Finding the Hit Single Formula

Predicting Success on the Billboard Hot 100 Charts

Jelica Bornath, Dec-2021

Galaxy Music Publishing Group

Problem Definition

- ## What we are trying to solve?
  - Focusing on a song's characteristics, and looking beyond external factors like marketing and timing, how can Galaxy Publishing Group continue to build their song catalogue with attributes that resonate with listeners?

- ## What does success look like?
  - Predicting a song's chart performance based on the song attributes and historically successful songs from the Billboard Hot 100

- ## What is the scope?
  - Song attributes from singles appearing on the Billboard Hot 100 chart across the last 10 years and each song's peak position on the chart

Key Outcomes

- The focus on the analysis and modelling was to predict if a song had the right attribute set to peak in the Top 10 of the Billboard Hot 100 Chart

- The Outcome - song attributes that make a Top 10 hit aren't differentiated enough from other songs in the chart to reliably predict the likelihood a song will peak in the Top 10.

- Recommendation – incorporate external factors like seasonality, marketing strategies, and artist reputation into the analysis and modelling for future exploration

Data Collection

- Billboard Hot 100 Chart data (2011-2021)
  - 4609 Unique Songs
- Spotify Web API song attribute data
  - 14 song attributes
- Target Feature – Peak Position
  - 548 songs peaked in the Top 10, these songs were evaluated against the remaining 4061 unique songs


Attributes

| **Genre**	**Key**	**Duration**
Song Genre (based on artist genre in Spotify API)	Key signature for the track (ie C, C# D etc.)	Length of the track
Time Signature	**Mode**	**Tempo**
Estimation of beats per measure	Modality of the track (ie major or minor)	Pace of a song in beats per minute (BPM)
Loudness	**Danceability***	**Speechiness***
Measure of loudness in decibels (dB)	Suitability for dancing, measured using tempo, rhythm etc.	Amount of spoken word in a song
Acousticness*	**Instrumentalness***	**Liveness***
Confidence level of whether a song is acoustic or not	Measure of a song instrumental components to vocal components	Detects presence of an audience
Valence*	**Energy***	*graded by Spotify Web API on scale from 0-1*
Measure of musical positiveness (ie happy song vs. sad song)	Perceptual measure of intensity and activity	

Exploratory Data Analysis

- When reviewed individually, song attributes vary across songs, but when peak position values are grouped into 20 sub-groups and averaged, variability decreases

- This was an early indication that song attributes alone may pose a challenge for predicting a song's success on the chart as there weren't strong correlations between the Peak Positions and attributes

Example of variability when songs are *analyzed individually*
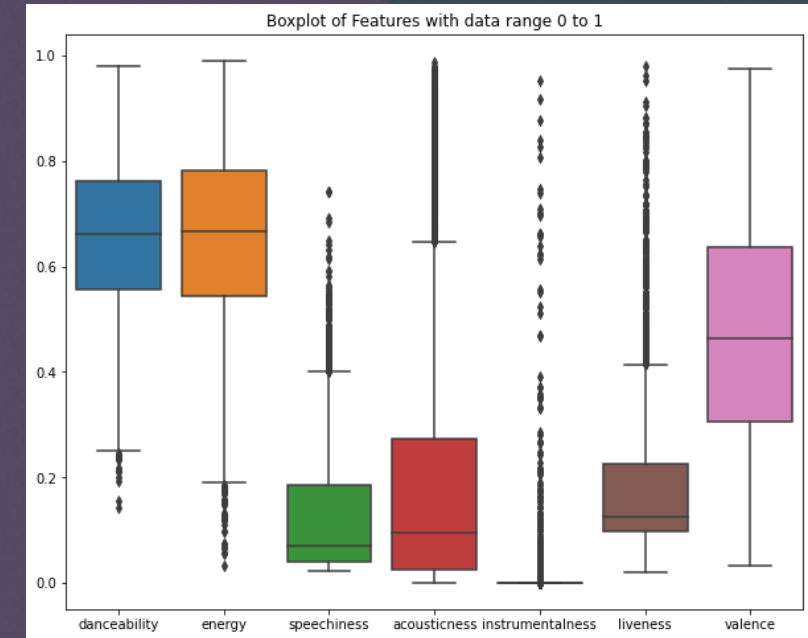
Example of variability when songs are *analyzed in Peak Position groups of 5*
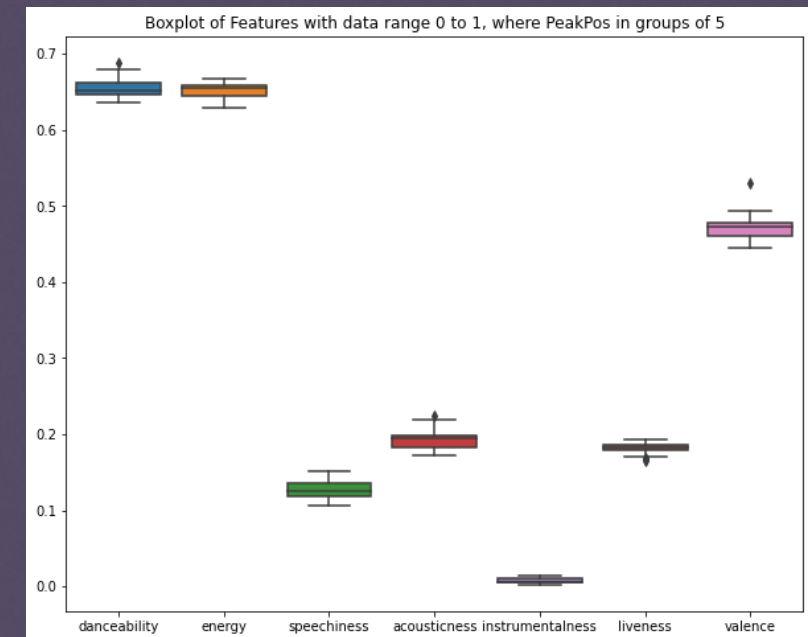
Approaching the Regression Problem

- First modelling approach was as a regression problem – could we predict the specific peak position of a song based on its attributes?

- Using Linear Regression and a Random Forest model, the reliability of these results between the training and test sets was very low

- Initial modeling indicated it might be possible to predict the peak position within +/- 25 positions, too unreliable given the max range of 100 positions on the chart
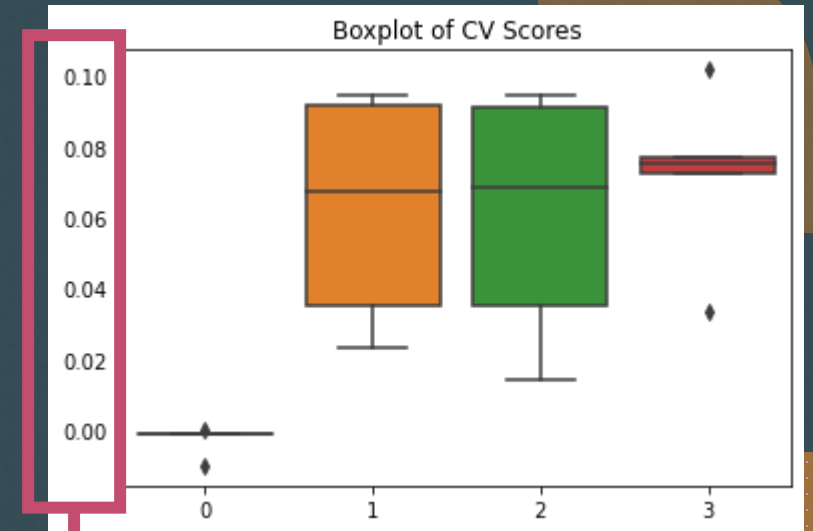
- *Very low cross-validation scores for the training and test sets across 4 regression models*

Approaching the Classification Problem

- Second modelling approach was as a classification problem – could we predict if a song will peak in the Top 10 based on its attributes?

- Modelling was performed using Logistic Regression, K Nearest Neighbors, and a tuned Random Forest model

- With the available data and model parameters, the model outcome was to assume **all songs** would peak between 11-100, rarely correctly classifying a top 10 song. This outcome doesn't help us answer the initial purpose of identifying a successful song

- Classification remains a better approach for this problem than regression, but more analysis and new data points would need to be considered to garner meaningful results
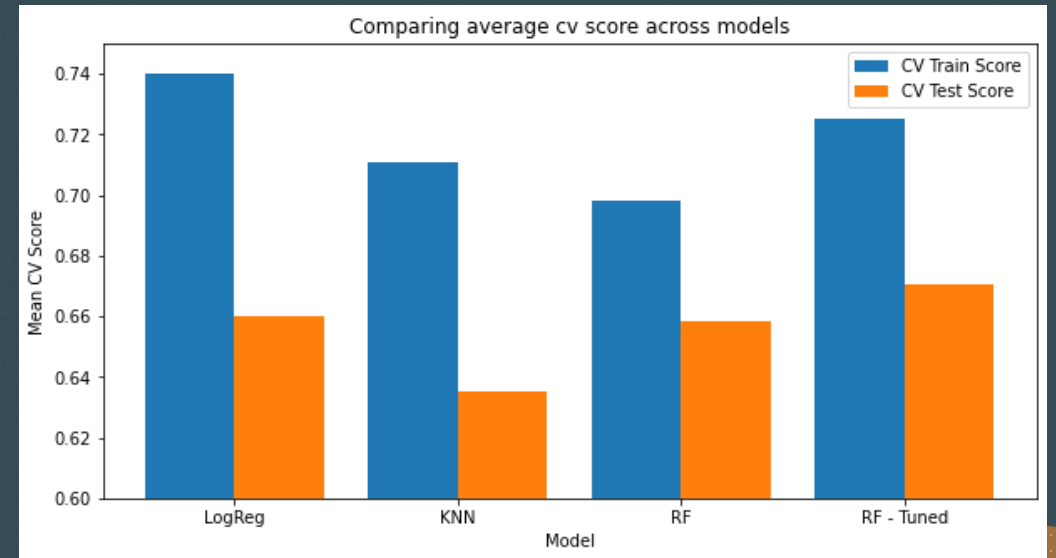
- *CV scores for the classification models were better but the model remains unreliable for predicting if a song will peak in the Top 10*

Recommendation

- Predicting if a song will perform well across the Hot 100 Chart is very difficult with song attribute data alone

- Recommendation is to pursue the Classification approach but consider incorporating external factors like song launch seasonality and the contribution of the credited artist's reputation
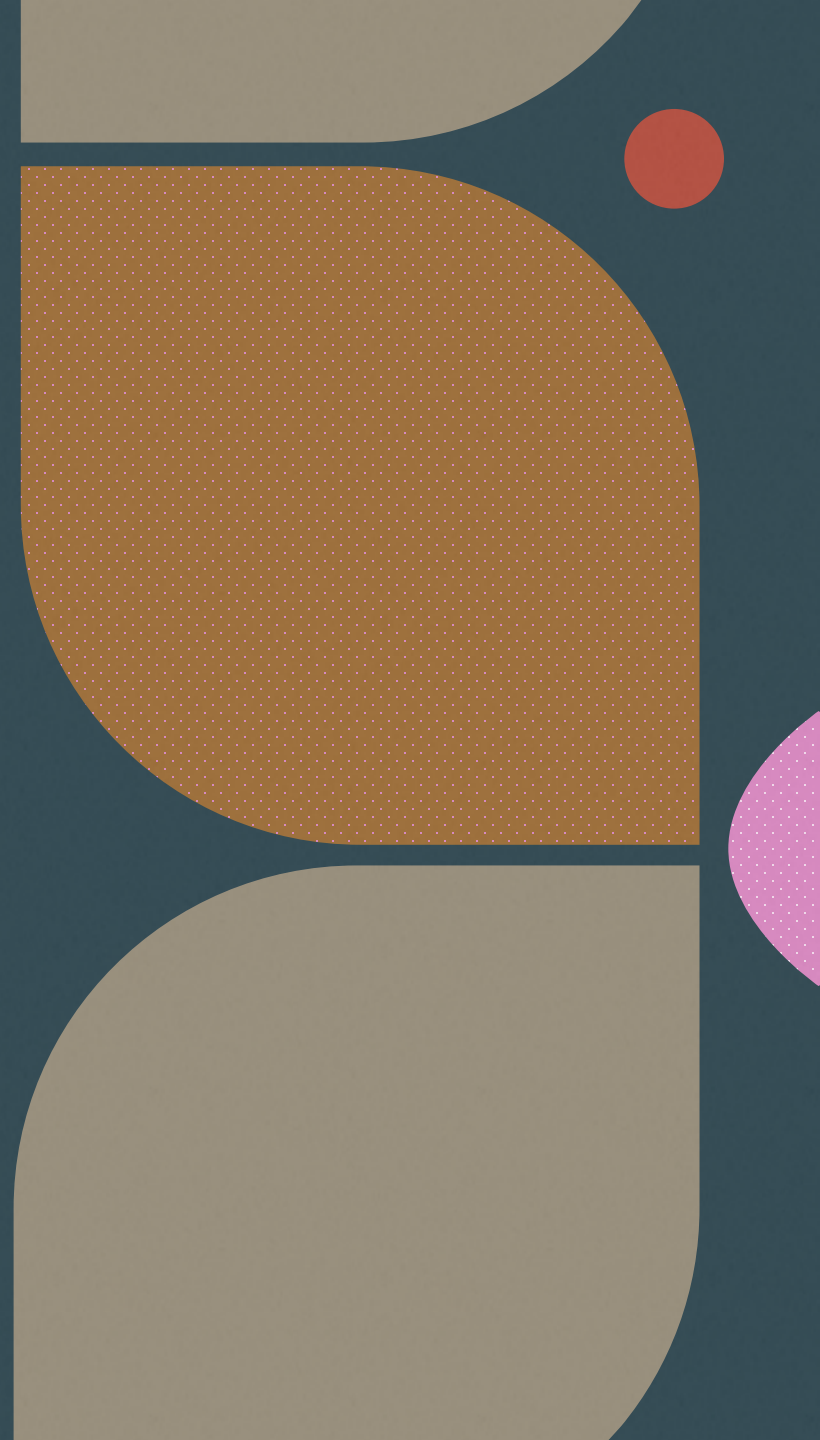
Additional Insight

- In the early phases of analysis, the number of weeks a song was on the chart was considered for the Target feature – How do results compare between a song's peak position vs. the longevity of a song on the charts?

- Unique songs were grouped into 5 week buckets (ie songs on the chart for 1-5 weeks, 6-11 weeks etc.)

- Results showed more variability than that of Peak Position but, in general, trends were similar and would result in similar challenges during the modelling phase

Example of variability when songs are *analyzed in Weeks On The Chart in groups of 5*

Boxplot of Features with data range 0 to 1, where Weeks on chart in groups of 5
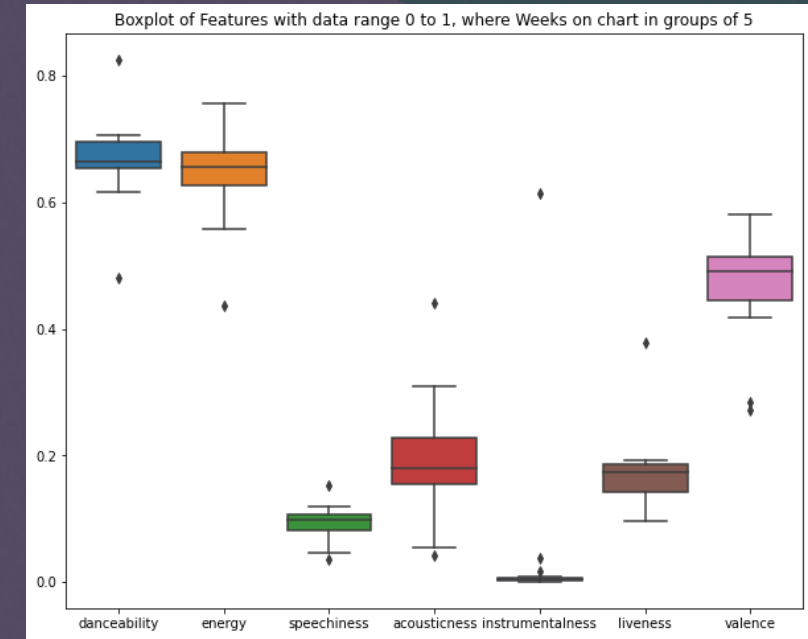
Example of variability when songs are *analyzed in Peak Position groups of 5*

(repeat from slide 5)

Boxplot of Features with data range 0 to 1, where PeakPos in groups of 5
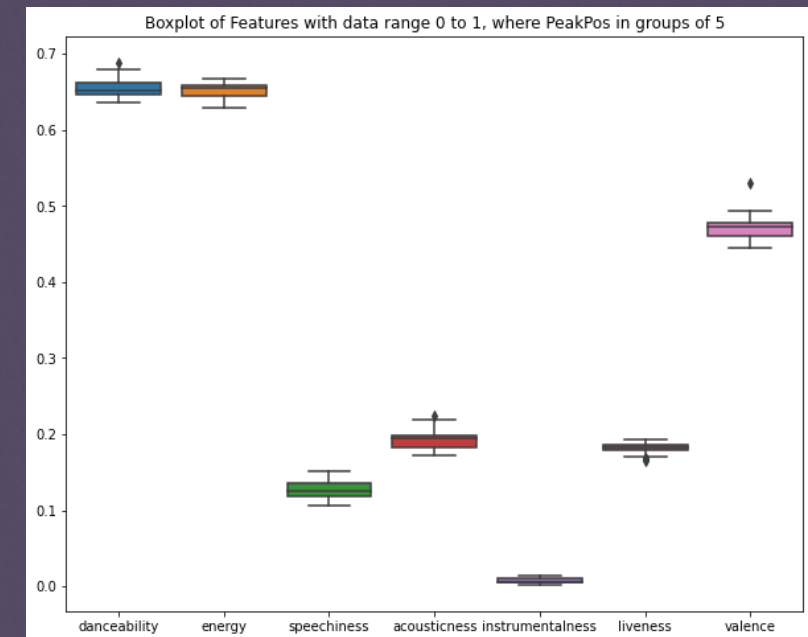
Summary & Conclusion

- What does success look like?
  - Predicting a song's chart performance based on the song attributes and historically successful songs from the Billboard Hot 100

- Following Data Collection, Analysis, and Modelling as a regression and classification problem, the results leave many questions around what makes a song successful on the charts

- While not a definitive answer to our initial question, these efforts are a foundation to further exploration as we try to hone in on the Hit Single formula