What goes into a record-breaking song?

ECO 395 Final Project

Soo Jee Choi, Annie Nguyen, Tarini Sudhakar

2023-04-10

I. Abstract

How do you make a hit song? Intuitively, a hit song should have a good beat, a catchy hook, and not-too-complicated lyrics. But we can do better than that, given the amount of data surrounding us. Spotify has been able to break a song into quantitative features such as energy, tempo, acoustics, and much more.

Using these features for songs released between 1985 to 2015, we predict whether a song will make the Billboard Top 100 in the USA. We run both a logistic regression and random forests model. Unfortunately, while both models have a high total accuracy, they are unable to predict well whether a song will be a hit. Random forests get 5% as the best sensitivity rate.

We also evaluate how features of a Top 10 Billboard hit change over this time period, including lyric analysis. We observe that PCA does manage to group songs by lyric and acoustic features representing the popular music genres of each decade. Our biggest contribution is compiling a dataset with songs from 1958 to 2015, containing Spotify audio features and lyric features for Billboard hit songs.

II. Introduction

Hit song science is the possibility of predicting whether a song will be a hit before it is distributed (see here. Knowing what kind of features go into a hit song help musicians, record labels, and music vendors generate larger revenues. It also increases the reach of an artist. For instance, if a record label knows that a song with particular features such as high tempo and repeated words, will have a better shot being popular, they would put more resources behind that. Similarly, a song with higher acoustic features will only appeal to certain audiences and not have mainstream appeal.

We compiled a dataset of songs from 1958 to 2015, both Billboard Top 100 and non-hits, with their audio features from Spotify. We had a total of 431,379 observations, with 3515 hit songs. We used the dataset built by Yamac Eren Ay on Kaggle as the master dataset. We added a dummy variable `billboard_hit` to indicate whether a song had been on the Billboard Top 100. For Billboard Top 100 songs, we also added lyric features from a publicly available data set by Github user `KevinSchaich`. You can find a detailed description of the dataset in Appendix 1.

We asked two questions:

1. Can we predict which song will become a Billboard Top 100 hit?
2. How have Top 10 Billboard hits changed over time?

For the first question, we followed the methodology laid down by Middlebrook and Sheik (2019). They build four prediction models for songs from 1985 to 2015: logistic regression, neural network, random forests, and support vector machine. We use random forests and logistic regression for our analysis.

Random forests combines multiple decision trees to improve predictive performance. Each decision tree is trained on a random subset of features and samples, and the final prediction is based on the majority vote of the individual trees. We will explore the use of random forests for hit song prediction and compare its performance with logistic regression.

Logistic regression models the probability of a binary outcome based on a set of input variables. In our case, the binary outcome is whether a song is a Billboard Top 100 hit, and the input variables are the various audio features extracted from the song.

For the second question, we evaluated how audio and lyric features have changed of Top 10 Billboard hits using Principal Component Analysis.

III. Predicting a Billboard Top 100 hit song

Random Forest We tested random forests for 10, 20, and 300 trees. We split our dataset into a training and test dataset in an 80-20 ratio. Given the size of the dataset, we used the foreach function on R to train multiple trees on different subsets of the data with a parallel backend from the "doParallel" package. At the cut-off level of 0.5, our random forest models all had a total accuracy rate of ~97%. But with the large number of non-song hits that we have, this is not realistice or a useful measure of accuracy. We want to look at the sensitivity or recall of the model, that is, how many songs have been correctly predicted as hits.

For our model with 10 trees, our in-sample sensitivity is 0.4254017 or 42% and out-of-sample sensitivity is 0.03981265 or 4%. Our out-of-sample RMSE was 0.0942759. What was more concerning was percentage of variance explained: -34.23. This indicated that our model was performing very poorly, and that we were better off our null model. Increasing the trees to 20 did not help that much. We got in-sample sensitivity of 0.4433818 or 44.3% and out-of-sample sensitivity is 0.04566745 or 4.5%. Out-of-sample RMSE was 0.09235341.

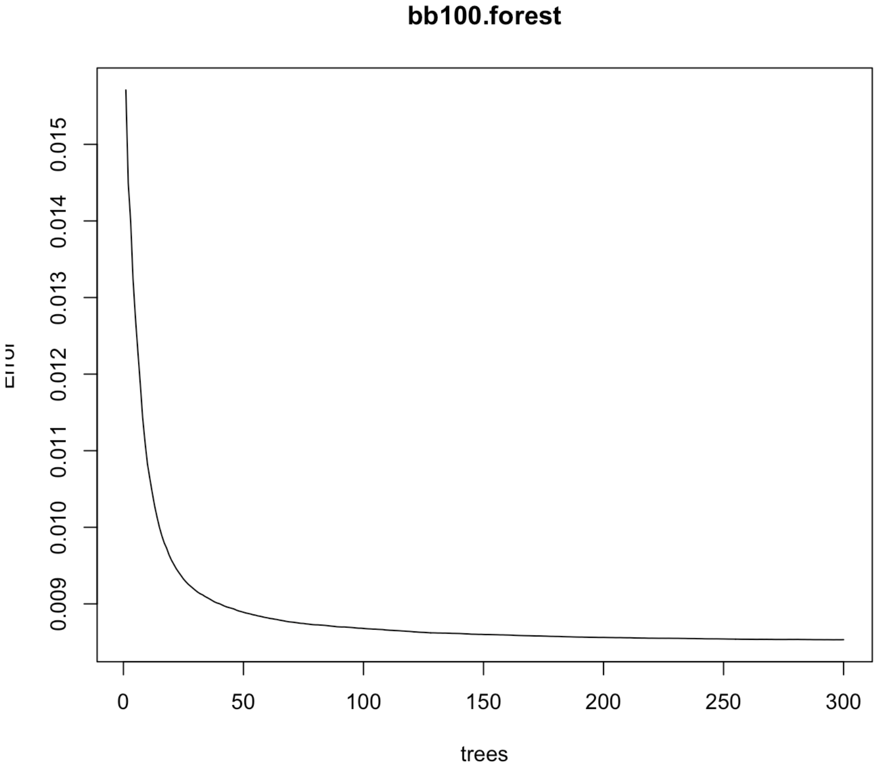


Figure 1: Random Forest with 300 trees

When we ran it for 300 trees, not much changed. We can see that the error dropped with increase in trees and percentage of variance explained increased to -6.46. But out-of-sample RMSE was again similar to the other models: 0.09063623. There was minute change in sensitivity, with in-sample increasing to 51.75% and out-of-sample increasing to 5.1%.

When we generated a variable importance plot, we can see that valence is the most important in classifying the data. Energy, speechiness, and danceability also rank high.

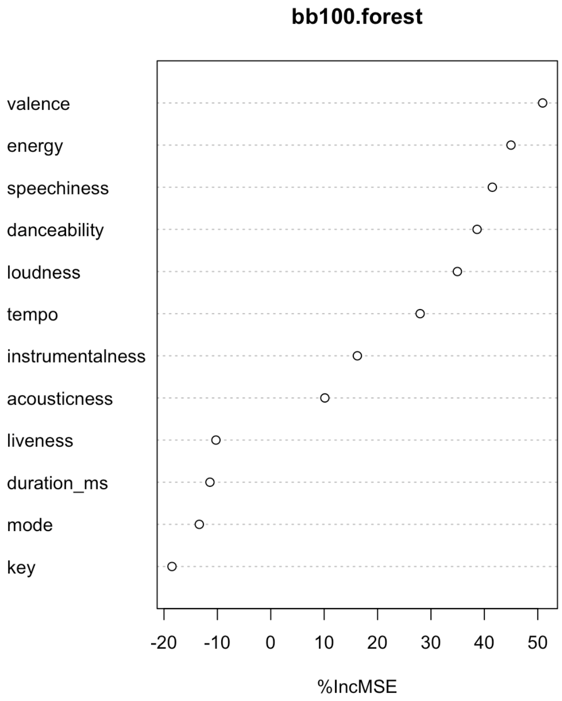


Figure 2: Variable Importance Plot

But given these overall results, our random forests model performed quite poorly. Due to computational constraints and little marginal increase in performance, we chose to not increase the number of trees in the model. We also modeled a single decision tree, where we got an out-of-sample RMSE of 0.09712917.

Logistic Regression

First, we build a logistic regression model to predict which song will become a Billboard Top 100 hit. `hit` is the binary variable in which 1 indicates that a song made it onto the Billboard 100 and 0 indicates that the song was not on the Billboard 100.

We will compare the out-of-sample performance of the following models:

1. Baseline Model: a small model that uses only the `duration_ms`, `danceability`, `energy`, `key`, `loudness`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valence`, `tempo`, and `time_signature` variables as features with no interaction terms.
2. Refined Model: a logistic regression model built using the forward selection process. The forward selection process considers all of the candidate variables listed above and interaction terms.

Model Building

Baseline Model: After running the baseline logistic regression model, we test for out-of-sample performance measures and see that it gives us an accuracy of 99.1%, an RMSE of about 5.295, and an AUC of 0.696.

```
##             acc   rmse_     auc
## Metrics 0.99199 5.29513 0.69613
```

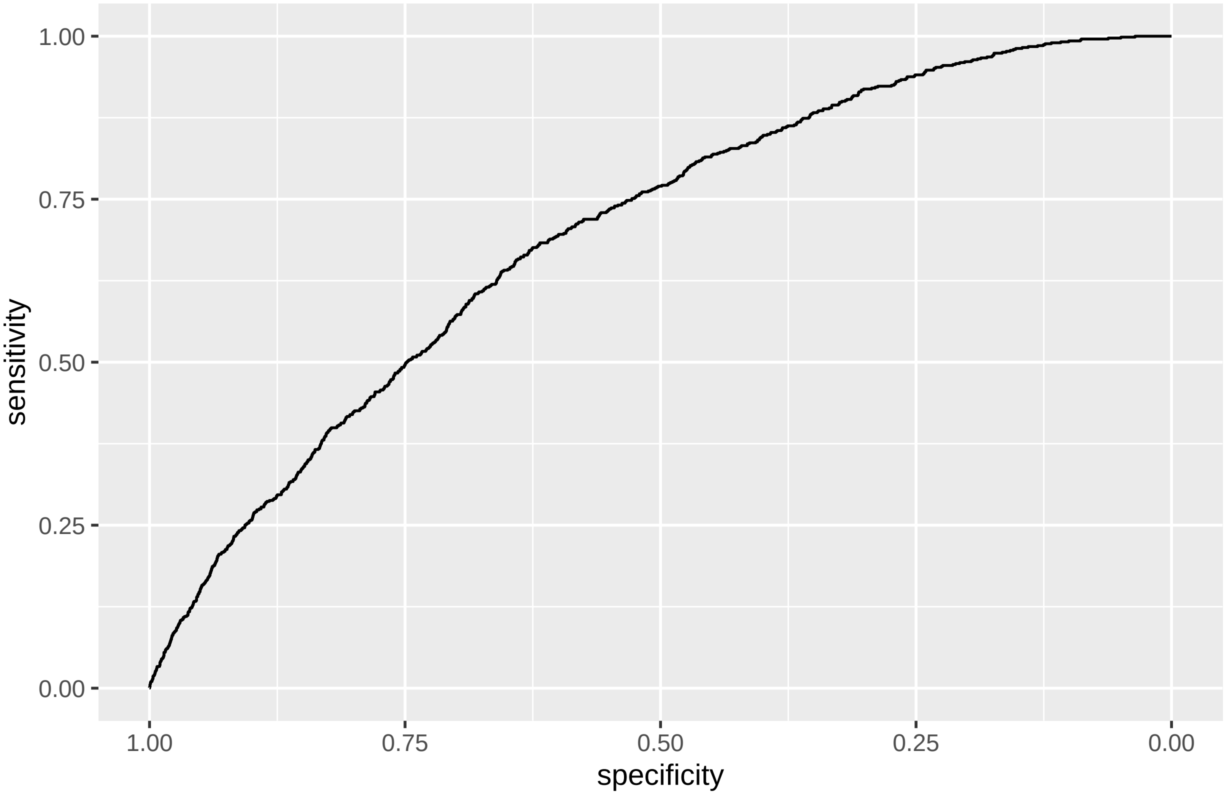
Refined Model: We build another logistic regression model using the forward selection process to get a model that includes the candidate variables and various combinations of interaction terms. Although accuracy remains the same and the RMSE score increases, we see that there is improvement in the AUC, which shows that the refined model does a better job predicting whether a song will be a Billboard hit or not.

```
##             acc   rmse_     auc
## Metrics 0.99199 5.76343 0.72086
```

Although accuracy is about as good as the baseline model and the AUC is improved, it is important to note that the baseline model is more parsimonious, and the trade-off between accuracy and parsimony is not substantial enough to justify using the refined model. In our case, we will stick to using the baseline model for analysis.

Model validation: The following is an ROC plot for the Baseline Model, using the test dataset:

Figure 3: ROC Curve



An ROC plots TPR vs. FPR. TPR is another name for sensitivity, while FPR is defined as (1-specificity), which explains why the numbers on the x-axis are flipped with 1 on the left and 0 on the right.

10-Fold Cross Validation In this step, we perform 10-fold cross validation for our baseline logistic regression model. For each fold, we will calculate the performance metrics, which includes accuracy, RMSE, and AUC values.

The results for each fold is as follows:

```
##              acc   rmse_     auc
## Metrics  0.99196 5.29181 0.69801
## Metrics1 0.99196 5.29143 0.69447
## Metrics2 0.99198 5.30657 0.68840
## Metrics3 0.99198 5.29341 0.69189
## Metrics4 0.99196 5.29624 0.70220
## Metrics5 0.99196 5.28496 0.69556
## Metrics6 0.99196 5.28076 0.70324
## Metrics7 0.99196 5.27873 0.72407
## Metrics8 0.99196 5.30642 0.70854
## Metrics9 0.99196 5.29560 0.69035
```

Average of performance metrics across 10 folds:

```
##        acc    rmse_      auc
## 1 0.991964 5.292593 0.699673
```

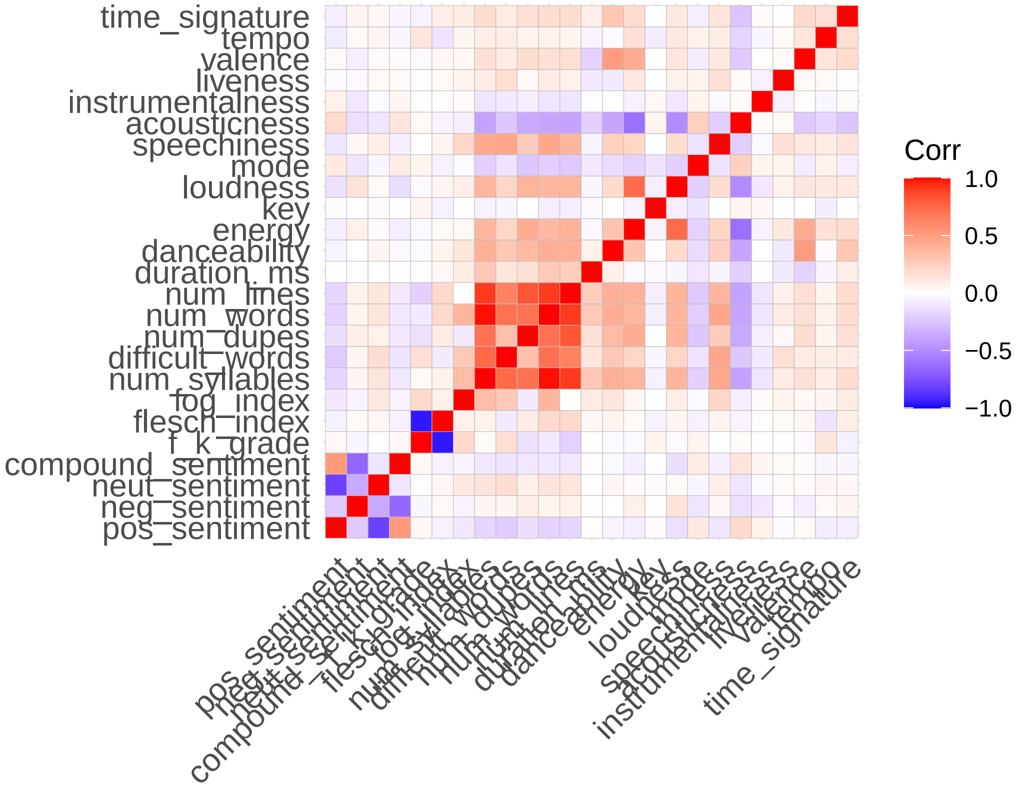
From the performance metrics, we see that accuracy stays consistent at 99% on average, RMSE is about 5.29, and the AUC is at about 0.7. Although the accuracy score is high and RMSE is consistently low, the AUC indicates that performance can be further improved with a better model.

IV. Identifying patterns in a Top 10 hit song

Principal Component Analysis (PCA) is a dimensionality reduction technique that can be used to identify patterns and trends in high-dimensional datasets. In the context of hit song prediction, we can use PCA to analyze how the audio and lyric features of top billboard hits have changed across decades.

First, we gathered a large dataset of songs released over several decades, including both top billboard hits. We then extract various audio features such as tempo, loudness, and danceability, as well as lyric features such as sentiment and complexity, from the songs using the Spotify Web API and natural language processing techniques. The complete data set contains the billboard top 100 hits from 1958-2015. Using the complete data set, we plot the correlation matrix for the acoustic and lyrical analysis variables, as presented below. From the matrix, as expected we see that acousticness and energy are negatively correlated while energy and loudness are positively correlated. Num_words, num_syllables, and num_lines are highly correlated, so only num_words was kept. Additionally, the various measures of sentiment (positive, negative, neutral, compound) are also highly correlated with each other, so compound_sentiment and neutral_sentiment were dropped.

Figure 4: Correlation matrix for all Billboard Top 10 hit



For each decade, we first center and scale our data. Centering the data makes sure that the clustering algorithm focuses on the patterns and differences in the data, rather than the absolute values of each variable. Scaling the data makes sure that the PCA and clustering algorithms treats each variable equally, regardless of its magnitude or unit of measurement.

Next, we use PCA to reduce the dimensionality of the dataset by projecting it onto a lower-dimensional subspace while preserving as much of the original variability as possible. This allows us to visualize the relationships between the different features and identify the most important factors that contribute to a song's success. We generate 10 principal components of the data for each decade. We observe that P10 explains the majority of the data, with P10 values of 84, 84, 83, 84, and 81 for the 1960s, 1970s, 1980s, 1990s, and 2000s, respectively. Each decade is known to be defined for different genre of music. For each decade, we examine at the first three principal components of the top 10 ranking songs for the decade to examine the acoustic and lyrical features of the songs during that era. The summary of the PCA results for each decade can be found in the Appendix 2.

The 1960s (see Table 1) were known for Rock 'N Roll, Motown, and The British Invasion. Various types of Rock 'N Roll were popular during this period such as Psychedelic Rock, Blues-Rock, and Progressive Rock. Popular singers/groups included The Beatles, Bob Dylan, The Monkees, and The Rolling Stones. We see that all three PCs are actually quite similar. They all include loudness, energy, liveness, a large number of word duplications, and negative sentiment. These may be capturing different types of Rock music popular in this period. The lack of variation in the PC1-3 results may be attributed to that fact that there were not a variety of different genres popular during this period.

```
## Table 1: 1960s
```

```
##            Category          PC1          PC2         PC3
## 1     pos_sentiment -0.068349051  0.309407857 -0.15376186
## 2     neg_sentiment  0.105443063  0.026350942  0.28276604
## 3         f_k_grade  0.006887580 -0.483835175  0.11285708
## 4      flesch_index -0.005798325  0.430678881 -0.16130754
## 5         fog_index  0.003020166 -0.316781745 -0.23741837
## 6   difficult_words  0.082015203 -0.425359076 -0.07652953
## 7         num_dupes  0.333740677  0.000358662  0.24474646
## 8         num_words  0.302966048 -0.233200305  0.02961088
## 9       duration_ms  0.046718337 -0.148042303  0.29883718
## 10     danceability  0.220778812  0.021491809 -0.42531120
## 11           energy  0.392510772  0.172665418  0.10729109
## 12              key -0.023755304 -0.210168567 -0.05896560
## 13         loudness  0.321837264  0.121098309  0.28674166
## 14             mode -0.153256244 -0.030028535 -0.17310836
## 15      speechiness  0.232700289 -0.065098128 -0.35205811
## 16     acousticness -0.323774164 -0.106048792 -0.23462042
## 17 instrumentalness -0.080263408 -0.004591959 -0.03122997
## 18         liveness -0.065527375  0.088793406  0.07099974
## 19          valence  0.372334094  0.061062096 -0.28122168
## 20            tempo  0.221556910 -0.096240819 -0.10186019
## 21   time_signature  0.290112551  0.004318784 -0.25060724
```

The 1970s (see Table 2) were known for Disco, Motown, Progressive Rock, Punk, and R&B. Disco was notably one of the biggets music trends of the decade. Popular singers/groups included Diana Ross, The Jackson 5, Elton John, and Marvin Gaye. PC1, PC2, and PC3 differ considerably from each other. PC1 is largely described by acousticness, instrumentalness, and positive sentiment, perhaps picking up the rise in R&B in this decade. PC2 is described with large number of duplicate words, negative sentiment, and acousticness, perhaps a grouping of punk music that was not large, but still present in the 1970s. And finally, PC3 is described by positive sentiment, valence, danceability, energy, and loudness. This is perhaps a clear nod to the great rise in disco during the 1970s.

```
## Table 2: 1970s
```

```
##            Category         PC1         PC2         PC3
## 1     pos_sentiment  0.05219442 -0.18518934  0.25077796
## 2     neg_sentiment -0.01564784  0.30441120 -0.09266793
## 3         f_k_grade  0.02874852 -0.48774425 -0.31553451
## 4      flesch_index -0.06453507  0.42884615  0.33898854
## 5         fog_index -0.13675195 -0.28764764  0.05359200
## 6   difficult_words -0.18853697  0.13672157 -0.46955081
## 7         num_dupes -0.29888258  0.31161323  0.03226087
## 8         num_words -0.33698348  0.23497514 -0.22228342
## 9       duration_ms  0.01350717  0.14825374 -0.14124454
## 10     danceability -0.34349715  0.01598261  0.18547523
## 11           energy -0.37733150 -0.19804515  0.10474721
## 12              key  0.13263799  0.04714126  0.25738588
## 13         loudness -0.24230563 -0.28536465  0.08266921
## 14             mode  0.15410039  0.09367324 -0.04439149
## 15      speechiness -0.15174898  0.08436471 -0.38214194
## 16     acousticness  0.33132246  0.11301044 -0.13648360
## 17 instrumentalness  0.09490405 -0.05803600  0.15787080
## 18         liveness -0.06402503 -0.06258826 -0.23465420
## 19          valence -0.36214811 -0.08402331  0.20596291
## 20            tempo -0.21223630 -0.07142630 -0.08210370
## 21   time_signature -0.22846509  0.06529056  0.07787142
```

The 1980s (see Table 3) were known for Pop, Heavy Metal, Glam rock, Hip-Hop, and Country. Popular singers/groups included Whitney Houston, Madonna, Dolly Parton, Michael Jackson, Lionel Ritchie, MC Hammer, and Bon Jovi. In the 1980's music was dramatically changed by the introduction of MTV (Music

Television) where a greater importance was placed on the appearance of musicians and gimmicks became commonplace. Hip-Hop also came into the mainstream during this decade. Again, PC1, PC2, and PC3 differ considerably from each other. PC1 is described by a larger number of words, energy, loudness, liveness, and speechiness, perhaps picking up the rise in hip hop during this decade. PC2 is described by neutral sentiment a larger number of words, liveness, acousticness, and instrumentalness. This grouping is more difficult to pin point, though it may be a point to country music or a grouping of pop music during this period. And finally, PC3 is described by negative sentiment, loudness, energy, higher tempo, and liveness. This may be describing a heavy metal and/or a rock grouping during this decade.

```
## Table 3: 1980s
```

```
##            Category         PC1          PC2         PC3
## 1     pos_sentiment -0.04290601 -0.228388823  0.06820728
## 2     neg_sentiment  0.06817927  0.021229427 -0.17630814
## 3         f_k_grade  0.15871452 -0.557711536 -0.02583334
## 4      flesch_index -0.17994392  0.561859043 -0.00527149
## 5         fog_index -0.12418088  0.049619354 -0.16644754
## 6   difficult_words  0.26344045  0.010363399 -0.01302170
## 7         num_dupes  0.35049279  0.179810015 -0.02575604
## 8         num_words  0.28661871  0.351432528 -0.08800739
## 9       duration_ms -0.04405311  0.078318989  0.14382514
## 10     danceability  0.27647594  0.034567621  0.44340168
## 11           energy  0.43406785  0.015772282 -0.18306741
## 12              key  0.01544486 -0.003629964  0.23014976
## 13         loudness  0.25597853  0.028921916 -0.31415597
## 14             mode  0.03402037 -0.080196427 -0.43504939
## 15      speechiness  0.26218214 -0.162157815  0.04749873
## 16     acousticness -0.31105141 -0.034262770 -0.14929773
## 17 instrumentalness -0.11464711  0.070284139  0.02439147
## 18         liveness  0.10976511  0.259848242 -0.33741193
## 19          valence  0.32405581  0.111350732  0.33480256
## 20            tempo  0.12927056 -0.187260388 -0.29063609
```

The 1990s (see Table 4) were known for Grunge, Rap, Hip-Hop, Bubblegum Pop, Contemporary R&B, and Country. Popular singers/groups included Nirvana, MC Hammer, Britney Spears, Backstreet Boys, N'Sync, Red Hot Chili Peppers, 2Pac, Notorious B.I.G., and Guns & Roses. Very clearly, the 1990's was an era filled with all different genre's of music. PC1 is described by acousticness, positive sentiment, longer length, and higher flesch_index scores (that is, the song lyrics are not of high difficulty to understand). This grouping may be an description of Contemporary R&B or Country. PC2 is described by loudness, energy, valence, and positive sentiment, perhaps pointing to pop songs during this period. And PC3 is described by positiveness, liveness, acousticness, and speechiness, which would describe a grouping of the Contemporary R&B present in the 90s.

```
## Table 4: 1990s
```

```
##            Category         PC1          PC2         PC3
## 1     pos_sentiment  0.25139747  0.064581333  0.05809538
## 2     neg_sentiment -0.12950421 -0.058217784 -0.04161303
## 3         f_k_grade -0.07389454  0.253963446  0.56166522
## 4      flesch_index  0.03447179 -0.340032746 -0.49153924
## 5         fog_index -0.12129877 -0.212705593  0.27007764
## 6   difficult_words -0.30776564 -0.141673516  0.23758463
## 7         num_dupes -0.31189376 -0.074467016 -0.05972697
## 8         num_words -0.34353869 -0.277009104  0.08190986
## 9       duration_ms  0.09560715 -0.294388826 -0.03004874
## 10     danceability -0.36152464 -0.028583114 -0.09519161
## 11           energy -0.31259364  0.345456969 -0.17474529
## 12              key  0.05053496  0.011871445  0.29910242
## 13         loudness -0.17776527  0.443918672 -0.14779956
## 14             mode  0.19033652  0.203771289  0.04252855
## 15      speechiness -0.31315725 -0.205931027  0.05555869
## 16     acousticness  0.29258271 -0.160275071  0.06202211
## 17 instrumentalness -0.05356797  0.199098193 -0.19940718
## 18         liveness -0.17040452 -0.179313962  0.21192900
## 19          valence -0.24259183  0.228267469 -0.13454254
## 20            tempo  0.01119284  0.164823224  0.15152688
## 21   time_signature -0.07846257 -0.003748324 -0.10890848
```

The 2000s (see Table 5) were known for Pop, Hip-Hop, R&B, and Rock. Popular singers/groups included Eminem, Kanye West, Nickelback, Beyoncé, Britney Spears, The Killers, and Linkin Park. The PCs all share energy and loudness features. However, PC1 is additionally described by a larger number of words, higher tempo, and more word duplications (repitition), perhaps an indication of the large presence of hip hop during this period. PC2 is additionally described by instrumentalness, liveness, valence, and negative sentiment, perhaps an indication of a section of Rock. And finally, PC3 is additionally described by liveness, higher tempo, more difficult words, and negative sentiment. This grouping is a but more difficult to pin down, though the grouping may again be referring to Rock, but perhaps a different variety (Rock was still very popular during this period with many different genres such as post-grunge, pop punk, post-hardcore, metalcore, and indie). But very clearly, the 2000s were defined by energy and loudness in their top hits.

```
## Table 5: 2000s
```

```
##            Category         PC1         PC2          PC3
## 1     pos_sentiment  0.02243519 -0.13844000 -0.115482316
## 2     neg_sentiment -0.06827033  0.14549017  0.010398357
## 3         f_k_grade -0.05310143 -0.44988150  0.432261140
## 4      flesch_index  0.14962079  0.35919882 -0.416882381
## 5         fog_index  0.23742758 -0.34914844  0.149184215
## 6   difficult_words  0.38078142 -0.21957422  0.049968850
## 7         num_dupes  0.25744430  0.15067650 -0.125630865
## 8         num_words  0.46320661 -0.07280825 -0.103349327
## 9       duration_ms -0.05998545 -0.05717810 -0.105295195
## 10     danceability  0.24908143 -0.11267685 -0.317897789
## 11           energy  0.18475914  0.37102496  0.376562161
## 12              key -0.01007690  0.20355358  0.115029867
## 13         loudness  0.17265207  0.30483651  0.366005471
## 14             mode -0.12250942 -0.22656232 -0.193385294
## 15      speechiness  0.36980591 -0.09820065 -0.009162254
## 16     acousticness -0.13652602 -0.21663981 -0.264715955
## 17 instrumentalness -0.12867192  0.06305902 -0.168736463
## 18         liveness -0.04872337  0.05065037  0.152979294
## 19          valence  0.36661640  0.05475899 -0.066604603
## 20            tempo  0.06090640 -0.08543920  0.096377975
## 21   time_signature  0.17403290 -0.13189635  0.020128643
```

V. Conclusion

Hit song prediction For the logistic regression, the baseline model's performance is comparable to the refined model's performance. The refined model has a slightly higher AUC, but there is no significant improvement in total accuracy or RMSE. Because the improvement in the model is not significant enough to justify losing parsimony from the baseline model, we conclude that the baseline model is the best choice for logistic regression modeling.

In the random forests model, increasing the number of trees led to marginal improvement in the model with large jumps in computational burden. The RMSE for random forests with 300 trees barely differed from that of a single decision tree. But it did get a sensitivity rate of 5%, whereas the logistic regression failed to predict any hit songs. We understand this is due to a major class imbalance, where we have several more observations of non-hit songs than we have of hit songs. Once we correct for that, we can tweak the random forests model by playing with maximum depth, Gini index and more.

Patterns in Top 10 hits While tastes for different genre of music has changed across decades, we see that certain characteristics, such as energy and loudness, continue to be important features of chart topping hits. We also observe variations in principal components that arise as different genres of music emerged through the decades. Overall, using PCA to analyze how the audio and lyric features of Top Billboard hits have changed across decades can provide valuable insights into the evolution of popular music and help inform future hit song predictions.

Appendix 1

Data Sources

The publicly available data set by Github user `KevinSchaich` analyzed the lyrics of Billboard's Top 100 songs from 1950-2015 using a variety of Natural Language Processing techniques. The description of the variables obtained from this data set are presented below:

- year: Release year of the song
- position: Position of Billboard's Top 100 for the given year
- title: Title of the song
- artist: Artist of the song
- pos_sentiment: Positivity association with lyrics. Value ranges between 0-1 inclusive, with 1 being 100% positive.
- neg_sentiment: Negativity association with lyrics. Value ranges between 0-1 inclusive, with 1 being 100% negative.
- neut_sentiment: Neutrality association with lyrics. Value ranges between 0-1 inclusive, with 1 being 100% neutral.
- compound_sentiment: The sum of positive, negative, and neutral scores which is then normalized between -1 (most extreme negative) and +1 (most extreme positive).
- f_k_grade: Flesch–Kincaid grade level of the song's lyrics. The Flesch-Kincaid Grade Level is equivalent to the US grade level of education.
- flesch_index: Flesch reading ease score of the song's lyrics. The Flesch reading ease score indicates the understandability of a passage with a number that ranges from 0 to 100. Higher scores indicate that the content is easier to read and understand.
- fog_index: Gunning-Fog readability index estimates the years of formal education a person needs to understand the text on the first reading.
- num_syllables: Number of syllables in lyrics
- difficult_words: Number of words not on the Dale–Chall "easy" word list
- num_dupes: Number of duplicate (repetitive) lines in lyrics
- num_words: Number of words in lyrics
- num_lines: Number of lines in lyrics
- genre_tags: song artist's associated genre tags

The positive, negative, neutral, and compound sentiment scores of each song's lyrics were gathered using Python's Natural Language Toolkit (NLTK) VADER model. Readability metrics for each song's lyrics were obtained using the textstat package for Python. Each song artist's associated genre tags were scraped using MusicBrainz API as well as the Musicbrainzng Python interface.

Due to data availability limitations, the data set provides approximately 80-90% coverage for all the songs on Billboard's list from 1950-2015.

We obtain audio features of each song in our billboard data set using Spotify's Web API. The audio features of each song are listed below:

- danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
- key: The key the track is in. Integers map to pitches using standard Pitch Class notation.
- loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Values typical range between -60 and 0 db.
- mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording, the closer to 1.0 the attribute value.
- acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- instrumentalness: Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
- liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
- valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- tempo: The overall estimated tempo of a track in beats per minute (BPM).
- duration_ms: The duration of the track in milliseconds.
- time_signature: An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).

Appendix 2

PCA 1960s

```
## Importance of first k=10 (out of 21) components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.0368 1.7334 1.44433 1.25109 1.21717 1.16063 1.10989
## Proportion of Variance 0.1975 0.1431 0.09934 0.07453 0.07055 0.06415 0.05866
## Cumulative Proportion  0.1975 0.3406 0.43995 0.51449 0.58503 0.64918 0.70784
##                            PC8     PC9    PC10
## Standard deviation     1.00614 0.95980 0.88960
## Proportion of Variance 0.04821 0.04387 0.03769
## Cumulative Proportion  0.75604 0.79991 0.83760
```

PCA 1970s

```
## Importance of first k=10 (out of 21) components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.0896 1.6086 1.5290 1.37228 1.24073 1.10958 1.04934
## Proportion of Variance 0.2079 0.1232 0.1113 0.08967 0.07331 0.05863 0.05243
## Cumulative Proportion  0.2079 0.3312 0.4425 0.53216 0.60546 0.66409 0.71652
##                            PC8     PC9    PC10
## Standard deviation     1.03131 0.91817 0.87106
## Proportion of Variance 0.05065 0.04014 0.03613
## Cumulative Proportion  0.76717 0.80732 0.84345
```

PCA 1980s

```
## Importance of first k=10 (out of 20) components:
##                            PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.8244 1.5448 1.37452 1.33312 1.26753 1.13748 1.08364
## Proportion of Variance 0.1664 0.1193 0.09447 0.08886 0.08033 0.06469 0.05871
## Cumulative Proportion  0.1664 0.2858 0.38022 0.46908 0.54941 0.61410 0.67282
##                            PC8    PC9    PC10
## Standard deviation     1.0315 0.9950 0.93986
## Proportion of Variance 0.0532 0.0495 0.04417
## Cumulative Proportion  0.7260 0.7755 0.81968
```

PCA 1990s

```
## Importance of first k=10 (out of 21) components:
##                            PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.2070 1.5573 1.4999 1.31277 1.18845 1.17275 1.03400
## Proportion of Variance 0.2319 0.1155 0.1071 0.08207 0.06726 0.06549 0.05091
## Cumulative Proportion  0.2319 0.3474 0.4546 0.53664 0.60390 0.66939 0.72030
##                             PC8     PC9    PC10
## Standard deviation     0.98877 0.86696 0.85528
## Proportion of Variance 0.04656 0.03579 0.03483
## Cumulative Proportion  0.76686 0.80265 0.83748
```

PCA 2000s

```
## Importance of first k=10 (out of 21) components:
##                            PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.8115 1.5715 1.4549 1.35752 1.29990 1.20228 1.09927
## Proportion of Variance 0.1563 0.1176 0.1008 0.08776 0.08046 0.06883 0.05754
## Cumulative Proportion  0.1563 0.2739 0.3747 0.46243 0.54289 0.61172 0.66927
##                             PC8     PC9    PC10
## Standard deviation     1.07008 0.97658 0.90718
## Proportion of Variance 0.05453 0.04541 0.03919
## Cumulative Proportion  0.72379 0.76921 0.80840
```