Project Report - Billboard Hot 100 Song Analysis

Jelica Bornath

Problem Definition

The label executives at Galaxy Music Publishing Group are looking to establish if there is a formula to a successful single in pursuit of continued growth for their artists. The focus of this exploration is on the song characteristics and how this impacts its performance as a hit. Beyond the marketing efforts and the timing for a song release, how can Galaxy continue to build their song catalogue with attributes that resonate with their listeners?

Success in this exploration will be to identify key attribute trends in songs on the Billboard Hot 100 and make a recommendation regarding which attribute combination is most likely to result in a successful Top 100 single for the group. The scope of this analysis was North American music trends from the last 10 years based on the Billboard Hot 100 list and song attributes made available through the Spotify Web API. Notable constraints to this exploration include the components of marketing and timing on a song's success - the purpose of the analysis will be on the song itself instead of the societal factors that play into a song's success.

Data Wrangling & Exploratory Data Analysis

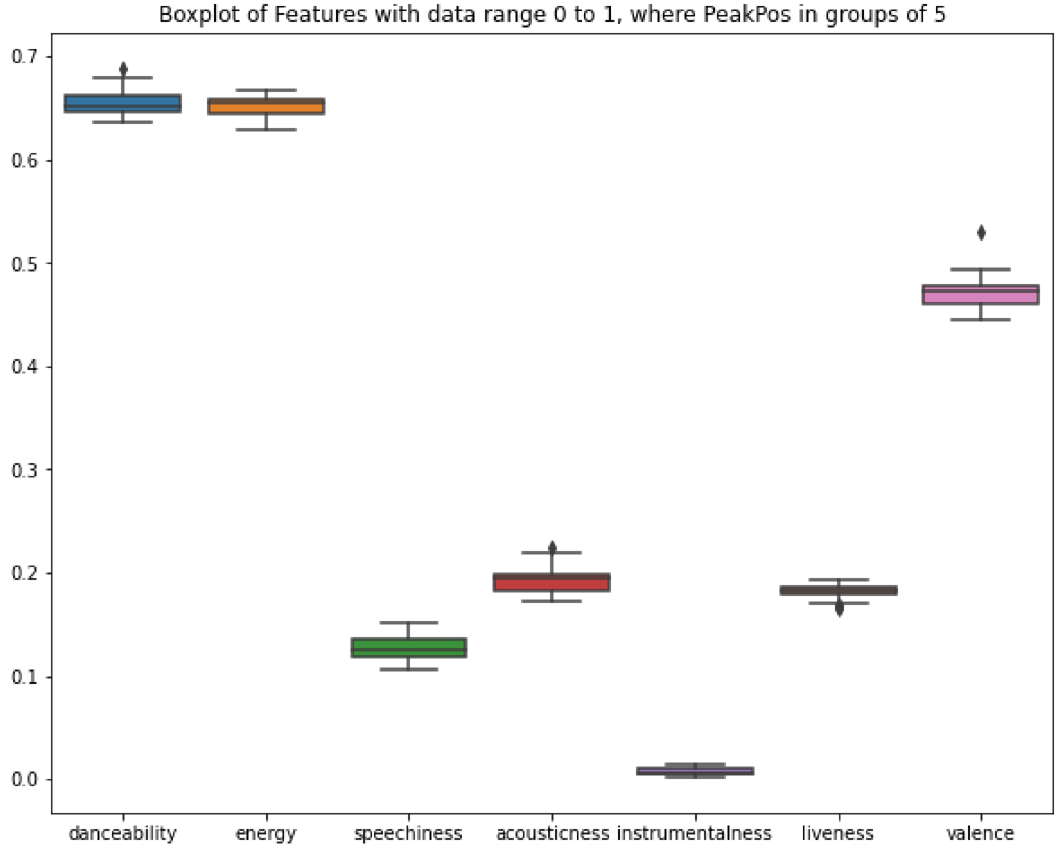
4832 unique songs on the Billboard Hot 100 Chart from the last 10 years were tied to the Spotify Web API data outlining song attributes including the following:

1) Genre - What genres are the song's artists associated with
2) Key - Key signature for the track (ie C, C#, D etc.)
3) Tempo - Pace of a song in beats per minute (BPM)
4) Duration - Length of the track
5) Time Signature - Estimation of beats per measure
6) Mode - Modality of the track (ie major or minor)
7) Danceability - Suitability for dancing, measured using tempo, rhythm etc.
8) Energy - Perceptual measure of intensity and activity
9) Loudness - Measure of loudness in decibels (dB)
10) Speechiness - Amount of spoken word in a song
11) Acousticness - Confidence level of whether a song is acoustic or not
12) Instrumentalness - Measure of a song instrumental components to vocal components
13) Liveness - Detects presence of an audience
14) Valence - Measure of musical postiveness (ie happy song vs. sad song)

The target feature for this exploration is the Peak Position of a song on the charts. Through the Data Wrangling process, 5% of the unique song data set was removed due to matching inconsistency with the Spotify Web API, resulting in 4609 unique songs for the EDA and subsequent modelling phases.

The results of the Exploratory Data Analysis process established that there weren't any strong trends between a song's Peak Position and its song attributes. This was explored by treating each song individually and also by grouping them into sections of 5 (Peak Position 1-5, 5-10, 10-15 etc.). This is partially depicted in Figure 1, where attributes graded on a scale of 0-1 by the Spotify Web API have relatively small distributions across the 20 sub-groups. This was the first indication it could be challenging to distinguish what made a song perform well based on its attributes.

Figure 1 - Boxplot of features with grading from 0-1, where PeakPos is in groups of 5.



For due diligence, the number of weeks that a song was on the charts was also explored, with similar inconclusive results.

The Genre representation was largely Pop, Hiphop and Rap, and the impact of this on the results was further explored in Preprocessing and Training after converting the categorical variable to a numerical one.

Preprocessing & Modelling

During this phase, there was a shift in target feature focus given the early EDA results and some poor initial modelling results with Linear Regression and the Random Forest Regression model. Instead of trying to predict the Peak Position of a song based on its characteristics, the focus shifted to predicting if a song had the right characteristic set to be in the Top 10. This isolated the characteristics of songs peaking in the Top 10 from those that did not and therefore turned a regression problem into a classification problem.

The results of this modelling were far more conclusive. Logistic Regression, K Nearest Neighbors and a Random Forest Classification model were tested, shown in Figure 3 and Figure 4.

Figure 3 - Accuracy scores for the Logistic Regression, K Nearest Neighbors, Random Forest and the tuned Random Forest models

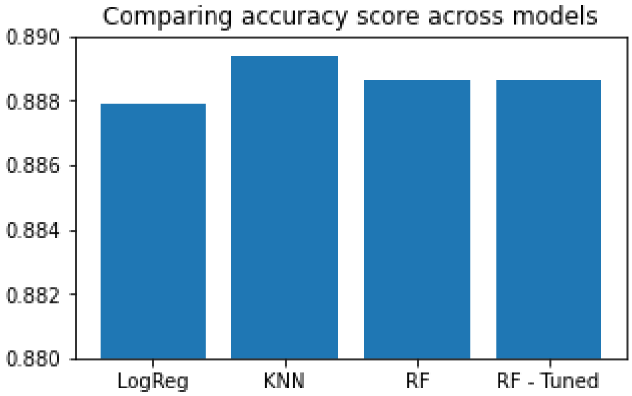
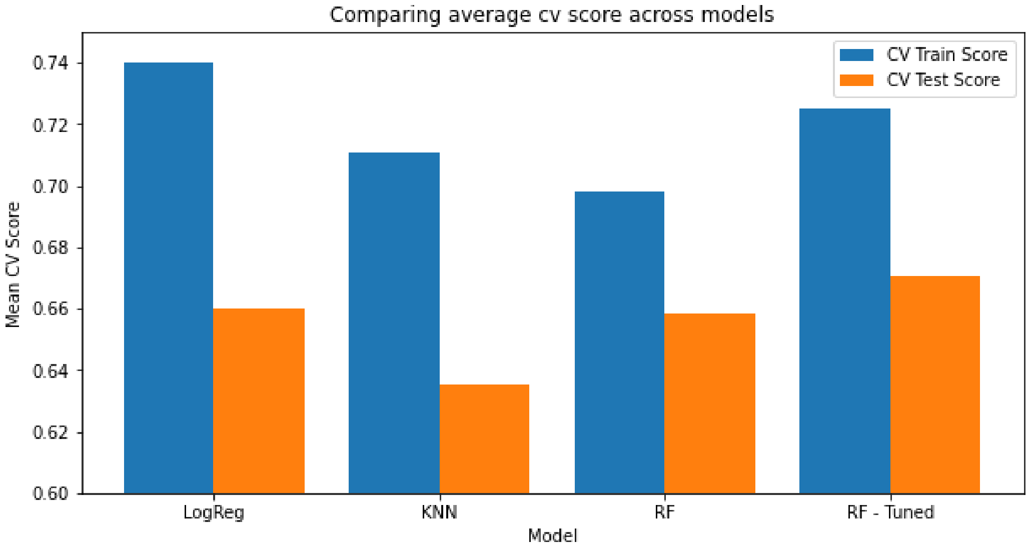
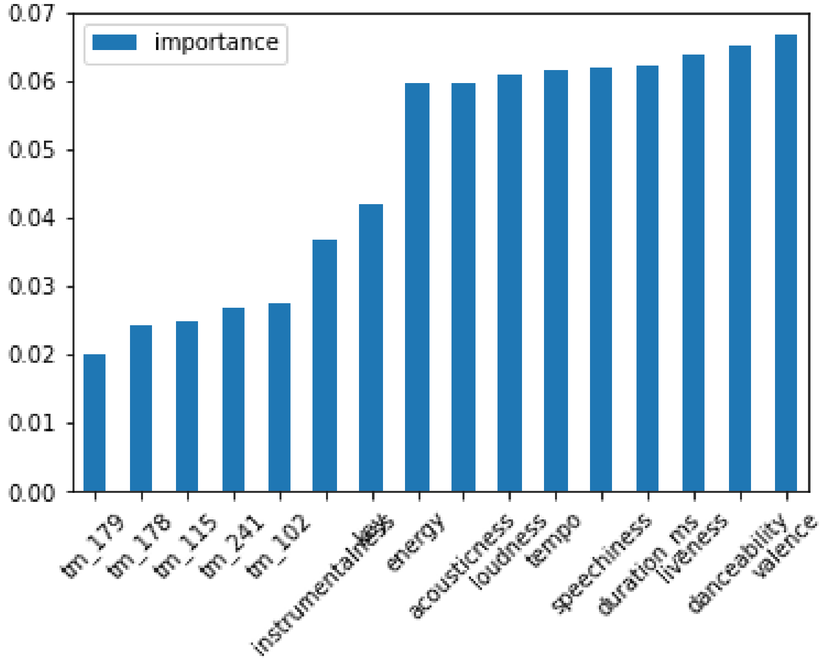


Figure 4 - Comparison of CV scores on test and training datasets across Logistic Regression, K Nearest Neighbors, Random Forest and the tuned Random Forest models



While K Nearest Neighbors has the highest accuracy, a tuned Random Forest model had better CV scores and the added benefit of clear feature importances. These feature importances are indicated in Figure 5. “tm_XXX” features refer to the Genres but attention should be given to the highest impact features, which include the song attributes themselves.

Figure 5 - Feature importances based on Random Forest Model



Based on these results, and an accuracy score of 0.88, the Random Forest Model is recommended for use. Using the model with a new song’s attributes can help predict if a new song is likely to peak in the top 10 or not, allowing us to further hone in on the song characteristics that lead to this result.

Future iterations of this model could look at the impact of release timing and seasonality on Peak Position as well as better utilizing the Genre data to identify artist combinations that could lead to a likely Top 10 hit.