survive*" features an interesting juxtaposition of low valence with high energy and danceability. Though not as danceable as "*I Will Survive,*" Bonnie Tyler's power-ballad "*Total Eclipse Of The Heart*," which went to Position 1 in 1983, similarly electrified a low-valence song with an energetic production. Bon Jovi's follow-up single, "*Livin*' On A Prayer," also went to Peak Position 1, and features a similar profile of high energy and liveness, contrasted with low valence and danceability.

Model Performance:

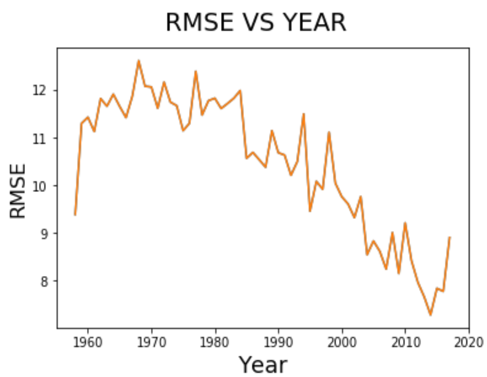


Fig 1

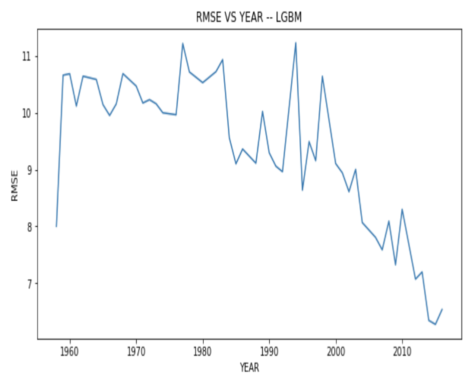


Fig 2

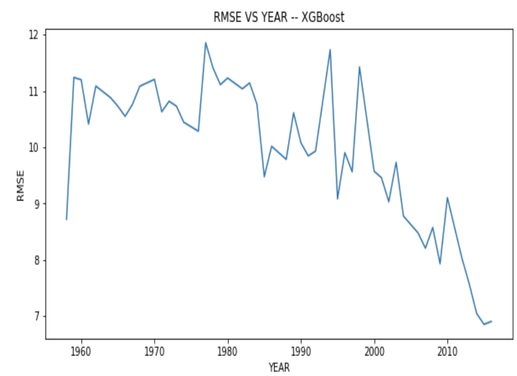


Fig 3

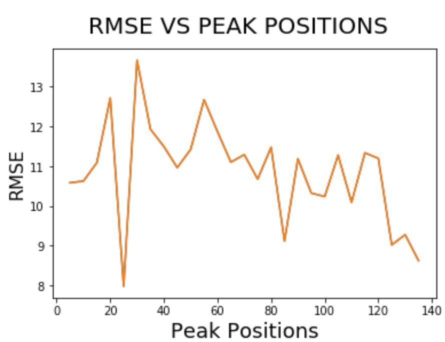


Fig 4

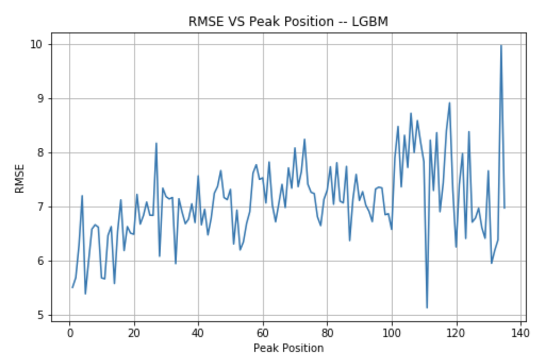


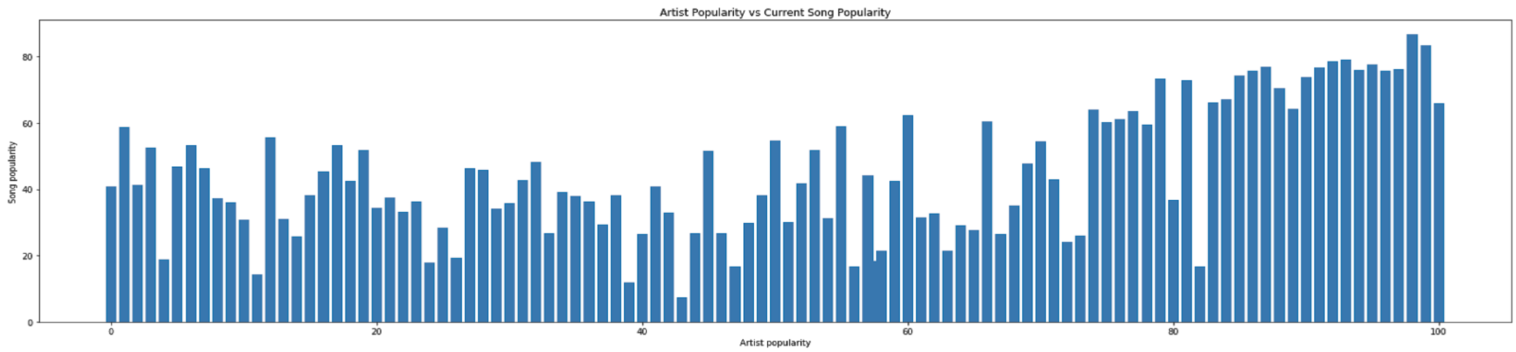
Fig 5

The plots in *Fig 1, Fig 2, Fig 3* display the variation of RMSE with year for the baseline, Light GBM and XGBoost model respectively. While there is an improvement on the RMSE scores for the advanced ML models, the models are more accurately able to predict popularity for current songs as compared to the older ones. The LGBM model result in *Fig 5* shows a considerable improvement on the regression baseline result in *Fig 4.* There is not a massive change in terms of the RMSE (between 6 and 8)among the first 100 peak positions. This is mainly because while our model correctly identifies the popular tracks, it's estimates' vary uniformly from the actual data thus giving an almost constant RMSE value. Slightly inconsistent data for the latter peak positions throws the RMSE values off from it's normalcy.

Analysis on result produced by the best model (Sniff Test):

In this section we try to correlate the insights we observe in real life with the results obtained by our model and try to justify the same.

Do songs of popular artists endure more?

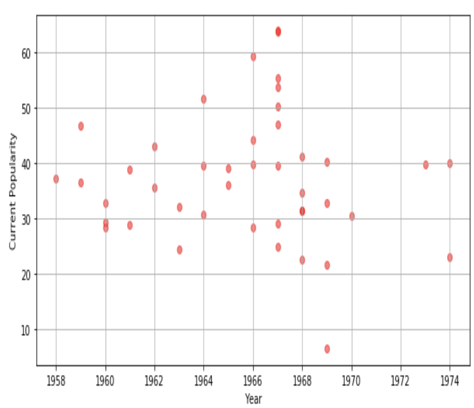


As we saw in the feature analysis section, artist popularity is a crucial feature(this was also observed in the sniff tests of our baseline model). With this plot, we can say that in general, songs by highly popular artists have endured more. However when we look at the lower range of numbers for artist popularity, we can see that the trend is somewhat less uniform. This could be explained by 'one hit wonders' as well as artists who were periodically popular.

Do older songs of artists endure more than their newer songs ?

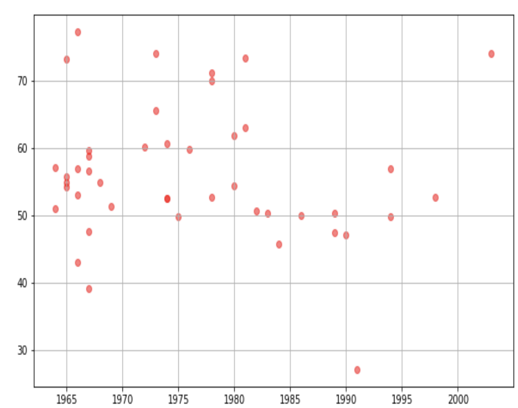
For songs in the billboard charts for famous singers, we have plot it's current popularity based on our metric and it's year of release.

Frank Sinatra

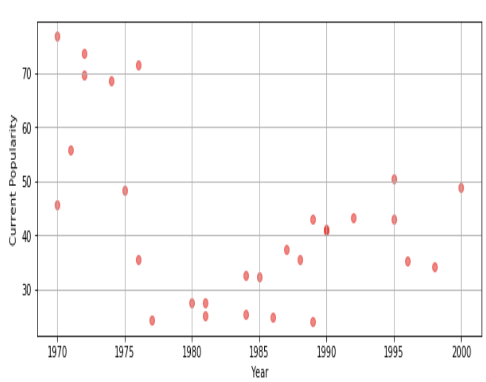


The Beatles

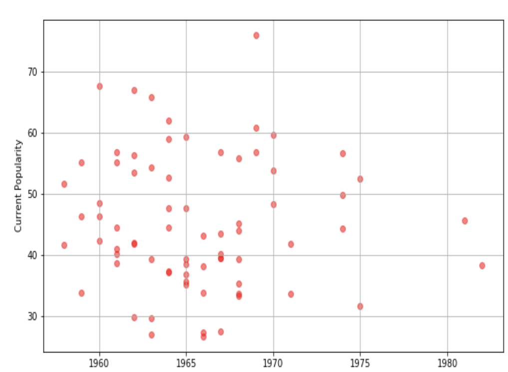
Rolling Stones



Elton John



Elvis



It is not easy to conclude if older songs of popular artists endure more than their recent songs. Sinatra achieved his peak mid career. When we look at the Stones or Elton John, most of their older works are indeed more popular than their latest works. Elvis on the other hand presents a completely different picture -- he has managed to produce a lot of consistent hits throughout his musical span. For the Beatles, they entered the charts in 1964 and just dominated the scenes. However, a lot of their songs are currently not as popular as some of the perennial hits. In general, few specific albums of an artist get more popular than the rest and this can be seen from the graphs.

Top Songs - Data vs Model

For the year 1970, we decided to explore if the top songs predicted by the LGBM model is similar to what the current popularity data shows. The model was trained on songs of all years excluding 1970, which then formed the test set. The results were as follows :

Model

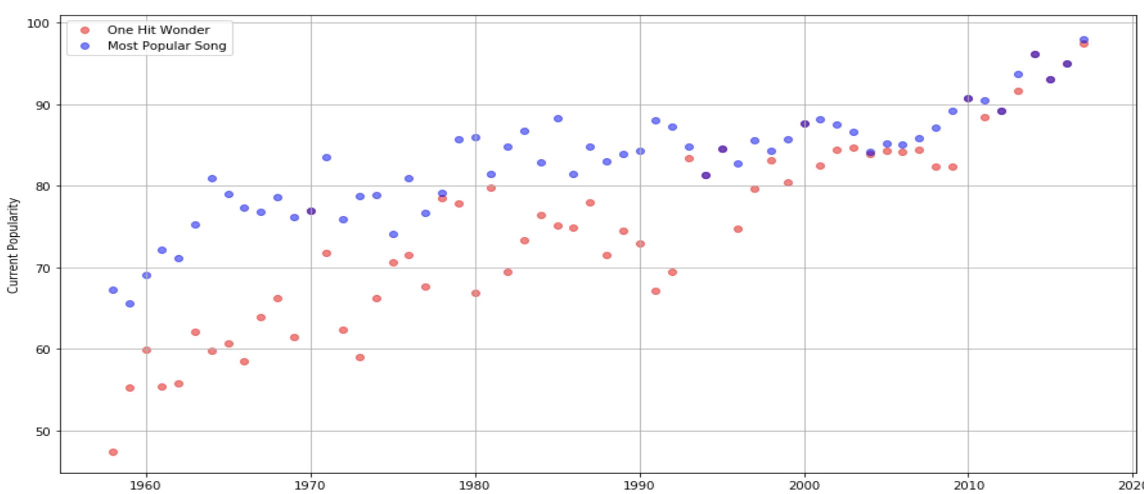
Song	Artist
Flesh And Blood	JOHNNY CASH
Immigrant Song	LED ZEPPELIN
Paranoid	BLACK SABBATH
Cecilia	SIMON & GARFUNKEL
Cracklin' Rosie	NEIL DIAMOND
ABC	JACKSON FIVE
The Love You Save	JACKSON FIVE
Your Song	ELTON JOHN
Tears Of A Clown	SMOKEY ROBINSON & THE MIRACLES
Wigwam	BOB DYLAN

Data

Song	Artist
In The Summertime	MUNGO JERRY
Your Song	ELTON JOHN
Black Magic Woman	SANTANA
Mississippi Queen	MOUNTAIN
ABC	JACKSON FIVE
Cecilia	SIMON & GARFUNKEL
All Right Now	FREE
Immigrant Song	LED ZEPPELIN
Yellow River	CHRISTIE
Lola	KINKS

The model has managed to identify a lot of the top popular songs but it fails to pick those up where the Artist Popularity isn't very high. These are basically the outliers -- one hit wonders. According to our model, artist popularity plays a major role in determining the song popularity and that is why it picks songs of those artists which have achieved top songs in the past. However a thing to note here is that the songs picked by our model were indeed very popular - just not in the top 10 for 1970.

One Hit wonders



A one hit wonder is essentially when an artist has had only one hit song. In our dataset, we look for a song by an artist which charted on the billboard but in the subsequent years, no other songs by that artist made the cut. Among these artists, we then choose the song(or you can call artist) whose current popularity is maximum. This results in giving us the "one-hit-wonder" artist for that particular year and the song which shot him to fame.

As we clearly see, in most cases, the current popularity of one hit wonders is less than that of the most popular song from that particular year which implies that one hit wonders do not endure as well as other songs by popular artists. For some years, the popularity values come close, but that could be for a variety of reasons.

List of popular One Hit Wonders as per our model(songs released before 2000)

We do not consider songs after 2000 as we feel the timespan is too short to classify a song as a one hit wonder.

	Title_x	Artist_x	Entry_Date_x	Entry_Position_x	Peak_Position_x	Total_Weeks_x	Popularity_x	Youtube viewcount_x	danceability_x	energy_x	key_x	loudness_x	mode_x	spee
2	Theme From 'A Summer Place'	PERCY FAITH	1/4/60	101	1	22	43	7668022	0.326	0.326	0.0	-15.144	1.0	0.02
12	In The Summertime	MUNGO JERRY	7/3/70	103	3	14	62	115442691	0.754	0.449	4.0	-14.013	1.0	0.06
17	Love Hurts	NAZARETH	10/25/75	109	104	3	54	44858962	0.477	0.347	7.0	-11.847	1.0	0.02
18	Don't Go Breaking My Heart	ELTON JOHN & KIKI DEE	7/3/76	66	1	20	67	22819643	0.743	0.858	5.0	-7.790	1.0	0.04
19	Gonna Fly Now (Theme From "Rocky")	BILL CONTI	4/23/77	84	1	20	54	21292043	0.438	0.812	0.0	-3.700	1.0	0.03
20	You're The One That I Want	JOHN TRAVOLTA & OLIVIA NEWTON-JOHN	4/1/78	65	1	24	58	229759312	0.743	0.774	0.0	-5.916	1.0	0.09
21	Boogie Wonderland	EARTH, WIND & FIRE & EMOTIONS	5/12/79	69	6	16	70	86753759	0.802	0.756	2.0	-10.791	0.0	0.03
23	Under Pressure	QUEEN & DAVID BOWIE	11/7/81	80	29	15	70	135700749	0.667	0.705	2.0	-7.657	1.0	0.04
26	Somebody's Watching Me	ROCKWELL	1/28/84	73	2	19	71	57487301	0.767	0.712	1.0	-4.128	0.0	0.03
28	Live Is Life	OPUS	1/25/86	88	32	16	59	86795903	0.633	0.952	9.0	-4.274	0.0	0.03

What we think we did right

- We were already considering peak position as a part of our original popularity metric in the baseline model. The change in approach suggested by the TA further helped us see why this would be beneficial.
- Appending YouTube views to current popularity metric. Using just spotify rank would be a biased metric because their normalization. We did not find any concrete information on the normalization process.
- By performing sniff tests on our baseline model we identified what additional features would help us improve the accuracy of our predictions. The feature analysis part more or less validates our intuition.
- Chart the popularity trends based on Peak Position across years .
- Perform an analysis on the results of our best model(sniff tests). This helped us seek out trends and patterns for the questions Professor has put on Piazza.
- Fiddle with multiple models and focus on their feature importance breakdown. Plot graphs comparing the over/under-performing songs with respect to median trends of important features.
- Read about the history of over and under performing songs identified with respect to our model and analyse why they did so and if their endurance or lack of it can be *qualitatively* justified.
- Split the entire dataset into year-wise/peak-position-wise sets to get a better analysis.

Conclusions and Future Work

Throughout our project, we have given considerable importance to enhancing our work based on the analysis of results of our models. We feel that this has helped us build a good model iteratively. The endurance of a given song depends on a multitude of factors. Starting from the most basic features and adding new features to try and capture the outliers predicted by our models has helped us understand what aspect a particular feature encapsulates. It is difficult to capture the real world fluctuations of endurance of songs in its entirety but our model does a decent job of covering most cases. The analysis on our best model aligns with the patterns we see in the real data to a great deal.

Further work that could be done:

- We haven't yet been able to explore how the shift in music taste over the years affects the song endurance. This would be an interesting aspect to explore.
- Another improvement : evaluating how adding a genre field(categorical) would affect the model accuracy.

References

[1] Billboard Website, Billboard Data

[2] Spotipy, Python API for extracting Information from Spotify

[3] Youtube API, Python API for extracting Information from Youtube

[4] UMD, Universal Music Database