CS109a Final Project
Clara Chen, Caleb Shi, Catherine Stanton, Nathan Zhao

What Causes a Hit: Predicting Song Success Based on Song Elements

Section 1: Background and Introduction

Music is one of the most historic modes of human expression and communication. Indeed, the creation of auditory impulses for expression is thousands of years old, and genres of music are related to — and often defined by — the time in which they were produced. Since the onset of the music-sharing devices and then the internet, ranking the popularity of songs — with the most successful being termed "hit" songs — has become its own industry. Billboard Magazine began publishing the Hot 100 Chart[1] in its current form in 1955: the list tracks weekly sales, streaming, and airplay.[2] The songs on the Hot 100 chart are determined by several "component" charts focused on radio airplay, single sales, streaming songs, and digital songs.

Especially as streaming services have gained popularity, the music industry has been inundated with large sets of data provided by streaming giants like Spotify. With that data comes the ability to ask interesting questions; a valuable one to industry executives and enthusiasts is, **What musical attributes in a song can most accurately predict whether or not it will be listed on the Billboard Hot 100 Chart?** While an individual executive might have their own hypotheses on what leads to success, the host of data from streaming services provides a quantitative and more objective way to predict song success[3].

For our purposes we will use "The Spotify Hit Predictor Dataset (1960-2019)" from Kaggle[4]. This dataset includes track names, artists, song features (such as danceability, energy, duration, etc.) for each track as shown on Spotify, and also whether the track was on the Billboard Hot 100 Chart during its decade of release. Given the features, we initially hypothesize that danceability, liveness, and valence will most accurately predict whether or not a song is listed on the Hot 100 Chart, based on our references[5] [6] [7] and intuition about what songs have been most popular in the past 50 years.

The original data on Kaggle is split into 6 files with each representing a decade from the 1960s-2010s. We concatenated the six files into one dataframe, with an added predictor for the decade in which the song was released. The combined dataset includes 41106 observations and 19 predictor variables. Then, we conducted an exploratory data analysis with visualizations and subsequently fit a naive model. We fit two main types of predictive models: logistic regression and decision trees, which are best suited to the categorical nature of our response variable. To further interpret these models and explore our data, we generated a feature importance graph and conducted principal component analysis. In order to properly evaluate the models, we split the data into training and test sets, with 80% of the observations used for the training data.

Section 2: Exploratory Data Analysis

Motivation

Before we began creating models, we created and analyzed some basic visualizations of the trends in how each numerical predictor variable relates to status on the Billboard Hot 100 Chart.

I: Visualizations

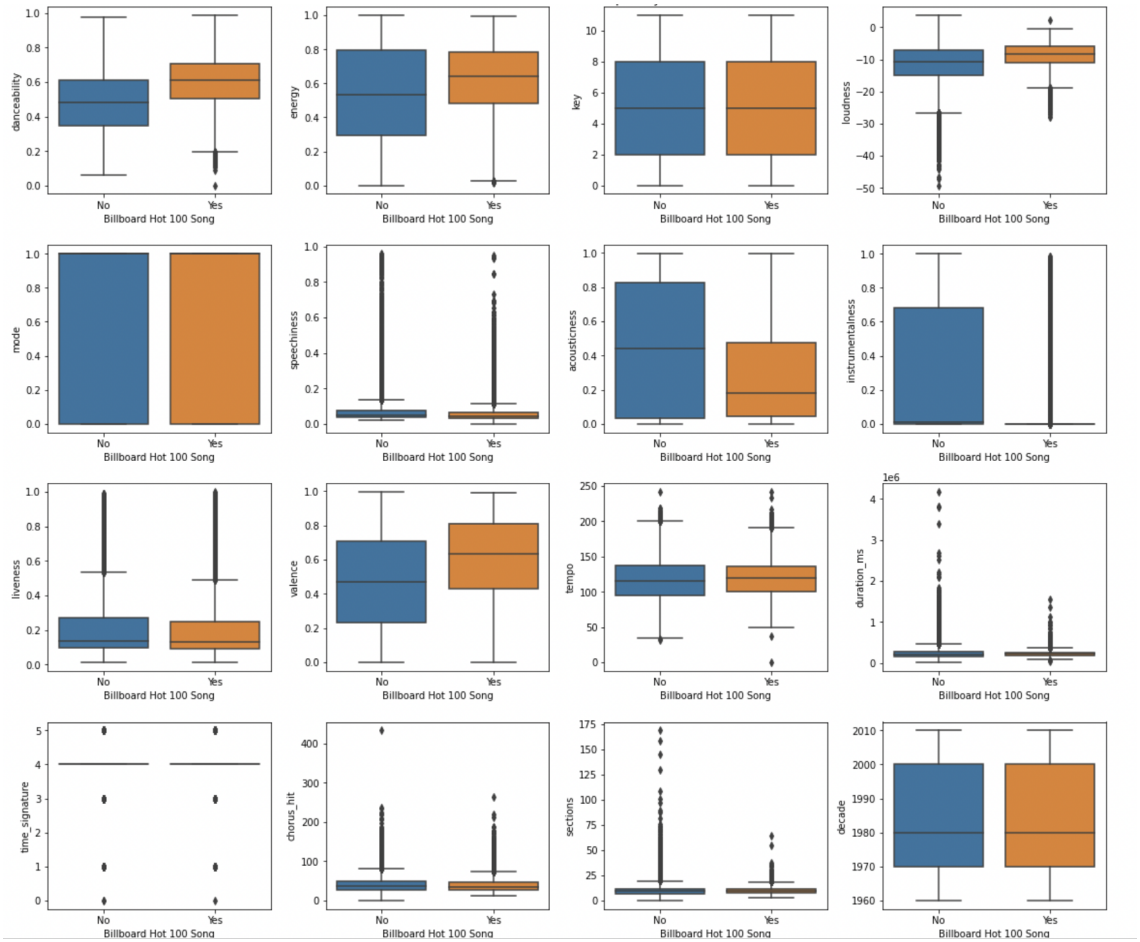

Figure 1: Boxplots of each numerical predictor variable, which are all song features, against the target variable, which is the binary value of whether a song was on the Billboard

For most of the predictors, the boxplots between popular and unpopular songs are similar (Figure 1) . Variables that show the most visible differences between successful and non-hit songs are danceability, loudness, instrumentalness, valence, and duration. According to the plots for these variables, songs that make it to the billboard tend to have higher danceability, loudness, and valence and lower instrumentalness and duration. However, there are many outliers in several of the boxplots, making it difficult to determine the shape of the distribution from box plots alone.

The last boxplot, showing the distributions of decades, indicates that the decade-wise datasets are balanced between songs on the Billboard Hot 100 Chart and songs not on it because the median, 25th percentile, and 75th percentile decades are the same for songs in each category. This tells us that the dataset is balanced between the classes, which is helpful for training models. However, this could be a limitation because, in real life, there will not be an equal number of songs on the chart and off at any point in time (there will be many more songs off the chart since only 100 songs make the chart). It's difficult to correct for this issue when analyzing our data because removing songs could be biased against some of our many features (i.e. we may inadvertently remove a lot of songs that share some feature(s)), discounting the importance of those features in the final model.

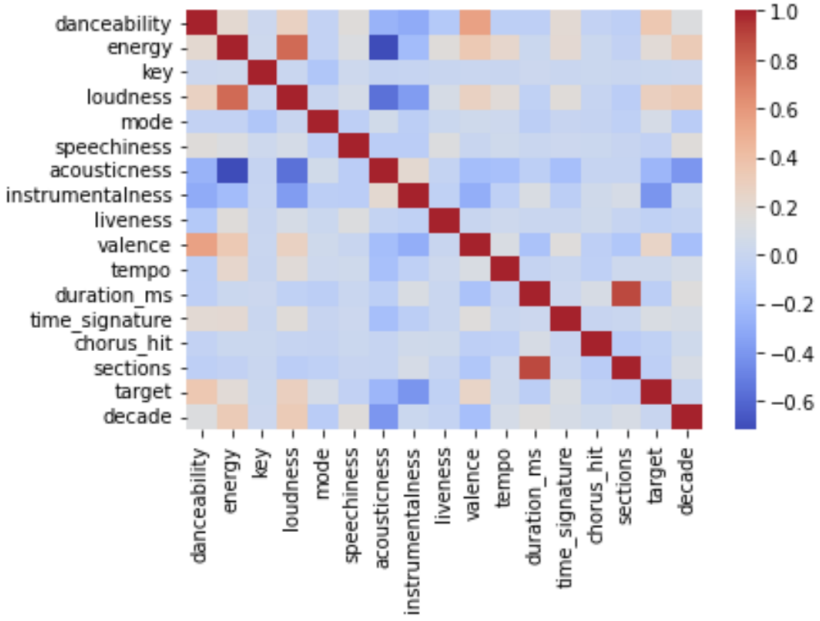

Figure 2: *Correlation matrix between all numerical variables*

We also fit a correlation matrix (Figure 2) to see which variables in the dataset were most strongly related to each other. In general, the colors are very muted, meaning that there isn't too much collinearity between variables. There are a few exceptions such as duration and section, or loudness and energy, which have positive correlations with each other. On the other hand, acousticness and energy exhibit strong negative correlations with each other. The strongest relationship between a predictor variable and our response is instrumentalness, and the relationship is negative.

The last element of our exploratory data analysis is a plot tracking change in mean and median values for each numerical predictor value over the decades between 1960 and 2010 (Figure 3). This is useful because it helps us decide which features may have contributed to a song's success more heavily in a given decade. For example, the mean and median of valence are high in the 1960s through the 1980s, but then they fall off (i.e. songs published in the 1990s and later had

lower valances). On the other hand, the loudness levels of songs published between the 1960s and 1980s was low, and grew considerably between the 1990s and 2010s. This could indicate that, in the 1960s through 1980s, high valence was a predictor for a song's success and high loudness was not. But during the 1980s, it's possible that the musical taste of the population changed to favor loudness, which would explain why songs started having less valence and more loudness on average. The changes in mean and median of numerical song properties over time could reflect artists changing style over time to keep up with trends in music on the Billboard Hot 100 Chart, or it could reflect the emergence of new artists with new features to contribute. Regardless, it would be interesting to compare this plot with the feature importance graphs generated later in our analysis.

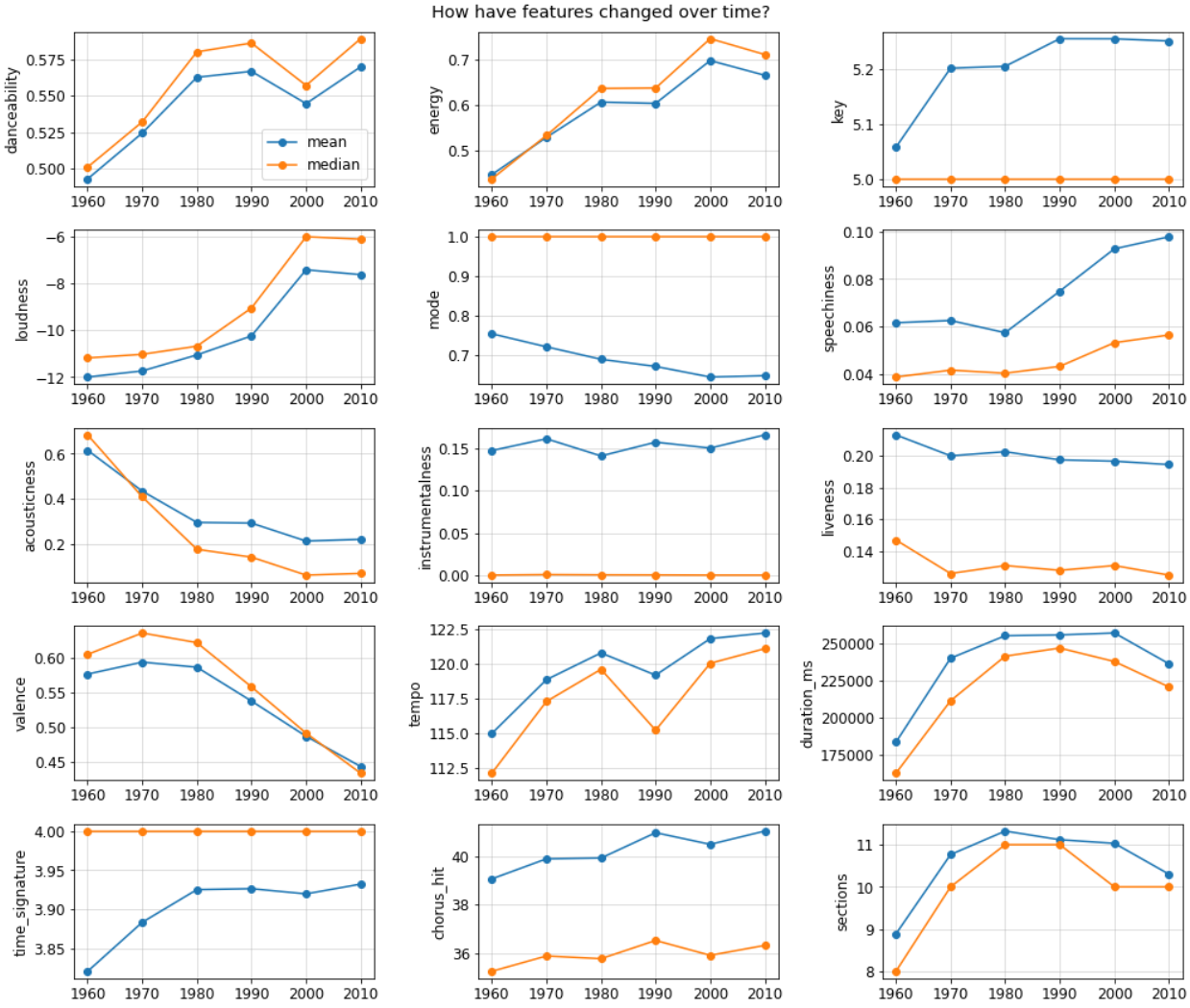

Figure 3: Mean and median song feature values over time

II: Naive Model

The proportions of songs on the Billboard Hot 100 Chart in both the train and test set are close to 50%. Specifically for the training set, 49.95% of the tracks were on the chart. Therefore, a naive model classifying all tracks in the majority category would classify all tracks as not being on the

Billboard Hot 100 Chart. In the test set, 50.22% of tracks were on the chart. So in the test set, the naive model would have a 49.78% accuracy rate. This will serve as a baseline comparison for more complex models in our analysis.

Section 3: Logistic Regression

Motivation

As outlined in our background section, being on the Billboard Hot 100 Chart or not is a categorical variable, since it takes on a value of True/False, Yes/No, 0/1. Therefore, a logistic regression model with two categories is a natural choice for a first model.

I: Basic Model

We begin by implementing a logistic regression model with no penalty using all 16 numerical predictors[1]. Based on the classification accuracy scores, including no penalty in the regression fails to improve the prediction of hits beyond the naive model, as the train score was 0.5054 and the test score was 0.4973. Given the large number of predictors within our dataset, we then considered whether imposing a penalty would improve the predictive ability of our model.

The default penalty for LogisticRegression() is the L2 norm, or ridge regularization. Implementing this penalty on the regression model does not improve the predictive accuracy of the model, as the train and test scores remained the same. Notably, the coefficients of the regression also did not change. This could be due to the coefficients of the no penalty model being already small and were therefore not penalized by the regularization. Alternatively, the regularization parameter may not have been suitable since we used the default value for this baseline model. Next, we used an L1 penalty, or lasso regularization. In order to do so, we changed the solver to 'liblinear' as required in the documentation. Including the L1 penalty significantly increased the predictive ability of the model, as the train score was 0.7340 and the test score was 0.7332. Additionally, the coefficients of the model became more extreme, with certain predictors emerging as having larger effects on the odds of a song being a hit.

After L1 regularization, the danceability of a song seems to be a major driver of if the song will be a hit. With a regression coefficient of 3.609, it seems that the danceability of the song significantly improves the probability of a song being a hit, which intuitively holds. Interestingly however, the energy of the song seems to be a negative predictor of hits, creating a bit of a contradiction with danceability considering that both measures consider similar variables (tempo, loudness, etc.). Further, three predictors also seem to be negatively correlated with hits: speechiness, acousticness, and instrumentalness. Speechiness measures how close to natural speaking the track is, so the negative correlation makes sense given that audiobooks and talk shows do not qualify for the Billboard Top 100. Acoustic tracks tend also not to be hits based on the regression as are tracks that are most instrumental.

[1] See figure 7 for the numerical predictors.

II: Cross Validation

In order to see if we could improve upon our lasso-like logistic regression model, we next applied cross-validation to refine our regularization hyperparameter. Using five-fold cross validation on our lasso-like logistic regression model, we found that our model has a training prediction accuracy of 0.7351 and a test prediction accuracy of 0.7340. Comparing this to our original L1 model, we see that the model's predictive ability only improves slightly with cross validation. Looking at the coefficients of the regression model, we see that coefficients generally remain similar even with the cross validation.

	model	train_score	test_score
0	no penalty	0.5054	0.4973
1	L1 penalty	0.7340	0.7332
2	L2 penalty	0.5054	0.4973
3	L1 penalty with 5-fold CV	0.7351	0.7340

Table 1: Training and test scores for logistic regression models.

Seeing that cross validation does not improve our base logistic regression model significantly, we will now turn our attention to other classification models. If we were to continue with the logistic regression model, we would look at pairs of predictors and formulate visualizations/decision boundaries to further analyze how these predictors affect whether a song is a hit or not.

	model	intercept	danceability	energy	key	loudness	mode	speechiness	acousticness
0	L1 penalty	16.51401	3.60942	-1.74171	0.01049	0.11177	0.41394	-2.97993	-1.48508
1	L1 penalty with 5-fold CV	20.81269	3.67248	-1.76374	0.01031	0.11537	0.41243	-2.96394	-1.52656

instrumentalness	liveness	valence	tempo	duration_ms	time_signature	chorus_hit	sections	decade
-3.39982	-0.24830	0.11662	0.00250	-2.660920e-07	0.16158	-0.00183	-0.01153	-0.00823
-3.40598	-0.25751	0.06414	0.00257	-1.971800e-06	0.16030	-0.00189	-0.01260	-0.01037

Table 2: Logistic regression coefficients of numerical predictors

Section 4: Decision Tree

Motivation

Fitting a single decision tree would allow us to classify songs as on the Billboard Hot 100 Chart or not, based on a series of splitting criteria (each song could traverse the decision tree and be categorized based on its values for certain predictors). While the logistic regression can be difficult to visualize in high dimensions, the decision tree remains relatively easy to interpret even with high depths.

Furthermore, in this case, a single decision tree makes more sense than an ensemble model, such as bagging, because in an bagged model, we would risk fitting models on sample data that includes the same song more than once and could therefore be biased towards unique features of that song.

I: Single Decision Tree Model

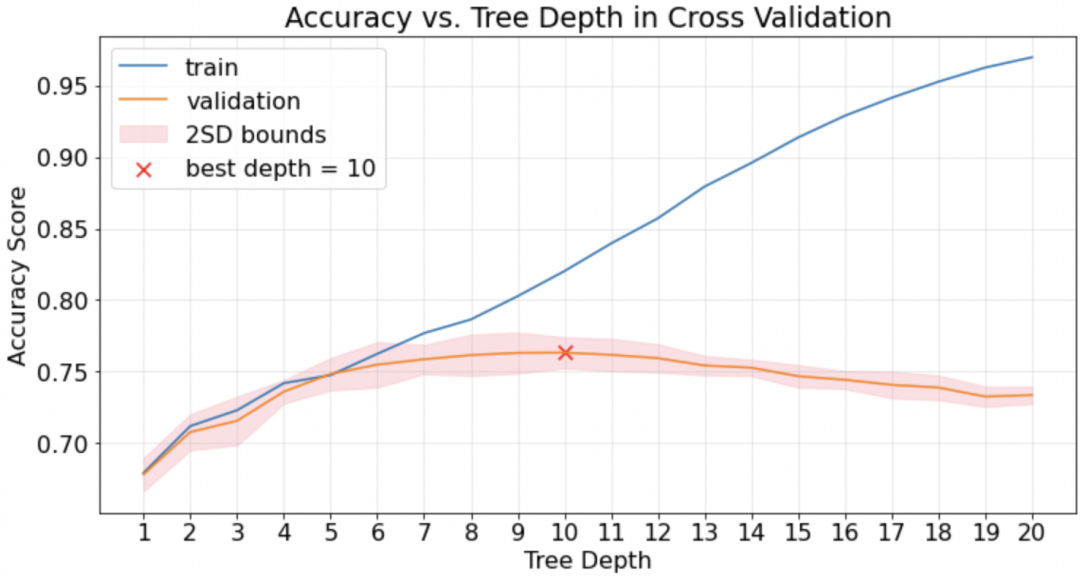

Figure 4: Accuracy vs. tree depth based on cross validation results

We fit a decision tree of depth 10 on the entire training dataset with the depth being chosen based on cross validation (Figure 4). The training accuracy of this model was 82.0%, and the test accuracy was 76.6%, which is over 25 percentage points better than the naive model and a few percentage points better than the logistic regression. From the confusion matrix created on the test data (Figure 5), it seems that the types of errors are not balanced. The model makes more errors when classifying songs that were not on Billboard than when classifying songs that did. This suggests that songs that aren't popular enough to reach the Billboard still share many similar features as songs that do.

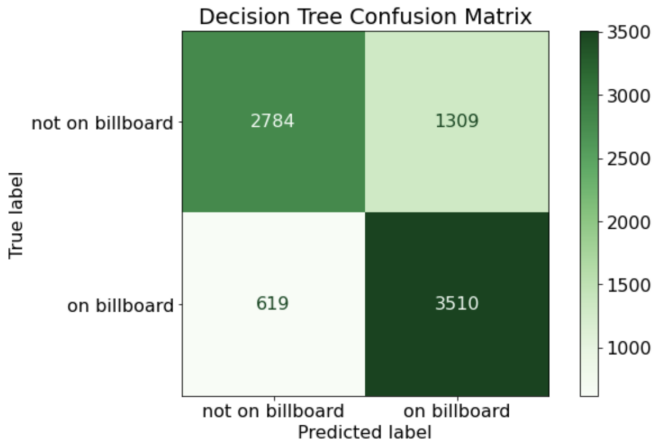

Figure 5: Confusion matrix of single decision tree with cross-validated depth predicting on test data

To get a better sense of what features contribute the most to popularity, we displayed the top of the tree (Figure 6) and created feature importance graphs (Figure 7). Instrumentalness was used for the first split of the tree, indicating that it is most predictive of which class a song will fall into. This is consistent with the correlation matrix in our exploratory data analysis where instrumentalness was the variable most correlated with the target variable of reaching the Billboard. Danceability, acousticness, and speechiness are also used in the first few splits of the tree, reflecting their importance as well. Given these top splits, the feature importance ranking is not surprising. For both mean decrease in impurity (MDI) and permutation importance, the ranking of features is very similar, with the same six variables - including decade, duration, and the four variables from the top splits of the tree - at the top of both feature importance charts.

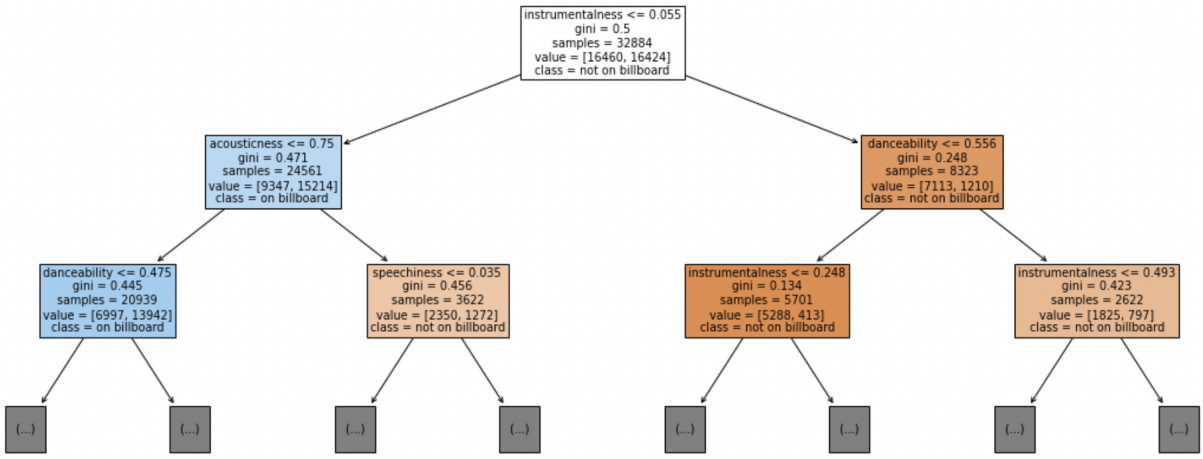

Figure 6: Visualization of single decision tree model up to depth 3

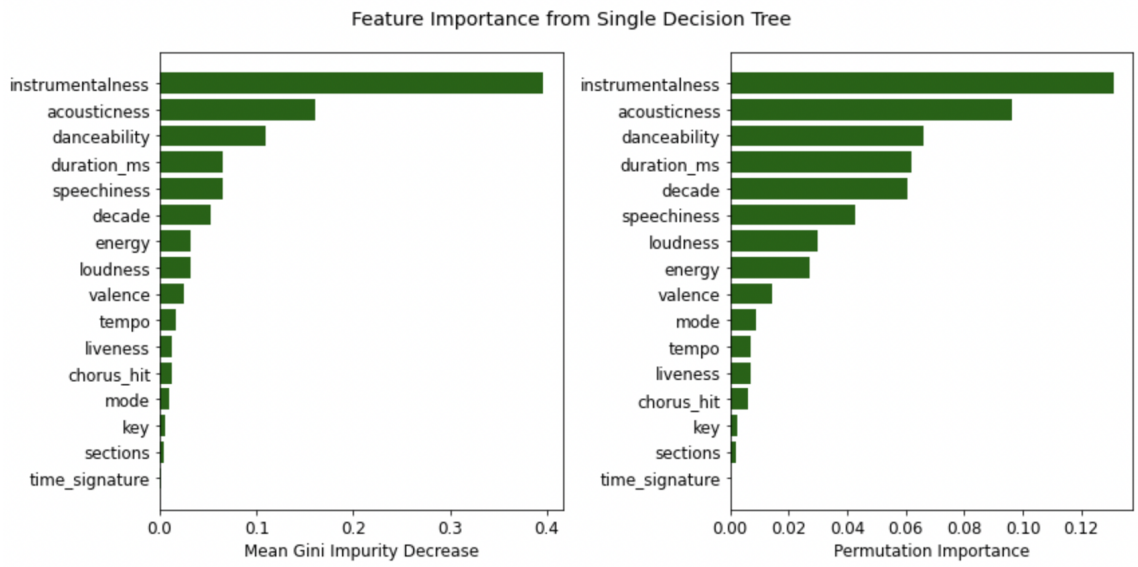

Figure 7: Feature importances for the single decision tree based on MDI and permutation importance methods

II: Decision Trees by Decade

As seen in our exploratory data analysis, the average values of many song features have changed from decade to decade. This suggests that features contributing to popularity may have also changed over time. To test this hypothesis, we fit individual decision trees for each decade. The proportion of songs from each decade was fairly similar between train and test sets (Figure 8), so we continue to use the same train-test split rather than creating a new stratified train-test split.

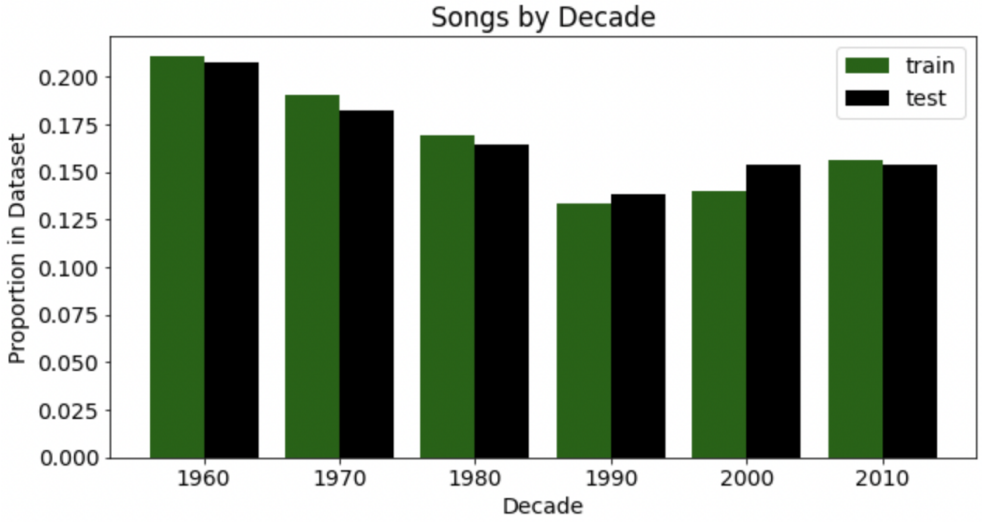

Figure 8: Comparison of decade proportions for songs in the train and test set

For each decade, a decision tree with maximum depth chosen through cross validation was fitted on songs in the training data from that decade. Including the decision tree fit on all training

observations, we created 7 decision trees, for which the train and test accuracies are summarized in table 1. The decade-specific trees performed similarly to the overall decision tree in terms of accuracy. However, the trees for more recent decades display some signs of overfitting as there is a nearly 6 percentage difference between the train and test accuracies for 2000 and 2010.

	decade	train_accuracy	test_accuracy
0	all	0.820156	0.765507
1	1960	0.773501	0.751465
2	1970	0.787458	0.726484
3	1980	0.822206	0.750555
4	1990	0.838769	0.812335
5	2000	0.873915	0.814873
6	2010	0.862405	0.808208

Table 3: Train and test accuracies for all decision tree models

Based on the decision trees for each decade, feature importances (MDI) were extracted for instrumentalness, acousticness, danceability, duration (milliseconds), and speechiness - the features that ranked at the top for feature importance of the single decision tree fit on all decades. The importance of these features was plotted over time (Figure 9), and we observed that the features that most determine whether a song will be on the Billboard has, in fact, changed between decades. Instrumentalness has always ranked as the most important feature, but has become increasingly important in the last two decades. Similarly, with the exception of 2010, danceability has been increasingly important for determining whether or not a song will be on the Billboard, matching our initial hypothesis. By contrast, acousticness and speechiness have generally decreased in their importance from decade to decade. Given the trends from previous decades, it seems that in the future, the single most important feature that determines a song's ability to be selected for the Billboard Hot 100 is its instrumentalness. We've previously found instrumentalness to be negatively correlated with reaching the Billboard Hot 100, so it seems that songs that have more instrumental music instead of vocals are becoming less popular over time.

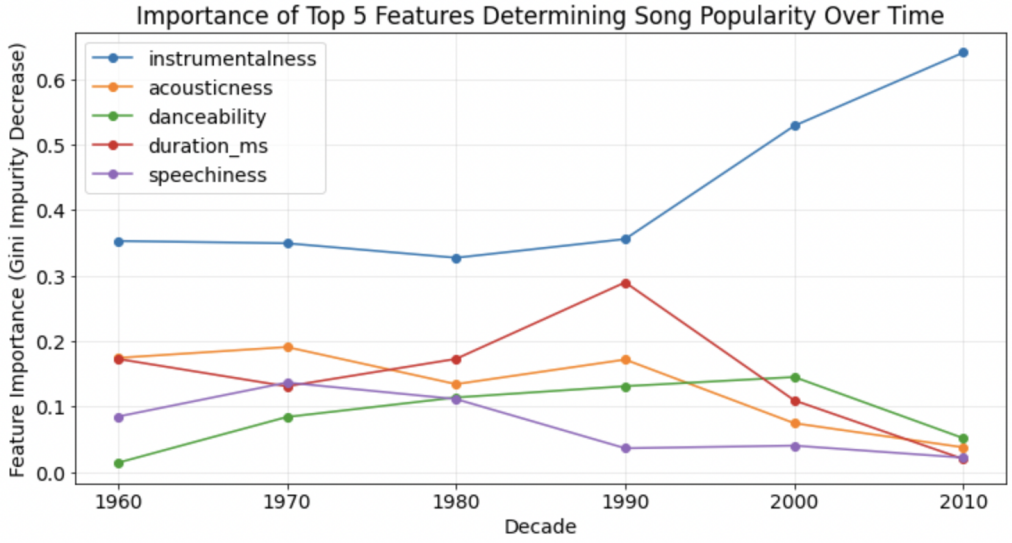

Figure 9: Feature importance over time for the top 5 features in figure 7

Comparing these feature importances with the graph of features over time created in our exploratory data analysis, it seems that the average trend in song features does not necessarily relate to features that are most important for reaching the Billboard Chart. For example, the instrumentalness of songs have not changed much on average even though that feature has become more and more important. There are some similarities for other variables. Songs have become more danceable and less acoustic over time, which matches the trends in feature importance as well.

Section 5: Principal Component Analysis

Motivation

Given that our data has many predictor variables, we found it relevant to reduce the dimensionality. Linear combinations of the first few principal components would help us determine which variables explained the greatest amount of variance between songs that succeeded in being placed on the Billboard Hot 100 Chart and those that failed. By projecting the data onto the first few principal components, we were able to gauge similarity between songs in terms of their principal features, and whether they clustered by genre in the visualization of our principal component analysis.

Using the Spotify API, we gathered and appended a genre variable to our entire dataset. Genre was tricky to work with because of the way it's stored by Spotify. Rather than assigning genres to individual songs, Spotify assigns genres to artists and albums. But the Spotify API has the potential to "recommend" songs based on a given genre. Thus, to generate genres associated with songs, we queried Spotify for its song recommendations for each of its 126 genres, and then matched the recommended songs to those in our Kaggle dataset. We found 924 songs that were common between Spotify's recommendations and our existing data. A reason that we could not find more is that the API can only recommend a maximum of 100 songs for each given genre and incorporates randomness in the songs it chooses.

When plotting the data from our principal component analysis, we used colors to differentiate genres. We chose to limit our visualization to the 10 most frequent genres for clarity. This plot let us see whether songs of the same genre were similar to each other in terms of the first few principal components. This is a particularly interesting observation because it lets us gauge the extent to which Spotify's genre labels are based on the first few principal components of our song data and the extent to which they are based on other features that we cannot access.

I: Principal Component Analysis Visualization

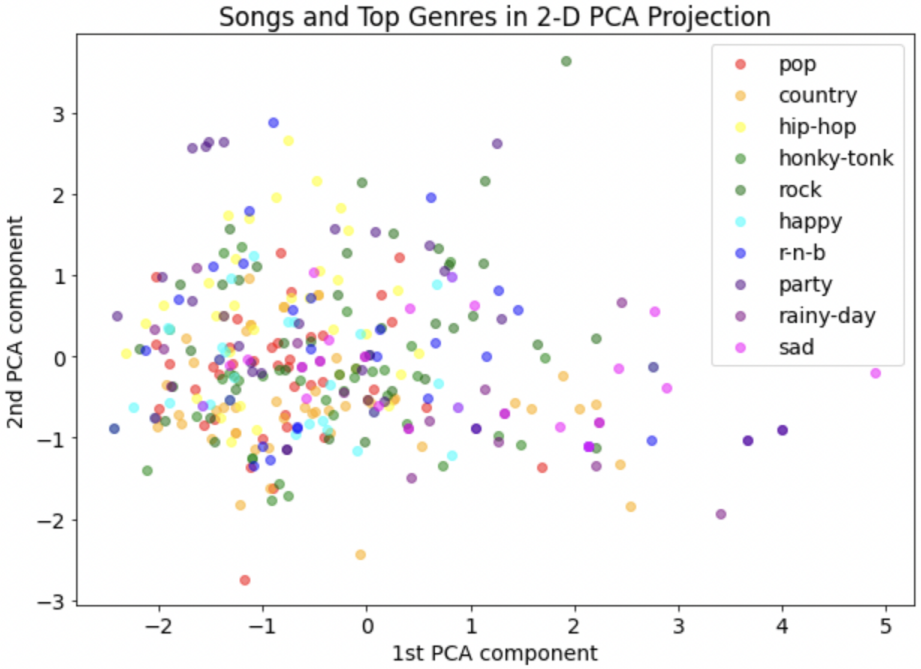

Figure 10: Principal component analysis projected onto the first two principal components, color-coded by the top 10 musical genres

The table to the right displays the feature variables and their coefficients in terms of the first principal component. The features are sorted in order of ascending coefficients, but the coefficient magnitudes are most important to interpret. We see that loudness, energy, and acousticness have the highest magnitudes, meaning that they explained most of the variance between songs in the dataset. Some of the variables that were most predictive of popularity in previous models, including speechiness and danceability, did not seem to contribute to as much variance in the data. This suggests that although there is some correlation, features that are most different between songs are not always those that determine whether a song will be popular or not.

	Feature	Coefficient
3	loudness	-0.473046
1	energy	-0.449011
9	valence	-0.260600
0	danceability	-0.234141
15	decade	-0.221885
12	time_signature	-0.204480
5	speechiness	-0.115810
10	tempo	-0.113735
8	liveness	-0.032460
2	key	-0.030258
13	chorus_hit	0.046383
4	mode	0.084399
14	sections	0.188836
11	duration_ms	0.200890
7	instrumentalness	0.256840
6	acousticness	0.425749

Table 4: Features and First Principal Coefficients (ascending) used in the Principal Component Analysis.

The 2D PCA projection (Figure 10) lacks clustering between songs of the same genre. In general, the songs are pretty evenly spread out across the projection and do not form distinct groups. Although the first principal component spans a slightly wider range, the lack of a clear trend in the points shows that there isn't too much difference between the variance explained by the two components. The equity in the distribution of different songs could be a reflection of our small set of songs used in the PCA, a result of the fact that we chose to use only the top 10 most

popular genres on Spotify, or a result of the fact that Spotify's genre-categorization is based on playlist titles and descriptions rather than song features [8]. The result is not completely surprising because songs of the same genre usually have high originality (both so that artists can express themselves and avoid copyright violations).

Section 6: Discussion and Conclusion

Across both logistic regression and decision trees, we found that the variables danceability, speechiness, acousticness, and instrumentalness were most predictive of whether a song would reach the Billboard Hot 100 Chart. The first variable was positively correlated with popularity and the latter three were negatively correlated. This was interesting in terms of the principal component analysis in which we found that loudness, energy, and acousticness explained most of the variance in our dataset. These findings somewhat align with our initial hypothesis, as we expected upbeat songs with high danceability would be more likely to reach the Billboard Chart, but other variables that we thought would be significant (liveness and valence), were not as important based on our analysis.

There exist inherent limitations in our models and analysis. First, the most prolific music-ranking chart, the Billboard Hot 100 Chart, is merely a proxy for popularity and opinions in the music industry are split as to whether it ought to be the standard-bearer. Second, as mentioned above, Spotify does not use song features to conduct genre classification, merely relying on titles and descriptions. From a musical standpoint, many would argue that not every title and description can sufficiently capture the musical genre of a piece. Finally, given the intricate ties between music and the zeitgeist, genre is inherently political and limited, as it does involve making subjective judgments as to the *category* of a given song.[2]

There are many avenues for future work. First, we would be eager to analyze based on other proxies aside from the Billboard Hot 100 Chart, including Spotify's Global Top 100 Chart and analogous charts from Apple Music and other streaming platforms. Of course, we would have to consider the potential bias of streaming platform rankings (i.e. they are incentivized to rank songs higher by artists that they promote more often). Second, we're inclined to explore other sources for music attributes apart from Spotify, including other streaming platforms, particularly given Spotify's methodology for determining, say, genre. We have yet to find a strong alternative through our research, but perhaps rigorous genre identification is an avenue that the music industry should invest in order to fully leverage the benefits of data in predicting song success.

[2] An example of the genre's political nature is in the controversy surrounding Lil Nas X's *Old Town Road*. In 2019, Billboard removed the song from its "Hot Country" chart, classifying it instead solely as "Hot R&B/Hip-Hop." [9]

References

[1] “Billboard Hot 100.” *Billboard*, Billboard, 13 Sept. 2022, https://www.billboard.com/charts/hot-100/#.

[2] Molanphy, Chris. “How the Hot 100 Became America's Hit Barometer.” *NPR*, NPR, 1 Aug. 2013, https://www.npr.org/sections/therecord/2013/08/16/207879695/how-the-hot-100-became-americas-hit-barometer.

[3] Thompson, Derek. “The Dark Science of Pop Music.” *The Atlantic*, Atlantic Media Company, 30 Nov. 2014, https://www.theatlantic.com/magazine/archive/2014/12/the-shazam-effect/382237/?utm_source=copy-link&%3Butm_medium=social&%3Butm_campaign=share.

[4] Ansari, Farooq. “The Spotify Hit Predictor Dataset (1960-2019).” *Kaggle*, 25 Apr. 2020, https://www.kaggle.com/datasets/theoverman/the-spotify-hit-predictor-dataset.

[5] “The Scientific Formula for Predicting a Hit Song.” *New Atlas*, 2 May 2015, https://newatlas.com/predicting-hit-songs/20939/.

[6] Askin, Noah, and Michael Mauskapf. “Cultural Attributes and Their Influence on Consumption Patterns in Popular Music.” *SpringerLink*, Springer International Publishing, 1 Jan. 1970, https://link.springer.com/chapter/10.1007/978-3-319-13734-6_36.

[7] Khalil, Hoda, et al. “Can Big Data Really Predict What Makes a Song Popular?” *The Conversation*, 11 Nov. 2022, https://theconversation.com/can-big-data-really-predict-what-makes-a-song-popular-189052.

[8] “Help - How Do Genre Charts Work? – Spotify for Artists.” *Help - How Do Genre Charts Work? – Spotify for Artists*, https://artists.spotify.com/en/help/article/how-genre-charts-work.

[9] Leight, Elias. “Lil Nas X's 'Old Town Road' Was a Country Hit. Then Country Changed Its Mind.” *Rolling Stone*, Rolling Stone, 6 June 2019, https://www.rollingstone.com/music/music-features/lil-nas-x-old-town-road-810844/.