Final Report
Do Popular Songs Endure

Kishan Varma (111323214), Aditya Srivastava (111424337), Sanaz Sheikhi (111733554)
Slot - 12/11 12 PM,Group 5

[1]Department of Computer Science, Stony Brook University
Stony Brook, NY 11790

{kvarma,adisrivastav,ssheikhi}@cs.stonybrook.edu

1. Challenge

Considering 1960's popular music, it seems that a surprising number of the most enduring songs, as measured by current sales or airplay, were not at the top of the charts when they came out.We have the following two challenges.

- The first one is to measure to what extent is this observation is true by analysis of pop charts and other data.
- The second challenge is building a model to predict the long term popularity of a recording.

2. Background

2.1. What is a hit song?

We have learned that there's a stark difference between a "chart-topping hit" and a "hit-selling single." To some, a chart-topping hit is determined by the BDS charts, while some will simply gauge a song by where it ranks on Billboard.com. When it comes to determining what a hit-selling single is, some will only use Nielsen's SoundScan reports. Generally, for determining a hit song is looking at how well it performed on the Billboard and Nielsen SoundScan reports. The Billboard data set was open source and we easily got it but Nielsen's data set needs payment so we put it aside.

Music editor of the award-winning magazine said "I think there are a lot of great records that, from a qualitative perspective, should sell, but when you go and look at SoundScan, you're mystified that it only sold 2,000 copies." He went on to say, "This is proof that selling music is not purely about the quality of the record. It's really about marketing, radio, YouTube, and everything else falling into place all at once, which is almost impossible to do." So, it would be safe to say that radio and YouTube charts play a role in the overall success of a song, but are not the only factors that determine the true hit potential or status of a recording. We must include timing, the artist's likability[1].

2.2. How a song is ranked?

The Billboard Hot 100 has been a famous source for song popularity rankings from 1958. There is a great tendency for artists and record companies to be able to predict the path of their songs along the Billboard rankings.

The Billboard Hot 100 has a formula. They track airplay on about a thousand terrestrial radio stations (AM/FM radio in the USA). Digital download song sales and physical singles sales (data from Nielsen SoundScan) is factored in. Radio airplay and sales are the ingredients that carry the most weight in the formula. Just added to the Hot 100 formula is the streaming data that makes up Billboard's new On-Demand Songs Chart (song plays from MOG, Muve Music, Rdio, Rhapsody, Slacker and Spotify), as well as plays on non-demand radio streams from Rhapsody and Slacker. The Hot 100 chart rankings formula now also includes plays on video request service Akoo and audio on-demand streams from MySpace and Guvera.

All these methods earn a song points which are weighted in a points system, and added together. More points equals higher rank in the Hot 100 [2][3][4]. Each week Billboard issues the Hot100 chart containing the position of 100 top songs. Fig. (1) indicates a part of Hot100 chart with its method in brief.

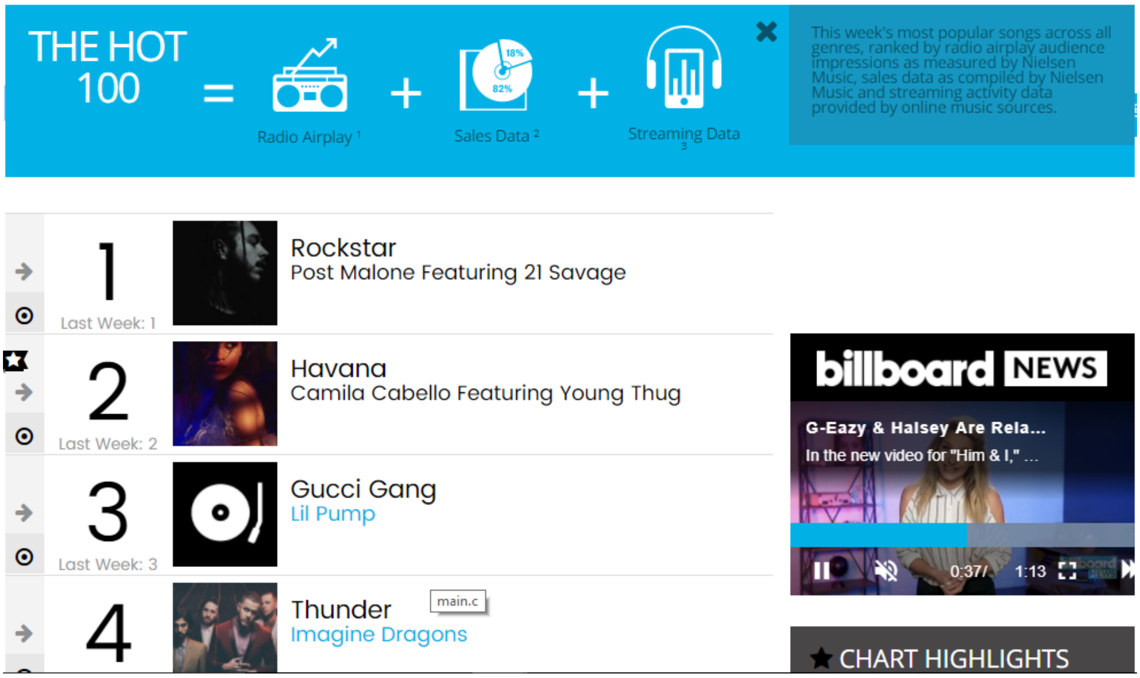

Figure 1. Billboard's Hot 100 Chart

2.3. Factors affecting rank of a song

Features of a song controls its ranking. Features include the artist who performed the song and genre of the song. People often tend to listen to songs released by a popular artist. Hence, reputation of the artist makes the song to enter hit charts easily. Similarly, the listening taste of audience change over a period. This listening taste becomes a trend for that year. For ex- Rock was in trend in earlier time than it is now, resulting in more songs released of rock in that period giving a large number of hit songs in charts from Rock genre. Hence, we can conclude that song artist and the famous genre in the predicting time are the two main factors the ranking of the song will be depending on.

3. Literature review

Predicting the ranking and popularity of songs are commercially valuable. Many companies are willing to pay a lot in order to know which styles or artist are more popular to invest on them. In the following we briefly review some related academic works.

3.1.

Koenigstein et al [5] investigated to predict Billboard rankings using Data-Mining and P2P networks. They investigated the relations between music file sharing and sales (both physical and digital) using large Peer-to-Peer query database information. They compared file sharing information on songs to their popularity on the Billboard Hot 100 and the Billboard Digital Songs charts, and showed that popularity trends of songs on the Billboard have very strong correlation to their popularity on a Peer-to-Peer network. They then showed how this correlation can be utilized by common data mining algorithms to predict a song's success in the Billboard in advance, using Peer-to-Peer information.

3.2.

Shapiro et al [6] analyzed metadata and audio analysis features from a random sample of one million popular tracks, dating back to 1922, and assess their potential on the Billboard Hot 100 list. They compared the results of Bayesian Additive Regression Trees (BART) to other decision tree methods for predictive accuracy. Through the use of such trees, they determined the interaction and importance of different variables over time and their effects on a song's success on the Billboard chart.

3.3.

Cibils et al [7] used Rigide Regression to predict the (n + 1) th position in the Billboard Hot 100 charts given the first n positions and musical data. Used this prediction as the next entry to predict the (n + 2) th entry, and keep repeating this regression until the song drops out of the charts. They claim that their results display a robust prediction engine that leverages both non-musical and musical features to predict the full path of a song across the Billboard charts.

4. Data Sets

Our Data sets consists of features of songs in two time periods. The old time period is from 1958 to 1970 and the present time period from 2005 to 2017.

4.1. Billboard API

The Billboard Hot 100 is the music industry standard record chart in the United States for singles, published weekly by Billboard magazine. Chart rankings are based on sales (physical and digital), radio play, and online streaming. billboard.py is a Python API for accessing music charts from Billboard.com. The data lists we took advantage of in the Billboard is the yearly top 100 chart rankings of songs released in the period from 1958 to 1970. As Billboard provides yearly charts, we converted these chart rankings into a chart ranking of 12 years by using the inverse point score method.

$$score = \sum_{n} (101 - i)$$

Where n is the number of weeks the song was on the chart.

4.2. Last.fm [8] and Youtube API[9]

Last.fm is a popular music website. Using a music recommender system called ”Audioscrobbler”, Last.fm builds a detailed profile of each user’s musical taste by recording details of the tracks the user listens to, either from Internet radio stations, or the user’s computer or many portable music devices. This information is transferred (”scrobbled”) to Last.fm’s database either via the music player itself (including, among others Spotify, Deezer, Tidal, and MusicBee) or via a plugin installed into the user’s music player. The data is then displayed on the user’s profile page and compiled to create reference pages for individual artists. The data list we took advantage of in the last.fm is the play count of the songs. Similarly, the data list we took advantage of in the YouTube is the view count of the song. We have finally added the play count and view count to get a cumulative count of a song implying its popularity in the present time.

4.3. pygn.py

pygn (pronounced ”pigeon”) is a simple Python client for the Gracenote Music API, which can retrieve Artist, Album and Track meta data with the most common options. The data list we took advantage of in pygn is the genre of the song, its mood and its tempo.

4.4. Grammy.com[10]

We used Grammy award won by an artist as one of the important decisive factor for our model. So we scraped out get the number of awards won by each artist from year 1958.

We successfully generated a data set of 8408 songs which appeared in Billboard Hot 100 chart from year 1958 to 1970 along with there features.

5. Features

5.1. Inverse Score

The Inverse Score of a song is calculated as the sum of the inverse scores at the various positions the song has appeared in the Billboard hot100 dataset.This is the method Billboard uses to calculate the top 100 songs at the end of a year.To calculate the inverse score at a given position of a song, we have used the below formula :

$$InverseScore = 101 - Position$$

$$TotalInversescore = \Sigma Inverse_Score$$

Now the total inverse score is sum of the inverse score of the song, artist pair,because it is observed some songs with same name are sung by different artists.

5.2. Old Popularity

Old popularity is a measure of how popular the song is in old period according to billboard.From the previous feature,we get the inverse scores of each song,artist pair.Each such pair is given a rank by arranging their inverse scores in descending order.This means that the song with highest inverse score will be ranked the least.

5.3. Number of Songs sung by the Artist

It has been observed people tend to remember the artist or the band that has many songs in their account.In fact,they do so much of contribution to music when their piece of art has really done well and been appreciated a lot. On these lines we have calculated the total number of songs an artist/band has sung in the period we aim to calculate the popularity and predict the future from them. This count is the number of songs and does not include repeated songs.

5.4. Artist Rank

From the previous feature we get the number of songs sung by the artist,sorting these counts in decreasing order gives us the rank of the artist in terms of their total music charts that have appeared in billboard.

5.5. Genre

The Genre of all the songs in our data is retrieved using pygn api, which is mentioned in the Datasets section of the report.

5.6. Genre Category

Genre of the song has been converted to a categorical int variable.

5.7. Lastfm Play Count

Lastfm is the one music organization which has given us the playcount of the song. Using the Lastfm api we have retrieved the lastfm playcount of each song till date since lastfm was established.This can be treated as the future count since the difference between years is almost 57 (1960 - 2017).

5.8. Lastfm Rank

Arranging the above Lastfm playcounts in the decreasing order has given us the rank of the song,thus we name it as Lastfm Rank.

5.9. Youtube Play Count

In this modern era of internet,it is observed that most of the people want to see songs along with hearing. This freshens the memory of the listener and also gives good visual entertainment. Youtube, a product of Google is one good and open platform for watching videos. It has such immense dataset that old to very old songs can be found here. Thus, we have calculated the youtube playcount using the youtube API, which only gives the counts of the legal and copyright protected sources. Hence, this is a clean,reliable and a legal resource to include.

5.10. Youtube Rank

Arranging the above Youtube playcounts in the decreasing order has given us the rank of the song,thus we name it as Youtube Rank.

5.11. Total PlayCount

It should be considered that when people watch songs, they do listen them too.Hence to go a step ahead we have added the the playcount from Lastfm and Youtube, hence named this count as Total Playcount

5.12. new_popularity

Arranging the above Total playcounts in the decreasing order has given us the rank of the song,thus we name this rank as new_popularity.

5.13. Grammy Award Winners

It is been observed that Grammy Awards are the acclaimed awards of the music industry.We have included the factor of ’win’ in our dataset.The artists who have won the awards are treated with more respect than others.

6. Observation

In this section, we will discuss about some interesting observation found after manipulating the complete data set.

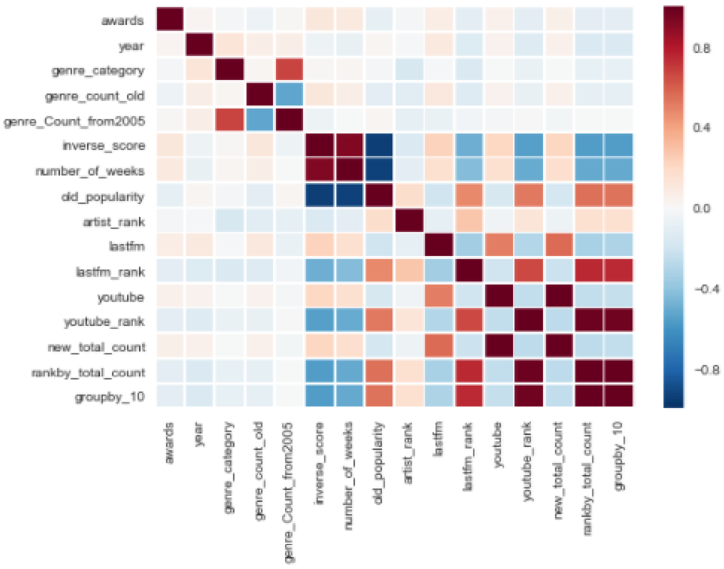

Figure 2. Correlation Matrix

genre	genre_category
Alternative & Punk	0
Classical	1
Electronica	2
Jazz	3
Other	4
Pop	5
Rock	6
Soundtrack	7
Traditional	8
Urban	9

Figure 3. Genre classification in categories

genre_Count_from2005	genre_count_old	genre_category
5730	58	0
13	18	1
1296	33	2
81	298	3
1250	226	4
13616	1235	5
1990	3655	6
208	42	7
12088	811	8
19436	1891	9

Figure 4. Genre Counts

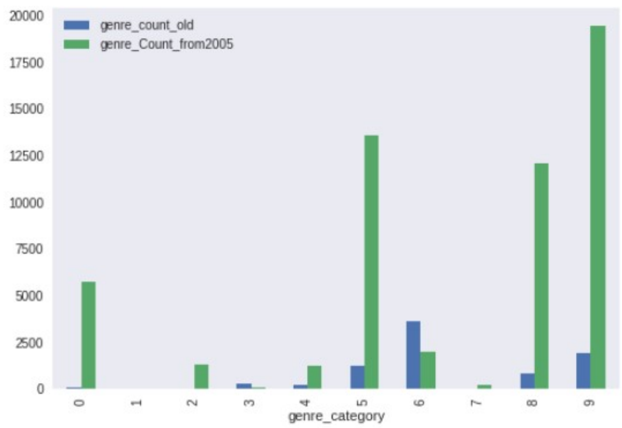

Figure 5. Bar Graph

We can see from the correlation matrix from Figure 2,that `old_popularity` is highly correlated with new popularity of a song. `new_total_count` is correlated with number of awards an artist has won, release year of a song, number of songs of a genre type, number of weeks it was in the chart, inverse-score and rank of an artist.

We also observed a significant amount of change in number of songs produced for a particular genre from period 1958-1970 to 2005-2015. This implies that the listening habit of audience has changed over the period. This will surely have an impact over the popularity of songs released in 1960s when they are compared in 2015.This can be observed in figures 3,4,5.

A noteworthy observation from Figure6, Most of the artists ranked between 1000 and 2500 have the cumulative inverse_scores of their songs less than 10000. On the other hand artists whose ranks are better than 1000 have cumulative inverse_score above 10000. More score implies more better the song. Obviously, the better the artist is ranked, the higher cumulative inverse_score of the songs is. So, our assumption on impact of artist rank on success of song is proved and we chose artist ranks as one of the features to train our prediction model based on.

7. Evaluation

Our main concern is to prove the following claim:
A surprising number of the most enduring song now were not at the top of the charts when they came out. By analysis of pop charts and other data, measure to what extent is this observation is true.

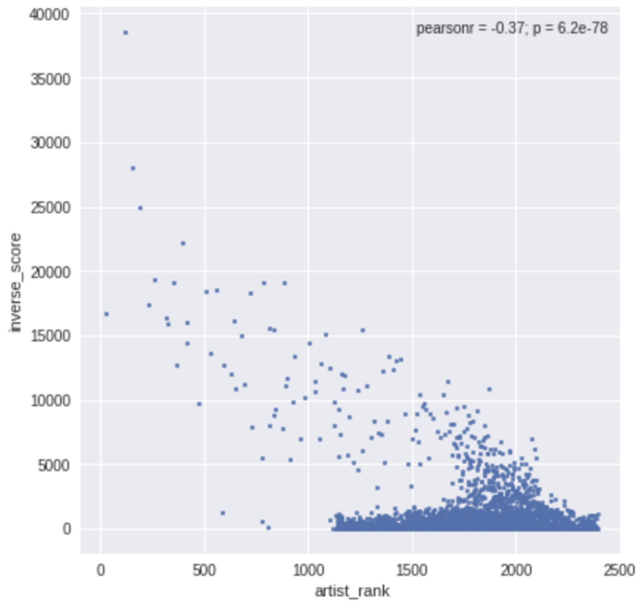

Figure 6. Scatter Plot of Artist Rank Vs Inverse Score Of Song

We try to predict new popularity of song using various models and will then prove our claim. In our regression models we will use Mean Square Error to evaluate its performance. For our classification model we will utilize F1-score and Mean Square Error to evaluate its performance.

8. Baseline Model

In this section, we have briefly explained our baseline models along with initial prediction results.

- **First baseline model**
  In our baseline model, each genre is divided into buckets of fixed size sorted by ”new popularity”. The model searches which genre the song belongs to and finds out in which bucket it falls. After successfully reaching the correct bucket, the method searches for the first entry of the song's artist in the bucket and returns the corresponding new popularity as the predicted value. If the artist is not found then we return the minimum of the new popularity rank from that bucket.
  Adding Some Example

- **Second baseline model**
  Based on our previous baseline model we noticed that ”genre” can have great impact on popularity (rank) of songs. Also, we investigated the change in ”genre” of old and new songs and considering Fig.5 we observed that genre taste has changed a lot. For example in 1960 many songs were ”Rock” while from 2005 on most of the songs were ”Urban”. So, we added ”genre” as a feature to our linear regression model as second baseline model and the results are as follows.

Table 1. Error Statistics

Model name	R Square	Mean Square Error
Baseline Model (Linear Regression)	-0.000978	6085898.4657

9. Advanced Model

We observed that popularity of the artist also has an impact over song's popularity. Now what decides popularity of a singer? Number of songs he has sung till the time and number of acclaimed awards he has won accounts to his popularity. So, in our advanced models, we have included artist popularity as feature. We have indexed each artist with respect to its popularity in decreasing order. We performed Ridge regression and Lasso Linear model on our dataset to predict new popularity of a song.

Both Ridge and Lasso Regression might appear to work towards a common goal, the inherent property and practical use cases differ substantially. They work by penalizing the magnitude of coefficients of features along with minimizing the error between predicted and actual observations. These are called regularization techniques.

9.1. Ridge Regression[11]

Ridge Regression performs L2 regularization, i.e adds penalty equivalent to square of the magnitude of coefficients. Minimizing objective = Least Square Objective + α * (sum of square of coefficients)
Here alpha is the parameter which balances the amount of emphasis given to minimizing Least Square Objective vs minimizing sum of square of coefficients. For our model, we have taken alpha as 0.5.

9.2. Lasso Linear Mode[12]

Lasso regression performs L1 regularization, i.e adds penalty equivalent to absolute value of the magnitude of coefficients. Minimizing objective = Least square objective + α * (sum of absolute value of coefficients).
Here alpha works similar to that of ridge and provides a trade-off between balancing Least Square Objective and magnitude of coefficients.

Table 2. Error Statistics

Model	Mean Square Error (Rank)	Mean Absolute Error
Ridge Regression	3761513.29	1582.25
Lasso Regression	3806372.24	1600.45

Figure 7 shows prediction result for Ridge and Lasso Regression Model.

We have based our predictions on the time span of around 50 years(1960-2017).It should be observed from the calculations carried out in the project that checking the exact rank/position of a song after so many years is not a good idea. Hence, based on these lines we strive to convert the problem to a classification problem where in we predict the grade of a song rather than the rank of a song. It would be interesting and worth mentioning to know how the songs can be graded. The songs are graded as per any common music chart would grade them(TOP10,TOP20 and so on).Here due to the variety

song	artist	year	genre_category	old_popularity	number_of_weeks	artist_rank	Ridge_regressi	Lasso_reg	new_popularity
We Can Make It Baby	The Originals	1970	6	5909	6	435	4214.995834	4214.989	4171
(I'm So) Afraid Of Losing You Again	Charley Pride	1969	8	5969	6	390	4515.834657	4515.825	4704
Broken Heart	The Fiestas	1962	6	6220	6	824	5340.041963	5340.04	6607
Fools Rush In	Rick Nelson	1963	6	1041	13	49	2476.694237	2476.697	1266
The Whistling Organ	Dave 'Baby' Cortez	1959	6	4849	8	1002	5271.623157	5271.625	5854
Traces/Memories Medley	The Lettermen	1969	6	3641	8	92	3172.511647	3172.51	6530
Wonderful You	Jimmie Rodgers	1959	6	4109	7	67	4676.215235	4676.223	7675
Viva Las Vegas	Elvis Presley	1964	6	7747	1	1	6089.797447	6089.8	309
Shapes Of Things	The Yardbirds	1966	6	1519	11	207	2391.415767	2391.419	1131
Ramrod	Duane Eddy His Twangy Guita	1958	6	3283	8	155	4468.46429	4468.474	5965
Merry-Go-Round	Marv Johnson	1961	9	5318	6	218	4819.705072	4819.703	4420
Don't Mess Up A Good Thing	Fontella Bass & Bobby McClur	1965	4	2522	11	909	3547.02592	3547.029	4583
Crying In The Chapel	Adam Wade	1965	5	6924	3	164	5723.170126	5723.172	7763
I Say Love	The Royal Guardsmen	1968	6	5772	4	282	4454.911353	4454.912	7707
Fingertips - Pt 2	Little Stevie Wonder	1963	9	292	15	561	2045.414525	2045.41	1584
Where The Blue Of The Night	Tommy Mara	1958	6	6162	4	1205	5907.966785	5907.973	7266
These Eyes	Jr. Walker & The All Stars	1969	4	1278	13	87	2161.659515	2161.659	3853
Whenever A Teenager Cries	Reparata And The Delrons	1965	6	4268	9	872	4108.100708	4108.097	3433
Chills And Fever	Ronnie Love And His Orchestr	1961	9	6356	4	2101	5898.489522	5898.488	5333
Moulty	The Barbarians	1966	6	7010	4	870	5425.205059	5425.201	2321

Figure 7. Prediction Result

and largeness of the dataset we have used the measure of percentage and classified songs as TOP10%,TOP20%,and so on. So now we have 10 bins with around 800 songs in each bin with respect to there new popularity, that is the popularity today. Now we predict what grade the song must get with respect to the input features.

9.3. Logistic Regression[13]

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable(in which there are only two possible outcomes). In our case, we will be using multinomial logistic regression. It is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels. Thus, it is an extension of logistic regression. Using logistic regression, we are predicting the probability of each song to which grade it belongs in period from 2005 to 2017.

9.4. Multinomial Naive Bayes[14]

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. It assumes that the presence of a partiular feature in a class is unrelated to the presence of any other feature. For Example- a fruit may be considered as apple if it is red, round and about 3 inches in diameter. Even if these features depend on each other or upon the existence of other features, all of these properties independently contribute to the popularity that this fruit is an apple and that is why it is known as 'Naive'. In our model we will be using Multinomial naive bayes because we have sample represent the frequencies with which certain events have been generated by a multinomial .

9.5. KNeighbours Classifier[15]

The algorithm for the k-nearest neighbor classifier is among the simplest of all machine learning algorithms. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all the computations are performed, when we do the actual classification.The neighbors are taken from a set of songs for which the class (for k-NN classification) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.The examples of the more frequent class's songs tend to dominate the prediction of the new example, because they tend to be common among the k nearest neighbors due to their large number.

9.6. Decision Tree Classifier[16]

The goal here is to create a model that predicts the value of a target variable based on several input variables. Each interior node of the tree corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represent a value of the target variable given values of the input variables represented by the path from the root to leaf. A decision tree is a simple representation for classifying examples.

Table 3. Model Performance Evaluatoin

	f1 score	Mean Squared Error
Logistic Rgression	0.17	9.43
K Nearest Neighbours	0.15	11.26
Multinomial Naive Bayes	0.14	10.84
Decision Tree Classifier	**0.16**	**8.27**

From the model performance comparison table (Table 3) , we observed that the Decision Tree Classifier has the combination of least f1-score and the least mean squared error.Hence, we choose the Decision Tree Classifier as the prediction model for this problem.

10. Final Predictions and Conclusion

song	artist	old_popularity	new_popularity	old_Grade	new_Grade	Predicted_Grade
Peter Gunn	Duane Eddy His Twangy Guitar And The Rebels	2690	7877	4	10	8
Back Up Train	Al Greene & The Soul Mate's	2828	7129	4	9	10
I Wanna Be Loved	Ricky Nelson	1920	6463	3	8	9
If Dreams Came True	Pat Boone	2168	6871	3	9	8
Bring It Up	James Brown And The Famous Flames	2934	7904	4	10	10
Precious, Precious	Jackie Moore	5743	2608	7	4	4
Viva Las Vegas	Elvis Presley	7747	309	10	1	2
Some Kind Of Wonderful	Soul Brothers Six	7723	2746	10	4	4
I Take A Lot Of Pride In What I Am	Dean Martin	6134	2374	8	3	3
Mom And Dad's Waltz	Patti Page	5166	2593	7	4	3
Mack The Knife	Bobby Darin	2	141	1	1	1
Tighten Up	Archie Bell & The Drells	252	287	1	1	1
So Much In Love	The Tymes	366	833	1	2	1
Yesterday	The Beatles	869	89	2	1	1
Stop! In The Name Of Love	The Supremes	685	696	1	1	2
Don't Let The Sun Catch You Crying	Gerry And The Pacemakers	991	710	2	1	1
Camel Back	A.B. Skhy	8396	7305	10	9	10
Surfer Street	The Allisons	7841	7741	10	10	10
Every Night, Every Day	Jimmy McCracklin	7718	7100	10	9	10
The Last Dance	The McGuire Sisters	8306	8187	10	10	10
Hot Cakes! 1st Serving	Dave Baby Cortez	7447	8150	9	10	10
What Now My Love	Groove Holmes	7639	7962	10	10	10

Table 4. Prediction Results

Our final rank prediction for song is in the included Table 4. predicted_Grade is the grade predicted by our model to fit the song in its appropriate bin. We can see that song "Some Kind of Wonderful" performed by artist "Soul Brother Six" of genre "Pop" has improved from its old ranking grade, i.e from grade 10 to grade 1. It is now among the top 10% of songs. Our prediction also shows it in bin 2, i.e in top 20% of songs. Now, why we see such a trend? By analysis of our result we see that, genre count in old period was 1235 which has become 13616 in present period. The artist rank(cumulative of number of songs sung and number of awards won) for the song is mediocre. Hence, in combination of these features, the rank of song has improved. Similarly, we see that several songs with poor old_grade have improved in future. This justifies our claim that most enduring songs now were not at the top of the charts when they came out. Reason for such behaviours is due to the features of song.

From inspecting our results, it is apparent that some songs appear to have been misclassified in the bins. These unexpected behaviour is because data available are scattered. According to the behaviour of features of a song, it is predicted in a grade which is far away from its correct bin because it may be the case that correct data for the song is present at lastfm or youtube without appropriate specification. As we have seen while fetching data, Lstfm or YouTube has either title of song or name of song in a different format. This makes our search query inefficient and we get some noise values.

11. References

[1] http://blog.sonicbids.com/from-the-pen-to-the-charts-how-a-hit-song-happens
[2] https://www.customchannels.net/billboard-hot-100-music-chart
[3] https://www.billboard.com/biz/billboard-charts-legend
[4]https://www.billboard.com/articles/columns/ask-billboard/5740625/ask-billboard-how-does-the-hot-100-work
[5] Noam Koenigstein, Yuval Shavitt, Tomer Tankel, Ela Weinsberg, Udi Weinsberg, "A framework for extracting musical similarities from peer-to-peer networks", Multimedia and Expo (ICME) 2010 IEEE International Conference on, pp. 1433-1438, 2010, ISSN 1945-788X
[6] Heather Shapiro, "A Bayesian Approach to Understanding Music Popularity", Undergraduate Honors Theses and Student papers, Duke University, https://dukespace.lib.duke.edu/dspace/handle/10161/9747,2015
[7] Cristian Cibils, Zachary Meza, Greg Ramel, "Predicting a Song's Path through the Billboard Hot 100", Thechniqal report, Stanford University,2015
[8] https://en.wikipedia.org/wiki/Last.fm
[9] https://developers.google.com/youtube/v3/
[10] "WINNERS RECORDS OF THE YEAR" - https://www.grammy.com/grammys/awards/winners-nominees/138
[11] Ridge Regression https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf [12] Lasso Regression https://en.wikipedia.org/wiki/Lasso_(statistics) [13] Logistic Regression https://www.medcalc.org/manual/logistic_regression.php
[14] Naive Bayes Classifier https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
[15] K-nearest Neighbors Algorithm https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm [16] https://en.wikipedia.org/wiki/Decision_tree_learning