Battle of the Bands: Trying to Identify Contributions to a Collaboration

CAUSE Webinar

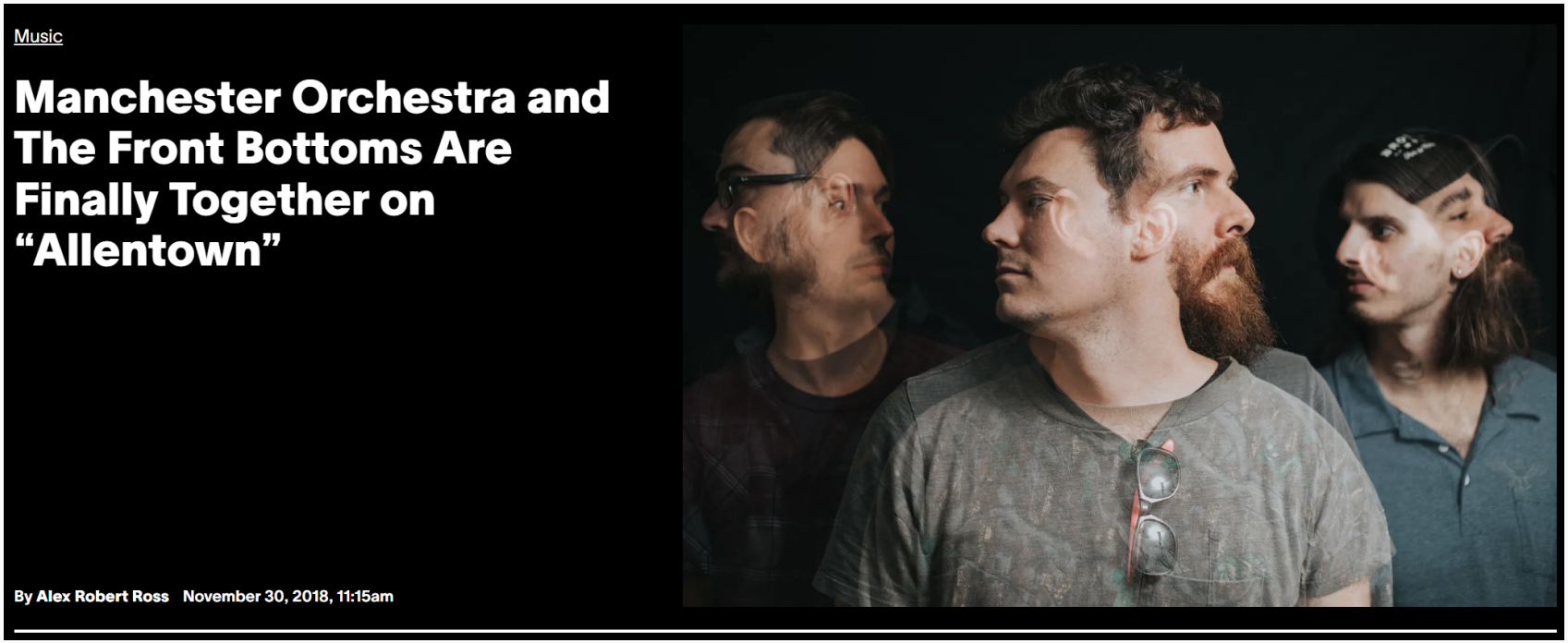
William Cipolli, Nicole Dalzell, Roy Bower, Ciaran Evans

2026-05-26

Why This Data Set Matters

- Hard to find real world data sets for teaching at a variety of levels
- Students rarely see data that must be created (not downloaded)
- Realistic messiness: missing values, duplicates, inconsistent metadata
- Single data set that supports intro $\rightarrow$ advanced courses
- Allows students to track progress
- Multiple models: logistic, multinomial, LASSO, trees

Our Application



Source: Ross (2018)

Motivating Questions

- Three bands collaborated to produce a single track, "Allentown"
- Can we use statistics and data science to determine:
  - which band the **sound** most resembles?
  - which band the **lyrics** most resemble?

The Problem (And The Opportunity!)

Problem

- We need a data set with:
  - information about the sound qualities and lyrics of the combined track
  - information about the sound qualities and lyrics for all tracks
- No such data set exists

Opportunity

- We create activities to show students at a variety of levels how to:
  - create the needed data
  - clean and process it
  - use modeling to answer the research questions

Learning Goals For Students

- Identify ethical constraints in data collection
- Create features from raw media (sound, lyrics from text)
- Clean/merge
- Fit/interpret logistic and multinomial models
- Communicate results to a general audience

Where We Teach

Will

- Colgate University: small liberal arts college
- Courses used: undergraduate Data Analysis
- Students per class: 30 or less

Roy

- Furman University: small liberal arts college
- Courses used: undergraduate Statistical Methods
- Students per class: 24 or less

Nicole and Ciaran

- Wake Forest University : medium-sized liberal arts university
- Courses used: graduate GLMs, undergraduate Text Analysis
- Students per class: 30 or less

What We Have: Tracks

- **The Front Bottoms Releases (61 tracks):** The Front Bottoms; Talon of the Hawk; Rose; Back on Top; Needy When I'm Needy; Going Grey; Ann; and End of Summer (Now I Know)
- **Manchester Orchestra Releases (77 tracks):** You Brainstorm, I Brainstorm, but Brilliance Needs a Good Editor; I'm Like a Virgin Losing a Child; Fourteen Years of Excellence; Mean Everything to Nothing; Simple Math; Cope (Deluxe Version); Hope; A Black Mile to the Surface
- **All Get Out Releases (42 tracks):** All Get Out; The Season; Movement; Nobody Likes a Quitter; No Bouquet

What We Have: Lyrics

Teaching Opportunity 1: Data Creation

- While we have all these tracks, we do not have information about the sound or the lyrics in a way that we can use for modeling
- How do we get it?
- Students may think about AI
- However, there are legal and ethical issues with using AI to create features from copyrighted music
- Example: Genius for lyrics (ethical issues) and Spotify API (no longer allowed due to AI)
- Storytime: What literally happened as we were writing the paper...

Ethical Pitfalls (Student Discussion Prompt)

Ask students:

- why can't we scrape lyrics from Genius?
- why did Spotify shut down API access for ML?
- what counts as "fair use" in data science?
- what are the risks of using AI-generated features?

Our Solution For Sound Features: Essentia

Teaching Angle: How can raw audio be converted into structured data?

- We use Essentia (Bogdanov et al. 2013; Alonso-Jiménez et al. 2020) to extract meaningful information from the spectrograms of each track
  - **Music Extractor:** loudness, silence rates, beats per minute, danceability, key, mode, etc. (67 variables)
  - **Models:** approachability, engagement, arousal, valence, happy, sad, etc. as predicted using models built from existing databases. (14 variables)

Our Solution For Lyric Features: LIWC

Teaching Angle: Text analysis foundations without legal issues

Using Linguistic Inquiry and Word Count (LIWC) (Boyd et al. 2022) and Bing Lexicon:

- psychometric dictionaries about a writer's:
  - psychology
  - social concerns
  - writing style
- 120 variables
  - e.g., positive/negative words, psychometric categories, etc.
- legal to share resulting data

The Result

- A large data set with many variables
  - 180 rows
  - 204 columns (201 features)
- A great space for student discussion about AI, data creation, and data privacy

Teaching Opportunity 2: Cleaning

Teaching Angle: Cleaning data can be challenging

Students must:

- handle missing values
- merge data
- handle duplicate tracks (same lyrics, different sounds!)
- transform features

Teaching Opportunity 2: Cleaning

Educators can:

- choose how clean or messy to make the data
- curate the data set to reflect students' experience

This makes the data set a good option for various learning goals

Teaching Opportunity 3: Modeling

Teaching Angle: Same data can be used to answer multiple research questions in different ways.

0. Empirical Evaluation
1. Basic analyses
    - where do the bands differ?
2. Logistic Regression
    - only two of the three bands have performing credits
    - predict performing artist (binary)
3. Multinomial Regression
    - all three bands have writing credit
    - predict writing artist (three classes)

What We Created: Activities

Activities for:

- Data Collection and Cleaning
- Exploratory data analysis
- Logistic regression
- Multinomial regression
- Model diagnostics (e.g., quantile residuals)
- Cross validation to assess predictive ability
- Feature selection (e.g., LASSO)

What We Created: Shiny Apps

Apps: https://shiny.colgate.edu/apps.html

- **Exploratory Data Analysis (EDA)**
- **t-test** or **ANOVA** (For comparing bands)
  - non-parametric alternatives (Mood's Median, Kruskal Wallis)
- **Logistic Regression**
- **Multinomial Regression**

Example: Logistic Regression (Sound)

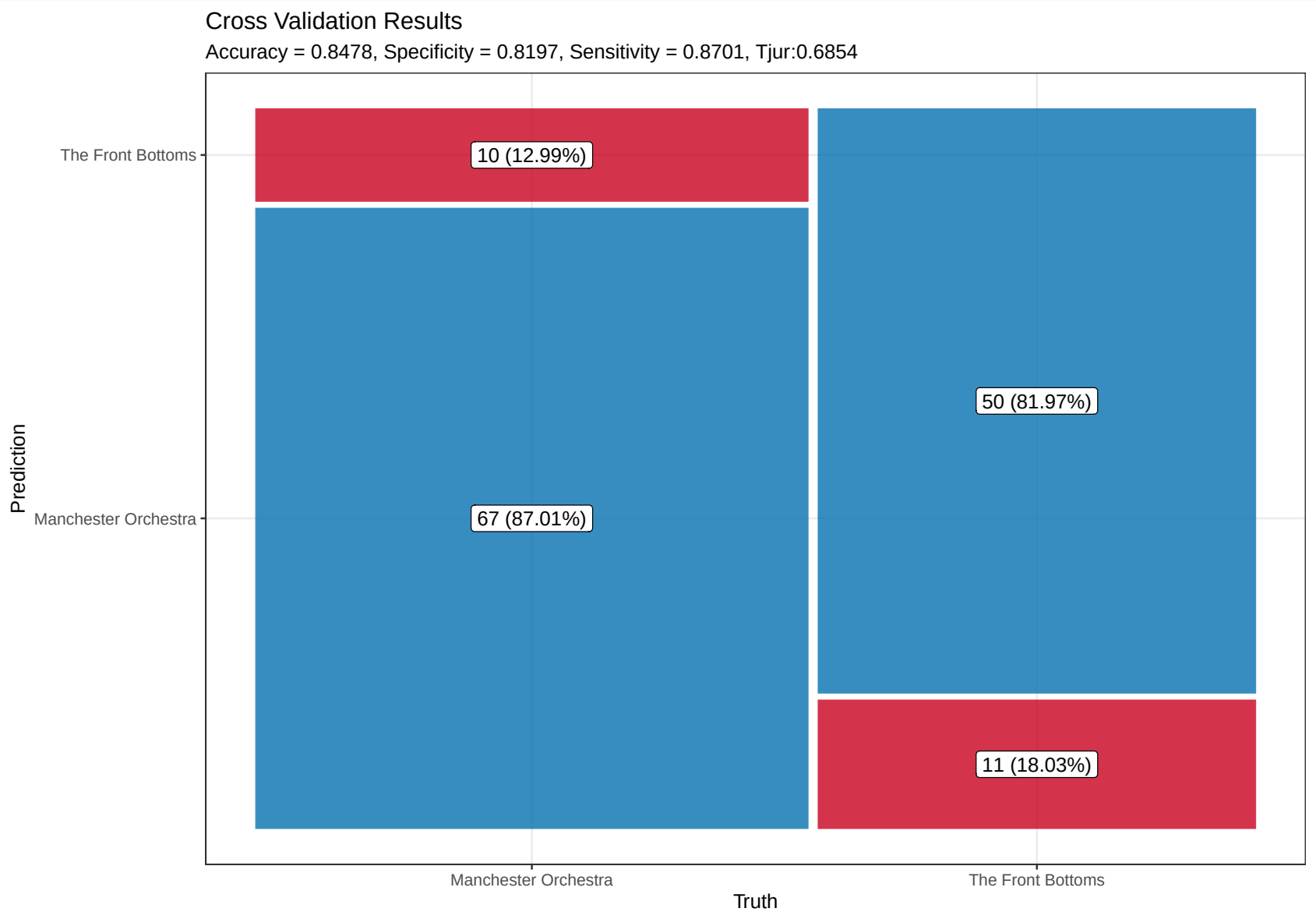
Students explore:

- which sound features separate TFB v. MO
- how to interpret logistic regression diagnostics
- how to interpret coefficients, p-values, etc.
- how to evaluate model fit
- etc.

Results (Sound Model)

	Estimate	SE	Z	p-value
(Intercept)	1.10	0.45	2.44	.01
overall_loudness	-1.83	1.26	-1.45	0.15
spectral_energy	-1.27	0.57	-2.24	.02
pitch_salience	0.35	0.50	0.69	0.49
tempo	0.37	0.40	0.92	0.36
dissonance	-2.48	1.21	-2.05	.04
danceability	0.43	0.44	0.98	0.33
tuning_frequency	1.06	0.43	2.45	.01
duration	0.79	0.40	1.96	.05
valence	-0.57	1.18	-0.49	0.63
arousal	0.28	0.78	0.35	0.72
acoustic	-1.73	0.72	-2.40	.02
timbreBright	1.56	0.48	3.23	$<.0001$
approachability	-1.43	0.70	-2.06	.04
engagement	-1.07	0.54	-1.98	.05

Results (Sound Model)



Example: Logistic Regression (Sound)

- The conditions for using the logistic model are met and the model fits well

What about Allentown?

1. Predict using sound features for Allentown
2. The model predicts that Allentown is an MO track

$$\hat{P}(Y = \text{Manchester Orchestra}) = 0.9344$$

Did The Model Get It Right?

Teaching Angle: Students realize the results make sense

Interview with the artists revealed:

- Andy Hull of Manchester Orchestra worked out the melody/music
- Brian Sella of The Front Bottoms then helped develop the chorus

The model's prediction of Manchester Orchestra matches the interview

What About The Lyrics?

If we want to model the lyrics, we now have all three bands in play, so Y is multinomial

- Using our features from LIWC, can we predict who the lyrics most "read like"?
- For this, we use multinomial lasso

Example: Multinomial Regression (Lyrics)

Students explore:

- which lyric features separate the **three** bands
- how to interpret multinomial diagnostics
- how LASSO reduces features
- how to interpret multinomial coefficients
- how to compare predicted probabilities
- how to evaluate model fit
- etc.

Results (Lyric Model): 120 Features Reduced To 29

term	FB	MO	AGO	term	FB	MO	AGO
WC	.0009	-.0013	.004	relig	-0.1916	0.1087	.0829
Tone	.005	-.057	.0007	want	-.0081	.0044	.0037
fnct	.0006	.007	-.0076	curiosity	-.0493	.0816	-.0324
you	-.0085	.0009	.0076	allure	-.0167	-.0224	.0391
shehe	.0062	.0033	-.0095	visual	-.0686	0.1385	-.0699
conj	.0703	.0307	-.1011	auditory	-.0006	.0273	-.0267
negate	-.0040	.0025	.0015	feeling	.1063	-.0995	-.0068
quantity	-.0239	-.0561	.0800	focuspast	-.0582	.0795	-.0213
achieve	-.0160	.0150	.0010	netspeak	.0985	-.0537	-.0448
certitude	-.0386	.0342	.0045	Comma	.0151	-.0038	-.0113
emotion	.0143	-.0162	.0019	Apostro	-.0041	.0001	.0040
Culture	.3465	-.1260	-.2205	OtherP	.0512	-.0269	-.0243
politic	-.4229	-.0675	.4904	positivewords	.0031	-.0043	.0012
tech	.0529	-.0184	-.0345	negativewords	.0055	-.0263	.0208
leisure	.0922	-.0287	-.0635				

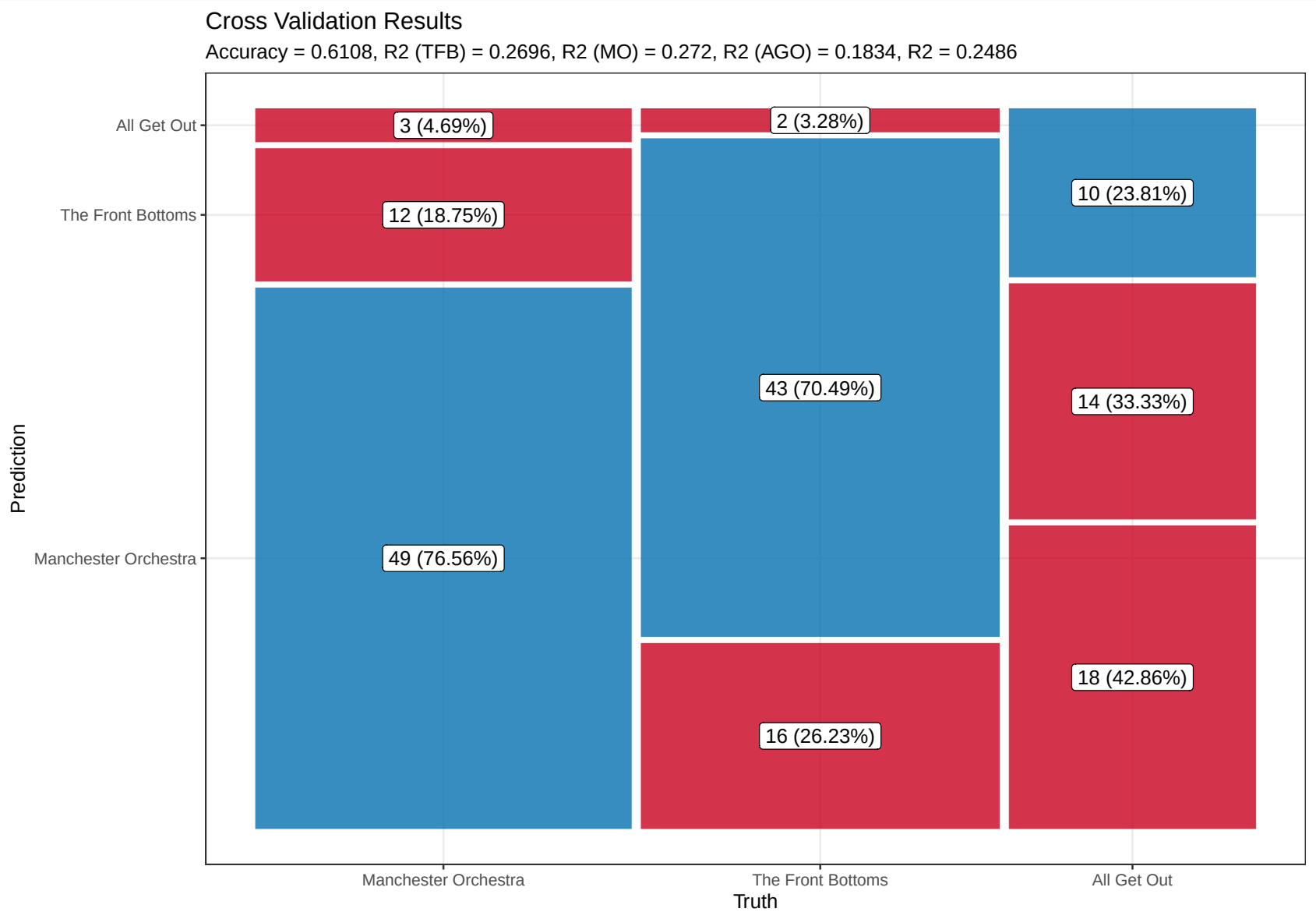
Teaching Opportunity: Talking To A Client

Teaching Angle: Students explain and communicate modeling results

Students practice explaining:

- MO has the most negative lyrics (tone) and least emotion
- AGO uses the most political language
- TFB uses the most tech-related words
- etc.

Results (Lyric Model)



Example: Multinomial Regression (Lyrics)

- The conditions for using the multinomial model are met and the model fits well

What about Allentown?

1. Predict using lyric features for Allentown
2. The model predicts Allentown is an AGO track (but is less sure than with the sounds analysis)

$$\hat{P}(Y = \text{The Front Bottoms}) = 0.1709$$
$$\hat{P}(Y = \text{Manchester Orchestra}) = 0.2760$$
$$\hat{P}(Y = \text{All Get Out}) = 0.5531$$

Did The Models Get It Right?

Teaching Angle: Students realize the results make sense

Based on interviews, we know:

- AGO wrote the first four lines of Allentown
- MO developed the melody/music
- TFB helped to develop the chorus

Our models:

- Sounds $\longrightarrow$ MO
- Lyrics $\longrightarrow$ AGO
- Both match the real-world story

Summary

This data set teaches:

- ethical data collection
- feature engineering
- cleaning/merging
- modeling/interpretation
- communication
- real-world uncertainty

And it's fun!

Thank You For Listening!

Questions?

References i

Alonso-Jiménez, Pablo, Dmitry Bogdanov, Jordi Pons, and Xavier Serra. 2020. "Tensorflow Audio Models in Essentia." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 266–70. IEEE.

Bogdanov, Dmitry, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Herrera Boyer, Oscar Mayor, Gerard Roma Trepat, et al. 2013. "Essentia: An Audio Analysis Library for Music Information Retrieval." In *Britto a, Gouyon f, Dixon s, Editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil. ISMIR; 2013. P. 493-8.* International Society for Music Information Retrieval (ISMIR).

References ii

Boyd, Ryan L, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. "The Development and Psychometric Properties of LIWC-22." *Austin, TX: University of Texas at Austin*. https://www.liwc.app/.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ross, Alex Robert. 2018. "Manchester Orchestra and the Front Bottoms Are Finally Together on 'Allentown'."

References iii

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.