Do popular songs endure?

Progress Report

Introduction

We have focused our efforts on building and analyzing a baseline model that predicts the song popularity as a function of the song's original billboard ranking and the current ranking(in our case, spotify and YouTube data). The set of initial features is minimal to avoid clutter and observe if there is some sort of decay evident in the popularity trend of a song over the years. Since the number of features is less and the output is basically some sort of decay over time, we expected our model to perform badly for older songs which may have peaked due to freak events(basically, outliers. Our sniff test is an evidence of the same). The general idea upto this point has been to analyze the outliers of the baseline model to see if we can capture certain trends. We understand that we cannot accurately predict the endurance of a song unless we can predict its current popularity.

In order to stay true to the instructor's emphasis on first building a correct baseline model, we have avoided adding additional features such as tempo, beats per minute, valence etc. This progress report entails the reasoning behind our choice of metrics and our analysis of the model predictions explained via a sniff test. We actually put in quite a bit of effort into our sniff test, and it turned out to be fun!

Data Scraping

For our baseline model, we extracted the data from the following sites :

- Billboard.com[1]: We scraped the 100 most popular songs for each year from 1958 to 2017 from billboard[1]. Another dataset we made included top 100 songs for each week starting from 1958. We use this to get the artist name, song name, entry date and peak position achieved.
- Universal music data (UMD)[4]: In order to get the number of weeks a particular song stayed on the billboard, we used the UMD data.
- Spotify API (spotipy[2]): Spotipy is a lightweight python API for Spotify's web API, which provides access to Spotify's music database. 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.

Estimation of current popularity

Initially, we had decided to use the spotify popularity as the current popularity. We found a few issues with the same :

- 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 these shortcomings, we decided to add YouTube view count as a contemporary measure of popularity. Our current popularity score is a product of the following :

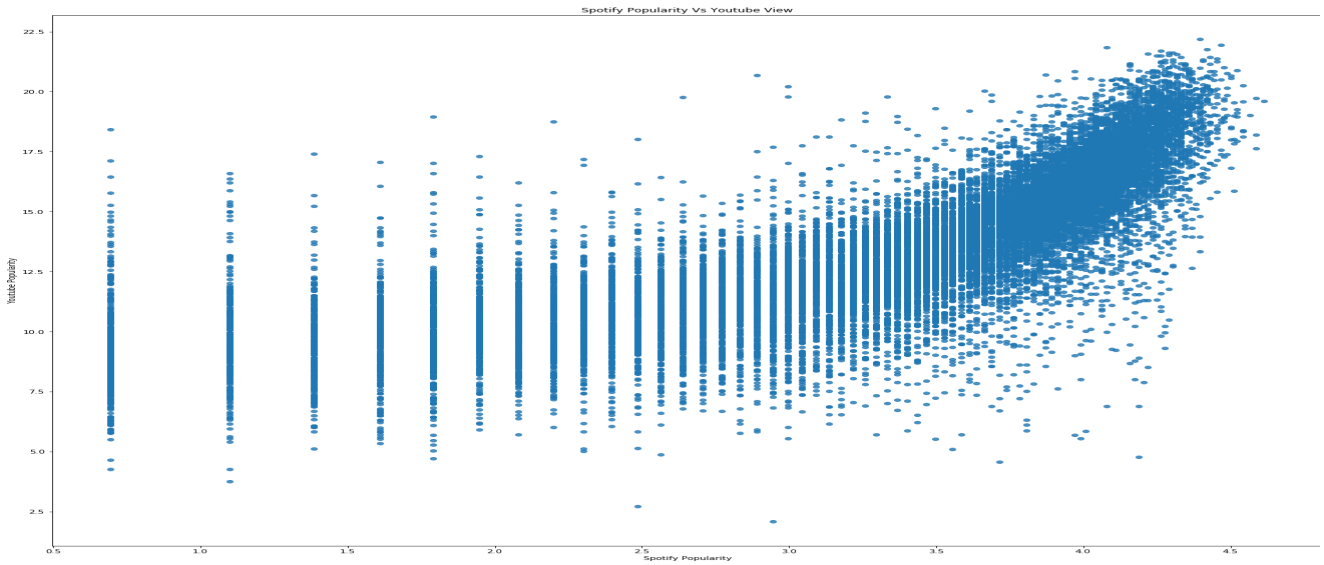
- Log of Spotify Popularity: Log of spotify scores, ranging between 0 to 100 with 100 being the most popular (based on the number of downloads and stream counts)
- Log of YouTube view count: Log of number of views on the YouTube video searched as mentioned above.

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. There are many obvious outliers that we will detail in our description of the sniff test.

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

Scatterplot showing a positive correlation between the popularity measures



Features

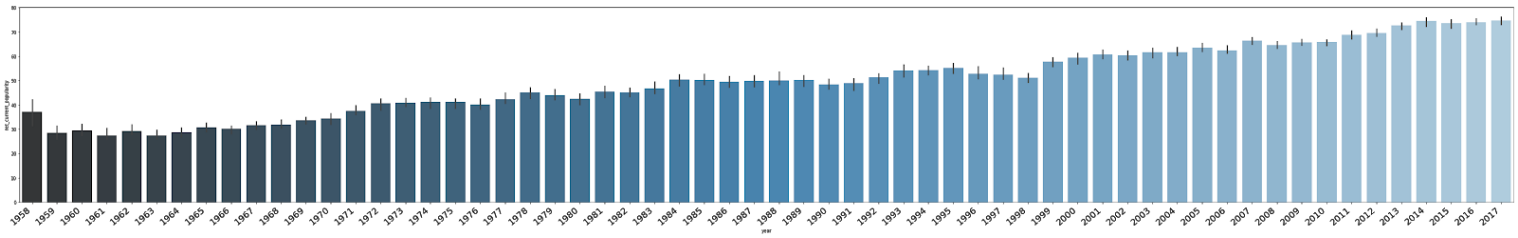
As mentioned in the introduction, in order to correctly analyze the baseline model that would predict the current popularity based on the original popularity and the time of release, we have limited our features to :

- 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: Spotify rank for how popular the artist is
- Original Popularity: Popularity at the time of release. Normalized between a value of 0 - 100 with the most popular songs having values closer to 100 and least popular songs having values closer to 0
- Current popularity : Product of logarithm of Spotify popularity score and youTube view count with most popular songs being closer to 100 (in our case, Shape of you by Ed Sheeran : 98.03) and least popular songs being closer to 0 (in our case, The best years of my life by Eddie Floyd : 2.96)

Exploratory data analysis

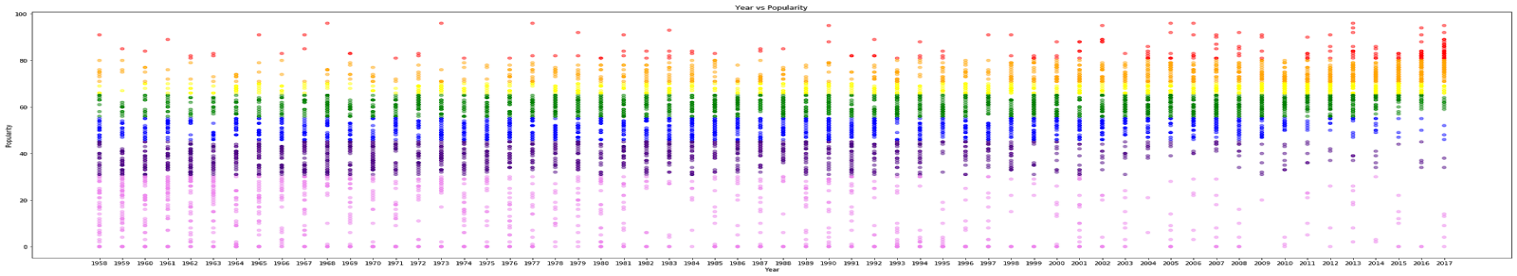
Our first and foremost analysis involves the median current popularity of songs released in a particular year. The trend is mostly inline with what we assumed - songs which released earlier in time have relatively lower current popularity, power law in play. However, there are a lot of songs which will defy this general trend and our work intends to find these over-performing and underperforming songs.

Median current popularity of songs from 1958 - 2017



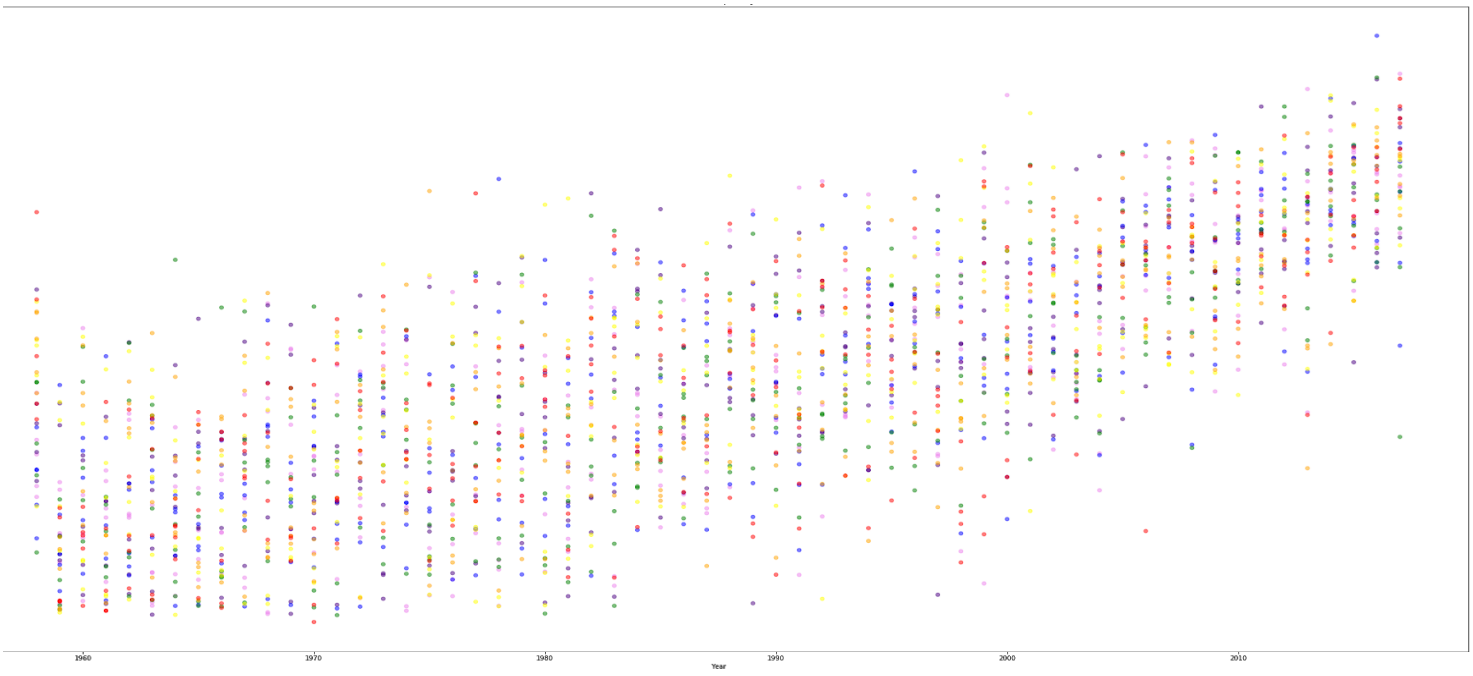
We came up with the next plot based on Professor Skiena's post on Piazza. While the Professor pointed out that we had erroneously assigned colors based on popularity rather than rank, one insight we could retrieve from this wrong plot was that fewer songs of earlier years are able to break into the top bracket of popularity (possibly due to our metric being just spotify rank which is normalized - this led us to reconsider our current popularity metric) - red. However, the presence of a few songs from yesteryears at high popularity values confirms the fact a lot of these songs have endured , peaked late or both.

Plot of current spotify popularity for top 100 songs of each year with a rainbow hue with scores <30 assigned violet and those above 80, red - the wrong rainbow which nevertheless led to some insights.



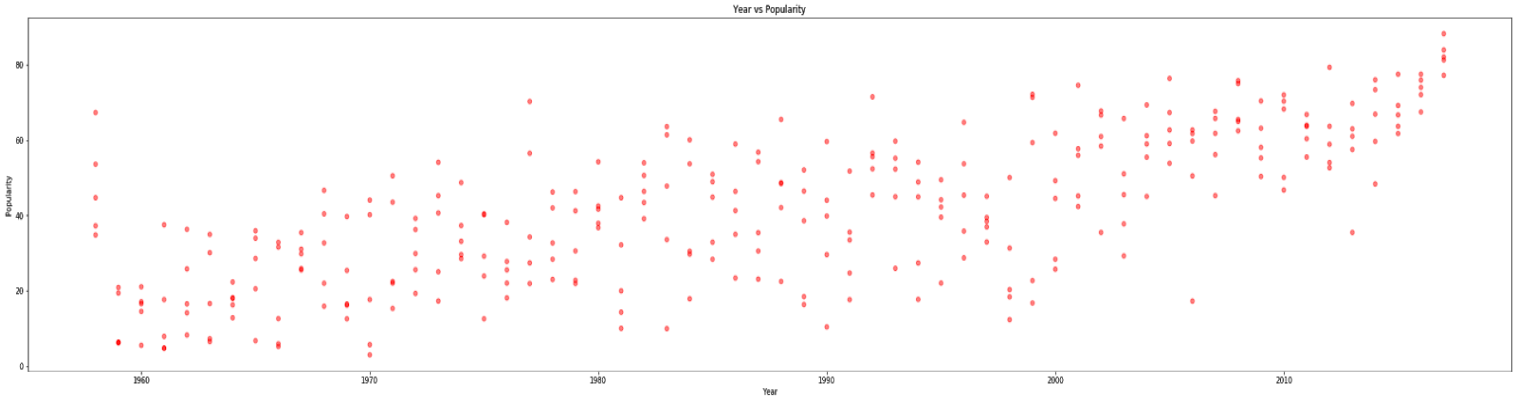
We were further advised to plot these points a little differently - use colors to represent the original popularity(peak position). Unfortunately, the plot seems a little cluttered to draw any valid conclusions from it. So, we decided to break the plot down into songs which were most and least popular to see if we could detect some sort of a power law.

Plot of current spotify popularity for top 35 songs of each year with the color representing it's original popularity. Violet represents the bottom 5 popular songs while Red represents the top 5 popular songs



On plotting how currently popular the most highly ranked songs over each year are, we see that more song of the recent years that were ranked highly are more popular. This collaborates the data plotted in the bar graph above. This shows us that it os not necessary that a popular song might endure over time. As the graph shows, quite a few highly ranked songs have a low current popularity score.

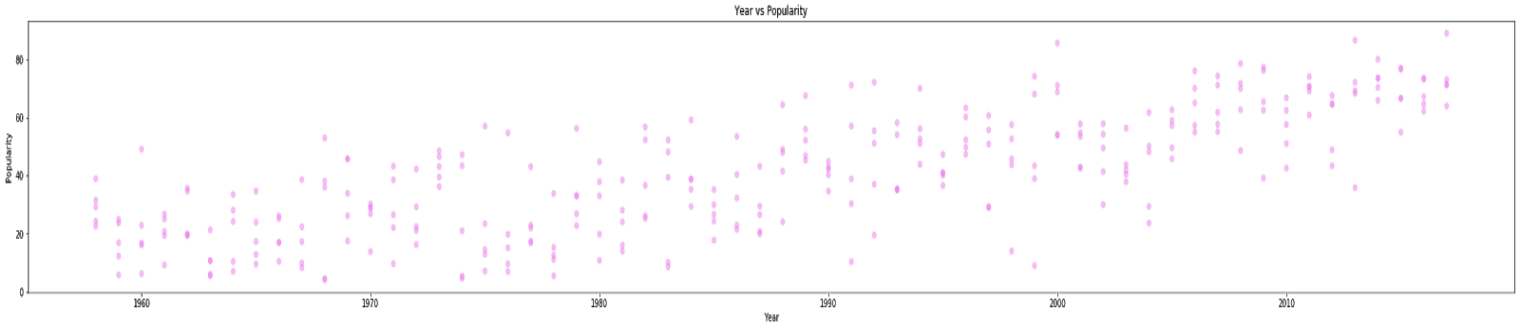
Plot of the current popularity of the most popular songs of each year



On plotting how currently popular the lower ranked songs over each year are, we still see that lower ranked songs over recent years still tend to have a higher popularity than their older peers (which could again be attributed to skewed online presence and lesser decay time). The interesting aspect to observe here is how lower ranking songs have a higher current popularity which conveys how certain songs have endured despite not being too popular initially

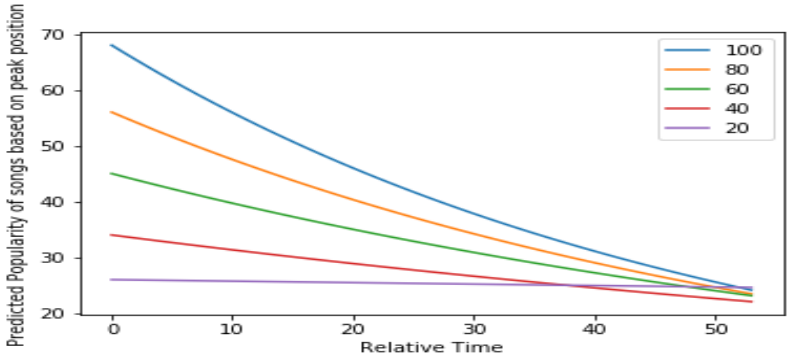
.

Plot of the current popularity of the relatively lesser popular songs of each year



The next plot captures the ideal median decay rate for songs with various peak position ranges as per our baseline model. The original median decay trend observed was more or less similar to the predicted one. This is how the song popularity would ideally wane but we know that isn't the case. Access to some sort of intermediate data over the years would help us get a more realistic view of the waning popularity.

Plot of popularity predicted by model based on peak position with respect to year



Baseline Model - Linear regression model

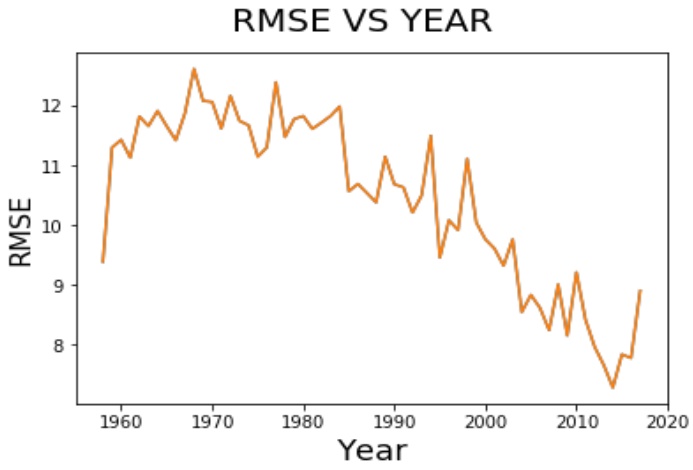
Our baseline model is a simple linear regression model. The motivation behind this is to find the intuition behind other features except the original popularity and time of release that may help in improving the prediction of the current popularity.

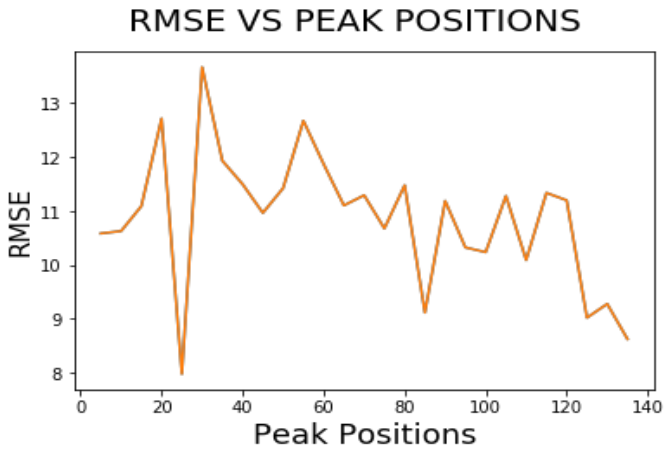
- We split our dataset into 80% training data and 20% testing data. Furthermore, we split our data into multiple sets both year-wise and decade-wise (to evaluate average error over the years, to perform sniff tests). We wanted to see if we could pinpoint particular cases where our model predictions fail miserably.
- A regression model tries to fit the data along a curve. As our model predicts the current popularity based mainly on the above two features, it fails to accurately predict the popularity for older songs that are currently famous. We believe the reason for this is that the regressor decays the original popularity for the song over time in absence of any other features that might boost the representation of a song's endurance over the years (such as danceability, tempo etc)

Error Statistics

- Split the dataset based on the year of release and for each sub-dataset, have an 80-20 split to find the RMSE value for the model over different years.
- The accuracy for model increases over time. The baseline model faces obvious issues in predicting over and under-performing songs. For songs that follow a general decay curve, the predicted popularity is more or less near the defined current popularity values but in cases where a song peaked significantly in later years, the model fails to capture the increase in popularity.
- Another reason for better accuracy over the latter years would be that the songs have been on YouTube since their inception and the number of views would closely correlate to their popularity. Moreover, after 2012, billboard also started considering YouTube views as a metric for their song rankings.
- The error for peak positions was averaged over 5 positions, 0-5, 5-10, and so on. We could not infer anything meaningful from this particular plot.

Plots of Rmse Vs Year and Rmse Vs Peak Billboard Positions





Analysis of actual-decay vs predicted-decay rates

The next set of graphs plot the actual vs predicted median decay over the span of a set of decades. We used the formula P = M * c^y to compute these as for most songs, the popularity does undergo an exponential decay. As mentioned above, this obviously fails for over-performing songs. As mentioned earlier, we realize that no song would follow an ideal decay but these plots reinforce our error statistics where for the recent years, our prediction accuracy increases due to the reasons outlined in the previous section.

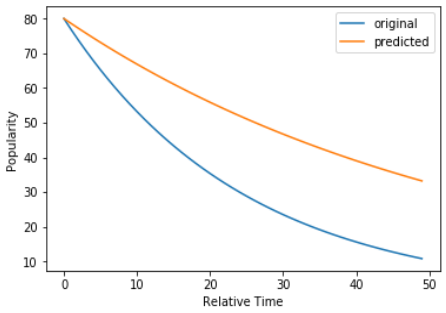


Fig 1(Years: 1960-1970)

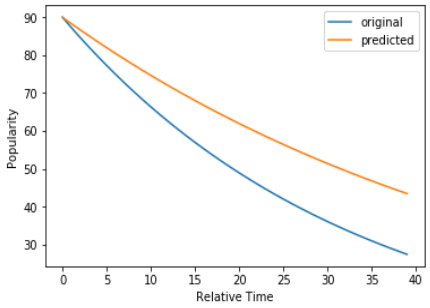


Fig 2(Years: 1970-1980)

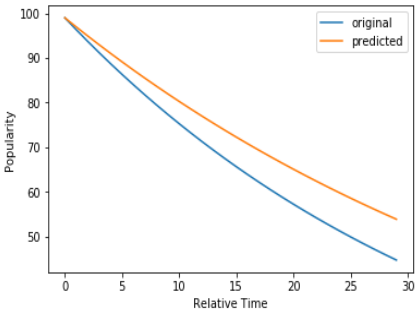


Fig 3(Years: 1980-1990)

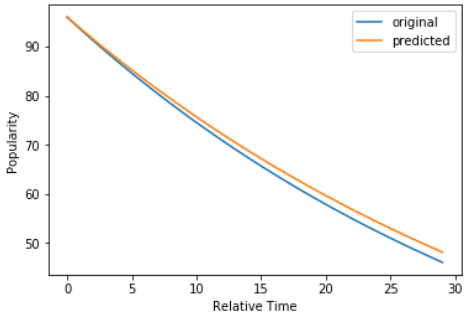


Fig 4(Years: 1990-2000)

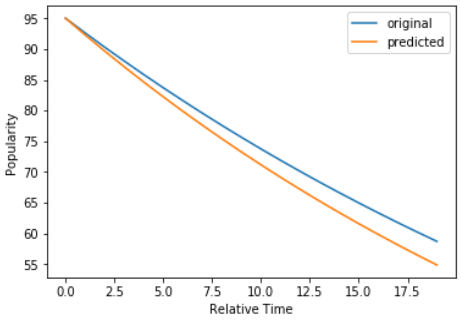


Fig 5(Years: 2000-2010)

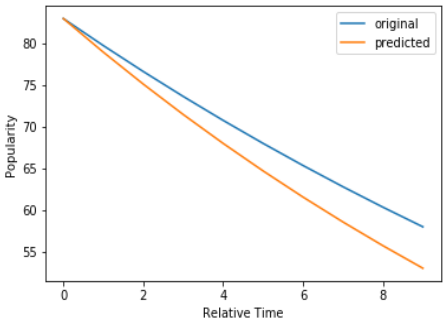


Fig 6(Years: 2010-2018)

Sniff test: over-performing and underperforming songs for each decade

In order to identify these songs, we compared the current popularity defined using spotify and YouTube and the popularity that our model predicted. Below listed are the 10 over-performing and underperforming songs for each decade starting from 1960 to 2000. We also tried researching a bit about the songs to see if there was any reason or any record available to showcase a late bloom or quick decline of the song's popularity.

1960 - 1970

Overperforming songs

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.*

	Title	Artist	Entry_Date	Total_Weeks	artist popularity	log_youtube	log_spotify	net_current_popularity	original_popularity	year	Artist_Nam
20632	After Laughter Comes Tears	7217	10/10/64	1	43.000000	15.534576	3.850148	59.810411	2	1964	WENDY RE
11088	L-O-V-E	4441	9/26/64	4	57.778352	16.737838	4.219508	70.625436	55	1964	NAT KING
15245	Bustin' Surfboards	6749	11/17/62	3	0.000000	14.747400	3.218876	47.470049	34	1962	THE TORN
6825	Louie Louie	6389	9/18/65	3	51.000000	15.942793	4.110874	65.538812	30	1965	THE KINGS
6108	Tip-Toe Thru' The Tulips With Me	6893	4/27/68	2	46.000000	16.786372	3.828641	64.269000	28	1968	TINY TIM
18783	(We're Gonna) Rock Around The Clock	590	6/15/68	2	57.778352	16.622999	3.988984	66.308876	18	1968	BILL HALE COMETS
6818	Break On Through (To The Other Side)	6187	4/8/67	1	78.000000	16.642589	4.262680	70.942028	11	1967	THE DOOR
18894	Days Of Pearly Spencer	1550	6/1/68	1	31.000000	14.456230	3.663562	52.961288	2	1968	DAVID McV
7096	Everybody's Talkin'	4561	8/10/68	5	63.000000	16.325401	4.110874	67.111663	23	1968	NILSSON
15379	Please Come Home For Christmas	1128	12/22/62	1	40.000000	15.296959	3.583519	54.816944	28	1962	CHARLES

Underperforming songs

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.

	Title	Artist	Entry_Date	Total_Weeks	artist popularity	log_youtube	log_spotify	net_current_popularity	original_popularity	year	Artist_Name	predicted_popularity	relative_pop
6567	Bits And Pieces	6164	3/28/64	12	57.778352	14.981808	0.693147	10.384598	132	1964	THE DAVE CLARK FIVE	40.553757	-30.169159
20182	Court Of Love	6765	9/21/68	9	57.778352	13.074257	0.693147	9.062384	111	1968	THE UNIFICS	39.417649	-30.355265
1745	In The Still Of The Night	4753	3/15/69	7	65.000000	10.115772	0.693147	7.011719	72	1969	PAUL ANKA	37.510060	-30.498341
14308	The Lonely Bull (El Solo Torro)	6742	10/20/62	15	57.778352	14.350278	0.693147	9.946855	130	1962	THE TIJUANA BRASS featuring HERB ALPERT	40.447006	-30.500152
20102	Forgot To Remember	2263	11/15/69	2	82.000000	9.308102	0.693147	6.451885	32	1969	FRANK SINATRA	37.059293	-30.607409
15307	Won't Find Better (Than Me)	6485	12/20/69	11	57.778352	9.629577	0.693147	6.674714	79	1969	THE NEW HOPE	37.417229	-30.742515
20007	Little Honda	6331	9/5/64	13	57.778352	12.766100	0.693147	8.848786	127	1964	THE HONDELLS	40.427442	-31.578656
4020	Do It Again A Little Bit Slower	3244	4/15/67	13	57.778352	10.582181	0.693147	7.335009	118	1967	JON AND ROBIN AND THE ON CROWD	40.968135	-33.633126
17228	When The Lovelight Stars Shining Through His Eyes	5816	11/23/63	12	70.000000	11.586678	0.693147	8.031273	113	1963	SUPREMES	42.119863	-34.088591
3973	Release Me (And Let Me Love Again)	2019	4/1/67	15	61.000000	11.332482	0.693147	7.855078	132	1967	ENGELBERT HUMPERDINCK	44.176557	-36.321478

1970 - 1980

Overperforming songs

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 (Please listen to this one if you haven't)

	Title	Artist	Entry_Date	Total_Weeks	artist popularity	log_youtube	log_spotify	net_current_popularity	original_popularity	year	Artist_Name	predicted_popularity	relative_pop
20481	Wuthering Heights	3382	11/4/78	1	64.0	17.565250	4.127134	72.494148	28	1978	KATE BUSH	34.801016	37.693132
353	Love Hurts	4459	10/25/75	3	61.0	17.619034	4.007333	70.605340	32	1975	NAZARETH	33.621244	36.984095
16852	Don't Stop Me Now	4992	2/17/79	4	87.0	19.344702	4.290459	82.997658	50	1979	QUEEN	46.472288	36.525370
16236	Daddy Cool	799	1/22/77	5	68.0	18.387024	4.174387	76.754560	71	1977	BONEY M	41.125422	35.629138
6444	Solsbury Hill	4840	4/30/77	5	33.0	16.294675	4.025352	65.591796	68	1977	PETER GABRIEL	30.642609	34.949186
14491	Highway To Hell	113	10/13/79	10	85.0	19.454486	4.406719	85.730460	89	1979	AC/DC	51.156873	34.573587
312	Dream On	134	10/20/73	9	80.0	18.183265	4.330733	78.746870	77	1973	AEROSMITH	44.667859	34.079012
9299	In The Summertime	4389	7/3/70	14	53.0	18.564285	4.143135	76.914333	133	1970	MUNGO JERRY	43.469091	33.445242
9385	Ain't No Sunshine	600	7/17/71	16	73.0	19.296708	4.330733	83.568897	133	1971	BILL WITHERS	50.158170	33.410727
1142	Sympathy	5050	4/11/70	3	32.0	14.322645	3.663562	52.471893	15	1970	RARE BIRD	19.996448	32.475444

Underperforming songs

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 ong 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)

	Title	Artist	Entry_Date	Total_Weeks	artist popularity	log_youtube	log_spotify	net_current_popularity	original_popularity	year	Artist_Name	predicted_popularity	relative_pop
14005	Sugar Pie Guy	6379	9/28/74	17	57.778352	13.049801	0.693147	9.045433	89	1974	THE JONESES	43.701415	-34.655982
12160	The Next Hundred Years	163	12/10/77	9	57.778352	8.481980	0.693147	5.879261	87	1977	AL MARTINO	40.916090	-35.036829
14806	What Am I Crying For	1641	10/21/72	13	57.778352	8.303009	0.693147	5.755208	97	1972	DENNIS YOST AND THE CLASSICS IV	41.061228	-35.306020
9469	I'm Gonna Let My Heart To Do Walking	5816	5/29/76	14	70.000000	11.769589	1.098612	12.930215	96	1976	SUPREMES	48.261163	-35.330948
11770	Skin Tight	6506	8/31/74	12	57.778352	13.642298	0.693147	9.456121	123	1974	THE OHIO PLAYERS	45.805886	-36.349765
4036	Look What You've Done For Me	157	4/1/72	12	71.000000	14.210742	0.693147	9.850136	132	1972	AL GREEN	48.471186	-38.621050
4032	The Lion Sleeps Tonight (Wimoweh)	5223	1/1/72	17	57.778352	12.249869	0.693147	8.490962	133	1972	ROBERT JOHN	47.280330	-38.789368
11368	Keep On Truckin' (Part 1)	1935	8/25/73	19	51.000000	10.122101	0.693147	7.016106	135	1973	EDDIE KENDRICKS	45.839025	-38.822919
8682	Smoke From A Distant Fire	6639	6/18/77	18	57.778352	14.570714	0.693147	10.099649	127	1977	THE SANFORD / TOWNSEND BAND	49.965957	-39.866308
15454	Theme From "Close Encounters Of The Third Kind"	3196	12/24/77	14	79.000000	10.591622	0.693147	7.341553	123	1977	JOHN WILLIAMS	54.065905	-46.724352

1980-1990

Overperforming songs

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.*

	Title	Artist	Entry_Date	Total_Weeks	artist popularity	log_youtube	log_spotify	net_current_popularity	original_popularity	year	Artist_Name	predicted_popularity	relative_pop
12213	This Must Be The Place (Naive Melody)	6731	11/26/83	8	1.0	16.208891	4.025352	65.246489	74	1983	THE TALKING HEADS	26.083596	39.162893
11136	Just An Illusion	2748	7/3/82	4	51.0	16.650372	4.007333	66.723588	30	1982	IMAGINATION	33.661758	33.061830
8805	White Wedding	624	11/27/82	2	68.0	17.088737	4.110874	70.249641	28	1982	BILLY IDOL	37.737421	32.512220
11941	I.O.U.	2301	9/24/83	1	38.0	16.267217	3.737670	60.801484	28	1983	FREEEZ	28.825326	31.976158
14077	Back In Black	113	12/14/80	16	85.0	19.565920	4.394449	85.981440	99	1980	AC/DC	55.422749	30.558691
6354	Last Night A D.J. Saved My Life	2758	2/26/83	6	47.0	16.942765	3.761200	63.725130	35	1983	INDEEP	34.206311	29.518820
12220	Tell Me If You Still Care	6635	10/29/83	1	55.0	16.458042	3.912023	64.384240	29	1983	THE S.O.S. BAND	35.075191	29.309049
18295	Between The Sheets	6354	6/18/83	3	69.0	17.152549	4.077537	69.940159	35	1983	THE ISLEY BROTHERS	40.822462	29.117697
11007	Crazy Train	4687	7/4/81	5	74.0	16.185849	4.290459	69.444730	30	1981	OZZY OSBOURNE	41.023587	28.421143

Underperforming songs

Don Henley had a successful solo career but he will probably still be remembered as the maestro behind the Eagles. It's surprising that his solo - Not Enough Love in the World is in the underperforming league but that could be because people would search for Eagles more on Youtube and Spotify as compared to the name Don Henley.

	Title	Artist	Entry_Date	Total_Weeks	artist popularity	log_youtube	log_spotify	net_current_popularity	original_popularity	year	Artist_Name	predicted_popularity	relative_pop
11778	The Way You Do The Things You Do / My Girl	1488	8/31/85	11	77.000000	10.073483	2.302585	23.195053	116	1985	DARYL HALL & JOHN OATES	56.408361	-33.213308
4810	If You Wanna Get Back Your Lady	6556	3/26/83	5	67.000000	8.378161	1.386294	11.614597	69	1983	THE POINTER SISTERS	45.211239	-33.596642
7293	Not Enough Love In The World	1790	5/25/85	17	68.000000	5.529429	3.496508	19.333691	102	1985	DON HENLEY	53.731422	-34.397731
11137	The Message	2527	8/21/82	15	57.778352	13.991143	0.693147	9.697921	74	1982	GRAND MASTER FLASH AND THE FURIOUS FIVE feat M...	44.847118	-35.149196
7208	Free Me	5267	7/6/80	10	57.778352	12.334123	0.693147	8.549362	83	1980	ROGER DALTREY	43.856390	-35.307027
17851	Somethin' 'Bout You Baby I Like	2470	5/17/80	10	65.000000	9.130756	1.098612	10.031161	94	1980	GLEN CAMPBELL & RITA COOLIDGE	47.031284	-37.000123
11612	In The Name Of Love	5033	9/1/84	10	57.778352	11.574782	0.693147	8.023028	78	1984	RALPH MacDONALD & BILL WITHERS	45.295674	-37.272646
12261	Hippy Hippy Shake (From "Cocktail")	6295	10/22/88	14	57.778352	11.659335	1.098612	12.809089	91	1988	THE GEORGIA SATELLITES	50.566906	-37.757817
3541	Run, Run Away	5603	4/7/84	17	57.778352	14.817802	0.693147	10.270918	116	1984	SLADE	52.465422	-42.194504
8175	Take It Easy	258	5/31/86	17	57.778352	14.062781	0.693147	9.747577	112	1986	ANDY TAYLOR	52.053598	-42.306022

1990-2000

Overperforming songs

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.

	Title	Artist	Entry_Date	Total_Weeks	artist popularity	log_youtube	log_spotify	net_current_popularity	original_popularity	year	Artist_Name	predicted_popularity	relative_pop
16820	Still Got The Blues	2371	2/16/91	3	60.000000	17.521986	4.060443	71.147027	39	1991	GARY MOORE	42.091852	29.055175
8407	Jeremy	4801	8/12/95	9	57.778352	18.115174	4.262680	77.219188	57	1995	PEARL JAM	48.535729	28.683459
10101	November Rain	2566	6/27/92	20	57.778352	20.627271	4.234107	87.338063	133	1992	GUNS N'ROSES	59.929764	27.408299
12900	Still D.R.E.	1837	11/27/99	7	84.000000	19.904343	4.304065	85.669590	66	1999	DR. DRE & SNOOP DOGG	58.375275	27.294314
653	You Get What You Give	4509	5/1/99	2	62.000000	17.706833	4.189655	74.185518	41	1999	NEW RADICALS	47.293774	26.891744
660	Ruff Ryders' Anthem	1760	2/20/99	1	76.000000	18.226010	4.204693	76.634769	42	1999	DMX	50.739631	25.895138
9698	Boom Boom Boom Boom	7122	8/7/99	6	66.000000	18.270248	4.248495	77.621063	52	1999	VENGABOYS	51.840695	25.780368
20325	Suavemente	1986	11/21/98	2	69.000000	18.072897	4.174387	75.443271	52	1998	ELVIS CRESPO	49.726371	25.716900
6218	U Can't Touch This	3941	4/28/90	17	62.000000	19.772979	4.262680	84.285878	128	1990	M.C. HAMMER	58.591744	25.694134
20914	7 Seconds	7383	10/8/94	4	59.000000	17.329329	4.043051	70.063365	38	1994	YOUSSOU N'DOUR & NENEH CHERRY	44.594630	25.468735

Underperforming songs

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.

	Title	Artist	Entry_Date	Total_Weeks	artist popularity	log_youtube	log_spotify	net_current_popularity	original_popularity	year	Artist_Name	predicted_popularity	relative_pop
18175	It Doesn't Matter	7059	9/19/92	3	57.778352	9.514658	0.693147	6.595059	48	1992	TYLER COLLINS	43.900258	-37.305200
20302	I Miss The Homies	4114	9/6/97	20	59.000000	14.053411	1.609438	22.618093	111	1997	MASTER P featuring PIMP C AND THE SHOCKER	60.643907	-38.025814
14575	My Love Is A Fire	1815	10/13/90	16	63.000000	10.594958	1.386294	14.687731	115	1990	DONNY OSMOND	56.552830	-41.865099
13629	Don't Go	3632	9/27/97	15	57.778352	8.600063	1.386294	11.922218	74	1997	LE CLICK featuring KAYO	53.788312	-41.866093
8948	Sweet Potato Pie	1776	4/9/94	19	57.778352	13.175758	1.098612	14.475050	109	1994	DOMINO	57.592508	-43.117458
509	The Crying Game (From "The Crying Game")	834	3/13/93	17	45.000000	2.708050	2.484907	6.729252	121	1993	BOY GEORGE	53.714823	-46.985571
8911	Funky Y-2-C	6569	7/2/94	17	57.778352	12.823005	0.693147	8.888230	96	1994	THE PUPPIES	56.111792	-47.223562
12284	Living In A Danger	121	10/22/94	20	67.000000	15.774545	0.693147	10.934081	116	1994	ACE OF BASE	61.297389	-50.363308
6376	Getto Jam	1776	11/27/93	20	57.778352	13.560546	0.693147	9.399454	129	1993	DOMINO	59.849602	-50.450148
4925	The Most Beautiful Girl In The World	5995	3/5/94	26	57.778352	12.301406	0.693147	8.526685	133	1994	THE ARTIST	63.956503	-55.429819

Conclusions

After analyzing the results of our basic model, it is evident that most songs decay as expected over time but a simple regressor would be unable to capture and accurately predict popularity for highly enduring songs. A mathematical equation that we expect would do the same would also be inadequate. Most over-performing songs follow a pattern where they peaked due to the re-emergence of a particular genre/artist or based on the artist popularity. However, one hit wonders would again not fit into this category. A much necessary aspect to improve the popularity prediction would be to try and capture general trends in music over the years that would help quantify a song's popularity further. One way to do this would be to shift focus to features such as the genre of a song, valence etc.

What we think we did right :

- Considering billboard rankings as original popularity measure. Rather than formulating originality scores based on normalized or absolute metric, we feel the peak billboard ranking of a song would be an ideal measure.
- 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.
- Focus on understanding why a basic model would not effectively capture any song that does not follow exponential decay, mainly, highly enduring songs, songs that peaked later, songs from a genre that might not have prevailed etc
- 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/decade-wise sets to get a better analysis.

Future work (What we could improve upon) :

So far, we have focused our efforts on exploratory analysis and evaluating the results of our baseline model rather than running behind reducing RMSE for the model. Hereon, we plan on focusing more on :

- Evaluating additional features that can be added to improve model accuracy. Focus mainly on features that would help quantify songs with high endurance i.e the songs our model identifies as over-performers. Weigh these features based on the observations seen in the baseline model(in particular, [trying to capture data that cause the high popularity).
- Analyzing how a song's genre affects it's popularity (we suspect this would be a very important feature but haven't had time to explore it yet).
- Experimenting with multiple machine learning models to get the best possible predictions.
- Coming up with a more formal and quantitative endurance metric which would allow us to report the most enduring songs over each decade.
- Try collecting intermediary data for song popularity over the years such as sales data to compare against the decay observed in model predictions.

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
[5] Jeremy ,Story of Jeremy