Stats 405 Final: Song Popularity

Executive Summary

The Fighting Lonzos: MAS 405 S2022

Kaustubh Deshpande, Omar Moore, Jonathan Shan,
Alex Veroulis, Anders Ward

June 7, 2022

1 Objective

For our project, we examined the Billboard Top 100 Songs for every day from January 2016 to 2022 and aggregated statistics for each individual song over this time period, where there were 3611 unique songs that entered the top 100. At the outset, our objective was to predict the amount of time that a given song would stay in the Billboard Top 100, regardless of rank. This way, we could better understand why certain songs may be more popular than others, and we could observe what characteristics of a song would most affect its popularity. In this scenario, we qualified the response variable in two separate ways, which required two different modeling approaches.

2 Approaches

Our first approach involved taking the longest streak of consecutive weeks in the Top 100 for each individual song over the entire period and using different models to predict this longest streak. Specifically, we used song metrics from Spotify, such as loudness and speechiness, to predict the longest streak for every song. After trying several different models, where we used techniques like log transformations of variables in linear regression, MLP, PCA, and backward stepwise regression, we found that none of the models yielded a low RMSE because the response variable had outlier streaks that were far larger than most songs. As a result, the best of these models had an RMSE of roughly 8.5, which was not a satisfactory result. We were inspired to redefine the problem at hand with our second approach.

The second approach redefined the response variable as a three-level factor variable. The levels of this factor included short, medium, and long; to determine the labels of the songs, we took a look at the summary statistics of the streak variable from the first approach and used percentiles to form three groups. The short group included all streaks of at most 1 week, which was roughly the 33rd percentile. Then, the medium group included streaks greater than 1 week but at most 12 weeks, which was the 66th percentile. Finally, the long groups included streaks greater than 12 weeks. After labeling each song, we found that the groups were of similar size; the short group had 1269 songs, the medium group had 1170 songs, and the long group had 1172 songs. We proceeded to use ordinal logistic regression by using variables obtained from the stepwise regression from approach 1. Namely, these variables were qualities of each song: valence (emotional feeling of song), the log of liveness (the presence of an audience in the recording), the log of speechiness (presence of words in song), peak (the best ranking of the song in the entire period), sdposition (the standard deviation of the song's rankings), and loudness (how loud a song is in decibels). The exponentiated coefficients of the ordinal logistic regression model are seen below in Table 1, along with the bounds of the 95 percent confidence interval for each predictor's exponentiated coefficient.

Table 1: Ordinal Logistic Regression Coefficients

Variable	Estimate	Lower Bound	Upper Bound
valence	0.294	0.209	0.413
log(liveness)	1.332	1.183	1.501
log(speechiness)	1.672	1.541	1.815
peak	1.041	1.037	1.045
sdposition	0.916	0.906	0.925
loudness	0.913	0.884	0.943

With regard to the model's accuracy, we split the data into training and test sets, with 80 percent of the data going in the training set and 20 percent of the data going in the test set. After training the model on the training set of 2890 songs, we made predictions on the test set for what each song's category would be. In the end, we found this model performed reasonably well on the test set, as 65.3 percent of songs were classified correctly, which is even more impressive than usual since there are three possible categories as opposed to the usual two categories in standard logistic regression. Below, Table 2 is the three-way contingency table showcasing the results of the model, where the rows represent the predicted values and the columns represent the actual values.

Table 2: Contingency Table

Prediction	Long	Medium	Short
Long	154	95	0
Medium	74	94	30
Short	6	45	223

3 Conclusion

For the first approach, we found that despite our rigorous modeling efforts, the structure of the response variable hindered our performance, which motivated us to change the structure of the problem we were trying to solve. As a result, we made the response variable a three-level factor, which allowed us to find some interesting tidbits about our data. For example, in the ordinal regression model, two variables stood out. The log(speechiness) variable had an estimated coefficient of 1.672, which means that for every two-fold increase in speechiness, the predicted odds of the song being in the top 100 for a longer period of time increases by a factor of roughly 3.18. On the other hand, the valence variable had an estimated coefficient of 0.294, which signifies that for every unit increase in this variable, the predicted odds of the song being in the top 100 for a longer period of time decreases by approximately 70.6 percent.

As for the contingency table, we see that the model did a great job of predicting songs in the short category correctly, with an 81.4 percent success rate. The model also did quite well predicting songs in the long category correctly, with a 61.8 percent success rate and zero incorrect predictions that were short songs. However, the model struggled the most when predicting songs in the medium category correctly; the success rate here was only 47.5 percent. This result makes sense though, as this is the middle group of the three and can lead to some borderline cases that are difficult to differentiate between the three categories. A potential remedy for this issue in the future could be to split the medium group into two, but even this strategy will not be able to remove all ambiguity.