Can We Predict a Top Hit?

Alfredo Lorenzo Mendiola and Quinn Bottomly

28 December 2023

Introduction

Music and sound have been researched time and time again. A simple search in Google Scholar will yield plenty of reading for anyone interested in reading something different for a very long time. Music itself has been around for even longer. It is even said to be "as old as humanity itself. Archaeologists have found primitive flutes made of bone and ivory dating back as far as 43,000 years, and it's likely that many ancient musical styles have been preserved in oral traditions." (Andrews, 2023)

There are clear aspects to tracks that can make them hits as opposed to being forgotten all together. One large aspect is the artist that puts out the song. There are certain artists with loyal fan bases that will play every track that artist creates and then are artists with loyal haters who will avoid certain tracks at all costs. For example, "Swifties" will repeatedly support and play Taylor Swift's tracks but some Eminem fans will completely avoid MGK tracks due to their previous 'beefing' with each other. Musicians also know that there are a few beats, tempos, and other factors that can help a track become popular to the masses but is not always a guarantee.

Nowadays computational power and knowledge have greatly advanced from when the music industry began and continue to rapidly advance. If we could take the domain knowledge of artists and the computational power of machine learning (ML) and accurately predict what would make a track popular, we could essentially change the music industry forever. Whether this change would be a good or detrimental change is up to every individual person to decide. The music industry could become a 'cookie cutter' industry and ruin the creativeness aspect for musicians very quickly. This could also help other artists, who could navigate that space to give listeners something new to listen to, find a way that would change the 'cookie cutter' outline as new music is produced. This also provides the possibility that artists will make minor adjustments to their styles and sounds to match what is the current 'cookie cutter' outline but with their own flavor added to it. Essentially, the music industry would have to do what every human has had to do throughout history to survive. Adapt and overcome.

Research Question and Design

H0 It is possible to predict which songs will be Top Hits using the attributes that Spotify assigns to the tracks, at or above a 90% correct threshold. **H1** It is not possible to predict which songs will be Top Hits using the attributes that Spotify assigns to the tracks, at or above a 90% correct threshold.

The data for this will need to be from a consistent source YoY. A large amount of tracks will need to be added to the dataset to be the non-hits and have the same attributes recorded as the hits. A large ratio of non-hits to hits will need to be kept in the data to avoid too many hits to non-hits causing it to skew the data. This also helps to simulate the music industry since there are millions of songs but very few of those become hits. Even less become Top Hits and make it as top 50/100/150 for the year. I will be running 4 models, 2 Random forests (RF) and 2 XGBoost. The first RF will include all features and no pruning so that I can use it for feature importance for my other models. I will then run a second RF where I top down prune and remove features. Then, I will run an XGBoost and leave all features in it and see how the

models differ in F1 scores, accuracy and Type 1 and Type 2 errors. Lastly, I will use k folding to see how my XGBoost can improve and compare the models' F1 score to the others.

Data Overview

Once I knew what I wanted to answer as my research question, I began to look for data to use in my research design. I mainly focused on different Spotify playlists that could fit my needs. I found that Spotify had 2 playlist series that would fit what I was looking for, 'Top Hits of XXXX' and 'Top Tracks of XXXX'. I looked for specification on how each playlist was curated, but could not find anything that concretely said how these playlist tracks were chosen. It did seem to be a consensus that 'Top Tracks of XXXX' was only based on Spotify streams while 'Top Hits of XXXX' were for the overall US market. To include the most popular songs of a certin year, I wanted to ensure the hits were not just based on Spotify streams alone. Some sources did say that the 'Top Hits of XXXX' were chosen for their overall popularity in the market, including radio play and other mediums of use, and not just on their streams on Spotify. So this playlist series looked to fit my research design the best.

Once I found the playlists for 2011 to 2022 and recorded their unique playlist IDs, I pulled the data from the Spotify API in two steps. First, I used the playlist IDs in the spotifyr package function, get_playlist_tracks(), to get the tracks from the playlists. The data came organized and clean as expected. I only had to manipulate and prune the data to be in dataframes for my use. Second, I had to take the unique track IDs that I had gotten from the playlists and pulled individual track attributes from the Spotify API using the spotifyr package function, get_track_audio_features(). Once I had these two sets of data per year, I combined them using the unique track ID.

For my data that would be for the non-hits, I found a dataset in Kaggle Datasets named 'Spotify 1.2M+ Songs' by Rodolfo Figueroa. It was a large dataset that had all of the track attributes that I had been looking for so I used this for my non-hits with minimal cleaning. I only had to make the variables names and formats match across datasets.

For the cleaning of the data, I did a lot of unlisting and formating how chr variables would show up. I looked at ranges, for 'NA' and missing values, and values did not make sense and cleaned those to the best of my ability. To identify duplicate tracks, I used the unique track ID to eliminate the possibility of the track names being stored differently. I eliminated songs that were from the 'Top Hits of XXXX' playlists from the non-hits to ensure there were no hits in that dataset. I did not eliminate duplicate hits since I believe if a song appears as a hit in multiple years, it should carry a larger weight in the outcome.

For certain variables, I either created or transformed them. I created a variable that would be the indicator for whether a track was a hit or a non-hit called 'hit' with a 1 (True) or 0 (False) value. I also created a numerical variable to account for all of the unique artist combinations based on the artist_ids column for analysis. I then changed explicit from a T/F logical variable to a 1/0 numerical variable for use in the analysis.

When the data was cleaned and formatted correctly, I ended up with 25 variables (descriptions below) and 1,205,002 observations overall.

The following are the variable descriptions per Spotify API: **id** The Spotify ID for the track. **name** The name of the track. **album** The name of the album. In case of an album takedown, the value may be an empty string. **album_id** The Spotify ID for the album. **artists** The name of the artist. (was cleaned from an array by me) **artist_ids** The Spotify ID for the artist. (was cleaned from an array by me) **track_number** The number of the track. If an album has several discs, the track number is the number on the specified disc. **disc_number** The disc number (usually 1 unless the album consists of more than one disc). **explicit** Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown). **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. **key** The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C#/Db, 2 = D, and so on. If no key was detected, the value is -1. **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. **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. **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. "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. **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. **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). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. **duration_ms** The duration of the track in milliseconds. **time_signature** An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4". **year** The date the album was first released. (cleaned to be just the year for consistency reasons) **hit** indicator variable whether the song was in the "Top Hits of XXXX" playlists **year_hit** variable to see which "Top Hits of XXXX" playlists a track was in. 0 if it was not a hit.

After going through my data and verifying validity or possible issues, I removed the following for various reasons: **id** This was used to keep track of unique observations during cleaning and is no longer needed for analysis. **name** This was used to keep track of unique observations during cleaning and is no longer needed for analysis. **album** This was used to keep track of unique observations during cleaning and is no longer needed for analysis. **album_id** This was used to keep track of unique observations during cleaning and is no longer needed for analysis. **artists** I used artist_ids in place of this to account for any possible differences in spelling. **track_number** A track's number on an album may have been a big reason for popularity back in the day, but now albums do not have to be played in order so the validity is lowered if I used this metric. This variable also had the issue of collaboration albums being made by Spotify affecting its post release information so the data is not consistent. **disc_number** This variable would only be different from 1 if there were multiple disks in the album. It makes sense with how music was released pre-internet, but now there are only '1 disk' releases. This caused a majority of the data to be a value of '1'. **time_signature** Per the Spotify API, this variable should have a range of 3-7. The data had a range of 0-5. At first I thought the 0 could mean there was not recording of it and the 1-5 were scaled versions of the 3-7. I could not find a way to prove this so I removed it from my analysis. **year** A song releasing at a certain time does have an impact on its popularity but there were issues with the data collected. For a major part of the data, only a year was listed. I also noticed while cleaning up years that did not make sense (0, 1900, etc.) that if a song was reuploaded to a new album collaboration by Spotify, the year for the track would be updated to that year. So it was not an accurate year of release either. **year_hit** This was a variable where more than 90% of my data would have a 0 since the non_hits dataset were assigned 0 by me.

EDA Description Statistics

I began to look at my cleaned data with some statistics using summary() and describe(). I was looking for clear issues with the data and getting a general idea of the values and ranges.

The following are summary statistics on the 3 datasets, hits, non-hits, and the combined dataset (all_hits):

```
##   artist_ids           explicit       danceability        energy
##  Length:1265        Min.   :0.0000   Min.   :0.1880   Min.   :0.0519
##  Class :character   1st Qu.:0.0000   1st Qu.:0.5910   1st Qu.:0.5590
##  Mode  :character   Median :0.0000   Median :0.6800   Median :0.6790
##                     Mean   :0.3273   Mean   :0.6691   Mean   :0.6609
##                     3rd Qu.:1.0000   3rd Qu.:0.7620   3rd Qu.:0.7890
##                     Max.   :1.0000   Max.   :0.9650   Max.   :0.9720
##       key            loudness            mode          speechiness
##  Min.   : 0.00   Min.   :-21.107   Min.   :0.0000   Min.   :0.02320
##  1st Qu.: 2.00   1st Qu.: -7.076   1st Qu.:0.0000   1st Qu.:0.04080
##  Median : 5.00   Median : -5.668   Median :1.0000   Median :0.05830
##  Mean   : 5.22   Mean   : -5.979   Mean   :0.5913   Mean   :0.09737
##  3rd Qu.: 8.00   3rd Qu.: -4.510   3rd Qu.:1.0000   3rd Qu.:0.10800
##  Max.   :11.00   Max.   : -1.702   Max.   :1.0000   Max.   :0.53000
##   acousticness       instrumentalness      liveness         valence
##  Min.   :0.0000129   Min.   :0.000000   Min.   :0.0210   Min.   :0.0381
##  1st Qu.:0.0217000   1st Qu.:0.000000   1st Qu.:0.0926   1st Qu.:0.3240
##  Median :0.0876000   Median :0.000000   Median :0.1190   Median :0.4920
##  Mean   :0.1846842   Mean   :0.008729   Mean   :0.1696   Mean   :0.4963
##  3rd Qu.:0.2640000   3rd Qu.:0.000046   3rd Qu.:0.2050   3rd Qu.:0.6740
##  Max.   :0.9780000   Max.   :0.896000   Max.   :0.9790   Max.   :0.9720
##      tempo         duration_ms          hit
##  Min.   : 64.93   Min.   : 97393   Min.   :1
##  1st Qu.: 99.97   1st Qu.:190476   1st Qu.:1
##  Median :120.01   Median :211467   Median :1
##  Mean   :120.63   Mean   :214701   Mean   :1
##  3rd Qu.:135.19   3rd Qu.:233456   3rd Qu.:1
##  Max.   :205.86   Max.   :613027   Max.   :1
```

```
##   artist_ids           explicit        danceability       energy
##  Length:1203737     Min.   :0.00000   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:0.00000   1st Qu.:0.356   1st Qu.:0.2520
##  Mode  :character   Median :0.00000   Median :0.500   Median :0.5240
##                     Mean   :0.06859   Mean   :0.493   Mean   :0.5095
##                     3rd Qu.:0.00000   3rd Qu.:0.633   3rd Qu.:0.7660
##                     Max.   :1.00000   Max.   :1.000   Max.   :1.0000
##       key             loudness            mode          speechiness
##  Min.   : 0.000   Min.   :-60.000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 2.000   1st Qu.:-15.256   1st Qu.:0.0000   1st Qu.:0.03510
##  Median : 5.000   Median : -9.792   Median :1.0000   Median :0.04460
##  Mean   : 5.194   Mean   :-11.810   Mean   :0.6715   Mean   :0.08438
##  3rd Qu.: 8.000   3rd Qu.: -6.718   3rd Qu.:1.0000   3rd Qu.:0.07220
##  Max.   :11.000   Max.   :  7.234   Max.   :1.0000   Max.   :0.96900
##   acousticness     instrumentalness       liveness         valence
##  Min.   :0.0000   Min.   :0.0000000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0376   1st Qu.:0.0000076   1st Qu.:0.0968   1st Qu.:0.191
##  Median :0.3890   Median :0.0081000   Median :0.1250   Median :0.403
```

```
##  Mean   :0.4468   Mean   :0.2829256   Mean   :0.2016   Mean   :0.428
##  3rd Qu.:0.8610   3rd Qu.:0.7190000   3rd Qu.:0.2450   3rd Qu.:0.644
##  Max.   :0.9960   Max.   :1.0000000   Max.   :1.0000   Max.   :1.000
##      tempo         duration_ms           hit
##  Min.   :  0.00   Min.   :   1000   Min.   :0
##  1st Qu.: 94.05   1st Qu.: 174080   1st Qu.:0
##  Median :116.72   Median : 224347   Median :0
##  Mean   :117.63   Mean   : 248849   Mean   :0
##  3rd Qu.:137.05   3rd Qu.: 285853   3rd Qu.:0
##  Max.   :248.93   Max.   :6061090   Max.   :0

##    artist_ids        explicit        danceability        energy
##  Min.   :     1   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 42576   1st Qu.:0.00000   1st Qu.:0.3560   1st Qu.:0.2520
##  Median : 83159   Median :0.00000   Median :0.5010   Median :0.5240
##  Mean   : 83495   Mean   :0.06886   Mean   :0.4932   Mean   :0.5097
##  3rd Qu.:124940   3rd Qu.:0.00000   3rd Qu.:0.6330   3rd Qu.:0.7660
##  Max.   :166796   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##       key            loudness            mode         speechiness
##  Min.   : 0.000   Min.   :-60.000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 2.000   1st Qu.:-15.247   1st Qu.:0.0000   1st Qu.:0.03510
##  Median : 5.000   Median : -9.785   Median :1.0000   Median :0.04460
##  Mean   : 5.194   Mean   :-11.804   Mean   :0.6714   Mean   :0.08439
##  3rd Qu.: 8.000   3rd Qu.: -6.713   3rd Qu.:1.0000   3rd Qu.:0.07230
##  Max.   :11.000   Max.   :  7.234   Max.   :1.0000   Max.   :0.96900
##   acousticness    instrumentalness       liveness         valence
##  Min.   :0.0000   Min.   :0.0000000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0375   1st Qu.:0.0000075   1st Qu.:0.0968   1st Qu.:0.191
##  Median :0.3880   Median :0.0080000   Median :0.1250   Median :0.403
##  Mean   :0.4465   Mean   :0.2826378   Mean   :0.2016   Mean   :0.428
##  3rd Qu.:0.8610   3rd Qu.:0.7180000   3rd Qu.:0.2450   3rd Qu.:0.644
##  Max.   :0.9960   Max.   :1.0000000   Max.   :1.0000   Max.   :1.000
##      tempo         duration_ms           hit
##  Min.   :  0.00   Min.   :   1000   Min.   :0.00000
##  1st Qu.: 94.06   1st Qu.: 174111   1st Qu.:0.00000
##  Median :116.73   Median : 224320   Median :0.00000
##  Mean   :117.64   Mean   : 248813   Mean   :0.00105
##  3rd Qu.:137.05   3rd Qu.: 285773   3rd Qu.:0.00000
##  Max.   :248.93   Max.   :6061090   Max.   :1.00000
```

After looking at the summary statistics, I ran descriptive statistics of the 3 datasets, hits, non-hits, and the combined dataset (all_hits).

Descriptive Statistics Table for hits dataset:

```
## pdf
##   2
```

sd	median	trimmed	mad	min	max	r
4.694e−01	0.000e+00	2.843e−01	0.000e+00	0.0000e+00	1.000	1.000
1.323e−01	6.800e−01	6.756e−01	1.260e−01	1.8800e−01	0.965	7.770
1.656e−01	6.790e−01	6.715e−01	1.705e−01	5.1900e−02	0.972	9.201
3.649e+00	5.000e+00	5.152e+00	4.448e+00	0.0000e+00	11.000	1.100
2.118e+00	−5.668e+00	−5.785e+00	1.887e+00	−2.1107e+01	−1.702	1.940
4.918e−01	1.000e+00	6.140e−01	0.000e+00	0.0000e+00	1.000	1.000
9.076e−02	5.830e−02	7.752e−02	3.558e−02	2.3200e−02	0.530	5.068
2.252e−01	8.760e−02	1.394e−01	1.175e−01	1.2900e−05	0.978	9.779
6.262e−02	0.000e+00	7.940e−05	0.000e+00	0.0000e+00	0.896	8.960
1.262e−01	1.190e−01	1.476e−01	5.545e−02	2.1000e−02	0.979	9.580
2.250e−01	4.920e−01	4.936e−01	2.609e−01	3.8100e−02	0.972	9.339
2.704e+01	1.200e+02	1.188e+02	2.948e+01	6.4934e+01	205.863	1.409
3.994e+04	2.115e+05	2.121e+05	3.155e+04	9.7393e+04	613027.000	5.156
0.000e+00	1.000e+00	1.000e+00	0.000e+00	1.0000e+00	1.000	0.000

Descriptive Statistics Table for non-hits dataset:

```
## pdf
##   2
```

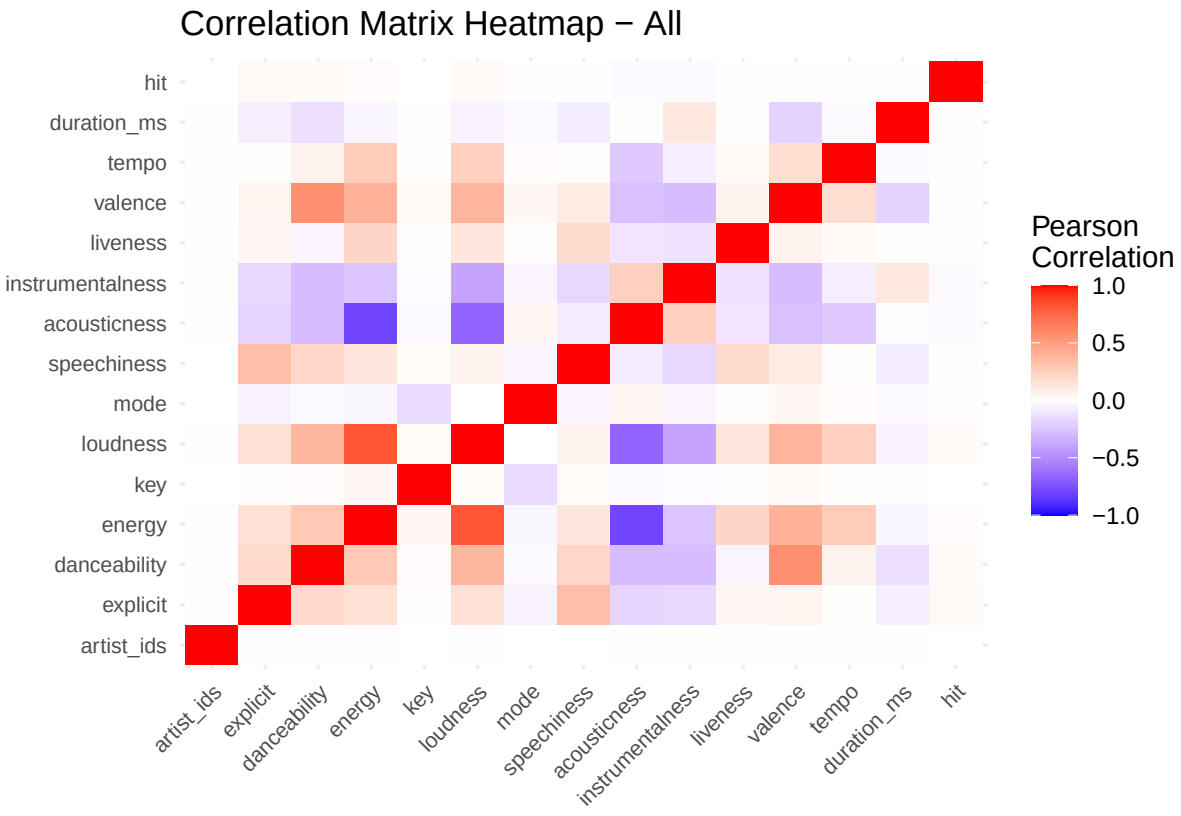
n	sd	median	trimmed	mad	min	max	ra
e−02	2.528e−01	0.000e+00	0.000e+00	0.000e+00	0	1.000	
e−01	1.897e−01	5.000e−01	4.953e−01	2.046e−01	0	1.000	
e−01	2.947e−01	5.240e−01	5.127e−01	3.810e−01	0	1.000	
e+00	3.537e+00	5.000e+00	5.158e+00	4.448e+00	0	11.000	
e+01	6.982e+00	−9.792e+00	−1.089e+01	5.518e+00	−60	7.234	
e−01	4.697e−01	1.000e+00	7.144e−01	0.000e+00	0	1.000	
e−02	1.160e−01	4.460e−02	5.602e−02	1.838e−02	0	0.969	
e−01	3.852e−01	3.890e−01	4.352e−01	5.604e−01	0	0.996	
e−01	3.763e−01	8.100e−03	2.373e−01	1.201e−02	0	1.000	
e−01	1.805e−01	1.250e−01	1.627e−01	6.331e−02	0	1.000	
e−01	2.705e−01	4.030e−01	4.170e−01	3.321e−01	0	1.000	
e+02	3.094e+01	1.167e+02	1.162e+02	3.216e+01	0	248.934	2
e+05	1.622e+05	2.243e+05	2.305e+05	8.154e+04	1000	6061090.000	6060
e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0	0.000	

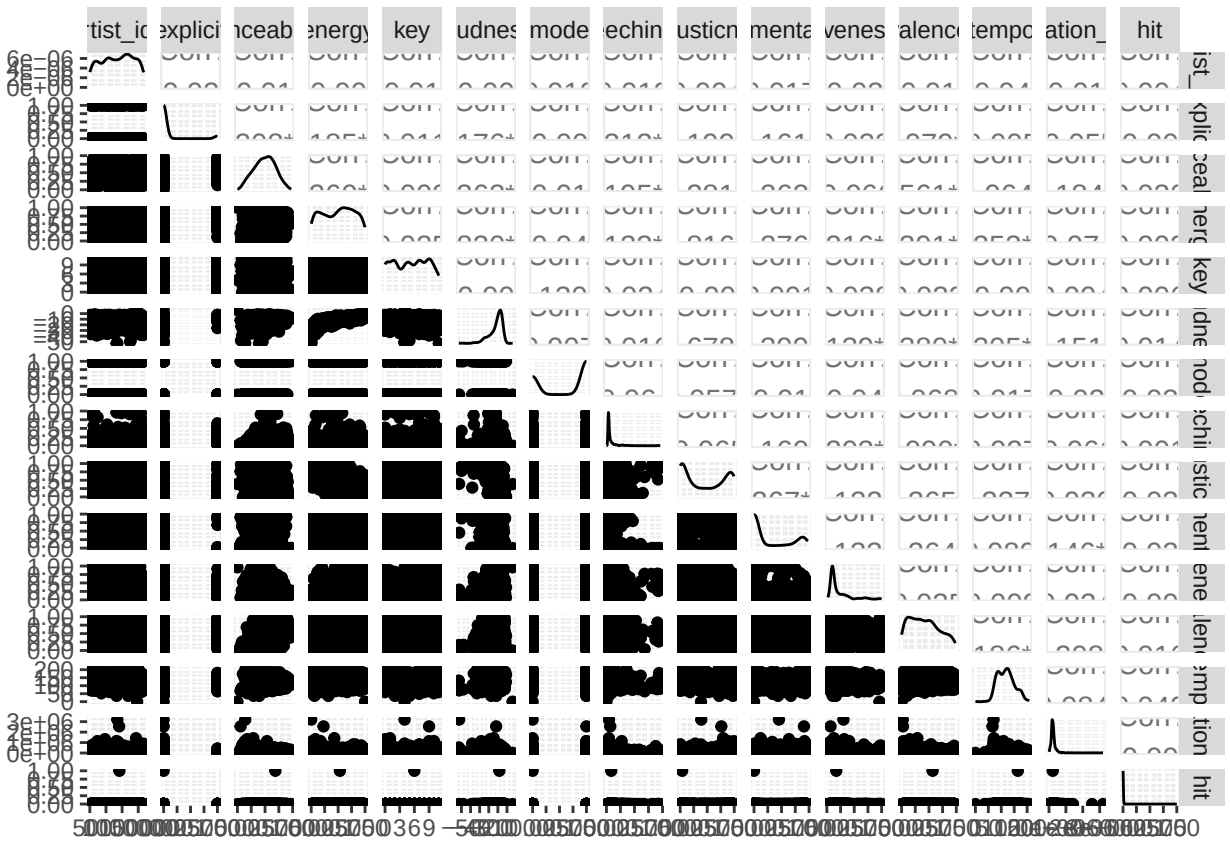
Descriptive Statistics Table for combined dataset (all_hits):

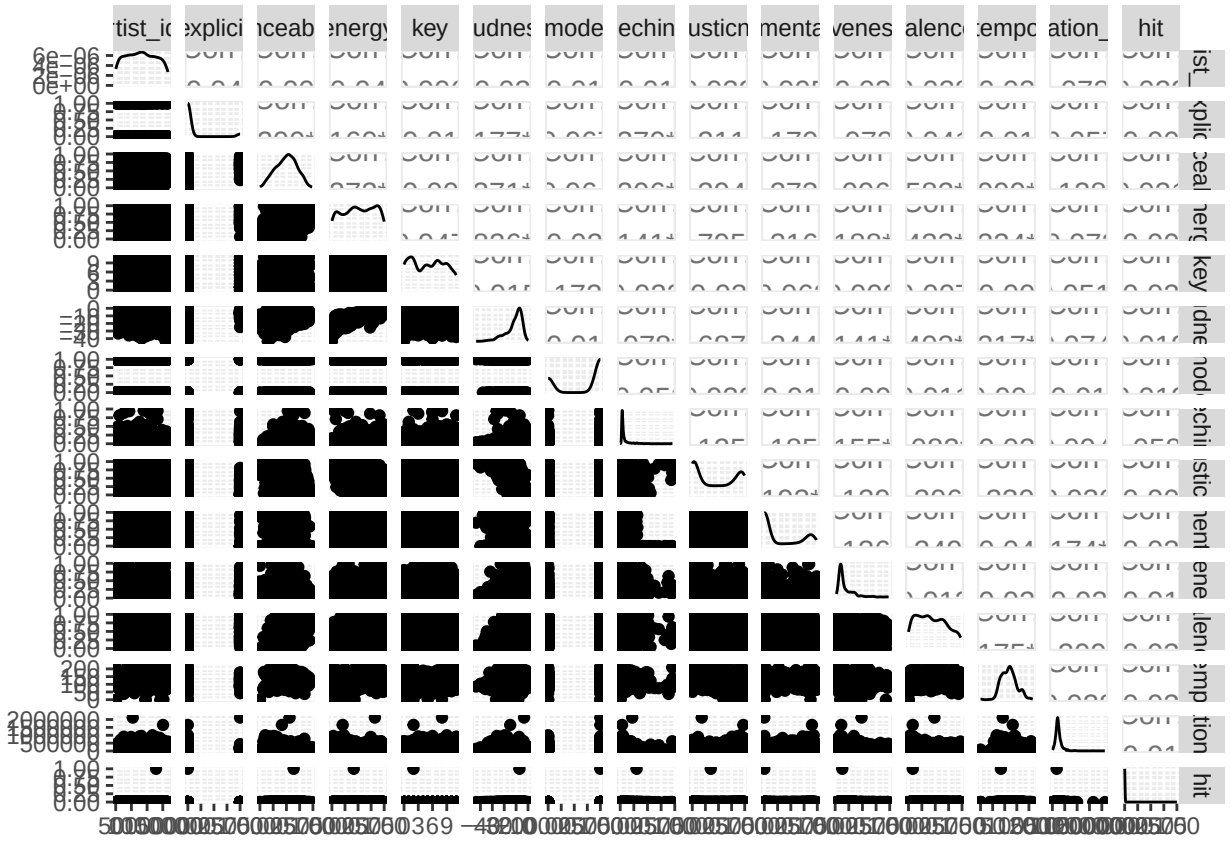
```
## pdf
##   2
```

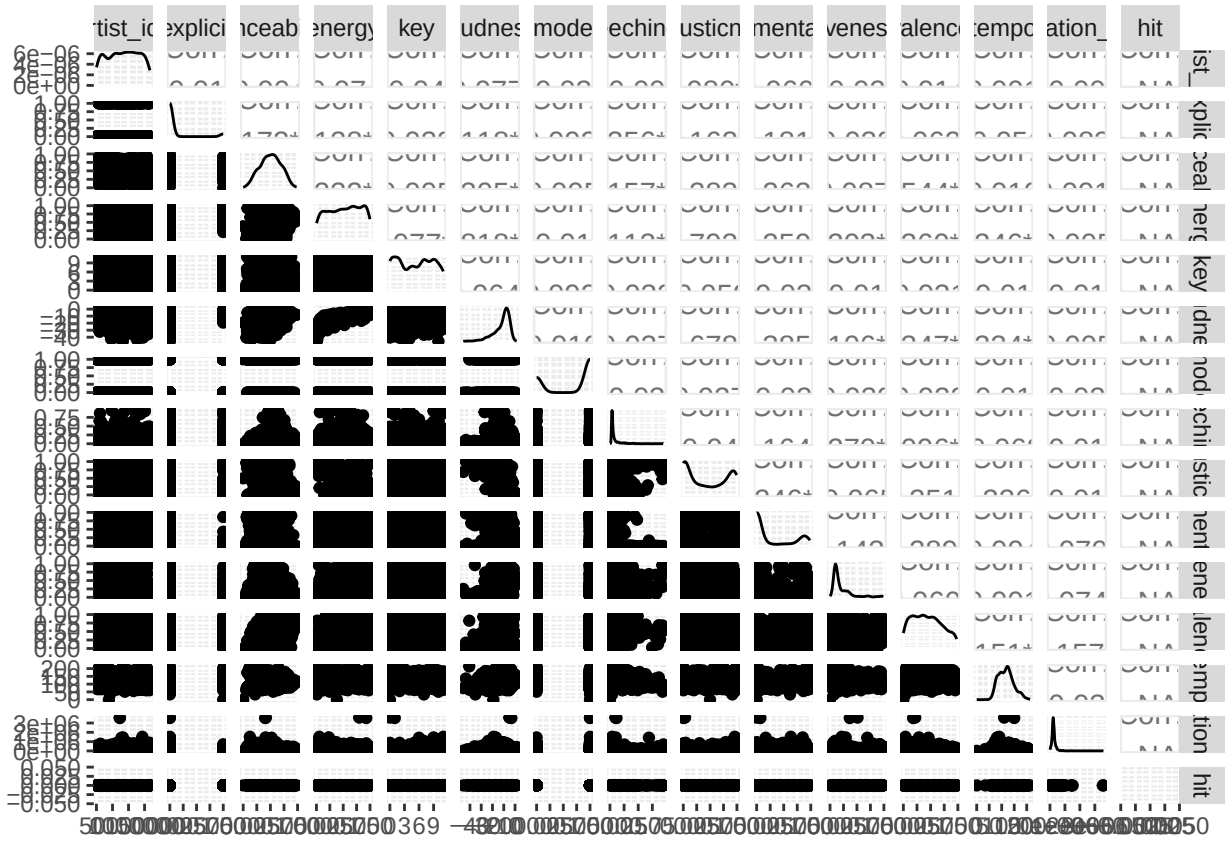
	sd	median	trimmed	mad	min	max	ra
-04	4.820e+04	8.316e+04	8.354e+04	6.098e+04	1	166796.000	1667
-02	2.532e−01	0.000e+00	0.000e+00	0.000e+00	0	1.000	1
-01	1.897e−01	5.010e−01	4.955e−01	2.046e−01	0	1.000	1
-01	2.946e−01	5.240e−01	5.129e−01	3.810e−01	0	1.000	1
-00	3.537e+00	5.000e+00	5.158e+00	4.448e+00	0	11.000	1
+01	6.982e+00	−9.785e+00	−1.088e+01	5.515e+00	−60	7.234	6
-01	4.697e−01	1.000e+00	7.143e−01	0.000e+00	0	1.000	1
-02	1.160e−01	4.460e−02	5.605e−02	1.838e−02	0	0.969	0
-01	3.852e−01	3.880e−01	4.348e−01	5.591e−01	0	0.996	0
-01	3.762e−01	8.000e−03	2.370e−01	1.186e−02	0	1.000	1
-01	1.804e−01	1.250e−01	1.627e−01	6.331e−02	0	1.000	1
-01	2.705e−01	4.030e−01	4.171e−01	3.321e−01	0	1.000	1
-02	3.093e+01	1.167e+02	1.162e+02	3.217e+01	0	248.934	24
-05	1.622e+05	2.243e+05	2.304e+05	8.146e+04	1000	6061090.000	60600
-03	3.238e−02	0.000e+00	0.000e+00	0.000e+00	0	1.000	1

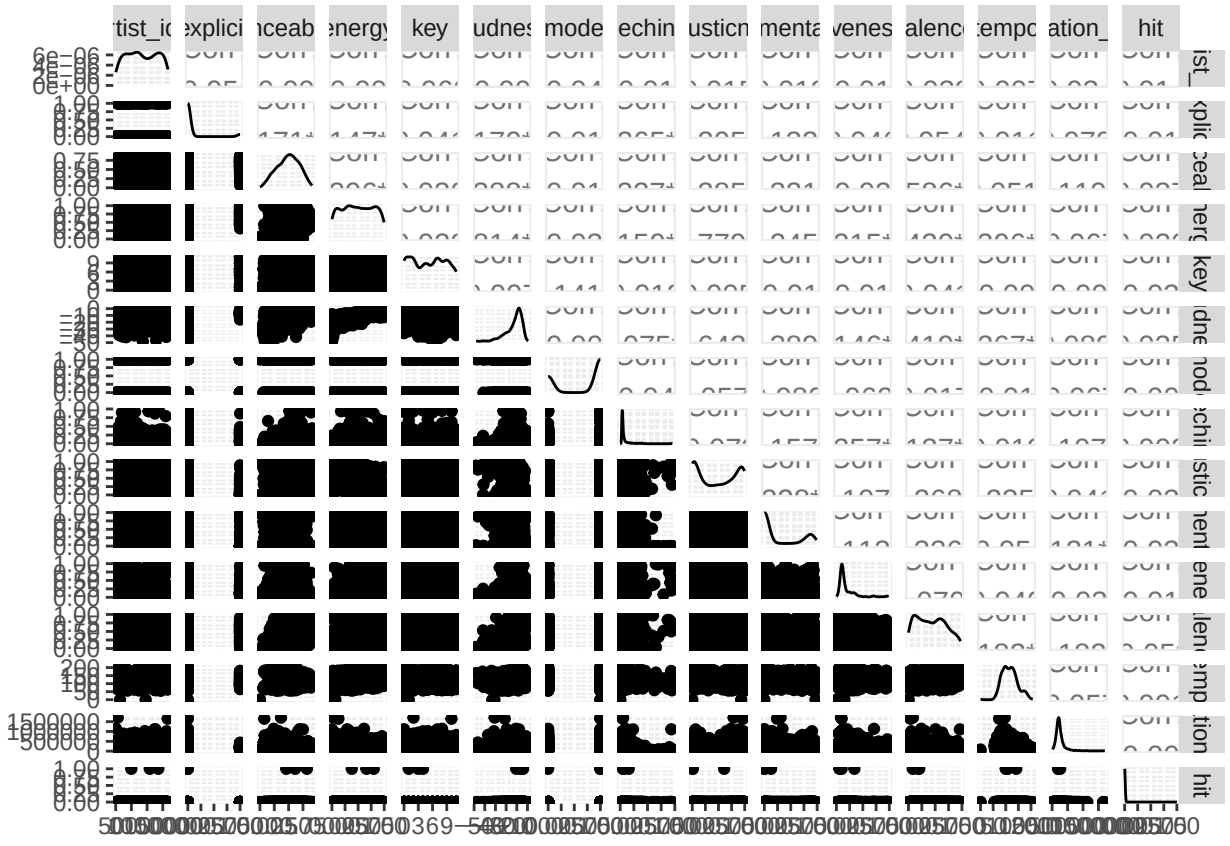
To look for relationships between variables I started off with a simple Correlation Heatmap for the combined dataset. I also wanted to create a scatterplot matrix to see how the variables interact with eat other, but I cannot create 1 scatterplot matrix for 1.2mil+ observations. So instead, I subset the data 10 times into 1200 observation datasets (rougly 0.1% of entire dataset) to see if there were consistent interactions. I did not look at each dataset individually since I need to see if there are any issues with the data as a whole. There might be different correlations in the hits dataset but I am interested in these correlations when combined with the non-hits as this is how I will be training and testing my models.

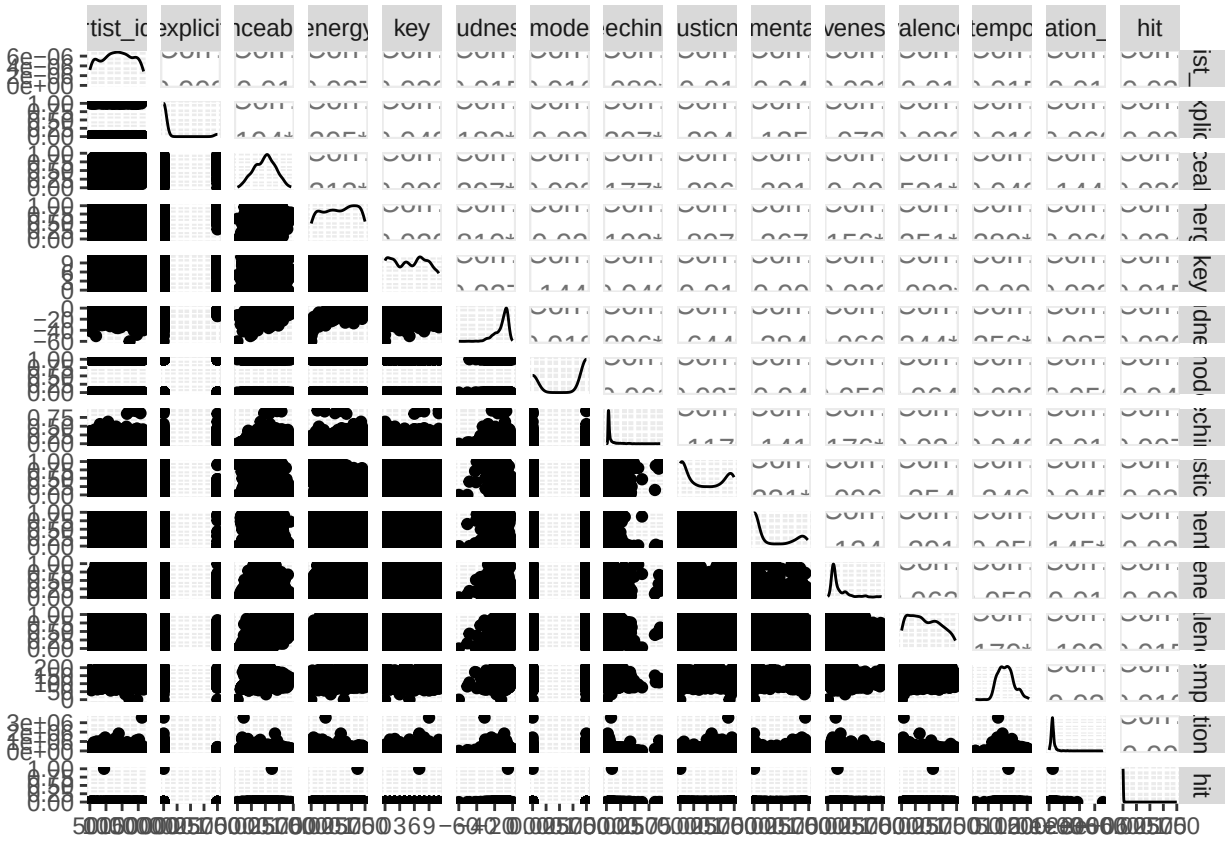

Correlation Matrix Heatmap − All
hit
duration_ms
tempo
valence
liveness
instrumentalness
acousticness
speechiness
mode
loudness
key
energy
danceability
explicit
artist_ids
artist_ids
explicit
danceability
energy
key
loudness
mode
speechiness
acousticness
instrumentalness
liveness
valence
tempo
duration_ms
hit
Pearson Correlation
1.0
0.5
0.0
−0.5
−1.0

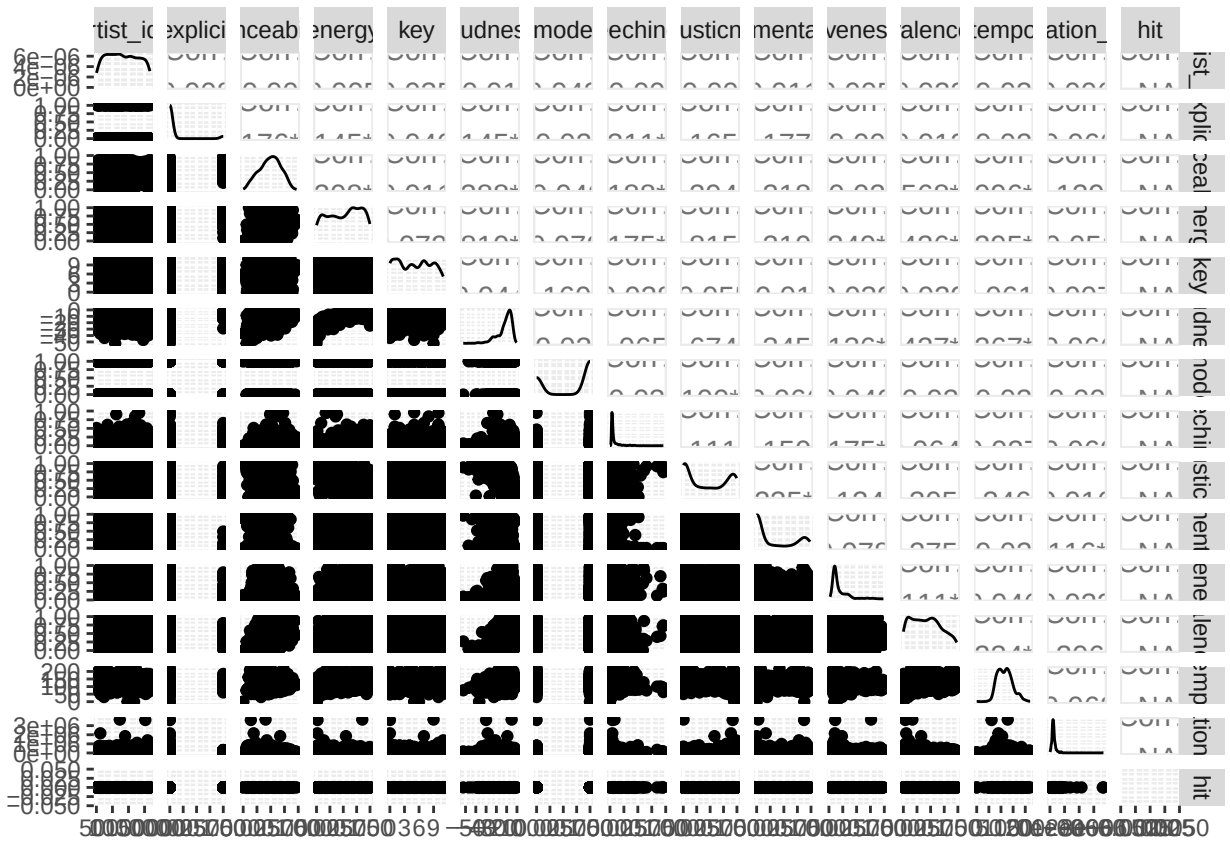

artist_id
explicit
danceability
energy
key
loudness
mode
speechiness
acousticness
instrumentalness
liveness
valence
tempo
duration_ms
hit

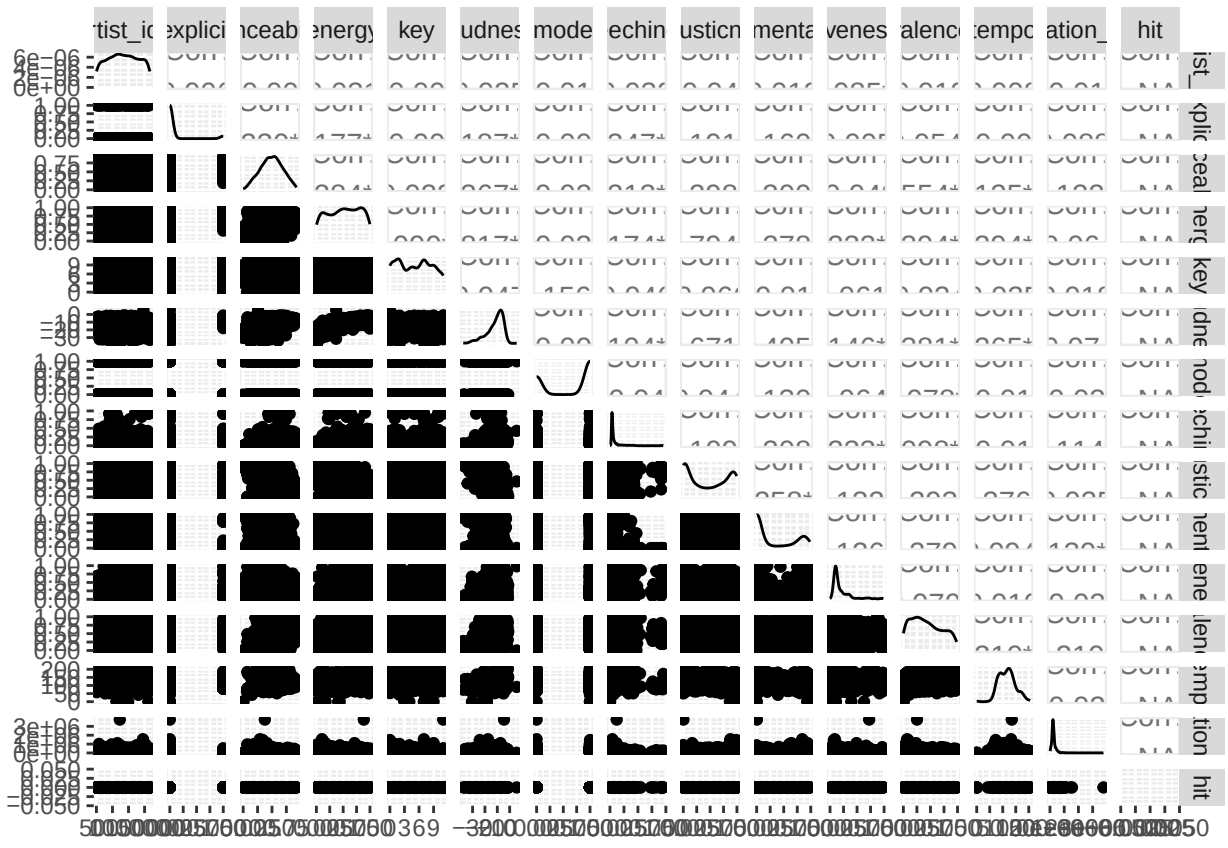

key
key
hit

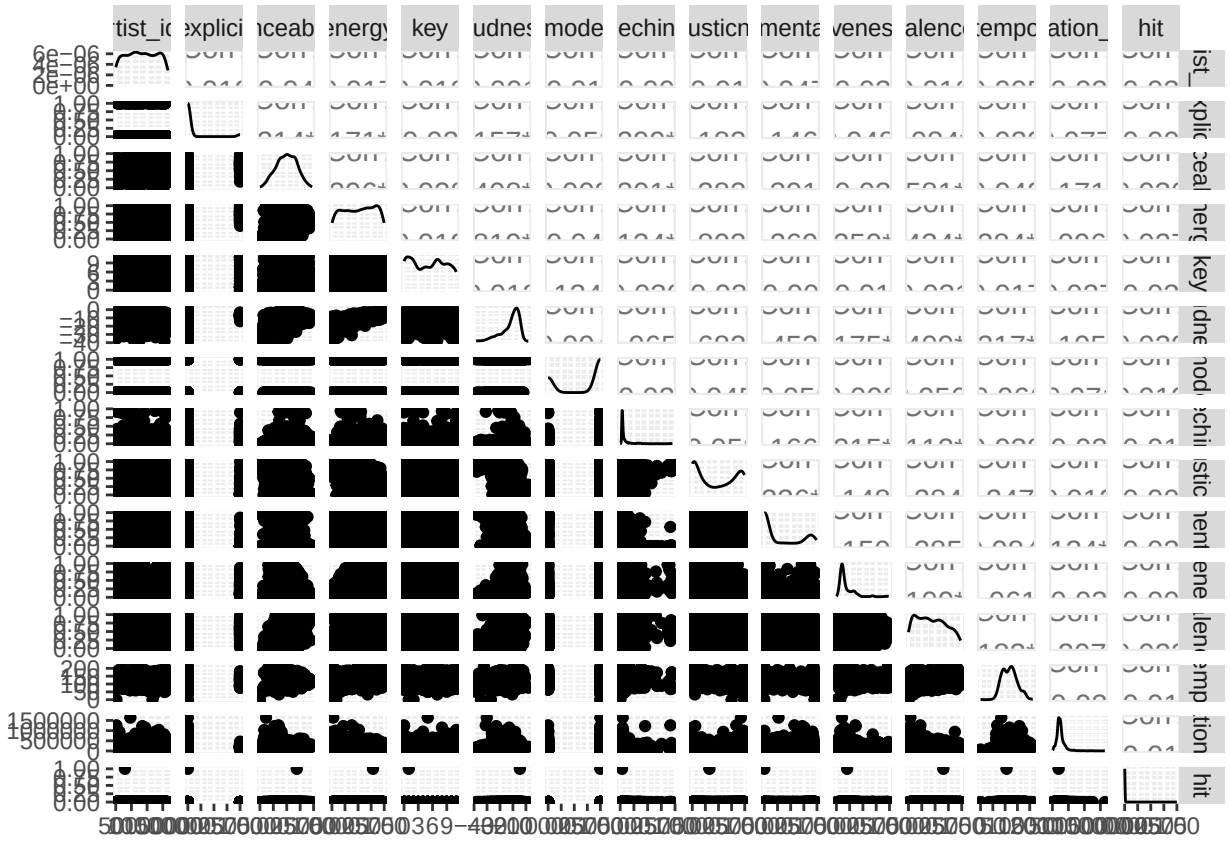

hit

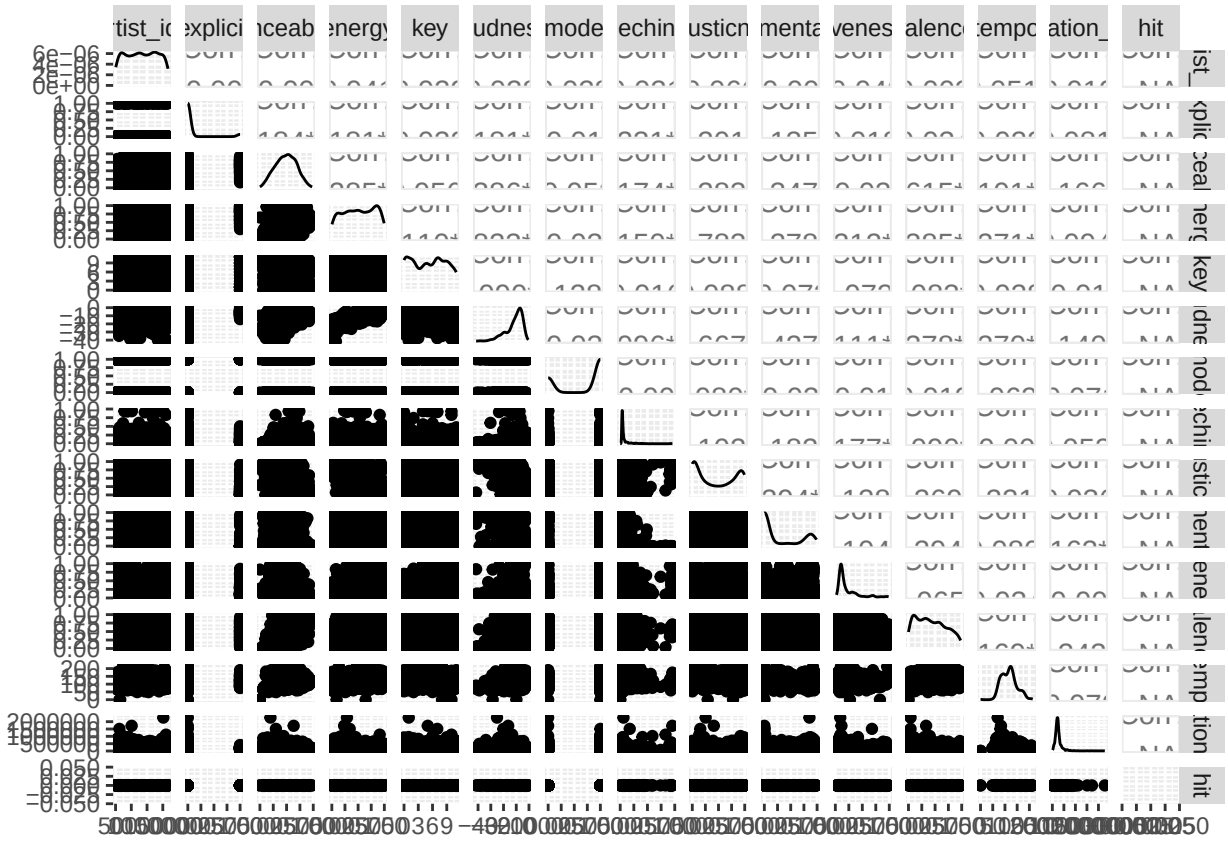

artist_id
explicit
danceability
energy
key
loudness
mode
speechiness
acousticness
instrumentalness
liveness
valence
tempo
duration_ms
hit

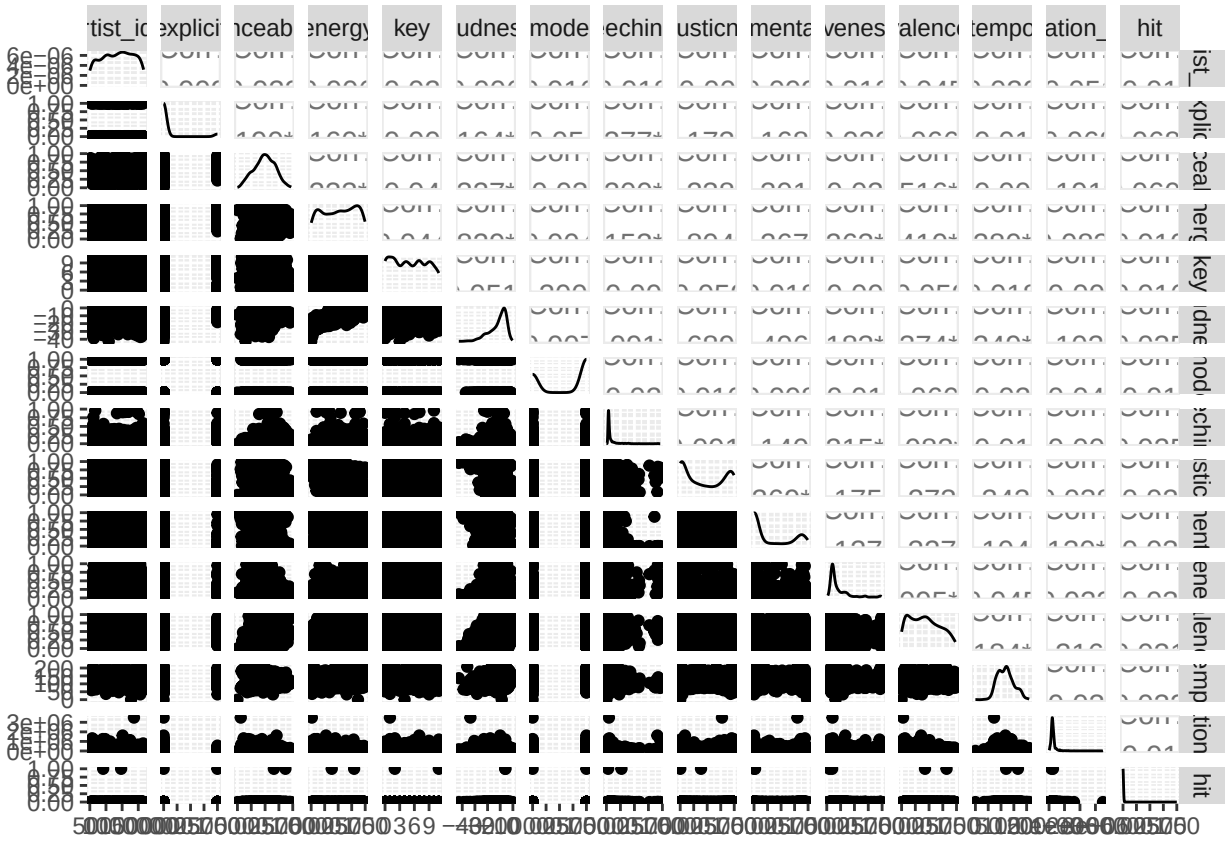

key
hit


hit


hit


key
hit


key
hit



Based on the Correlation Heatmap for the combined dataset, there are a few strong positive correlations and slightly more strong negative correlations. The variables with clear strong positive correlations are loudness and energy and a case can be made for valence and danceability. For the negative correlations, the variables acousticness and energy have a clearly strong correlation, as well as acousticness and loudness. Cases for strong negative correlations could also be made for the following variables, instrumentalness and danceability, instrumentalness and energy, instrumentalness and loudness, acousticness and danceability, valence and acousticness, valence and instrumentalness, and finally tempo and acousticness.

Looking at the 10 Scatterplot Matrices for the combined dataset subsets, some clear relationships arise. Loudness and energy are highly correlated and tightly grouped making them prime candidates for feature reduction. Duration_ms has clear outliers in some of the subsets but in others it looks like there could be a higher density at the lower end of duration_ms but still a spread across the way to the outliers. This means I cannot remove the outliers entirely. Other variables were binomial and have it show clearly in the matrices, such as explicit, hit, and mode.

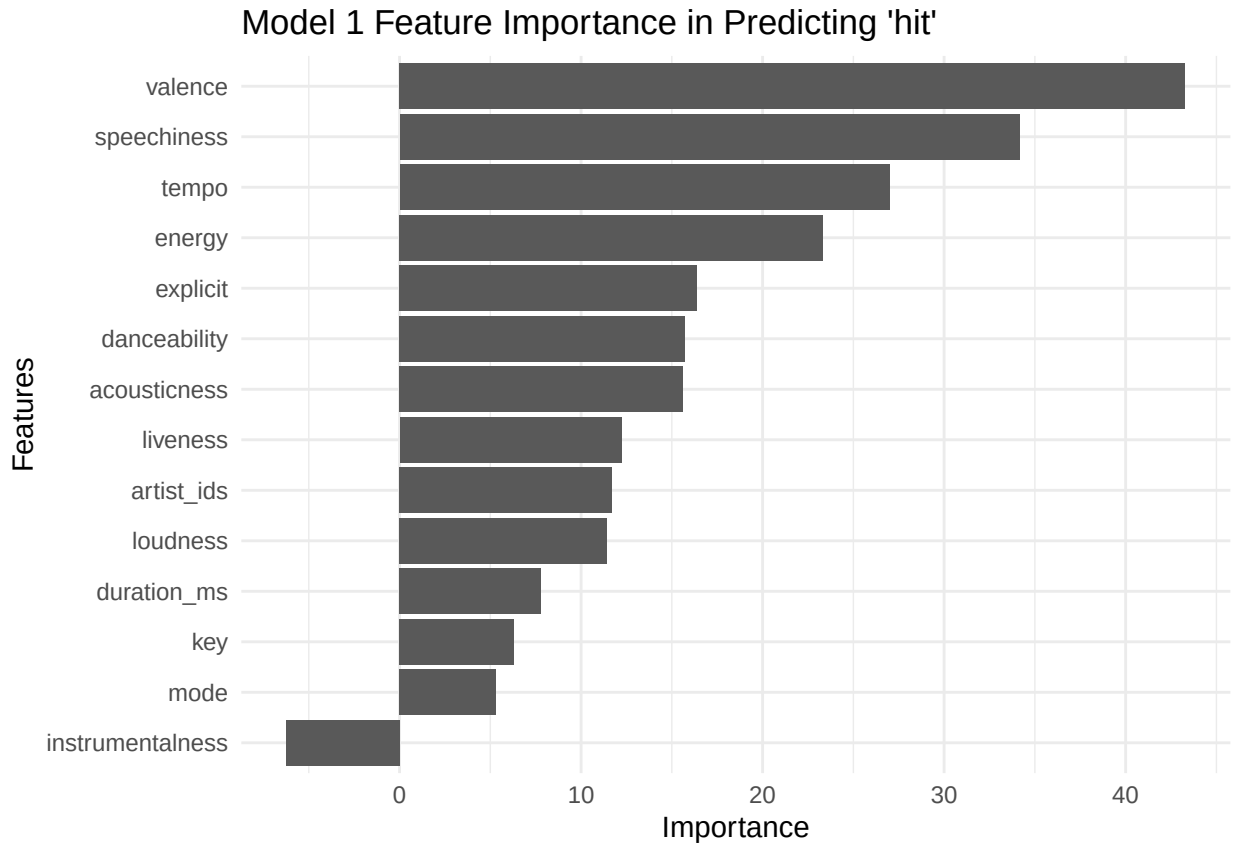
Based on the relationships I saw I want to do feature selection after running my first RF with all features. I will use the Feature Importance values from model 1, specifically the Mean Decrease Accuracy metric, to chose which features to eliminate. Also, for the relationship of Loudness and energy, I will plan to remove one of them based on the lower feature importance value.

Model Description

Model 1

Model 1 is a classification RF with all 14 IVs included as features. The hit is what the model is trying to predict. The combined dataset is split into a 80% training dataset and a 20% test dataset. I am running

150 trees with no pruning occurring. I will then pull the feature importance values from this model and use those to prune Model 2.




```
##           Actual
## Predicted      0      1
##         0 240760    212
##         1     12     17
```

```
## [1] "Accuracy: 0.9991"
```

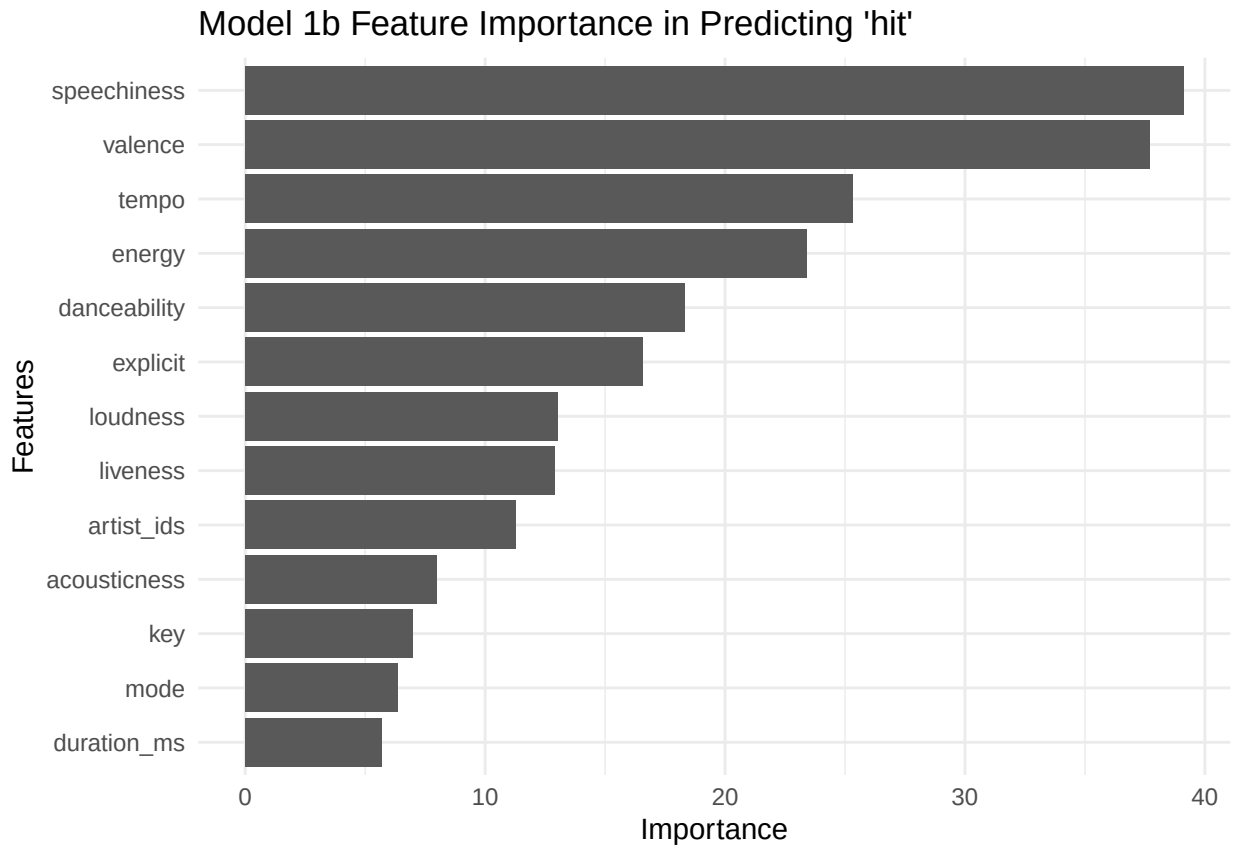
```
## [1] "Type 1 Error Rate: 0"
```

```
## [1] "Type 2 Error Rate: 0.9258"
```

```
## [1] "F1 Score: 0.1318"
```

Model 1b

After looking at the feature importance values of Model 1, I decided to rerun it and eliminate instrumentalness since it showed it had a negative effect on the model when included. No other changes were made and the same seed was used to ensure the 'randomness' was replicated.




```
##           Actual
## Predicted      0      1
##         0 240760    213
##         1     12     16
```

```
## [1] "Accuracy: 0.9991"
```

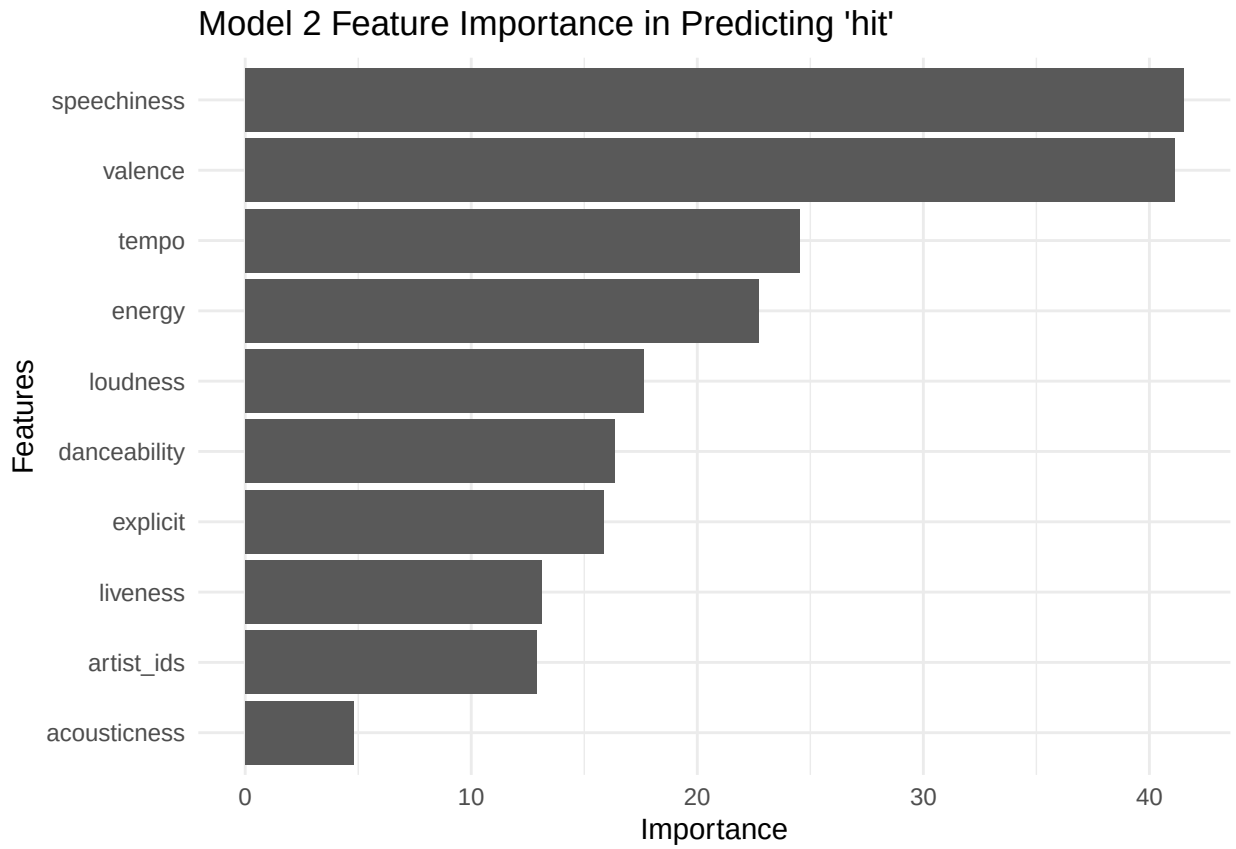
```
## [1] "Type 1 Error Rate: 0"
```

```
## [1] "Type 2 Error Rate: 0.9301"
```

```
## [1] "F1 Score: 0.1245"
```

Model 2

Model 2 is my second RF where I deploy top down pruning by limiting the depth of the trees to 5. I left it at 5 to prevent overfitting after I eliminated features with importance values less than 10% from Model 1. I removed duration_ms, key, and mode. The feature duration_ms was the one that the scatterplot matrices showed was most likely filled with plenty of outliers so this was a good feature to remove. Key did not look to be correlated above 0.01 with any other features so it is essentially a dummy variable. Mode was a binomial variable and only statistically significantly correlated to key. This RF was also run with 150 trees.



Actual Predicted 0 1 0 240762 214 1 10 15 [1] “Accuracy: 0.9991” [1] “Type 1 Error Rate: 0” [1] “Type 2 Error Rate: 0.9345” [1] “F1 Score: 0.1181”

Model 3

Model 3 is a XGBoost with all features in it. I chose to leave all features to see if it would predict better than Model 1 with all features. I again used a 80% training and 20% testing split for the dataset. I tried a few different max depths for the trees but found that 12 was returning the higher F1 score. I did a 100 rounds for the model.

[1] train-logloss:0.438588 [2] train-logloss:0.298061 [3] train-logloss:0.209603 [4] train-logloss:0.150488 [5] train-logloss:0.109655 [6] train-logloss:0.080846 [7] train-logloss:0.060250 [8] train-logloss:0.045361 [9] train-logloss:0.034521 [10] train-logloss:0.026595 [11] train-logloss:0.020755 [12] train-logloss:0.016439 [13] train-logloss:0.013252 [14] train-logloss:0.010805 [15] train-logloss:0.009037 [16] train-logloss:0.007706 [17] train-logloss:0.006672 [18] train-logloss:0.005863 [19] train-logloss:0.005231 [20] train-logloss:0.004735 [21] train-logloss:0.004290 [22] train-logloss:0.004016 [23] train-logloss:0.003770 [24] train-logloss:0.003558 [25] train-logloss:0.003402 [26] train-logloss:0.003271 [27] train-logloss:0.003183 [28] train-logloss:0.003039 [29] train-logloss:0.002975 [30] train-logloss:0.002920 [31] train-logloss:0.002867 [32] train-logloss:0.002776 [33] train-logloss:0.002704 [34] train-logloss:0.002641 [35] train-logloss:0.002529 [36] train-logloss:0.002479 [37] train-logloss:0.002322 [38] train-logloss:0.002266 [39] train-logloss:0.002216 [40] train-logloss:0.002168 [41] train-logloss:0.002137 [42] train-logloss:0.001999 [43] train-logloss:0.001936 [44] train-logloss:0.001894 [45] train-logloss:0.001858 [46] train-logloss:0.001758 [47] train-logloss:0.001718 [48] train-logloss:0.001635 [49] train-logloss:0.001595 [50] train-logloss:0.001573 [51] train-logloss:0.001547 [52] train-logloss:0.001538 [53] train-logloss:0.001514 [54] train-logloss:0.001497 [55] train-logloss:0.001484 [56] train-logloss:0.001431 [57] train-logloss:0.001415 [58] train-logloss:0.001389 [59] train-logloss:0.001296 [60] train-logloss:0.001238 [61] train-logloss:0.001184 [62] train-logloss:0.001161 [63] train-logloss:0.001155 [64] train-logloss:0.001129 [65] train-logloss:0.001062 [66] train-logloss:0.001051 [67] train-logloss:0.001043 [68] train-logloss:0.001019 [69] train-logloss:0.000950 [70] train-logloss:0.000906 [71] train-logloss:0.000886 [72] train-logloss:0.000861 [73] train-

logloss:0.000848 [74] train-logloss:0.000838 [75] train-logloss:0.000820 [76] train-logloss:0.000787 [77] train-logloss:0.000758 [78] train-logloss:0.000739 [79] train-logloss:0.000710 [80] train-logloss:0.000690 [81] train-logloss:0.000683 [82] train-logloss:0.000679 [83] train-logloss:0.000673 [84] train-logloss:0.000668 [85] train-logloss:0.000659 [86] train-logloss:0.000641 [87] train-logloss:0.000634 [88] train-logloss:0.000615 [89] train-logloss:0.000581 [90] train-logloss:0.000553 [91] train-logloss:0.000545 [92] train-logloss:0.000526 [93] train-logloss:0.000504 [94] train-logloss:0.000483 [95] train-logloss:0.000470 [96] train-logloss:0.000456 [97] train-logloss:0.000453 [98] train-logloss:0.000442 [99] train-logloss:0.000435 [100] train-logloss:0.000428 Actual Predicted 0 1 0 240753 213 1 19 16 [1] "Accuracy: 0.999" [1] "Type 1 Error Rate: 1e-04" [1] "Type 2 Error Rate: 0.9301" [1] "F1 Score: 0.1212"

Model 4

Model 4 is the second XGBoost that I ran. This one was combined with k folding to get an average F1 score and see if averaging across multiple XGBoost would result in a higher F1 score. I did 5 folds and set the tree depth to 6 to prevent too much overfitting. As with Model 3, I left all features in for Model 4 and did 100 rounds for the model.

[1] train-logloss:0.438557 [2] train-logloss:0.298009 [3] train-logloss:0.209543 [4] train-logloss:0.150449 [5] train-logloss:0.109635 [6] train-logloss:0.080877 [7] train-logloss:0.060345 [8] train-logloss:0.045563 [9] train-logloss:0.034862 [10] train-logloss:0.027092 [11] train-logloss:0.021426 [12] train-logloss:0.017299 [13] train-logloss:0.014283 [14] train-logloss:0.012086 [15] train-logloss:0.010459 [16] train-logloss:0.009267 [17] train-logloss:0.008391 [18] train-logloss:0.007741 [19] train-logloss:0.007250 [20] train-logloss:0.006875 [21] train-logloss:0.006568 [22] train-logloss:0.006346 [23] train-logloss:0.006169 [24] train-logloss:0.005986 [25] train-logloss:0.005878 [26] train-logloss:0.005793 [27] train-logloss:0.005720 [28] train-logloss:0.005637 [29] train-logloss:0.005566 [30] train-logloss:0.005505 [31] train-logloss:0.005461 [32] train-logloss:0.005432 [33] train-logloss:0.005386 [34] train-logloss:0.005338 [35] train-logloss:0.005324 [36] train-logloss:0.005286 [37] train-logloss:0.005214 [38] train-logloss:0.005161 [39] train-logloss:0.005133 [40] train-logloss:0.005080 [41] train-logloss:0.004988 [42] train-logloss:0.004884 [43] train-logloss:0.004809 [44] train-logloss:0.004797 [45] train-logloss:0.004781 [46] train-logloss:0.004743 [47] train-logloss:0.004686 [48] train-logloss:0.004653 [49] train-logloss:0.004596 [50] train-logloss:0.004534 [51] train-logloss:0.004496 [52] train-logloss:0.004485 [53] train-logloss:0.004451 [54] train-logloss:0.004432 [55] train-logloss:0.004350 [56] train-logloss:0.004308 [57] train-logloss:0.004295 [58] train-logloss:0.004243 [59] train-logloss:0.004186 [60] train-logloss:0.004159 [61] train-logloss:0.004142 [62] train-logloss:0.004105 [63] train-logloss:0.004093 [64] train-logloss:0.004074 [65] train-logloss:0.004055 [66] train-logloss:0.004046 [67] train-logloss:0.004018 [68] train-logloss:0.004008 [69] train-logloss:0.003959 [70] train-logloss:0.003914 [71] train-logloss:0.003889 [72] train-logloss:0.003858 [73] train-logloss:0.003828 [74] train-logloss:0.003780 [75] train-logloss:0.003732 [76] train-logloss:0.003674 [77] train-logloss:0.003643 [78] train-logloss:0.003610 [79] train-logloss:0.003584 [80] train-logloss:0.003578 [81] train-logloss:0.003543 [82] train-logloss:0.003523 [83] train-logloss:0.003476 [84] train-logloss:0.003442 [85] train-logloss:0.003423 [86] train-logloss:0.003410 [87] train-logloss:0.003404 [88] train-logloss:0.003343 [89] train-logloss:0.003289 [90] train-logloss:0.003252 [91] train-logloss:0.003208 [92] train-logloss:0.003144 [93] train-logloss:0.003122 [94] train-logloss:0.003087 [95] train-logloss:0.003052 [96] train-logloss:0.003006 [97] train-logloss:0.002993 [98] train-logloss:0.002971 [99] train-logloss:0.002949 [100] train-logloss:0.002930 [1] train-logloss:0.438561 [2] train-logloss:0.298016 [3] train-logloss:0.209551 [4] train-logloss:0.150457 [5] train-logloss:0.109651 [6] train-logloss:0.080895 [7] train-logloss:0.060360 [8] train-logloss:0.045575 [9] train-logloss:0.034872 [10] train-logloss:0.027088 [11] train-logloss:0.021427 [12] train-logloss:0.017301 [13] train-logloss:0.014283 [14] train-logloss:0.012085 [15] train-logloss:0.010470 [16] train-logloss:0.009291 [17] train-logloss:0.008409 [18] train-logloss:0.007752 [19] train-logloss:0.007267 [20] train-logloss:0.006908 [21] train-logloss:0.006605 [22] train-logloss:0.006323 [23] train-logloss:0.006129 [24] train-logloss:0.005995 [25] train-logloss:0.005866 [26] train-logloss:0.005756 [27] train-logloss:0.005677 [28] train-logloss:0.005624 [29] train-logloss:0.005530 [30] train-logloss:0.005483 [31] train-logloss:0.005429 [32] train-logloss:0.005387 [33] train-logloss:0.005319 [34] train-logloss:0.005279 [35] train-logloss:0.005248 [36] train-logloss:0.005211 [37] train-logloss:0.005194 [38] train-logloss:0.005171 [39] train-logloss:0.005160 [40] train-logloss:0.005097 [41] train-logloss:0.004996 [42] train-logloss:0.004905 [43] train-logloss:0.004840 [44] train-logloss:0.004771 [45] train-logloss:0.004726 [46] train-logloss:0.004672 [47] train-logloss:0.004622 [48] train-logloss:0.004601 [49] train-logloss:0.004554 [50] train-logloss:0.004545 [51] train-logloss:0.004511 [52] train-logloss:0.004462 [53] train-logloss:0.004443 [54] train-logloss:0.004429 [55] train-logloss:0.004374 [56] train-logloss:0.004309 [57]

train-logloss:0.004262 [58] train-logloss:0.004210 [59] train-logloss:0.004085 [60] train-logloss:0.004050 [61]
train-logloss:0.004031 [62] train-logloss:0.004024 [63] train-logloss:0.003986 [64] train-logloss:0.003955 [65]
train-logloss:0.003897 [66] train-logloss:0.003831 [67] train-logloss:0.003798 [68] train-logloss:0.003752 [69]
train-logloss:0.003700 [70] train-logloss:0.003676 [71] train-logloss:0.003652 [72] train-logloss:0.003647 [73]
train-logloss:0.003629 [74] train-logloss:0.003621 [75] train-logloss:0.003613 [76] train-logloss:0.003591 [77]
train-logloss:0.003513 [78] train-logloss:0.003510 [79] train-logloss:0.003504 [80] train-logloss:0.003502 [81]
train-logloss:0.003480 [82] train-logloss:0.003470 [83] train-logloss:0.003463 [84] train-logloss:0.003453 [85]
train-logloss:0.003438 [86] train-logloss:0.003414 [87] train-logloss:0.003405 [88] train-logloss:0.003362 [89]
train-logloss:0.003341 [90] train-logloss:0.003310 [91] train-logloss:0.003281 [92] train-logloss:0.003229 [93]
train-logloss:0.003204 [94] train-logloss:0.003177 [95] train-logloss:0.003167 [96] train-logloss:0.003146 [97]
train-logloss:0.003111 [98] train-logloss:0.003079 [99] train-logloss:0.003064 [100] train-logloss:0.003037
[1] train-logloss:0.438563 [2] train-logloss:0.298020 [3] train-logloss:0.209557 [4] train-logloss:0.150466
[5] train-logloss:0.109662 [6] train-logloss:0.080905 [7] train-logloss:0.060371 [8] train-logloss:0.045590 [9]
train-logloss:0.034889 [10] train-logloss:0.027104 [11] train-logloss:0.021437 [12] train-logloss:0.017310 [13]
train-logloss:0.014297 [14] train-logloss:0.012095 [15] train-logloss:0.010494 [16] train-logloss:0.009313 [17]
train-logloss:0.008424 [18] train-logloss:0.007769 [19] train-logloss:0.007278 [20] train-logloss:0.006912 [21]
train-logloss:0.006608 [22] train-logloss:0.006388 [23] train-logloss:0.006213 [24] train-logloss:0.006043 [25]
train-logloss:0.005936 [26] train-logloss:0.005847 [27] train-logloss:0.005764 [28] train-logloss:0.005672 [29]
train-logloss:0.005620 [30] train-logloss:0.005566 [31] train-logloss:0.005490 [32] train-logloss:0.005454 [33]
train-logloss:0.005404 [34] train-logloss:0.005362 [35] train-logloss:0.005302 [36] train-logloss:0.005230 [37]
train-logloss:0.005179 [38] train-logloss:0.005145 [39] train-logloss:0.005130 [40] train-logloss:0.005091 [41]
train-logloss:0.005048 [42] train-logloss:0.005019 [43] train-logloss:0.004986 [44] train-logloss:0.004969 [45]
train-logloss:0.004899 [46] train-logloss:0.004839 [47] train-logloss:0.004820 [48] train-logloss:0.004793 [49]
train-logloss:0.004773 [50] train-logloss:0.004702 [51] train-logloss:0.004675 [52] train-logloss:0.004628 [53]
train-logloss:0.004595 [54] train-logloss:0.004498 [55] train-logloss:0.004468 [56] train-logloss:0.004455 [57]
train-logloss:0.004420 [58] train-logloss:0.004412 [59] train-logloss:0.004377 [60] train-logloss:0.004300 [61]
train-logloss:0.004293 [62] train-logloss:0.004275 [63] train-logloss:0.004158 [64] train-logloss:0.004135 [65]
train-logloss:0.004105 [66] train-logloss:0.004080 [67] train-logloss:0.004072 [68] train-logloss:0.004045 [69]
train-logloss:0.003997 [70] train-logloss:0.003976 [71] train-logloss:0.003914 [72] train-logloss:0.003881 [73]
train-logloss:0.003848 [74] train-logloss:0.003786 [75] train-logloss:0.003723 [76] train-logloss:0.003694 [77]
train-logloss:0.003664 [78] train-logloss:0.003635 [79] train-logloss:0.003619 [80] train-logloss:0.003560 [81]
train-logloss:0.003549 [82] train-logloss:0.003546 [83] train-logloss:0.003526 [84] train-logloss:0.003515 [85]
train-logloss:0.003500 [86] train-logloss:0.003471 [87] train-logloss:0.003464 [88] train-logloss:0.003437 [89]
train-logloss:0.003409 [90] train-logloss:0.003370 [91] train-logloss:0.003360 [92] train-logloss:0.003340 [93]
train-logloss:0.003319 [94] train-logloss:0.003275 [95] train-logloss:0.003261 [96] train-logloss:0.003232 [97]
train-logloss:0.003203 [98] train-logloss:0.003150 [99] train-logloss:0.003124 [100] train-logloss:0.003085
[1] train-logloss:0.438555 [2] train-logloss:0.298007 [3] train-logloss:0.209540 [4] train-logloss:0.150444
[5] train-logloss:0.109631 [6] train-logloss:0.080872 [7] train-logloss:0.060334 [8] train-logloss:0.045548 [9]
train-logloss:0.034847 [10] train-logloss:0.027073 [11] train-logloss:0.021415 [12] train-logloss:0.017288 [13]
train-logloss:0.014274 [14] train-logloss:0.012081 [15] train-logloss:0.010470 [16] train-logloss:0.009275 [17]
train-logloss:0.008398 [18] train-logloss:0.007738 [19] train-logloss:0.007238 [20] train-logloss:0.006877 [21]
train-logloss:0.006587 [22] train-logloss:0.006365 [23] train-logloss:0.006196 [24] train-logloss:0.006035 [25]
train-logloss:0.005932 [26] train-logloss:0.005826 [27] train-logloss:0.005741 [28] train-logloss:0.005650 [29]
train-logloss:0.005584 [30] train-logloss:0.005546 [31] train-logloss:0.005496 [32] train-logloss:0.005393 [33]
train-logloss:0.005368 [34] train-logloss:0.005320 [35] train-logloss:0.005243 [36] train-logloss:0.005206 [37]
train-logloss:0.005167 [38] train-logloss:0.005127 [39] train-logloss:0.005082 [40] train-logloss:0.005059 [41]
train-logloss:0.005003 [42] train-logloss:0.004961 [43] train-logloss:0.004932 [44] train-logloss:0.004895 [45]
train-logloss:0.004832 [46] train-logloss:0.004819 [47] train-logloss:0.004795 [48] train-logloss:0.004762 [49]
train-logloss:0.004728 [50] train-logloss:0.004686 [51] train-logloss:0.004624 [52] train-logloss:0.004548 [53]
train-logloss:0.004539 [54] train-logloss:0.004522 [55] train-logloss:0.004487 [56] train-logloss:0.004439 [57]
train-logloss:0.004362 [58] train-logloss:0.004345 [59] train-logloss:0.004338 [60] train-logloss:0.004289 [61]
train-logloss:0.004223 [62] train-logloss:0.004174 [63] train-logloss:0.004141 [64] train-logloss:0.004009 [65]
train-logloss:0.003940 [66] train-logloss:0.003909 [67] train-logloss:0.003889 [68] train-logloss:0.003876 [69]
train-logloss:0.003868 [70] train-logloss:0.003849 [71] train-logloss:0.003770 [72] train-logloss:0.003735 [73]

train-logloss:0.003675 [74] train-logloss:0.003651 [75] train-logloss:0.003604 [76] train-logloss:0.003588 [77] train-logloss:0.003583 [78] train-logloss:0.003541 [79] train-logloss:0.003533 [80] train-logloss:0.003516 [81] train-logloss:0.003497 [82] train-logloss:0.003486 [83] train-logloss:0.003455 [84] train-logloss:0.003439 [85] train-logloss:0.003392 [86] train-logloss:0.003389 [87] train-logloss:0.003358 [88] train-logloss:0.003338 [89] train-logloss:0.003287 [90] train-logloss:0.003265 [91] train-logloss:0.003217 [92] train-logloss:0.003180 [93] train-logloss:0.003156 [94] train-logloss:0.003127 [95] train-logloss:0.003097 [96] train-logloss:0.003086 [97] train-logloss:0.003079 [98] train-logloss:0.003072 [99] train-logloss:0.003048 [100] train-logloss:0.003005 [1] train-logloss:0.438579 [2] train-logloss:0.298044 [3] train-logloss:0.209588 [4] train-logloss:0.150503 [5] train-logloss:0.109704 [6] train-logloss:0.080952 [7] train-logloss:0.060427 [8] train-logloss:0.045649 [9] train-logloss:0.034953 [10] train-logloss:0.027185 [11] train-logloss:0.021534 [12] train-logloss:0.017413 [13] train-logloss:0.014404 [14] train-logloss:0.012207 [15] train-logloss:0.010605 [16] train-logloss:0.009419 [17] train-logloss:0.008526 [18] train-logloss:0.007873 [19] train-logloss:0.007397 [20] train-logloss:0.007023 [21] train-logloss:0.006737 [22] train-logloss:0.006480 [23] train-logloss:0.006291 [24] train-logloss:0.006160 [25] train-logloss:0.006038 [26] train-logloss:0.005955 [27] train-logloss:0.005874 [28] train-logloss:0.005809 [29] train-logloss:0.005758 [30] train-logloss:0.005662 [31] train-logloss:0.005578 [32] train-logloss:0.005504 [33] train-logloss:0.005443 [34] train-logloss:0.005401 [35] train-logloss:0.005353 [36] train-logloss:0.005323 [37] train-logloss:0.005297 [38] train-logloss:0.005262 [39] train-logloss:0.005220 [40] train-logloss:0.005197 [41] train-logloss:0.005178 [42] train-logloss:0.005155 [43] train-logloss:0.005059 [44] train-logloss:0.005006 [45] train-logloss:0.004962 [46] train-logloss:0.004927 [47] train-logloss:0.004872 [48] train-logloss:0.004805 [49] train-logloss:0.004767 [50] train-logloss:0.004746 [51] train-logloss:0.004692 [52] train-logloss:0.004682 [53] train-logloss:0.004666 [54] train-logloss:0.004647 [55] train-logloss:0.004634 [56] train-logloss:0.004607 [57] train-logloss:0.004592 [58] train-logloss:0.004566 [59] train-logloss:0.004558 [60] train-logloss:0.004510 [61] train-logloss:0.004464 [62] train-logloss:0.004416 [63] train-logloss:0.004294 [64] train-logloss:0.004259 [65] train-logloss:0.004233 [66] train-logloss:0.004199 [67] train-logloss:0.004140 [68] train-logloss:0.004124 [69] train-logloss:0.004120 [70] train-logloss:0.004078 [71] train-logloss:0.004057 [72] train-logloss:0.004018 [73] train-logloss:0.003992 [74] train-logloss:0.003931 [75] train-logloss:0.003900 [76] train-logloss:0.003876 [77] train-logloss:0.003793 [78] train-logloss:0.003777 [79] train-logloss:0.003720 [80] train-logloss:0.003697 [81] train-logloss:0.003678 [82] train-logloss:0.003664 [83] train-logloss:0.003659 [84] train-logloss:0.003639 [85] train-logloss:0.003616 [86] train-logloss:0.003580 [87] train-logloss:0.003525 [88] train-logloss:0.003504 [89] train-logloss:0.003485 [90] train-logloss:0.003479 [91] train-logloss:0.003457 [92] train-logloss:0.003447 [93] train-logloss:0.003420 [94] train-logloss:0.003379 [95] train-logloss:0.003354 [96] train-logloss:0.003324 [97] train-logloss:0.003318 [98] train-logloss:0.003267 [99] train-logloss:0.003220 [100] train-logloss:0.003192 [1] "Average F1 Score: 0.9995"

Analysis

The F1 scores were 0.1318 for Model 1, 0.1245 for Model 1b, 0.1181 for Model 2, and 0.1212 for Model 3. The accuracy of the models was 0.9991 for Model 1, 0.9991 for Model 1b, 0.9991 for Model 2, and 0.9990 for Model 3. The precision for the models was 0.5862 for Model 1, 0.5714 for Model 1b, 0.6000 for Model 2, and 0.4571 for Model 3. The recall for the models was 0.0742 for Model 1, 0.0699 for Model 1b, 0.0655 for Model 2, and 0.0699 for Model 3. All of these are extremely close to each other and the Type 1 and Type 2 errors were no different. When taking all of these into account, Model 1 still looks to perform the best when it comes to prediction. Model 4 I ran after looking at my results from the first 4 models and it had an average F1 score of 0.9995 with a balanced accuracy of 0.5170. I am skeptical of this result as the only real difference between Model 4 and Model 3 is the 5 k-folds that I performed. I do not see k-folding improving my model's performance from an F1 score of 0.0699 to 0.9995. Due to the overall metrics and my skepticism of Model 4's metrics, I reject my null hypothesis. I was not able to find a model that could accurately predict if a song is a Top Hit above 90%.

Limitations and Issues

A clear limitation that arose early on was the size of my datasets. I had over 1.2 million observations and 25 variables at the beginning. This caused issues with being able to sift through the data for every outlier

and possible issue within observations. It also made it impossible to do multiple visualizations without sub-setting to extremely small subsets. These subsets are so small that I would have needed at least 100 to feel comfortable that most of trends in the data were being captured by the subsets. The size also put limitations on what I could do computational on a laptop. I am fortunate to have had a very powerful laptop because if I had not, it would have been more difficult than it already was.

An issue that I noticed when running the models is that the amount of non-hits to hits may have skewed the models to prefer to just predict a non-hit to keep overall accuracy higher. The ratio of 1 hit to 10 non-hits may have been too large for this project.

Another issue I realize now is that the data collection is flawed. I took songs that were in the 'Top Hits of XXXX' playlists and these were clearly all top hits. But, when taking the non-hits into account, there are most likely a large amount of tracks that were also hits but did not make the playlists due to the limitation of 100 or so tracks. These tracks in the non-hits would still show attribute values and trends similar to the top hits tracks and cause the non-hits variables to not have enough distance from the hits variables for the models to find. Every model had the issue of False Negatives being extremely high and I believe it is due to this reasoning. However, there are some tracks every so often that become massively popular and I believe these are the ones that the models accurately got as True Positives.

Future Work Possibilities

Per the Spotify API terms and conditions, nothing can be taken to produce a ML algorithm. I have used this just for personal training and interests and do not intend to work on this further. I do believe that Spotify has these terms in place because they are working on something in house. Spotify is well known for its custom playlists tailored to individual users and I believe they do not want the external competition. Just recently in December 2023, they released an AI called DJ that plays a personalized radio station for the individual user. This is exactly what I assumed was occurring in house at Spotify and I believe they are working on even more complex algorithms as well.

Resources

AB, S. (2023). About Spotify. Retrieved from For the Record: https://newsroom.spotify.com/company-info/ AB, S. (2023). Loud&Clear. Retrieved from byspotify: https://loudandclear.byspotify.com/ AB, S. (2023). Web API. Retrieved from Spotify for Developers: https://developer.spotify.com/documentation/web-api Andrews, E. (2023, August 8). What Is the Oldest Known Piece of Music? Retrieved from History: https://www.history.com/news/what-is-the-oldest-known-piece-of-music Antal, D. (2022, December 15). spotifyr. Retrieved from RDocumentation: https://www.rdocumentation.org/packages/spotifyr/versions/2.2.4 Maci. (2023, August 24). Spotify Facts. Retrieved from FACTS.NET: https://facts.net/spotify-facts/ Ward, T. (2017, July 31). How Do Songs End Up On Spotify Playlists Anyway? Retrieved from Forbes: https://www.forbes.com/sites/tomward/2017/07/31/how-do-songs-end-up-on-spotify-playlists-anyway/?sh=1623fc0cbbb7