STA 214: Project

Dr. Dalzell, Spring 2024

The Data

We have a client who is a serious music lover, and they have two favorite bands: The Front Bottoms and Manchester Orchestra. They are interested in building a model that can do two things:

- (1) They want to determine what song characteristics differentiate the two bands. In other words, what traits make a song "sound like" one band versus the other?
- (2) The two bands collaborated on a song called *Allentown*. The client wants us to use our model to determine if the song is more like a The Front Bottoms song or more like a Manchester Orchestra song. In other words, which band had the biggest impact in the collaboration, or was the song roughly equally representative of traits of both bands??

Our response variable for this analysis is the binary variable Y = `artist`, where Manchester Orchestra is treated as $Y_i = 1$ and The Front Bottoms is treated as $Y_i = 0$. The client gives us information on $n = 138$ songs by these two artists. *Allentown* is not included in this list, as we will be using this song for prediction later.

In addition to artist, the client gives us information on the following explanatory variables:

- **danceability**: a numeric measure of how danceable a song is; higher danceability score means the song is easier to dance to.
- **energy**: a numeric measure of how energetic a song is; higher energy score means more energy.
- **key**: A, C, etc.; the letter represents the key the song is in. Letters that are further along in the alphabet are higher keys, so G is higher than C.
- **loudness**: A relative measure of how loud the song is. Note: All values are negative, so larger numbers (closer to 0) are louder songs.
- **mode**: Major or Minor
- **speechiness**: a numeric measure of how many words are spoken during the song; ranges from 0 (no words) to 100 (highest density of words)
- **acousticness**: a numeric measure of how acoustic a song is; ranges from 0 (low acoustic) to 1 (high acoustic)

- **instrumentalness**: a numeric measure of how much vocal content is present in a song; the more negative the number is, the more likely the song is to have vocal content. The closer to 0 the value is, the less likely the song is to have vocal content. Note: This is usually measured from 0 (very likely to have vocal content) to 100 (likely contains no vocal content). This value has been logged for this project because the scale of the values was very small.
- **liveness**: a numeric measure of whether you can hear the audience in the background of the song, cheering or singing, etc.; 0 (no audience) to 100 (audible audience)
- **valence**: a numeric value from 0 to 1 of how positive a song is; higher valence means more positivity.
- **tempo**: a numeric measure for how fast the pace of the song is; higher scores means faster songs.
- **duration_s**: how long the song is in seconds

What you will be submitting

You will be submitting two documents for this project.

Formal Report

- This is the write up that will explain your work.
- More details on the sections needed for this report are included below.
- In your formal report, there must be no code and no unformatted code output showing. This includes warnings and other stray code output – hide it all.
- You will be graded on spelling, grammar, formatting, and writing, as well as your stats. Make sure you use spell check.

Code Appendix:

- This is a Markdown file with annotated code.
- The goal is that a person who reads your report, and wants to replicate your results, could access your code appendix and completely reproduce the results and figures in your formal report.
- One suggestion. It is easiest to simply work in Markdown to make your formal report. Annotate the code along the way, but use echo = FALSE on line 9 of your Markdown file to hide all the code. When you are done, knit the file to save your report. Then, submit both the .Rmd and PDF / html versions of the document!

The Report

Your formal report MUST contain the following sections, which must be labeled and in this order in your final report.

This project will be written as a formal report. You have been provided with data and research goal, and you are writing a report explaining your process and results.

Because this is a formal report, each section in your project must be clearly labeled. Within each section, write in paragraph form with complete sentences, and label any tables or graphs, as well as the axes of the graphs. You will be graded on spelling, writing, and formatting, as well as your content. Use spell check!!

Formatting Note: No code or any unformatted code output of any kind may be visible in your report. There is a video on Canvas as well as a resource page to help with some formatting! Remember to change your code to `echo = FALSE` in the first chunk of your RMarkdown file to hide your code.

Section 1: Exploratory Data Analysis

In this section, your task is to explore the relationship between each possible explanatory variable and the response variable `artist`. For each explanatory variable, create an appropriate plot to visualize the relationship between X and the response variable (or log odds of being a Manchester Orchestra song, as appropriate). Each graph must be labeled and formatted appropriately.

Writing Tip 1: Make sure you have some sort of transition sentence that explains to your reader what this section is about to do. Something like "In this section, we will explore the relationships in the data.", for instance, would be appropriate. **You need such a transition sentence before every section in your paper**. This improves readability and flow of your work.

Directly after each plot, you will be describing what this plot tells us about the relationship of interest.

- Does it look like there is a relationship between the two quantities being plotted? Describe the relationship if so.
- Do we need to consider any transformations? If so, go ahead and do that and decide which transformations, if any, are needed.
- If there are any outliers, comment on those. You do not need to comment if there are not any visually evident outliers.

Writing Tip 2: Whenever you include a plot in a report, you must have words explaining what information the plot is giving you. We never want to have plot with no explanations. Make sure that any plots in the report:

- Have x and y axis labels that you create (don't use the default ones that R gives you)
- Have a title (If this is the first plot in Section 1, it should be titled Figure 1.1).
- Are referred to by name in the text ("As shown in Figure 1.1,. . . ")

Formatting Tip 1: You will have a lot of plots, so we need to either stack or shrink them. To learn how to do this, look here: https://bookdown.org/dalzelnm/bookdown-demo/graphing-formatting-graphs.htmlstacking-graphs

The function for empirical log odds plots you should use is:

```
get_logOddsPlot <- function(x, y, xname, yname, chosentitle, formulahere){

  if(class(y) =="factor"){
    y <- as.numeric(y)-1
  }

  if(class(y) =="character"){
    y <- as.numeric(as.factor(y))-1
  }

  sort = order(x)
  x = x[sort]
  y = y[sort]
  a = seq(from = 1, to = length(x), by=.05*length(y))
  a = round(a)
  b = c(a[-1] - 1, length(x))

  prob = xmean = rep(0, length(a))
  for (i in 1:length(a)){
    range = (a[i]):(b[i])
    prob[i] = mean(y[range])
    xmean[i] = mean(x[range])
  }

  prob[prob == 0] = .001
  prob[prob == 1] = .99

  g = log(prob/(1-prob))

  dataHere <- data.frame("x" = b, "LogOdds" = g)

  suppressMessages(library(ggplot2))

  ggplot(dataHere, aes(x =xmean, y = LogOdds)) +
    geom_point() +
    geom_smooth(formula = formulahere, method = "lm", se = FALSE) +
    labs(x = xname, y = yname, title = chosentitle)
}
```

Section 2: Mode or Key?

Now that you have finished EDA between the explanatory and Y, the client wants us to explore whether we should include **mode** or **key** in the model. The client knows that musically, these two things should have a relationship, a certain keys are generally more likely to be major or minor modes. If the two explanatory variables are strongly related to one another, we should only include one in our model.

Our client first asks you to create a plot to explore whether or not mode and key are related for these data. Based on your plot, is the client right that these variables are related?

If the variables have a relationship, the client only wants to include one of the two in the model. State which of the two you would recommend the client include (mode or key) and justify why that choice produces a model is a better fit to the data.

Formatting Tip 2: Remember that we can use `knitr::kable` to professionally format tables. If I build a model called model1, I can use `knitr::kable( summary(m1)$coefficients)` to obtain the table.

Formatting Tip 3: If your R output is a single number, just type the number in the sentence. However, if you need a table like you do for a hypothesis test, remember that we can use `knitr::kable` to professionally format those tables, too. There must be NO raw R output in your project.

Section 3: Energy, Acousticness, and Loudness

Our client believe that the rating of a song's energy is likely related to acousticness and loudness. In other words, the client worries that these three variables are measuring similar things and may be related.

Check to see if your client is correct. There are a variety of valid ways to do this, just choose one.

The client then asks us whether it is enough to just include energy in the model, or if we have evidence that suggests that including acousticness, loudness, or both along with energy provides a model that is a better fit to the data. Build appropriate models to explore this and respond to your client's question: which combination of these three variables would you recommend including in the model? Show appropriate (formatted) output or numeric measures and justify your choice.

Section 4: Building a Model

Based on your results from the previous sections, your client would like you to build a model for $Y = artist$. The client asks you to use all explanatory variables, **aside from any you have already decided not to use in Section 2 or Section 3.**

Build the model requested by the client. You do **not** have to write out the fitted model, as that would be very long. However, you do need to show a formatted table with the coefficients.

The client asks you for some measure of model fit. Since we are not comparing models, the tool we generally use for that is the percent drop in deviance. Compute and interpret the percent drop in deviance.

Based on your fitted model, describe to your client what traits appear to be related to a song being from Manchester Orchestra rather than The Front Bottoms. Additionally, describe any traits that seem to be the same across the two bands.

Section 5: Making Predictions

In addition to understanding what trait might distinguish the two groups, recall that the client is also interested in predicting whether a particular song is more similar to a Manchester Orchestra song than a The Front Bottoms song. Because of this, the client would like an overall picture of how well our model can predict the artist of a song.

Using a threshold of .6, create a confusion matrix (professional formatted using `knitr::kable`) to explore the predictive abilities of your model. Once you have created the confusion matrix, the client would like you to use **at least three different numeric summaries** to describe the predictive abilities of the model.

Section 6: Allentown

The client would now like you use your model to estimate whether or not the new song, *Allentown*, is more like a Front Bottoms song or more like a Manchester Orchestra song. They provide you with the traits of the song:

- danceability: 0.404
- energy: 0.272
- key: C

- loudness: -10.575
- mode: Major
- speechiness: 0.0445
- acousticness: 0.615
- instrumentalness: -7.005368
- liveness: 0.13
- valence: 0.119
- tempo: 126.18
- duration (in seconds): 193.573

Based on these explanatory variable values, the client would like you to use your model from Section 4 to obtain a predicted probability that the song was written mostly by Manchester Orchestra (meaning find $\hat{\pi}_i$ for this song). State the predicted probability and explain whether this suggests the song is more like a Manchester Orchestra song or more like a The Front Bottoms song, or if the song seems to have roughly equally influence from each group.

Based on your results in Section 4 about what traits are related to a Manchester Orchestra song versus a The Front Bottoms song, does this predicted probability and what it suggests about the group that has the main influence in the song make sense? Explain your reasoning.

This section will serve as the conclusion to your project!

And you are done!! Congratulations!!