A Comprehensive Analysis of the Relationship Between Spotify and Billboard Rankings Using Linear Regression, Correlation Analysis, KNN Clustering, and Time-Series Modelling

Kaily Mejia (mejiak@rpi.edu)

Rensselaer Polytechnic Institute 110 8th St., Troy, NY, 12180 United States

Abstract

This paper attempts to build different data models to accurately predict a song's ranking on the Billboard Global 200 chart based on it's ranking on the Spotify Global 200 chart. Considering how streams account for the largest amount of points a song can get to count towards the charts and Spotify is the world's largest streaming service it seems like the relationship between the charts should be clear and measurable. The data used to measure this connection is from twenty-seven weeks of chart information starting from April 13th, 2023, to October 14th, 2023. The data was synthesized from both Spotify and Billboard data that is publicly available. However, the songs considered were only those at appeared on both charts during this period. Using clustering, linear regression, correlation, and time-series analysis a measurable relationship between the charts has emerged.

Motivation

Over the past twenty-five years, the music industry has seen incredible changes to their business model. The dawn of digital sales and streaming has completely changed the way people listen to music Quantifying the relationship between Billboard's Global 200 Charts and Spotify's Global 200 Charts can provide interesting insights into what makes a song popular in the current landscape of the music industry.

Data

To conduct this analysis, compiling the datasets from the two sources, Billboard and Spotify, was essential. Both charts are published weekly and are readily available online. Acquiring the data from Spotify was straightforward, the charts could be easily accessed and downloaded as data files for each week. A web-scraping approach was employed, each week of Billboard chart data was taken from the website and formatted as a datafile. To create a unified dataset, each corresponding week's data from both sources was joined based on song title, resulting in a comprehensive data file for each of the 27 weeks. The final dataset produced was from all the weekly data files, pairing the data from both Spotify and Billboard for a given week, being combined.

Linear Regression Model

The linear regression model serves as the foundational exploration into the linear association between the current week's ranking on the Billboard chart and the corresponding Spotify rank. The leading hypothesis was that the strongest predictor of a song's Billboard ranking would be it's ranking on Spotify's chart. Thus, the predictor variable is the Spotify rank, while the response variable is the Billboard rank. The model is constructed using the "**lm**" function in R, where the linear relationship is explored through a regression line fitted to the data.

```
call:
lm(formula = namedata$current_week ~ namedata$rank, data = ndtrain_set)

Coefficients:
  (Intercept)  namedata$rank
      27.1423         0.6837
```

Figure 4: Linear Regression Model Formulation

```
> ndlinregMAE
[1] 54.53782
> ndlinregRSMSE
[1] 66.57448
```

Figure 9: Linear Regression Error Calculations

KNN Model

Using KNN Clustering, both Billboard and Spotify ranks were considered for clustering to visualize clusters formed by K-means, providing insights into the inherent groupings of the song rankings. The hypothesis was that there would be groups the followed around the axis at which Spotify and Billboard ranking were equal. The choice of k, the number of clusters, is set arbitrarily at 20, as each week had 200 ranking slots it seems that 20 clusters that could show the trends in the data without being overfitted. Even though K was chosen arbitrary, the KNN clustering model underwent k-fold cross-validation, exploring different k values to identify the optimal parameter for maximizing accuracy.

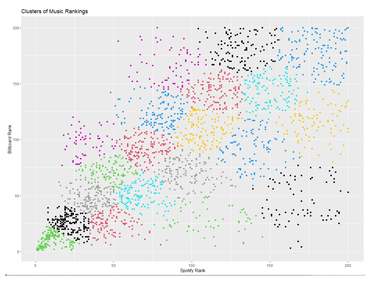

Figure 17: KNN Plot (K = 20)

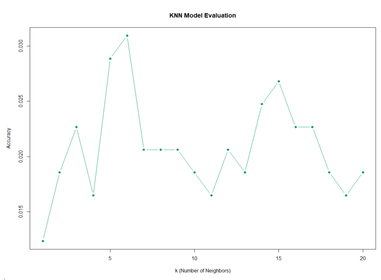

Figure 18: KNN Model Evaluation

Conclusion

In conclusion, the synthesis of linear regression, correlation analysis, clustering, and time-series modelling provided a multifaceted understanding of the Spotify-Billboard relationship. This project serves as a foundation for future explorations, encouraging a continuous refinement of models and approaches to unravel the ever-changing landscape of music chart dynamics. In subsequent exploration, it would be interesting to combine streaming data from sources outside of Spotify. The gaps that need to be filled are probably from other streaming sources.

Correlation Model

As an optimization of the linear regression model, the correlation model delves into multiple linear regression based on features with a strong correlation to the current week's ranking. The model utilizes the "**cor**" function to compute the correlation matrix and subsequently selects features with a correlation coefficient above the defined threshold. Features were then selected based on a correlation coefficient above 0.5, encompassing last_week, peak_pos, wks_on_chart, rank, peak_rank, previous_rank, weeks_on_chart, and streams. The features that were above the threshold were "current_week", "last_week", "peak_pos", "rank", and "previous_rank". Similarly, to the linear regression, the multiple linear regression employed the "**lm**" function.

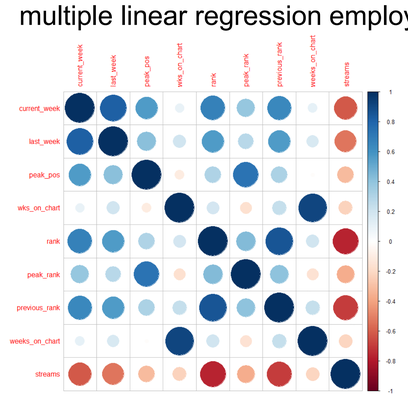

Figure 10: Correlation Matric Heat Map

```
> ndcorMSE
[1] 17.53078
> ndcorRSME
[1] 25.05107
```

Figure 16: Correlation Model Error Calculations

```
Call:
lm(formula = ndtrain_set$current_week ~ ., data = ndtrain_set[,
    features])

Coefficients:
  (Intercept)     last_week      peak_pos          rank  previous_rank
      5.41694       0.53962       0.32715       0.26809       0.03641
```

Figure 11: Multiple Linear Regression Model Formulation

Time – Series Model

The cross-correlation between Spotify and Billboard rankings over 27 weeks was analyzed through a time-series model. Using the function "**ccf**" to visualize the cross-correlation of the temporal relationship between Spotify and Billboard rankings. The chosen lag was based on the number of weeks' worth of data at 27. The statistical significance of the model was assessed through a t-test applied to the cross-correlation coefficients.

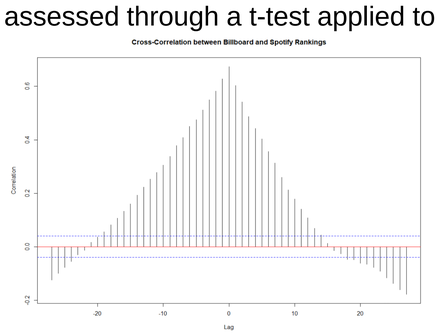

Figure 19: Cross – Correlation Between Billboard & Spotify

```
        One Sample t-test

data:  timecor$acf
t = 5.3604, df = 54, p-value = 1.766e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1090407 0.2393412
sample estimates:
mean of x
0.1741909
```

Figure 20: T - Test Results for Cross - Correlation Model

References

Billboard. (2023, December 2). *Billboard Global 200*. Billboard. https://www.billboard.com/charts/billboard-global-200/

Spotify. (2023, November 30). *Spotify Charts - Spotify Charts are made by fans*. Charts.spotify.com. https://charts.spotify.com/charts/view/regional-global-weekly/latest

Wikipedia Contributors. (2019, June 13). *Spotify*. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Spotify

Wikipedia Contributors. (2023, November 27). *Billboard Hot 100*. Wikipedia. https://en.wikipedia.org/wiki/Billboard_Hot_100#:~:text=The%20Hot%20100%20is%20ranked