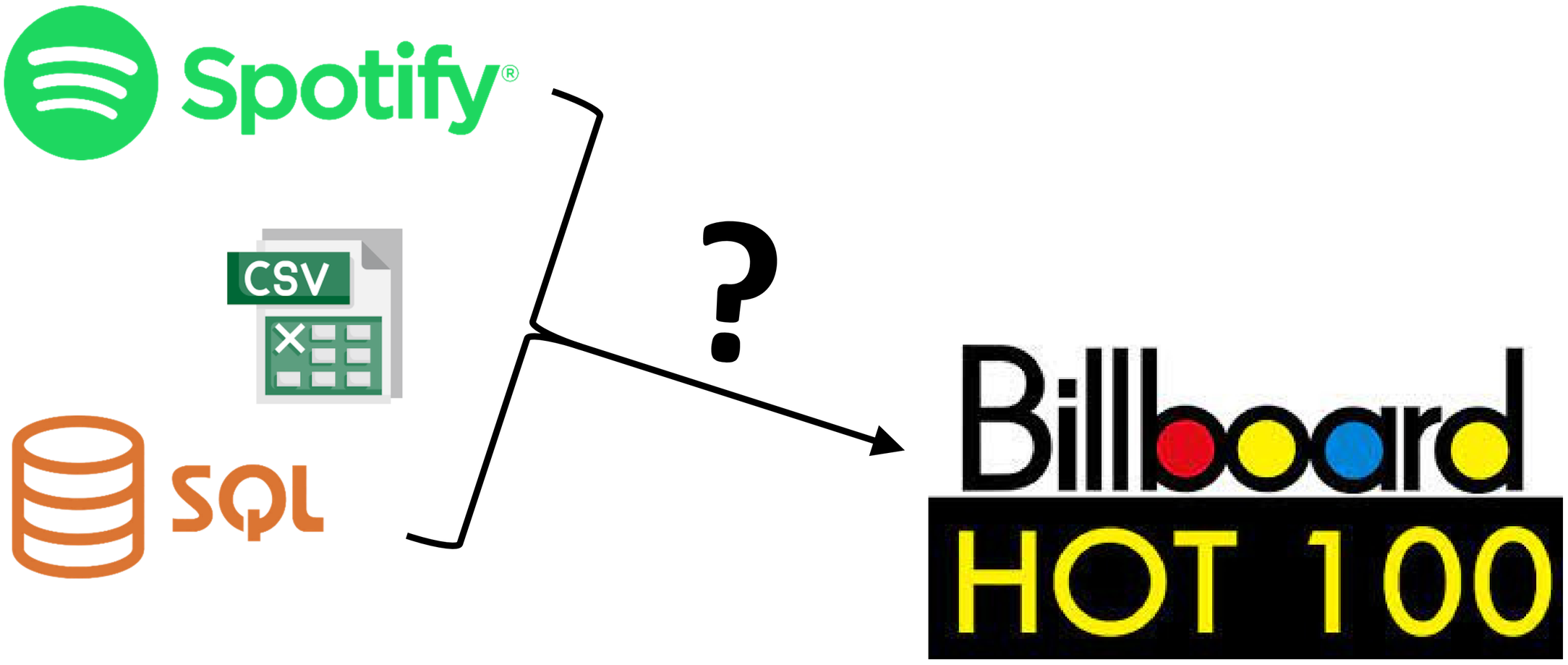
Comparing Song Audio Features to Rankings on The Billboard Hot 100

Kevin Carr 501150122
Supervisor: Ceni Babaoglu, PhD
Date: December 8, 2022

Ryerson University

The Central Problem

- Can we predict commercial success using audio feature data?



The Project Approach

1. Compile and clean the data
2. Exploratory data analysis
3. Remove outliers
4. Cluster into groups
5. Create classification models
6. Evaluate performance

The Data

Audio Feature	Type
acousticness	number <float>
danceability	number <float>
duration_ms	integer
energy	number <float>
instrumentalness	number <float>
key	integer
liveness	number <float>
loudness	number <float>
mode	integer
speechiness	number <float>
tempo	number <float>
time_signature	integer
valence	number <float>

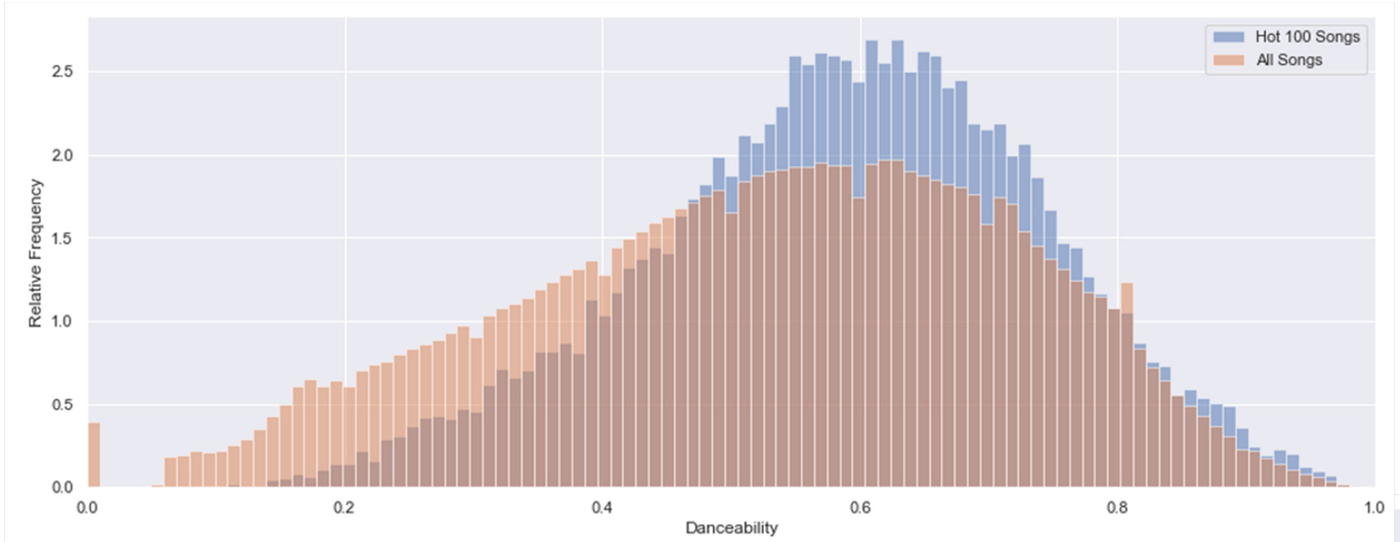


Figure 2. Danceability Histogram

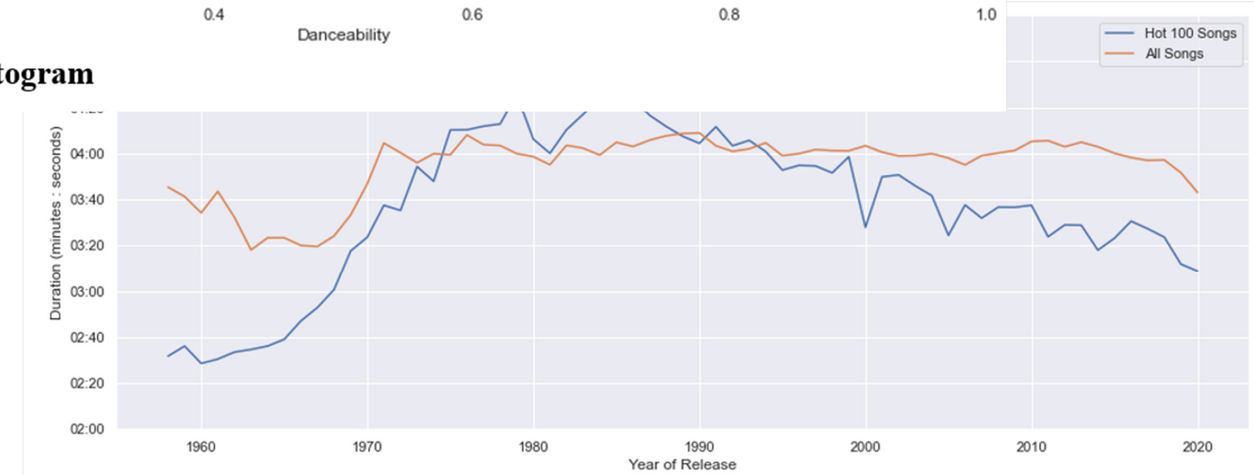


Figure 16. Duration History Comparing the Billboard Hot 100 with All Songs

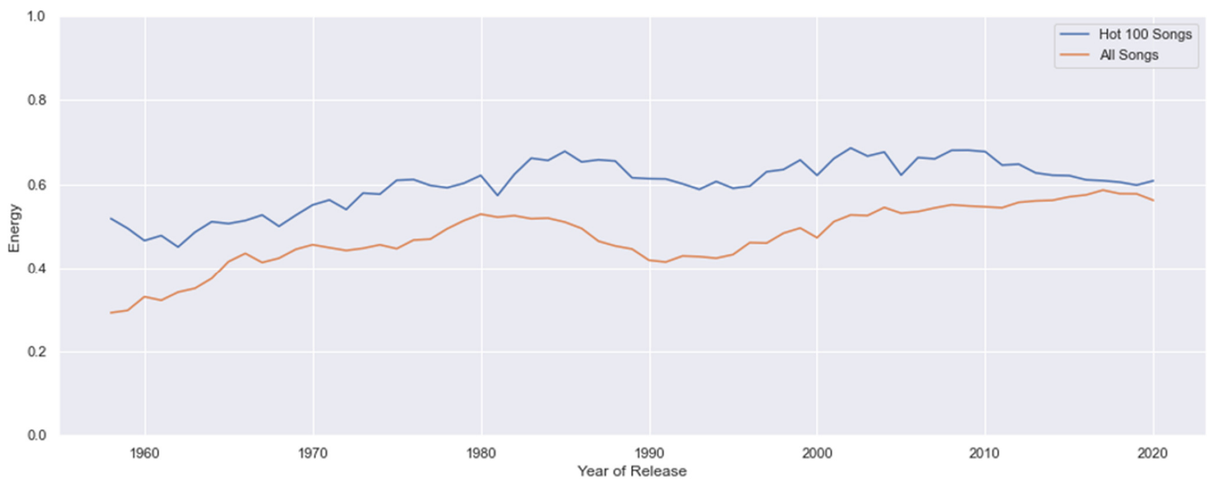


Figure 17. Energy History Comparing the Billboard Hot 100 with All Songs

The Data


speechiness -0.03 0.19 -0.05 0.06 -0.16 0.02 0.18 -0.02 -0.04 1.00 -0.01 -0.01 0.08 -0.01
tempo -0.22 0.07 0.02 0.26 -0.05 0.01 0.01 0.24 0.00 -0.01 1.00 0.12 0.16 0.00
time_signature -0.17 0.22 0.05 0.19 -0.05 0.01 0.00 0.18 -0.02 -0.01 0.12 1.00 0.14 0.01
valence -0.19 0.56 -0.16 0.35 -0.29 0.03 0.02 0.36 0.03 0.08 0.16 0.14 1.00 0.02
in_B100 -0.01 0.02 -0.01 0.01 -0.03 -0.00 -0.00 0.02 0.01 -0.01 0.00 0.01 0.02 1.00
acousticness
danceability
duration_ms
energy
instrumentalness
key
liveness
loudness
mode
speechiness
tempo
time_signature
valence
in_B100
−0.50
−0.75
−1.00

Clustering Approach

- Removed outliers
  - Non-musical tracks (e.g., “ringtone”)
  - Non-song music (e.g., “TV theme”)
- Clustered into 3 groups
  1. K-Means clusters, optimised by silhouette score
  2. K-Means clusters, equal to number of genres
  3. Manual clusters, grouped into genres

Classification Approach

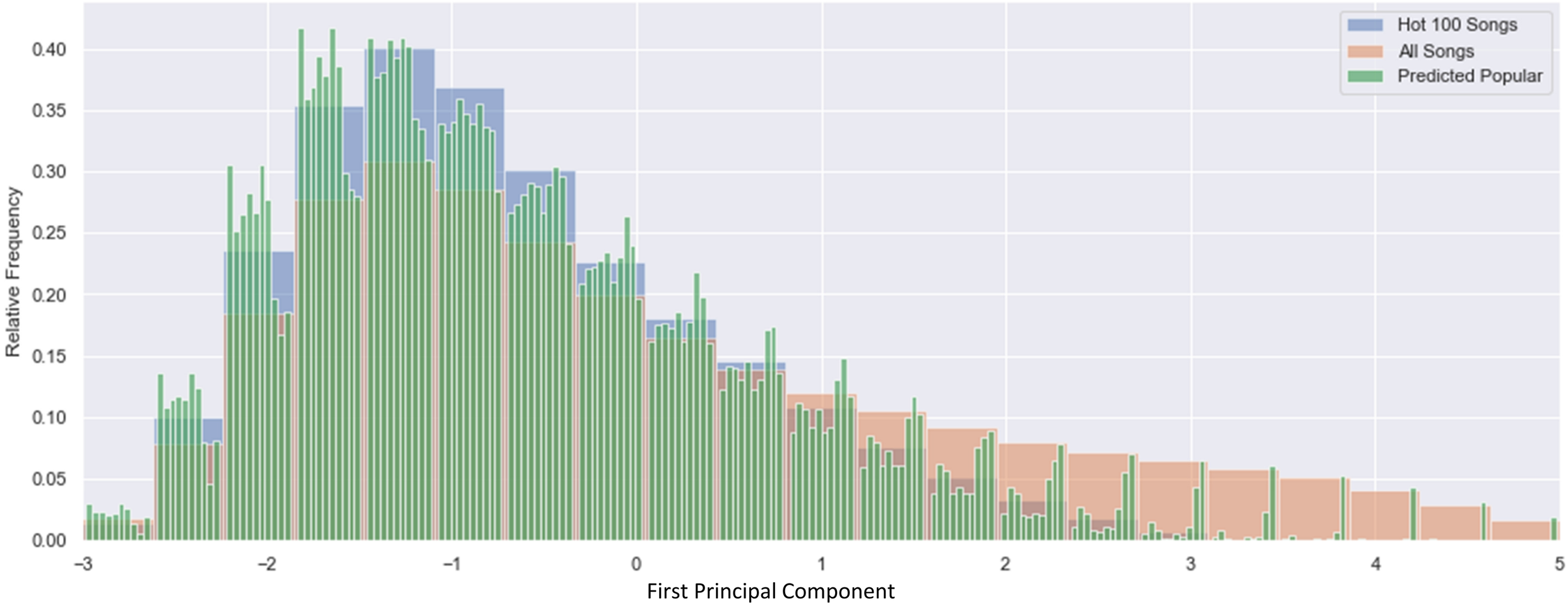
- Cross-validation
  1. Split data into 5 stratified folds
     - Consistent random seed for consistency between tests
  2. Train the model using 4 of the folds
     - Balanced using undersampling (using consistent random seed)
  3. Predict the class
     - Predict using the unbalanced test fold
  4. Repeat for each fold
     - Save all out-of-fold predictions for model evaluation

Classification Models / Scenarios

1. Logistic Regression – Default Hyperparameters
2. Decision Tree – Default Hyperparameters
3. K-Nearest Neighbours – Default Hyperparameters
4. Random Forest – Default Hyperparameters
5. AdaBoost – Default Hyperparameters
6. Logistic Regression – Tuned Hyperparameters
7. Decision Tree – Tuned Hyperparameters
8. Logistic Regression – Clustered By K-Means Version 1
9. Logistic Regression – Clustered By K-Means Version 2
10. Logistic Regression – Clustered By Genre

Findings

- Most models made predictions consistent with observed data

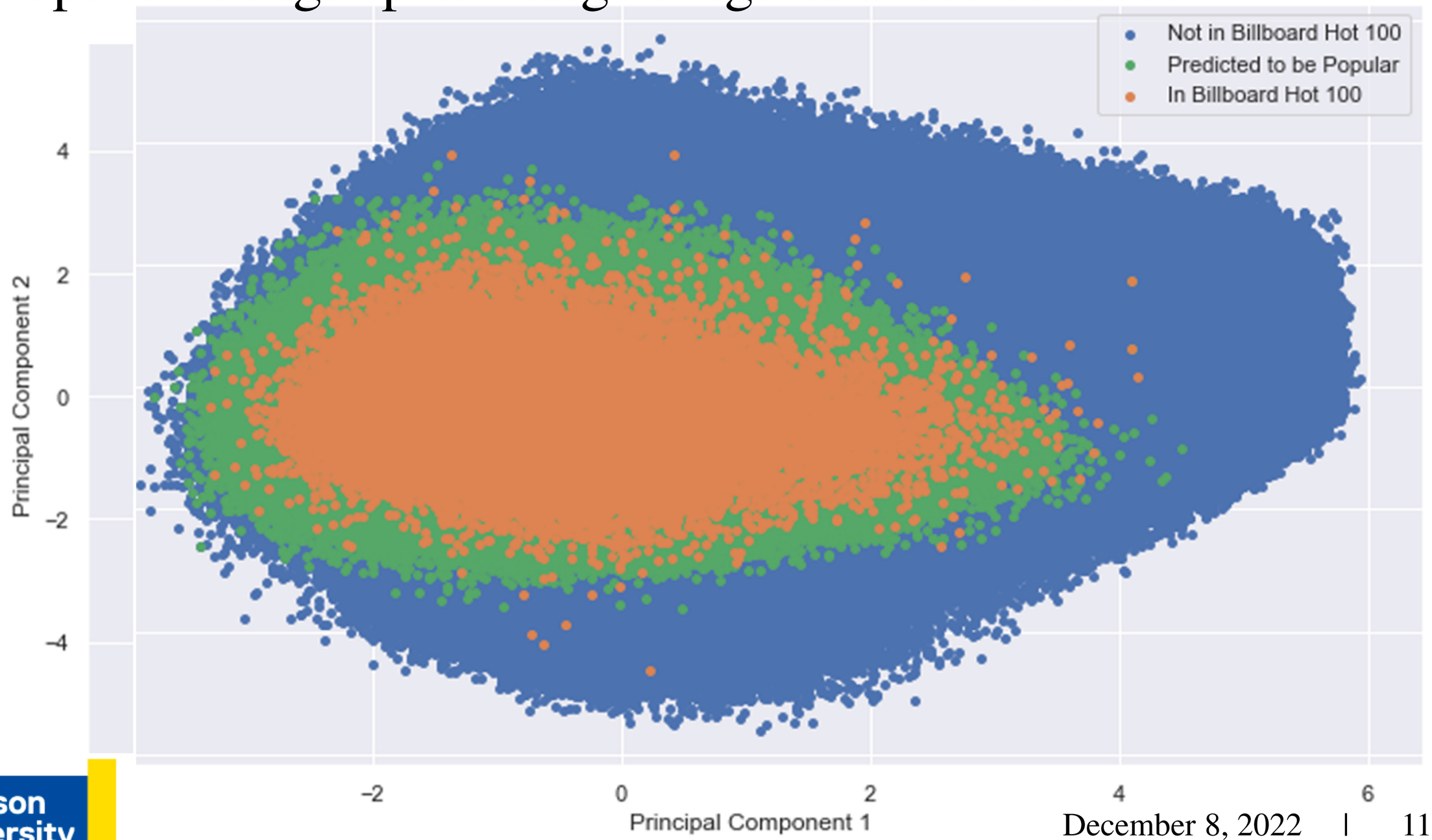


Findings

- Low precision was unavoidable
  - “Popularity” is highly unbalanced by definition
  - “Popular” songs span a large range of audio features

Findings

- “Popular” songs span a large range of audio features



Findings

- Clustering did not improve performance

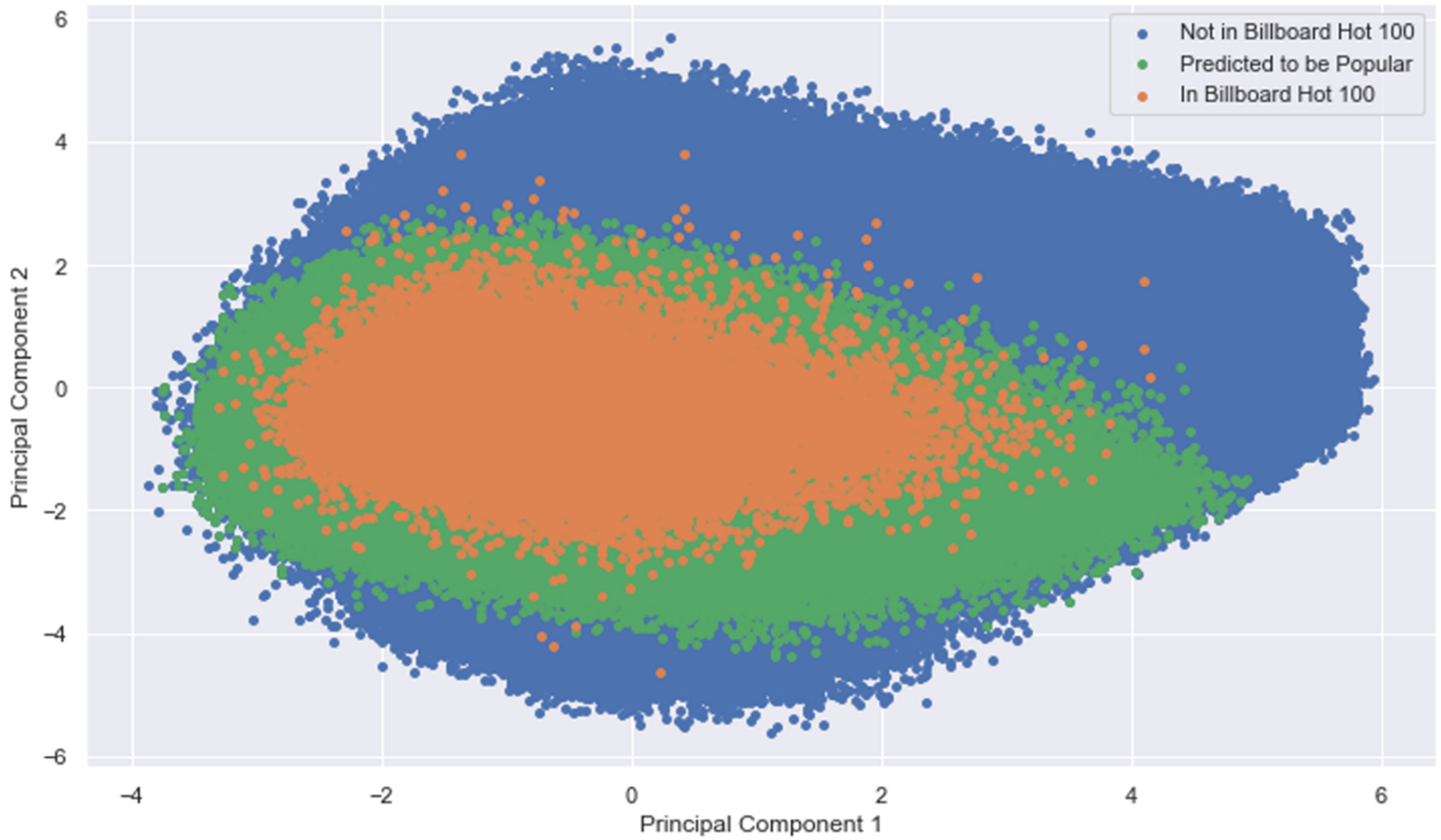
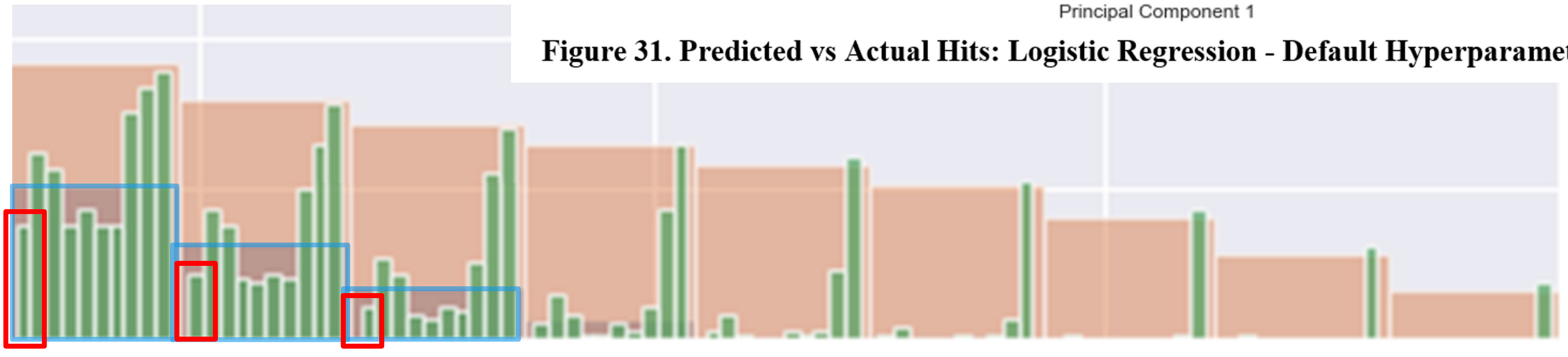


Figure 31. Predicted vs Actual Hits: Logistic Regression - Default Hyperparameters



Findings

- Clustering did not improve performance

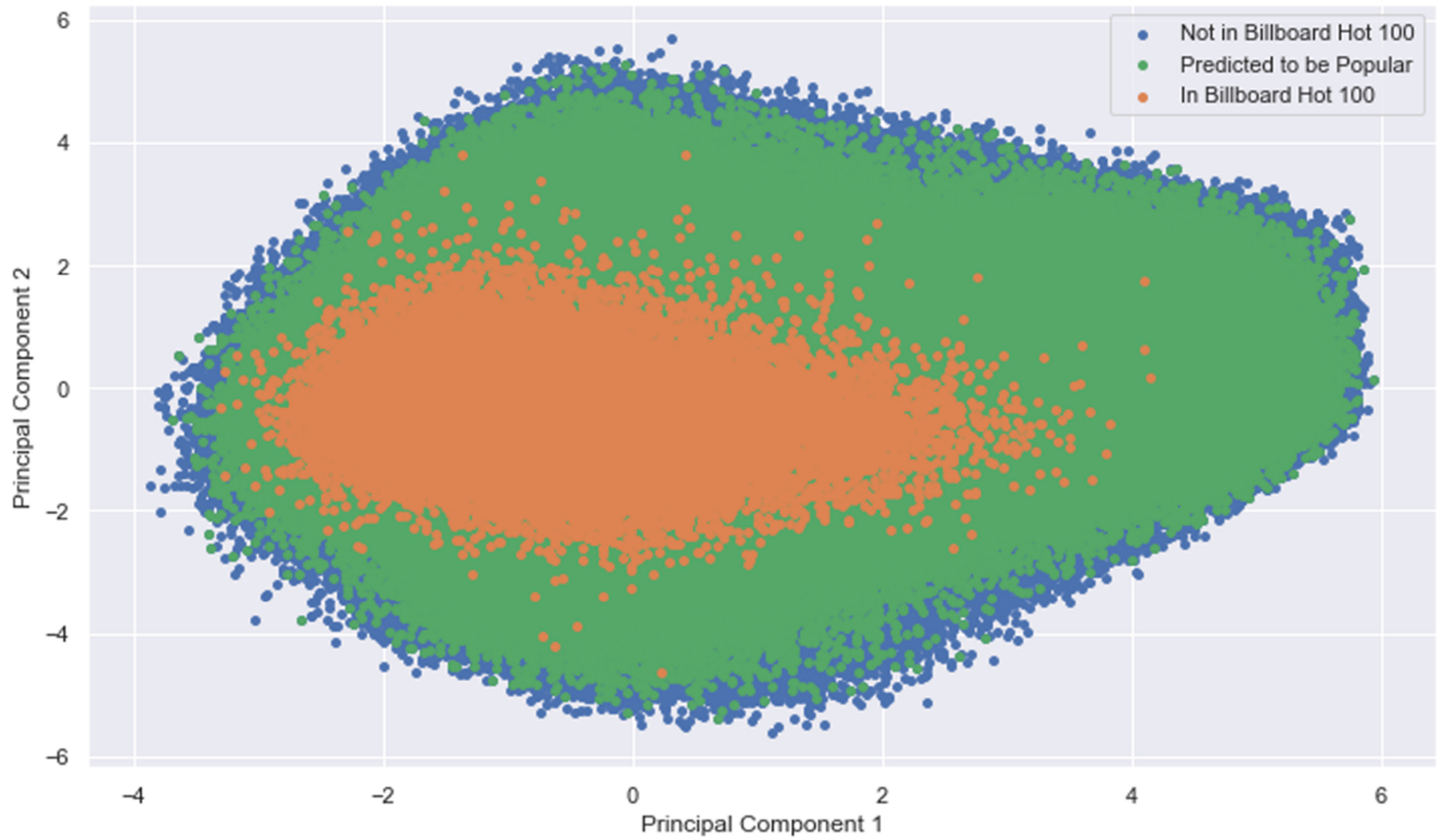
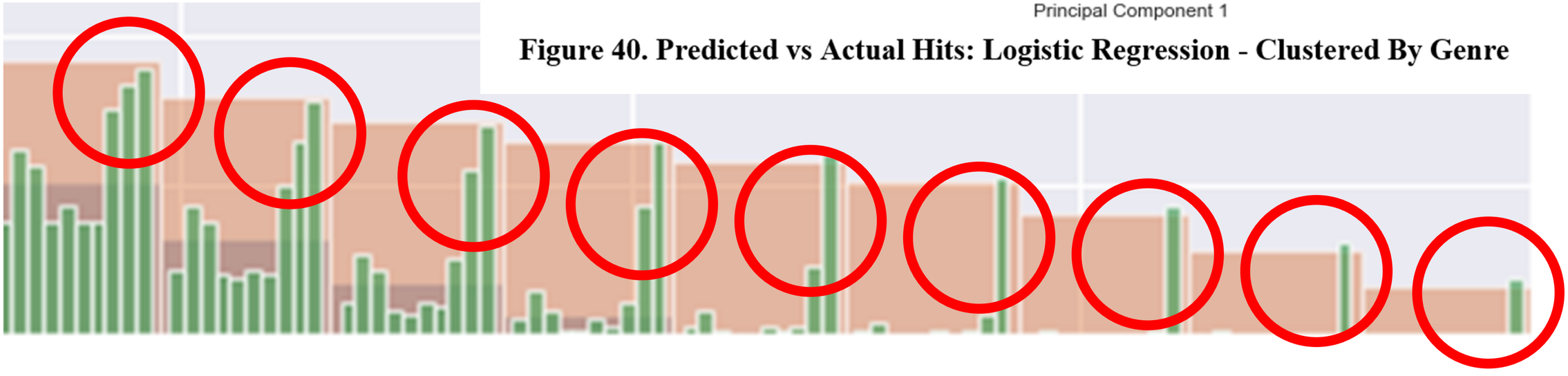


Figure 40. Predicted vs Actual Hits: Logistic Regression - Clustered By Genre



Discussion and Conclusions

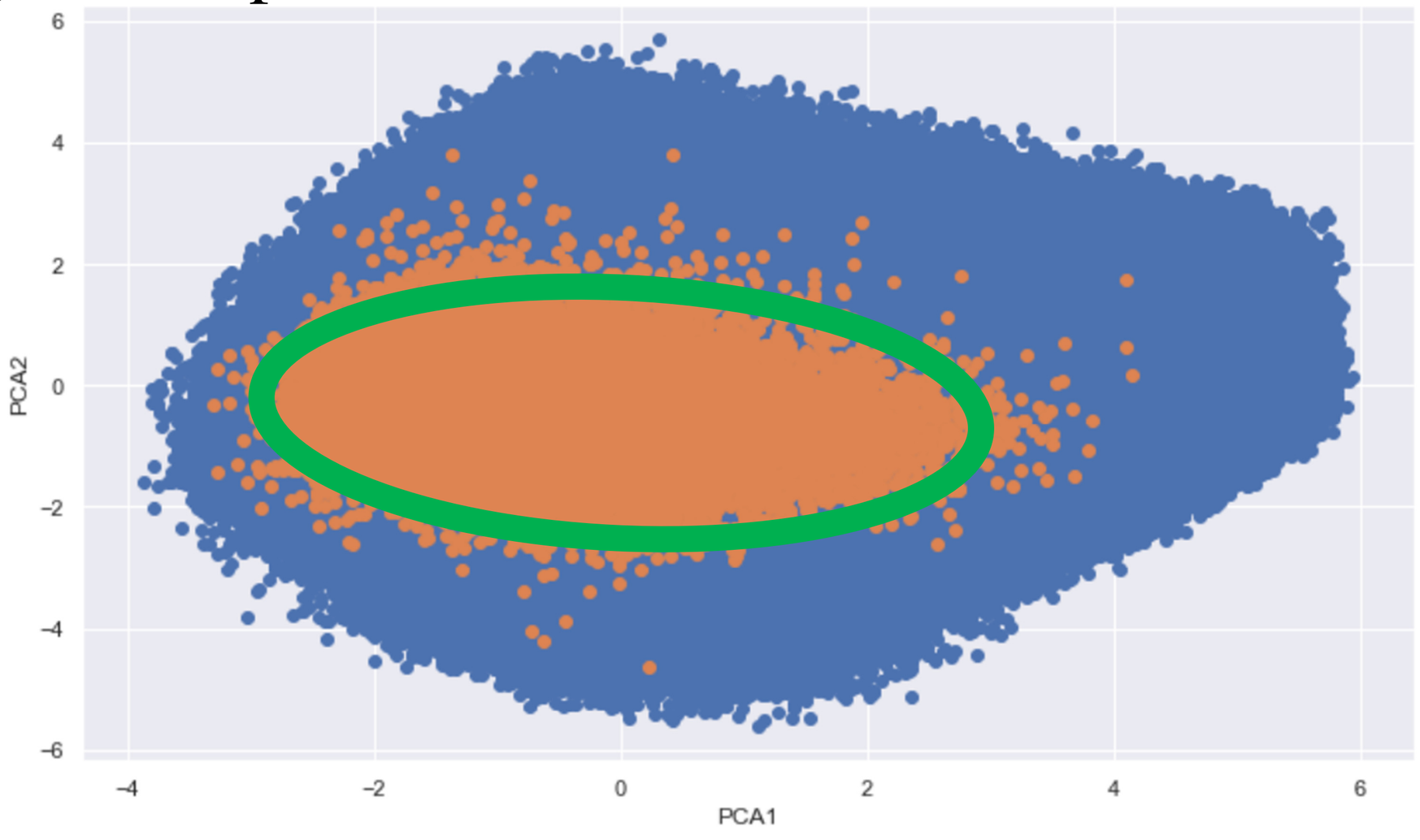
- The Random Forest performed best
- Most models had similar scores
- Undersampling was required to achieve non-trivial results
- Tuning high-precision models likely isn’t possible

Future Work

- An optimized Random Forest could perform well
  - This would unfortunately take a long time to tune
- A plugin for music software could be useful for musicians
  - Limitations include calculating audio features from raw audio
- PCA could be useful for improving accuracy and/or speed

Future Work

- An analytical PCA model could be used to predict whether a song has the potential for commercial success



Thank you!

Ryerson University

Any questions?