Detecting Stylistic Variation in Pop Production*

Michaela Drouillard

December 1, 2023

This study investigates the feasibility of predicting whether a song is produced by Jack Antonoff using Spotify's audio feature API. We developed two models: a logistic regression model and a random forest model. The logistic regression model, focusing on Antonoff's collaborators' discographies, demonstrates an accuracy of 76%, precision of 39%, and a recall of 76% when predicting whether a track has an Antonoff credit. The random forest model, which includes discographies of other notable pop producers, shows an accuracy of 61% when predicting tracks' producers, comparing favorably to a No Information Rate of 43%. A significant finding is that 'danceability', a metric developed from human-tagged data, emerges as a crucial predictor in both models. This research contributes to the broader understanding of computational analysis in cultural contexts, illustrating how a blend of human judgment and algorithmic processing can approximate musical styles

1 Introduction

It is widely acknowledged that style recognition in music is difficult. Style, which is hard enough to define without building computational representations of it, is often constructed using lower-level perceptual features such as a pitch and tempo. We are interested in understanding how to create an approximation for production style based on a data set built using the Spotify audio features API, which includes both crowdsourced features and features derived from machine listening techniques. To do this, we explore different audio feature metrics and identify which were the most import predictors in identifying pop producers' style. We focus on Jack Antonoff, as he is a prolific and widely used producer, who also has his own solo acts.

By way of background, Antonoff began his career as a guitarist in fun., and is the frontman in Bleachers. He has produced and co-produced music for artists including Lorde, Taylor Swift,

*Code and data are available at: https://github.com/michaeladrouillard/spotify_pop. We thank Rohan Alexander, Carl Wilson and Inessa De Angelis for valuable suggestions. For any comments or suggestions, please contact michaela.drouillard@mail.utoronto.ca

Florence and the Machine, The Chicks, Clairo, The 1975, Grimes, Zayn, Pink, Lana Del Rey, Olivia Rodrigo, and the Minions: The Rise of Gru soundtrack. His style is considered identifiable, and his collaborators are influential – as Andrew Marantz writes in the *New Yorker*: "When his band releases an album, the world responds politely. When he produces one by Lorde or Lana Del Rey or Taylor Swift, the world wobbles on its axis" (Marantz).

We obtained tracks from six artists with extensive discographies who have collaborated with Antonoff, both before and during their collaboration. We also obtained tracks from similarly prolific pop producers: Joel Little, Ariel Rechstaid, Max Martin, Greg Kurstin, Paul Epworth and Rick Nowels. Our data set contains 3738 tracks in total, which were split into a training and test sets.

We built two models to explore the data. The first model, a logistic regression, is employed to observe any patterns in audio feature scores across Antonoff's collaborations with artists. The visualizations and analysis offer a closer look at Antonoff's oeuvre, along with a portrait of the trends in female singer-songwriter contemporary pop (according to Spotify's API). It draws on a dataset of artists who have collaborated with Antonoff (1954 tracks total), with a binary variable to tag whether the song was an Antonoff production. The training set included 1465 tracks, and the testing set included 489 tracks.

The second model, a Random Forest (RF), trains on a data set that includes the discographies of other prominent pop producers to classify producers. This expands upon and complements the results of the logistic regression – where a logistic regression predicting an Antonoff-produced track in an artist's discography might actually be picking up on an artist releasing pop music, training a Random Forest on a bigger dataset with more producers captures both non-linear relationships in the data and accounts for intra-genre differences. In obtaining the strongest predictor variables in the Random Forest, we can better understand how it distinguishes between different pop producers' styles.

We find that, in both models, `danceability` is the strongest predictor of whether a song has been produced by Antonoff, and the strongest predictor in differentiating between producers generally. Additionally, when we look at how danceability relates to other features in the dataset, we find that the differences in these relationships are informative: the degree to which danceability's relationships with other features vary helps us most accurately classify tracks. In other words, using danceability along with its interactions with other dataset features gives us the most accurate predictions.

In an interview with Spotify principal engineer Glen McDonald, we learned that the danceability and valence features were built using crowdsourced methods. With the understanding that what we understand as "data-driven" in recommendation systems is often intertwined with human subjectivity – engineers' understandings of taste and taste-making also shape the recommendation systems that they build (Seaver) – our results suggest that it's possible to create representations of style by combining features derived from both automated audio analysis and more crowdsourced, subjective contributions. This research adds to a more informed

understanding of how style and tastes can been algorithmically rendered using Spotify's audio features.

The remainder of this paper is structured as follows: Section 2 details the data curation and extraction processes, including documentation on features from Spotify's audio features data set. Section 3 specifies the models used and Section 4 details their performance and most important predictors. Section 5 situates the study in the broader context of computational stylistics and recommendation systems, detailing how we can conceptualize "style" given the results and limitations of our study. We provide background on the Spotify features' provenance based on an interview we conducted with Glen McDonald, a principal engineer at Spotify.

2 Data

We collected data from the Spotify API, which we accessed using the `spotifyr` package in R R Core Team. For our first dataset, we acquired audio feature data on the complete Lorde, Taylor Swift, St. Vincent, Lana Del Rey, The Chicks, Florence and the Machine, and Bleachers discographies. HAIM, Marina and the Diamonds, Maggie Rogers, Sharon Van Etten, and Mitski were included to represent other contemporary pop artists with overlapping fan bases who have never collaborated with Antonoff. Of the artists who did collaborate with Antonoff, we chose artists who have produced multiple albums, and at least one album with Antonoff. This dataset contains 1954 tracks in total.

For our second dataset, we acquired the discographies of comparable contemporary pop producers' by scraping their respective discography Wikipedia pages. Tracks were included if producers were either the solo producer, the primary producer, or a co-producer (a limitation of this study being that we can not know or truly parse who takes ownership for what during creative collaborations, especially when our audio features use one number to describe an entire song). The producers include Epworth, Rechstaid, Max Martin, Nowels, Joel Little, and Greg Kustin. These producers were chosen based both on their popularity, and because some have collaborated with the same artists as Antonoff. We combined these tracks with the first datasets tracks, for a total of 3738 tracks from 828 different artists[1].

The data was manually validated, and live performances, karaoke editions, international versions or translations, and remix albums were removed. Where deluxe albums were available, original albums were deleted to avoid duplicate rows.

We created a binary variable, `is_antonoff`, which contains 1 if Jack Antonoff produced or co-produced the song, and 0 if it was produced by somebody else. The information underpinning this discography was drawn from the "Jack Antonoff Production Discography" Wikipedia page (Wikipedia).

[1] In the full, combined version of the dataset, there were tracks (for instance, by Mitski or Maggie Rogers) that were produced by producers other than the main producers included in our study. The producer columns for these tracks were marked as "other".

Table 1: Counts of Tracks Per Producer in Original Dataset

Producer	Number of Tracks
antonoff	326
epworth	321
kurstin	559
little	90
martin	778
nowels	42
other	1558
rechtshaid	64

Audio Features Dataset

Spotify provides variables for each song, some of which were developed by them, which enables comparisons across artists. The final audio features in our data set are: `artist_name`, `track_name`, `energy`, `danceability`, `key`, `loudness`, `mode`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valence`, `tempo`, `jack`.

Figure 1 depicts the range of scores for the tracks across all variables except for tempo, mode, and loudness, which are measured using different metrics. Spotify documents the `get_artist_audio_features` variables as follows (Spotify):

> **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.
>
> **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. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
>
> **instrumentalness**: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
>
> **speechiness:** Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the

closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

key: The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C /D , 2 = D, and so on. If no key was detected, the value is -1.

liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

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. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically 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.

tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

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).

By analyzing the cumulative distribution functions of various audio features, particularly when categorizing songs based on whether they were produced by Antonoff (Figure 2), distinct patterns emerge. Notably, while attributes such as liveness, instrumentalness, and speechiness exhibit relative homogeneity, there are slight disparities in features like danceability, energy, loudness, acousticness, and valence.

Examining the temporal evolution of artists' sounds, especially in correlation with their collaboration with Antonoff, provides insights into the broader trends in pop music. Each point in Figures 3-8 represent one track.

In Figure 3, we observe a downward trend in energy scores among artists collaborating with Antonoff. However, this trend appears consistent with the overall trajectories of these artists' works, suggesting broader shared trends in this area of pop, as exemplified by Lana Del Rey

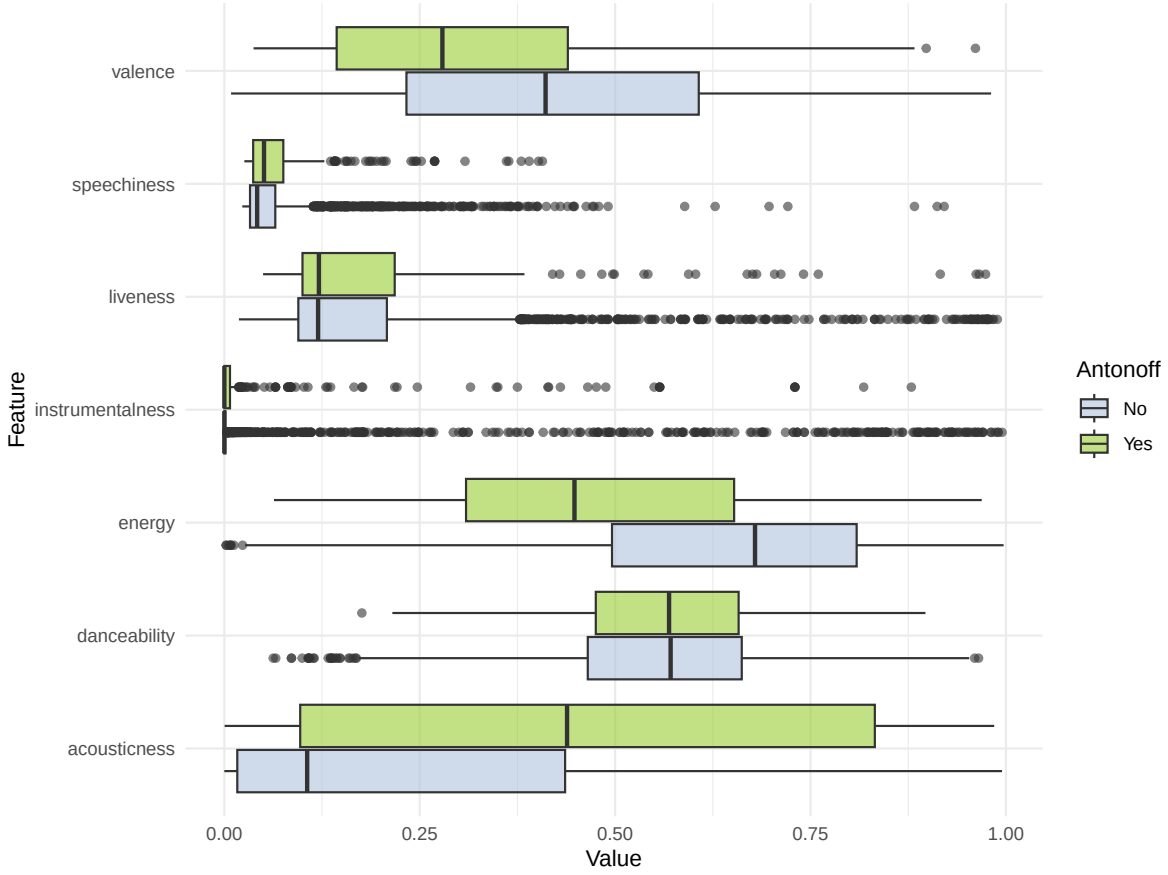

Figure 1: Comparing the Range and Mean Values of Each Variable From the Spotify Audio Features API for our Full Data Set

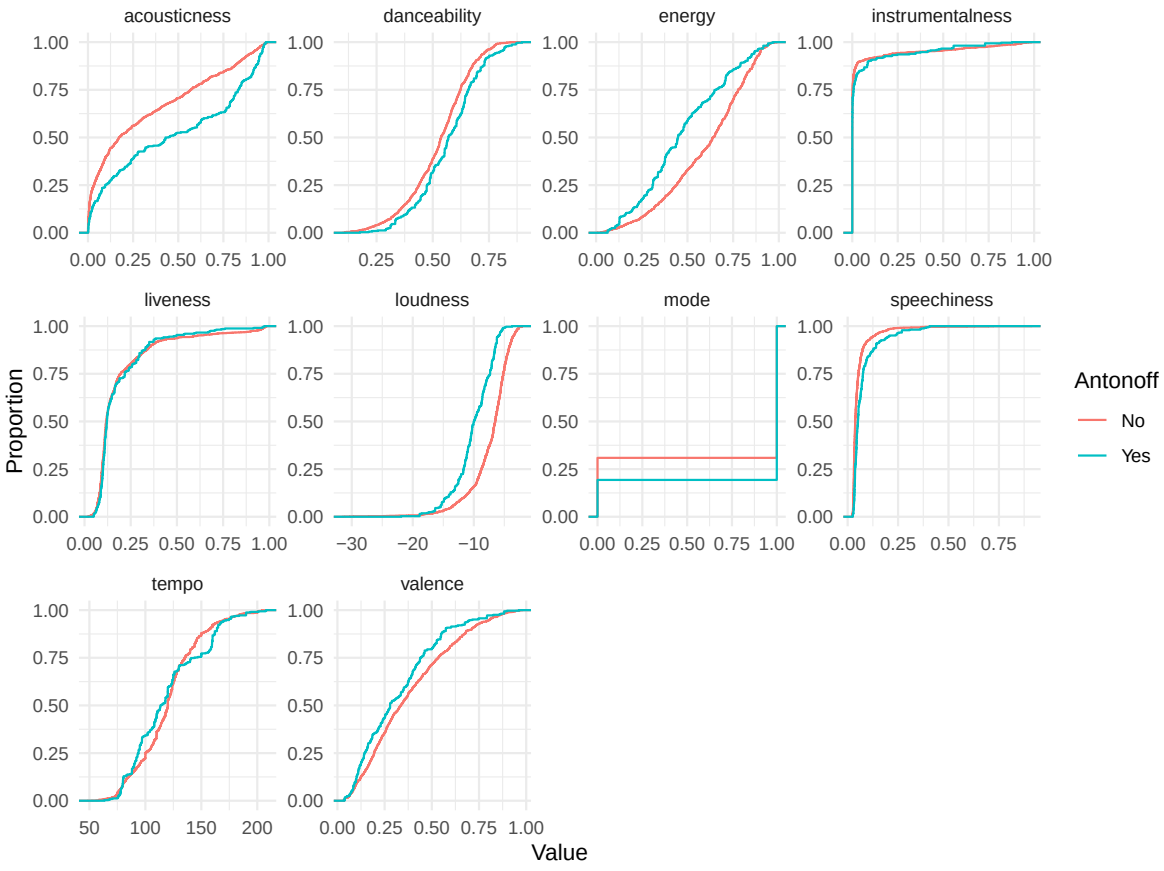

Figure 2: Cumulative Distribution Function of Each Variable From the Spotify Audio Features API for our Full Data Set

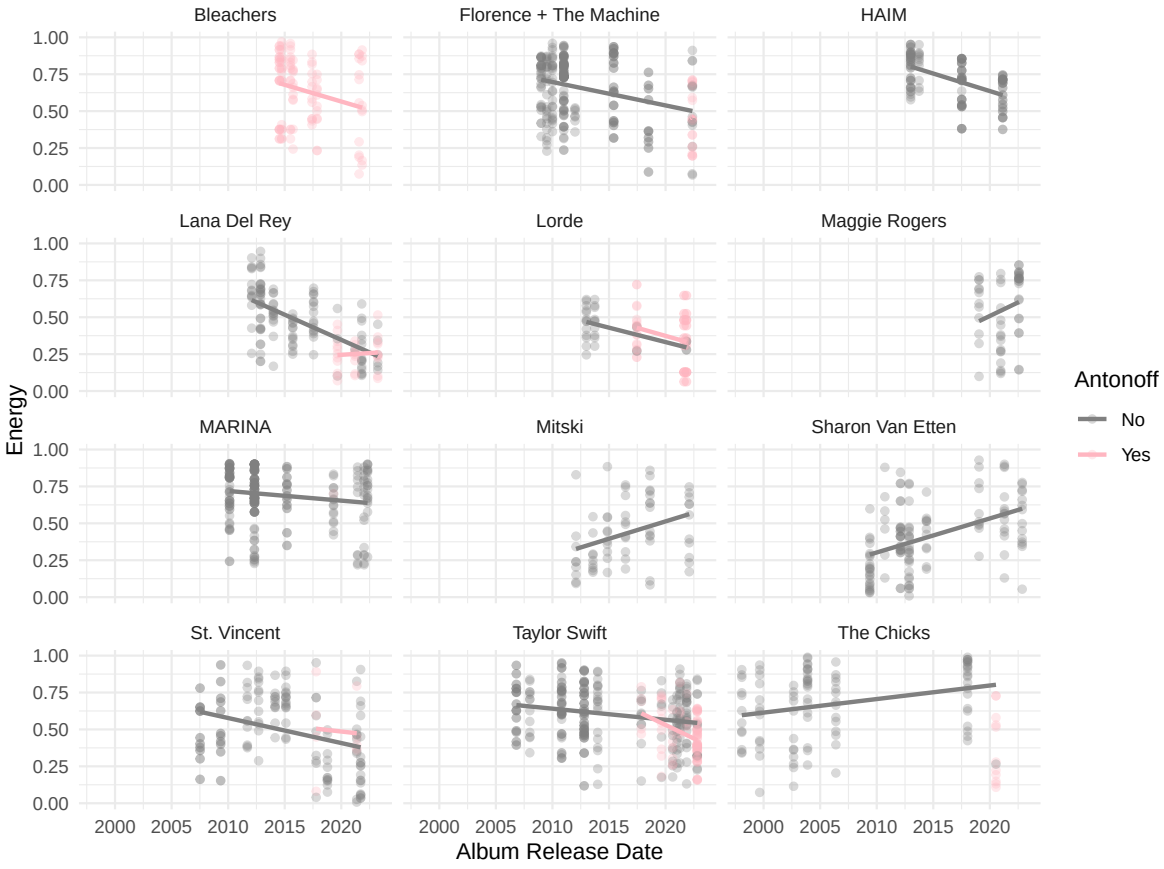

Figure 3: Energy Scores For Contemporary Pop Artists Over Time

and Taylor Swift. Notably, Sharon Van Etten, Maggie Rogers, and Mitski, display upward energy trajectories in their discographies.

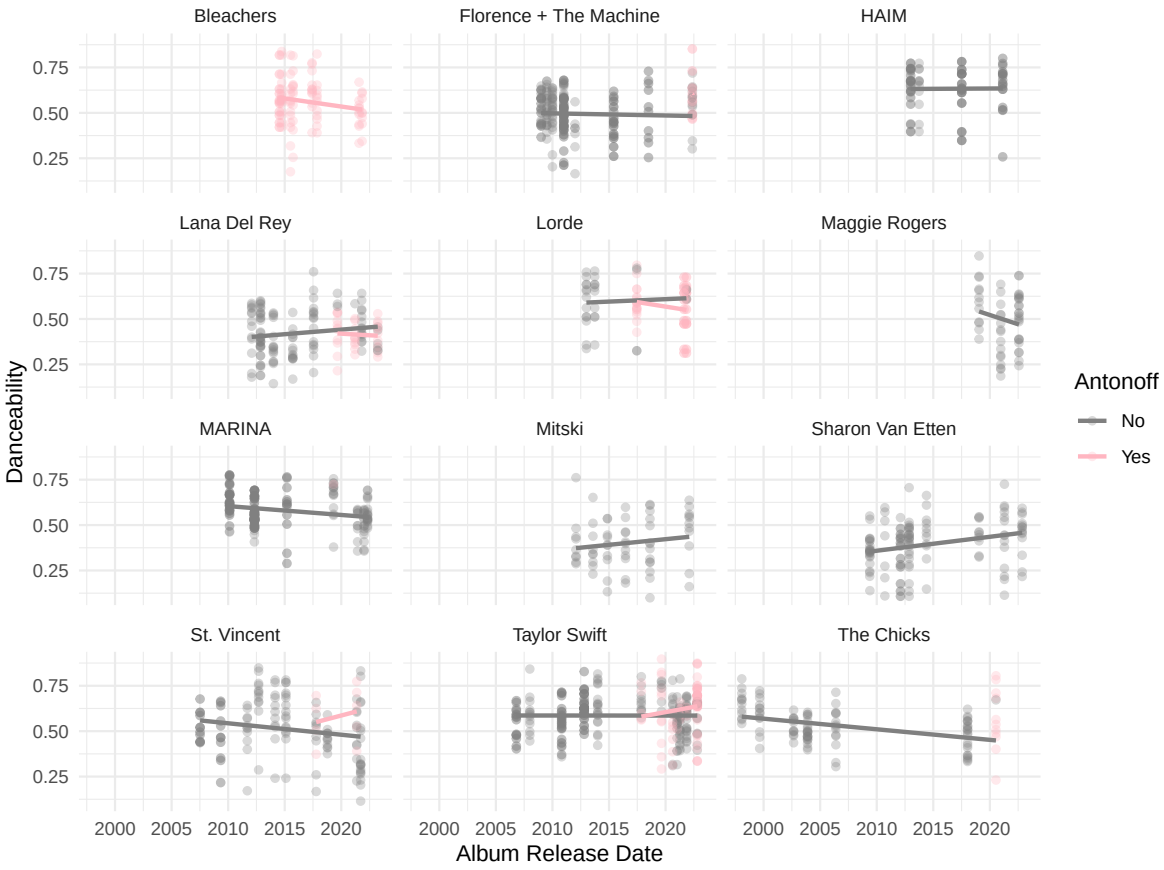

Figure 4: Danceability Scores For Contemporary Pop Artists Over Time

No significant trends are observable in danceability over time across the artists' collaborations with Antonoff, as demonstrated in Figure 4, although a difference in danceability scores across songs in every album, across all artists can be observed. This suggests that danceability scores across tracks in every album are consistently varied over time. Notably, The Chicks' earlier work, between 2002 and 2005, have the smallest amount of variance in danceability. This could be due to genre difference, since their earlier recordings, like the album "Home", were country.

Figure 5, which focuses on valence, reveals a notable shift in valence scores in Lorde's discography, where her tracks display a slight upward trend and a greater range in valence scores over time. Otherwise, similarly to danceability, the valence scores display significant variance across tracks in every album.

Figure 6 underscores a trend toward increased acousticness in the works of Taylor Swift, Lorde,

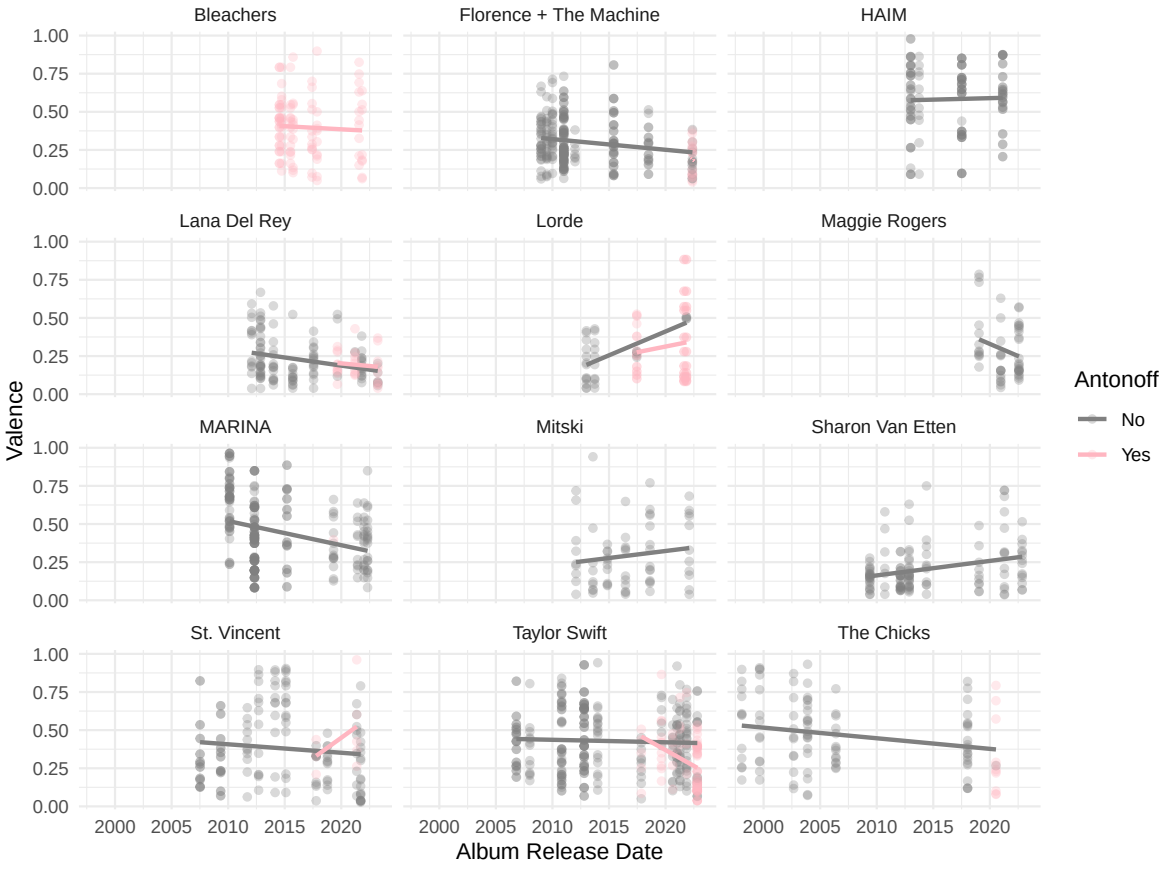

Figure 5: Valence Scores For Contemporary Pop Artists Over Time

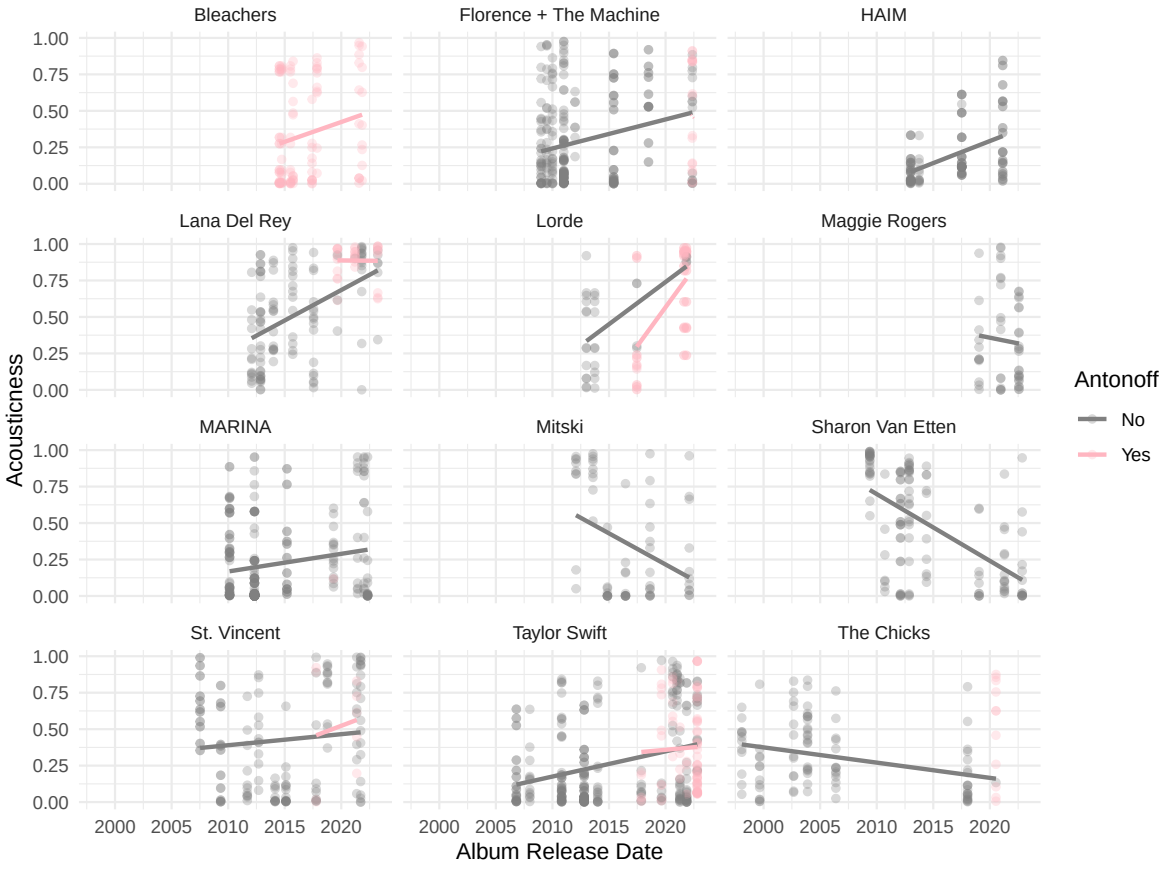

Figure 6: Acousticness Scores For Contemporary Pop Artists Over Time

and HAIM, contrasted with declining trends in artists like Mitski, Sharon Van Etten, and Maggie Rogers. Tracks produced by Antonoff also display an upward trend (such as Bleachers or Lorde), or appear in artists' discographies when an upward trend is already underway (such as Lana Del Rey, Taylor Swift, and Florence and the Machine).

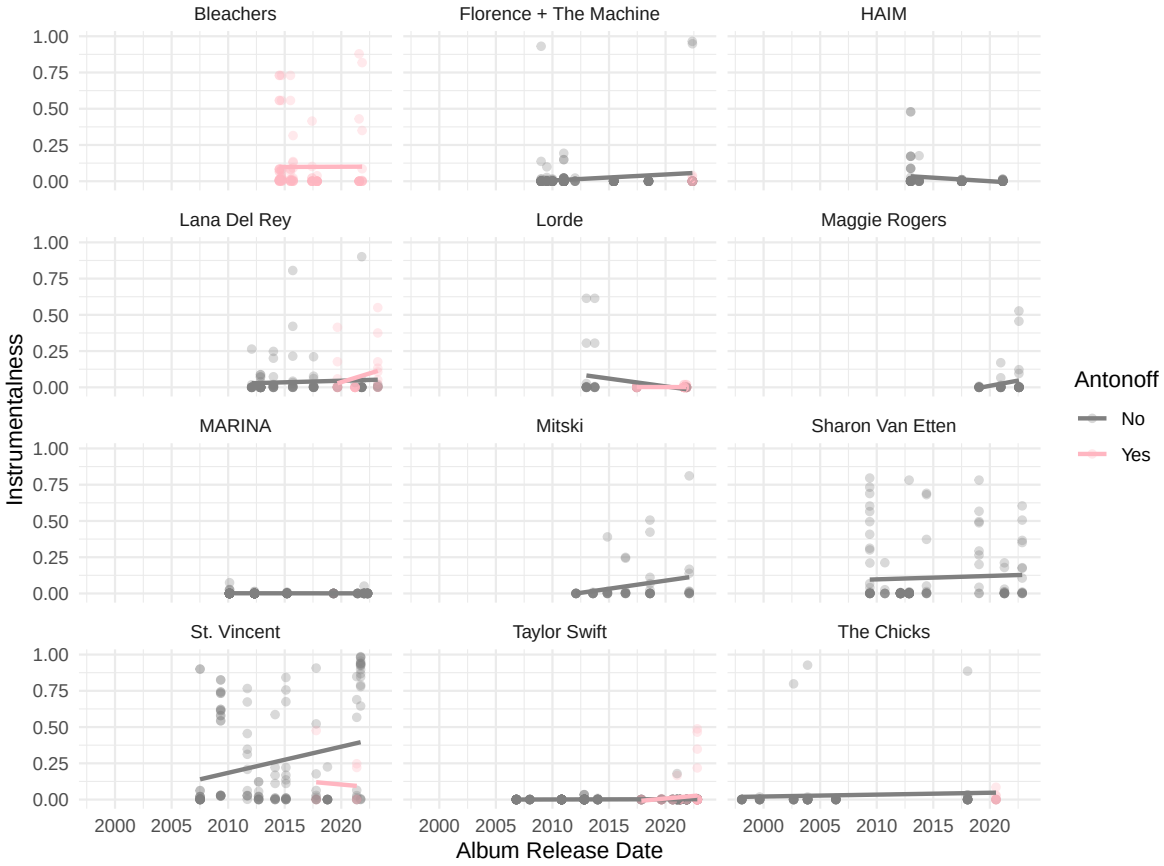

Figure 7: Instrumentalness Scores For Contemporary Pop Artists Over Time

Figure 7 exhibits no significant correlations, and scores are consistently low for most artists, excluding St. Vincent, whose tracks display both a consistently high degree of variance and a noticeable upward trend over time.

Finally, tempo trends, as shown in Figure 8, are generally stable across artists, with notable exceptions in collaborations involving Lana Del Rey and Taylor Swift. In Lana Del Rey's tracks, her collaborations with Antonoff see an upward trend in tempo. In Taylor Swift's tracks, her collaborations with Antonoff exhibit a downward trend.

We have observed subtle trends in variables associated with songs produced by Jack Antonoff, as evidenced by our cumulative distribution functions and minor trends in the evolution of artists' sound characteristics, when collaborating with Antonoff (particularly energy).

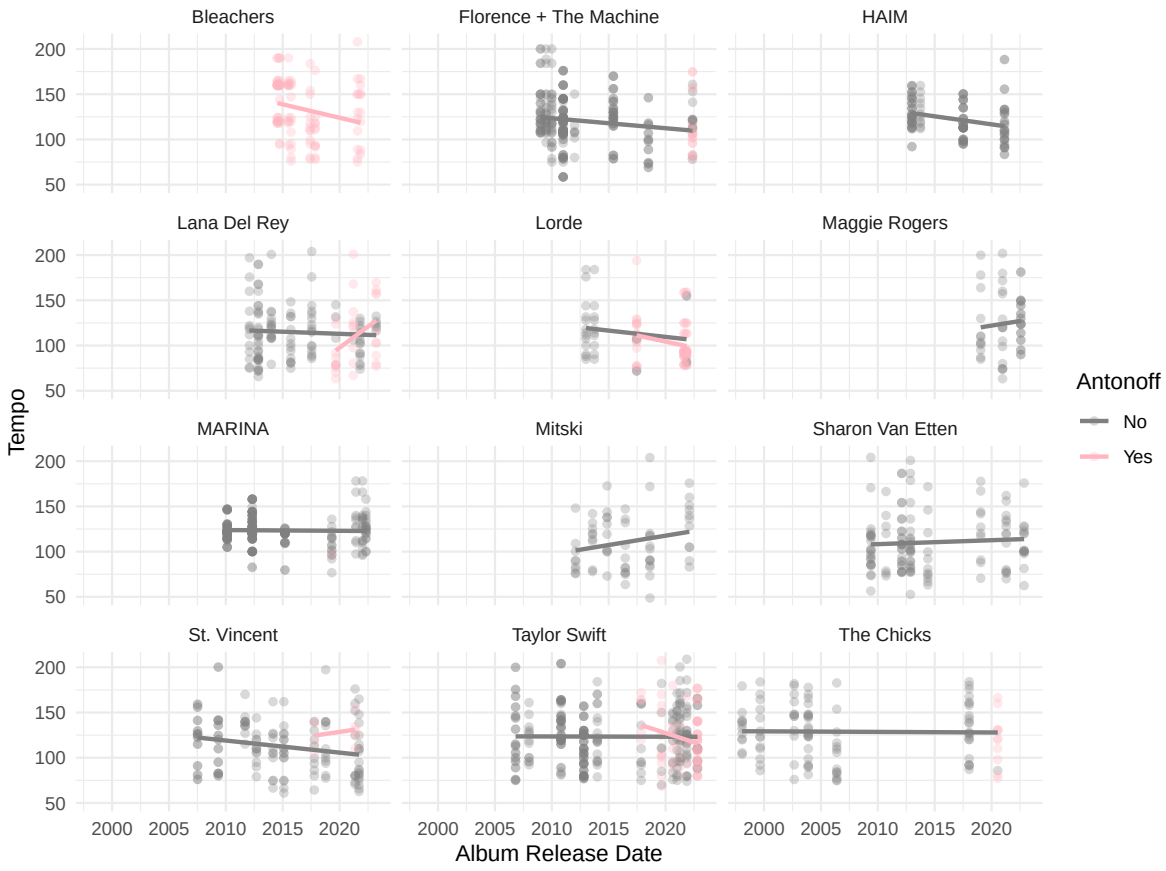

Figure 8: Tempo Scores For Contemporary Pop Artists Over Time

To be clear, we do not suggest that Antonoff's involvement is the cause for the change in artist's sound. Arguments of "Antonoffication" risk removing agency and artistic credit from (female) artists, and attributing both their critical success and failures to Antonoff (Wilson). In an industry primarily led by male producers, and using features so limited in their ability to describe anything about the process of creation, it would not be appropriate the use these trends to suggest that Antonoff's involvement is the cause behind changes in sound. The purpose of these visualizations is to observe the distribution of variables across artists, over time, with a focus on Antonoff.

This exploratory data analysis sets the stage for a) employing logistic regression to quantify the likelihood of Antonoff's involvement in a song based on these audio features, b) classifying tracks produced by a wider range of producers using a Random Forest, c) identifying the most influential variables in this classification to better understand how style can be inferred from the provided data set features.

3 Model

3.1 Logistic Regression

The aim of the logistic regression is to determine whether we can predict Antonoff's influence on artists' sound. We select a total of 1465 tracks from the dataset from 12 artists: Lorde, Taylor Swift, St. Vincent, Lana Del Rey, The Chicks, Florence and the Machine, HAIM, Marina and the Diamonds, Maggie Rogers, Sharon Van Etten, Mitski, and Bleachers. Marina and the Diamonds, Maggie Rogers, Sharon Van Etten and Mitski have not collaborated with Antonoff, but their tracks were added to contribute more data from the same taste pool to provide more context for the model.

The binary response variable, called `is_antonoff`, has a value of 1 if the song was produced by Antonoff, and a 0 otherwise. There are 10 predictor variables, all of which are continuous: **`danceability`**, **`energy`**, **`loudness`**, **`mode`**, **`speechiness`**, **`acousticness`**, **`instrumentalness`**, **`liveness`**, **`valence`**, and **`tempo`**.

The formula for this logistic regression can be written as:

$$\begin{aligned} Pr(y_i = 1) = &\text{logit}^{-1}(\beta_0 + \beta_1 \times \text{danceability} + \beta_2 \times \text{energy} \\ &+ \beta_3 \times \text{loudness} + \beta_4 \times \text{mode} + \beta_5 \times \text{speechiness} \\ &+ \beta_6 \times \text{acousticness} + \beta_7 \times \text{instrumentalness} \\ &+ \beta_8 \times \text{liveness} + \beta_9 \times \text{valence} + \beta_{10} \times \text{tempo}) \end{aligned}$$

where $\Pr(y_i = 1)$ is the probability of a song being produced by Antonoff, β_0 is the intercept, and $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}$ are the coefficients corresponding to the 10 predictor variables.

After some minor pre-processing using the `tidyverse` package (Wickham et al.) in R (R Core Team), we followed Julia Silge's downsampling methodology to build a model that accounts for imbalanced classes (Silge). We downsampled the majority class using the **`step_downsample()`** function from the **`themis`** package. Then, we used 5-fold cross-validation with stratification by the **`is_antonoff`** variable to estimate the performance of the model. The model was fit using `tidymodels` (Kuhn and Wickham).

3.2 Random Forest

Random forests (RF) build an ensemble of decision trees during the training phase, and then use majority voting for classification tasks. For each input of data, each decision tree in the forest makes a prediction, and the final classification is determined by the majority vote of these predictions. For example, if more trees in the ensemble predict a certain class for the input data than any other class, the RF model assigns that class to the input. Each tree in the ensemble is trained on a random sample of data taken with replacement, and only a random subset of features are considered when splitting at each node. This method ensures that each tree in the ensemble is unique, avoiding overfitting and capturing non-linear relationships and more complex patterns in the data.

The training process for the RF model can be summarized as follows (Hastie et al.) :

1. For each iteration ($b = 1$) to (B):
    a. Draw a bootstrap sample (Z^*) of size (N) from the training data.
    b. Grow a random-forest tree (T_b) to the bootstrapped data by recursively performing the following steps for each terminal node of the tree until the minimum node size (n_{min}) is reached:
        i. Select (m) predictors at random from the (p) predictors.
        ii. Determine the best predictor and split-point among the (m).
        iii. Split the node into two daughter nodes.
2. The ensemble of trees ($\{T_b\}_{b=1}^{B}$) constitutes the random forest model.

For a new instance (x), the RF model prediction ($\hat{C}^{RF}(x)$) is determined by the majority vote from the ensemble of trees:

$$\hat{C}^{\mathrm{RF}}(x) = \text{majority vote} \left\{ \hat{C}_b(x) \right\}_{b=1}^{B}$$

Where ($\hat{C}_b(x)$) is the class prediction of the (b)-th tree.

The categorical feature `producer`, which contains the names of the producer for each track, was one-hot encoded to create binary columns for each producer (i.e. `is_antonoff`, `is_elworth`,

etc.). The producer feature was then converted into a factor variable, and the dataset was split into training and test sets.

The RF was trained using the `ranger` package in R (R Core Team, Wright and Ziegler). We set the model to interpret variable importance using Gini impurity, which is a measure of misclassification. At every split in a decision tree, the algorithm tests how well each feature splits the data. The features that split the data the most accurately, and thus with the least impurity, would have the lowest impurity score. A variable's importance can be calculated by looking at how much the tree nodes that use that variable reduce impurity on average.

Gini impurity was chosen on the grounds that we are interested in not only classifying producers, but, more importantly, in understanding how the underlying data informs these decisions. By calculating the Gini impurity not only for the entire model, but specifically for each producer, we can better understand how the variables in our dataset describe the differences between producers. Without song structure, instruments, sound itself, or other more ephemeral components of style, these differences can be taken as a proxy for style.

4 Results

4.1 Logistic Regression

The model's accuracy is 0.757, and the ROC AUC score is 0.819. Based on the confusion matrix presented in Table 2, the precision score is 0.385, meaning that approximately 39% of songs predicted to be produced by Antonoff were actually produced by him. The recall of 0.756 indicates that about 76% of songs produced by Antonoff were actually correctly identified in the model.

Table 2: Confusion Matrix for Logistic Regression Results

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	62	99
Predicted Negative (0)	20	308

Table 3 and Figure 9 show variables' respective importance in determining whether a track was produced by Antonoff.

Variable importance, which refers to the impact of each predictor variable in explaining the variation in the response variable, can offer a data-driven picture of what might set a producer's sound apart.

`Danceability`, with a negative importance score of 4.142, had the highest level of importance in determining whether a song was produced or co-produced by Jack Antonoff. `Valence` was

Table 3: Importance Scores of Each Variable in Determining Antonoff's Tracks

Variable	Importance	Sign
danceability	4.628	NEG
energy	2.423	NEG
valence	1.542	POS
speechiness	1.412	NEG
instrumentalness	1.336	POS
liveness	1.117	NEG
mode	0.505	NEG
loudness	0.487	POS
acousticness	0.042	POS
tempo	0.005	NEG

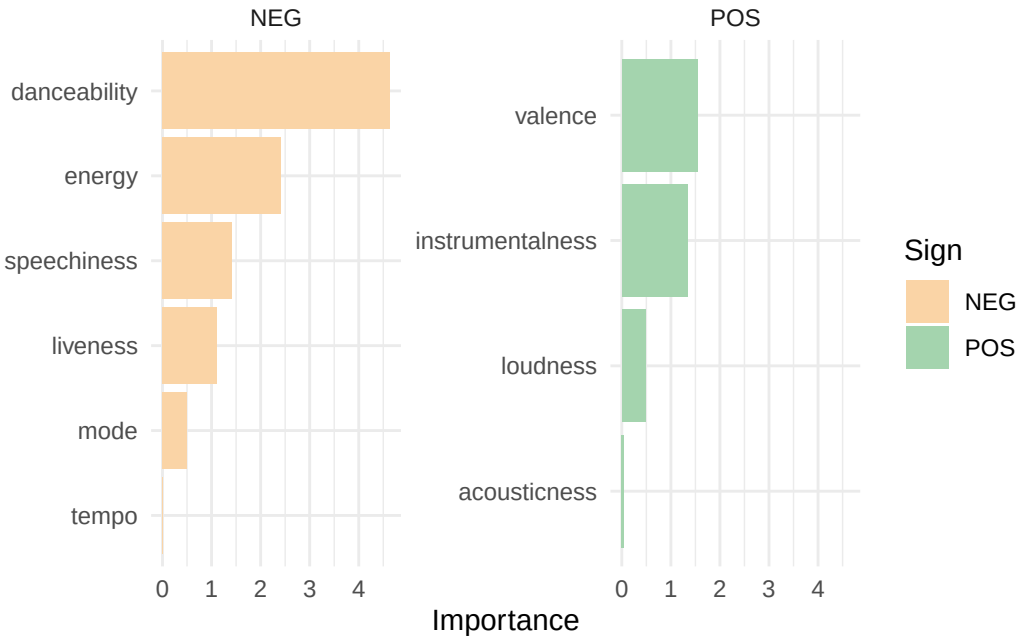

Figure 9: Importance Scores of Each Variable in Determining Antonoff's tracks

Table 4: Top 20 Most Danceable Antonoff-Produced Songs

Artist Name	Danceability	Track
Taylor Swift	0.897	I Think He Knows
Taylor Swift	0.875	Vigilante Shit
Florence + The Machine	0.852	Heaven Is Here
Bleachers	0.838	I Wanna Get Better - Demo Version
Taylor Swift	0.824	Cornelia Street
Bleachers	0.824	Hate That You Know Me - MTV Unplugged
Bleachers	0.818	Wake Me
Bleachers	0.814	Shadow
Taylor Swift	0.811	Paper Rings
The Chicks	0.805	Julianna Calm Down
Lorde	0.796	Sober
The Chicks	0.784	Texas Man
Bleachers	0.776	Hate That You Know Me
St. Vincent	0.774	Daddy's Home
Bleachers	0.769	All My Heroes
Taylor Swift	0.766	Look What You Made Me Do
Taylor Swift	0.751	Question...?
Taylor Swift	0.743	Lavender Haze
Taylor Swift	0.739	False God
Bleachers	0.732	Goodbye

the second most important, with a positive score of 2.559. The third, `instrumentalness` correlates positively with Antonoff tracks with a score of 2.309.

In the 20 songs in Jack Antonoff's discography with the highest `danceability` scores, Bleachers tracks account for 35% of this list, and Taylor Swift tracks account for 40% of this list (Table 4). St Vincent, The Chicks, Lorde, and Florence and the Machine are also featured.

4.2 Random Forest

Table 5: Overall Performance Statistics

Metric	Value
Accuracy	0.6111
95% Confidence Interval	(0.579, 0.6425)
No Information Rate	0.4274
P-Value (Acc > NIR)	<2.2e-16

Metric	Value
Kappa	0.4223

As presented in Table 5, the accuracy of the model (0.6111) being significantly higher than the No Information Rate (0.4274) means that it is doing more than just blind guessing based on class prevalence. A p-value below 0.05 suggests the model may be better than a naive approach.

The model's accuracy and Kappa statistics suggest a moderate level of performance. However, there is enough demonstrated statistical significance to suggest that there are underlying patterns in the data that differentiate producers.

The variables with the most variance in importance between each producer were danceability, loudness, and energy (Table 6).

The variance in variable importance scores are visualized in Figure 10. Similar to logistic regression, danceability stands out as a variable with a high degree of variance from producer to producer, meaning that it contributes to successful classification.

Relative to other producers, Antonoff's danceability has the least positive correlation with valence (0.311), the most negative correlation with acousticness (-0.073), and significant positive correlations with speechiness (0.336), loudness (0.366), and energy (0.274). In Figure 11, we can observe how the patterns of variables' correlations with danceability differ between producers.

Table 6: Variables Ranked by the Amount of Range

Variable	Min Value	Max Value	Range
energy	-0.779	0.821	1.600
loudness	-0.728	0.562	1.290
danceability	-0.484	0.553	1.037
speechiness	-0.445	0.300	0.745
acousticness	-0.449	0.271	0.720
mode	-0.295	0.289	0.584
instrumentalness	-0.235	0.330	0.565
valence	-0.212	0.202	0.414
liveness	-0.245	0.085	0.330

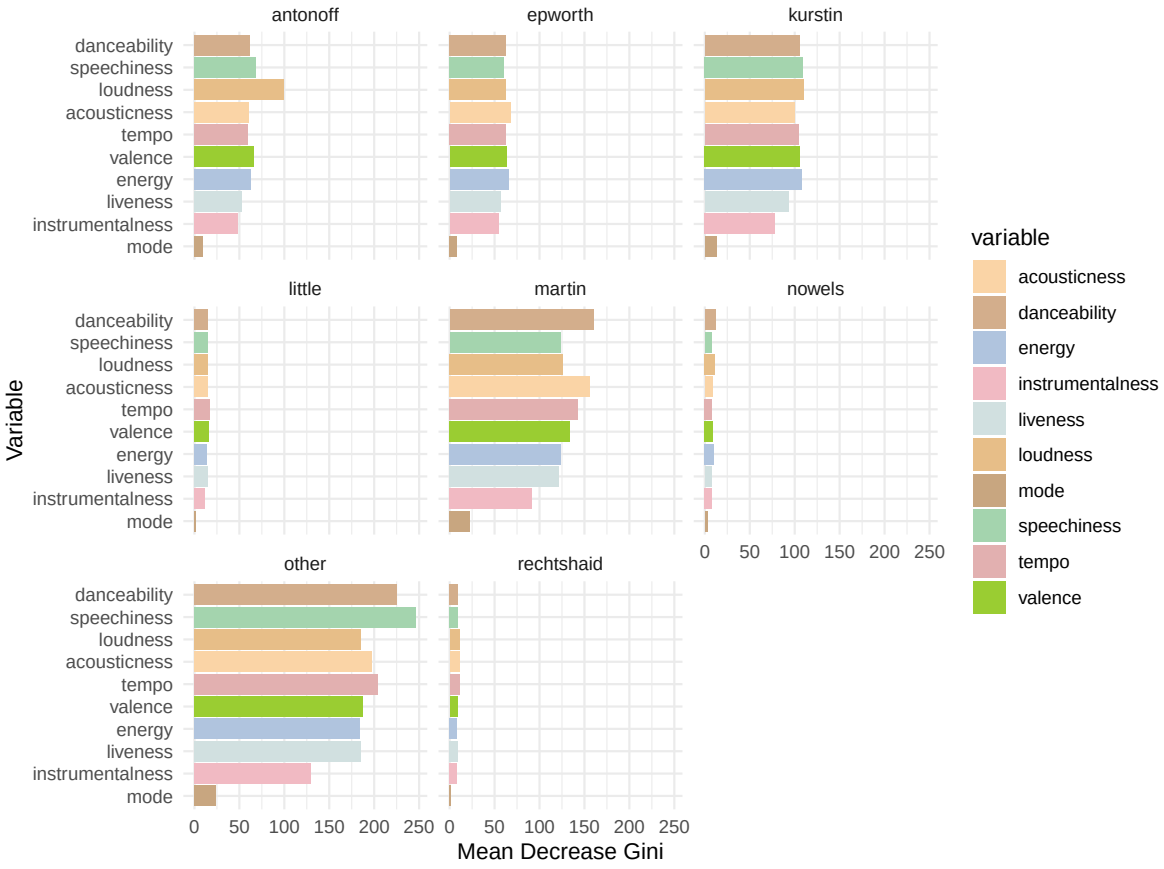

Figure 10: Variable Importance by Producer

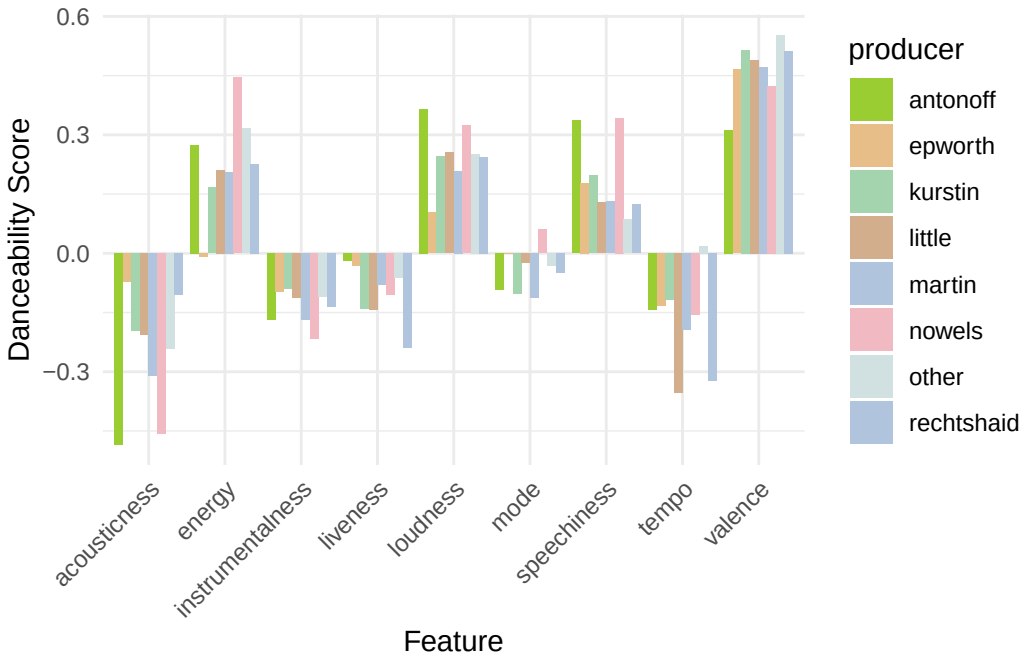

Figure 11: Danceability Scores Correlated with Other Features Across Producers

5 Discussion

5.1 Defining Style

Style is bizarre and complex. In order to define it, Roger Dannenberg differentiates between texture and style: ""Texture" usually refers to sound and the activity of making sound, while "style" is most often used to describe the general impression or intention provided by a texture." (Dannenberg). According to Dannenberg, texture is a composite of aspects of music like that we can hear on longer or shorter time scales, such as harmonic rhythms, or differences in pitches, or loudness, among many others. These aspects are more objectively measurable, while style is the "sound colour" embedded in the combination and patterns of these elements.

We can build on these lower level perceptual features to arrive at approximations of style. For instance, David Cope's Experiments in Musical Intelligence measures composers' style by detecting recurring patterns (Cope). First, recurring patterns are detected within a piece, which indicate the significant stylistic elements specific to that piece. Then, many compositions are analyzed by the same producer in order to find patterns specific to that producer. When the patterns specific to individual pieces are ruled out, a more latent "composer pattern" emerges. This composer pattern is a very basic representation of a composer's style, and can be further developed.

Our study follows a similar line of inquiry – the Spotify API mostly provides metrics related to lower level perceptual features (with danceability and valence being the obvious anomalies). However, a major limitation to our study is that it doesn't account for patterns within a track or the overall architecture of songs (or, for that matter, albums). Our dataset provides metrics that stand in for an entire piece, making a major part of the song inaccessible. On a similar note, we also attribute one producer to each track. In pop production, co-productions are frequent, and the general spirit of the genre is more collaborative, making the idea of detecting a latent "composer pattern" harder to differentiate from the collaborations they're embedded in.

Working with these limitations, danceability, loudness, and energy scores still emerge as key predictors in successful instances of classification. Whether or not we can attribute these stylistic differences solely to the named producer, it is a classification nonetheless.

Using Dannenberg's definitions of texture and style as a framework, we can understand each feature of the dataset as encoding different elements of texture. In applying statistical models to these features, the underlying patterns in the relationship between these textures can be taken as an approximation for style.

5.2 Our Results

For instance, our logistic regression model was not able to strongly predict whether a song was produced by Jack Antonoff using the Spotify API's audio features data. However, when classifying the tracks, `danceability` was the strongest predictor.

Our Random Forest results, which were moderately strong at classifying tracks, indicate that Antonoff's music, characterized by its `danceability` score, exhibits a significant positive correlation with factors such as `speechiness`, `loudness`, and `energy`, while showing a negative correlation with `acousticness`, and a less pronounced positive correlation with `valence` compared to other music producers (Figure 11). This profile—marked by speechiness, loudness, energy, and a blend of positivity with a touch of nostalgia—aligns with both our interpretations and critical assessments of Antonoff's sound.

Antonoff's tracks have been described as anthemic (Rosen), with a distinct vocal treatment that makes it sound, as Caleb Gamman, a Youtuber who went viral for his critique of Jack Antonoff's sound, told us, "the way it would sound to the person who's singing them" (Gamman).

> In the treble mix of the song, like in the very upper end, sort of above where the vocals are – he often really crushes that down. It's lower in volume than anything else, which is sort of a weird effect. When you hear someone speaking, you hear a lot of that, like, noise in their voice. When you yourself are speaking, you hear less of it, right? You hear more of your own voice bouncing around in your head. And so it sort of creates this effect of – **like he mixes his vocals the way it would**

> **sound to the person who's singing, them. Which is sort of strange.** He often has no vocals on that upper end, which is very unusual, and then often there are random little bits of noise happening up there which is the sort of thing he likes to do.

Perhaps, without further documentation on the `speechiness` feature's construction, clearing the noise and isolating the middle range of the vocals may make for stronger overall `speechiness` score in Antonoff's tracks. Antonoff also tends to collaborate with singer-songwriter artists, whose music is more lyric-oriented than dance hit pop.

On that note, even Antonoff-produced songs with the highest `danceability` scores aren't exactly dance hits, compared to production styles like that of Max Martin.

For instance, the highest scoring song in danceability, *I Think He Knows* by Taylor Swift, can pass as danceable, yet the next highest-scoring songs – Taylor Swift's *Vigilante Shit* or Florence and the Machine's *Heaven is Here* – don't resonate as dance tracks. But they aren't understated, ambient tracks either – they exhibit some sort of visceral quality. Like Gamman's description of Antonoff's vocal design sounding like how they would sound to the person who's singing them, these tracks seem like they're trying to get inside of your skin. But they don't have the rhythm of a dance track.

I Wanna Get Better by Bleachers typifies the Antonoff style as identified by our model. This song is more anthemic, resonating with a sort of neo-Springsteen aesthetic that Antonoff is often characterized as (or admitting to) being influenced by, than danceable (Pareles, Horn). It also does sound like it's trying to demand your attention, and the speechiness, loudness, and energy which are distinctly correlated with danceability in Antonoff's RF results, are all apparent. The redemptive, broken-but-getting-better tone can also be heard in the lyrics, but it still doesn't have the kind of rhythm for dancing – which maps onto how Antonoff's "valence" score isn't as a high as other pop artists in Figure 11.

Cornelia Street by Taylor Swift further exemplifies this near-danceable category. While melodically catchy, the lyrics themselves are wistful and longing. The vocal treatment, as described by Gamman, is there – Swift's vocals sound magnified, carrying the bridge of the song.

5.3 "Danceability" and "Valence" Features' Crowd-sourced Construction

We spoke with Glen McDonald, a Principal Engineer at Spotify who worked at The Echo Nest, a music intelligence start-up that was acquired by Spotify in 2014. While McDonald did not build the features, he has worked with them, and provided information on their provenance. McDonald revealed that `valence` and `danceability` were created by giving tracks to college interns and asking them to tag whether a song was positive or gloomy, or danceable or undanceable. Variables like `energy` or `instrumentalness`, on the other hand, were determined

primarily through machine listening techniques, with human subjectivity being applied to fine-tune the features[2] (McDonald).:

> "You could imagine writing a formula for energy that combines loudness and tempo and degree of harmonic variation or something. So that feature was [machine learning], but that one's more like a human helping a machine figure out a formula. Whereas valence is teaching the machine to try to reproduce a purely human thing. The same with danceability. I mean, danceability is whether a human can dance to it. The machine's not gonna dance, so the machine can't have any opinion on that. The computer could have an opinion on energy. And the computer can definitely have an opinion on loudness. So there's a spectrum from, loudness as purely analytical, and then energy is a little like loudness, with a little more subjectivity. And then danceability and valence are purely subjective."

McDonald confirmed that this process didn't account for lyrics, which made valence a particularly difficult variable to build. It could, in theory, pick up on aspects of vocal performance, but it didn't process anything about languages or words or the meanings of songs. Take an upbeat, happy-sounding Elliot Smith song with devastating lyrics: the machine might register it as happy, where human listeners understand that it's sad. Furthermore, two humans might even disagree on the song's valence: "Plus, we have the confounding factor of like, a song that seems happy today could be sad tomorrow because the singer was killed in a plane crash. The song didn't change, but our world changed and our reactions changed" (McDonald).

McDonald suggests combining energy and valence to create quadrant, which usually works "fairly well" to describe music (see Figure 12) (McDonald): "High energy and high valence could be generally happy, cheerful, upbeat. Low energy, low valence could be sad or downbeat. High energy, low valence is sort of angry. Low energy, high valence is serene or calming."

5.4 Conclusion

What does the danceability variable then signify in Antonoff's music? Our findings suggest that, similar to McDonald's quadrants, Antonoff's most "danceable" songs are better understood through the lens of correlated features, even though danceability emerges as the primary predictor. This crowd-sourced metric, in conjunction with machine-listening-derived features that capture songs' "textures", offers a unique insight into an artist's style, capturing an intuitive understanding of a song's tone and energy.

This approach not only mirrors the multifaceted nature of auditory perception, but also somehow enriches our model's capacity to capture the essence of an artist's style, or at least distinguish between producers in a limited data set. While crowdsourced metrics might initially

[2] A notable example of this being that bluegrass songs were given very high `speechiness` ratings. Because there were no banjos in the training data, the instruments were registered as human speech. The engineers had to go back and add more songs with banjos to the training data.

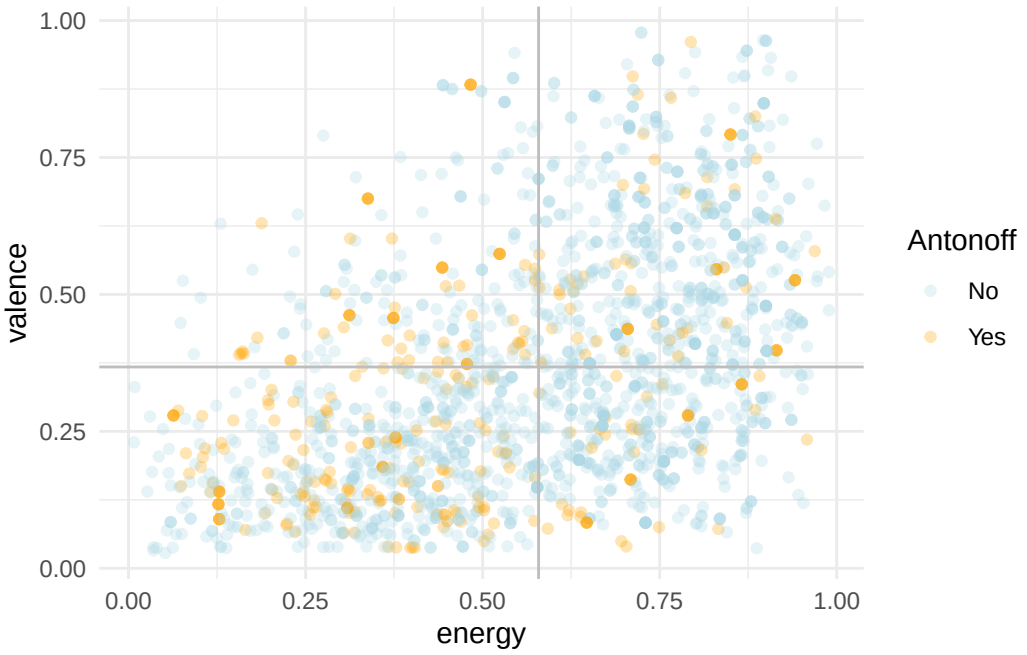

Figure 12: Valence and Energy Quadrant

seem rudimentary, this subjective aspect can play a crucial role in this process. The work of Martikainen et al. illustrates this point effectively, combining open-source audio tools with qualitative, crowd-sourced listener data to understand audio-based stylistic variation in podcasts (Martikainen et al.).

The role of human subjectivity in interpreting and experiencing music alongside computational representations of it becomes even more apparent when considering the broader scope of music recommendation systems. In Nick Seaver's anthropological study of music recommendation companies, Seaver writes that engineers "develop ways of thinking about musical preference and software, attempting to reconcile them with each other" (Seaver), referring to the fact that engineers' understandings of style and taste influence the designs of the music recommendation systems they build. Our study therefore contributes to a more nuanced understanding of recommendation systems as a whole, highlighting the balance between computational measures and human interpretation in the evolving landscape of music analysis and appreciation.

Future studies should reproduce crowdsourced danceability or valence variables to better understand the underlying user preferences that motivate these metrics. In doing so, we can build more robust ways to computationally model more ephemeral and subjective aspects of music such as style.

References

Cope, David. *Computers and Musical Style*. A-R Editions, 1991.

Dannenberg, Roger. "Style in Music." *The Structure of Style, ISBN 978-3-642-12336-8. Springer-Verlag Berlin Heidelberg, 2010, p. 45*, May 2010, https://doi.org/10.1007/978-3-642-12337-5_3.

Gamman, Caleb. *Interview Conducted by Michaela Drouillard on 23 March 2023*. 2023.

Hastie, Trevor, et al. *The Elements of Statistical Learning*. Springer, 2009, https://doi.org/10.1007/978-0-387-84858-7_15.

Horn, Olivia. "Jack Antonoff Doesn't Want to Just Take up Space." *The New York Times*, 2021, https://www.nytimes.com/2021/07/22/arts/music/jack-antonoff-favorites.html.

Kuhn, Max, and Hadley Wickham. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles*. 2020, https://www.tidymodels.org.

Marantz, Andrew. "Jack Antonoff's Pop-Music Collaboration Machine." *The New Yorker*, May 2022, https://www.newyorker.com/magazine/2022/05/23/jack-antonoff-pop-music-collaboration-lorde-taylor-swift.

Martikainen, Katariina, et al. *Exploring Audio-Based Stylistic Variation in Podcasts*. 2022, pp. 2343–47, https://doi.org/10.21437/Interspeech.2022-10871.

McDonald, Glen. *Interview Conducted by Michaela Drouillard on 11 April 2023*. 2023.

Pareles, John. "Wallops of Exuberance with Traces of Yearning." *The New York Times*, Sept. 2014, https://www.nytimes.com/2014/09/06/arts/music/jack-antonoff-tweaks-his-jersey-roots-in-bleachers.html.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2022, https://www.R-project.org/.

Rosen, Jody. "Jack Antonoff Is Only Making Music with Friends." *The New York Times Magazine*, 2020, https://www.nytimes.com/interactive/2020/03/11/magazine/jack-antonoff-profile.html.

Seaver, Nick. *Computing Taste: Algorithms and the Makers of Music Recommendation*. University of Chicago Press, 2022.

Silge, Julia. *To Downsample Imbalanced Data or Not, with #TidyTuesday Bird Feeders*. Julia Silge, 2023, https://juliasilge.com/blog/project-feederwatch/.

Spotify. *Get Track's Audio Features*. https://developer.spotify.com/documentation/web-api/reference/get-audio-features.

Thompson, Charlie, et al. *Spotifyr: R Wrapper for the 'Spotify' Web API*. 2022, https://CRAN.R-project.org/package=spotifyr.

Wickham, Hadley, et al. "Welcome to the Tidyverse." *Journal of Open Source Software*, vol. 4, no. 43, 2019, p. 1686, https://doi.org/10.21105/joss.01686.

Wikipedia. *Jack Antonoff Production Discography — Wikipedia, the Free Encyclopedia*. 2023, https://en.wikipedia.org/wiki/Jack_Antonoff_production_discography.

Wilson, Carl. "The Real Reason the Internet Hates Jack Antonoff." *Slate*, 2023, https://slate.com/culture/2023/11/taylor-swift-producer-jack-antonoff-1989-taylors-version.html.

Wright, Marvin N., and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software*, vol. 77, no. 1,

2017, pp. 1–17, https://doi.org/10.18637/jss.v077.i01.