2024

# Predicting the Popularity of Music Records

USING LOGISTIC REGRESSION

RASHMITHA SHAMANTHULA

SMITH SCHOOL OF BUSINESS | Toronto

## Introduction

This report outlines the development and evaluation of logistic regression model to determine whether a song will rank in the Top 10 based on various musical and artist-related features. The dataset used for this analysis includes multiple numeric variables that capture different aspects of songs, such as their tempo, loudness, and key, along with information about the artists.

## Initial Model:

Initial model was built by including all numeric variables available in the dataset. The main objective of performing this step is to establish a baseline for further refinement of model.
Performance Metrics:

- Test Accuracy: 72%
- Specificity: 70%
- Sensitivity: 81%
- Area Under the Curve (AUC): 0.84

**Analysis and Interpretations:**

Confidence in Time Signature, Key, and Tempo:

The variables key_confidence, tempo_confidence and timesignature_confidence have positive coefficients, which means the higher these values, the more likely song is to be in the Top10

Complexity Interpretation:

The model does not explicitly measure musical complexity but suggests that songs with lower confidence in key, tempo, and time signature—potential indicators of complexity—are less likely to achieve Top 10 status. In the context of this model, higher musical complexity is inferred to be less popular with broader audiences, who may prefer more predictable and straightforward musical structures.

## Data Exploration:

Significant data exploration has been done on the dataset to extract meaningful insights mentioned below.
1) There are no null values in the dataset
2) Almost 15% of songs in the dataset are in top10 so setting a threshold of 0.15 to classify whether the song belongs to top10 or not.
3) There are 1047 unique artists, so the model is going to be complicated if we create dummy variables for each artist.

4) There are some duplicate song IDs in the dataset
5) We have some highly correlated independent variables in our data set
for example: timbre_0_max and loudness, loudness and energy etc.,

## Feature Engineering:

**Data transformation**: To address the skewness in the distribution of certain continuous independent variables, appropriate data transformations such as log and cubic transformations have been performed on them

**Interaction Variable:** In the dataset we have strong correlation between 'loudness' and 'energy'. The general trend is if the music is louder, it tends to be more energetic. It would be interesting to see if they have any synergy effect. Hence created an Interaction variable 'energy_loudness'

**Popular Artists:** Popular artists tend to have more hits. Extracting the artists who tend to have high number of songs in Top10. Extract the top 5% artists who have the highest hits and create a new column (binary variable) 'popular_artists' and set the value=1 is the artist in the observation belongs to this top 5%. This variable helps in capturing the influence of an artist's popularity on a song's success.

**Keyword Analysis:** Songs that capture emotional aspects might influence a song's popularity. Creating a new variable 'has_keyword' to look for certain keywords such as 'forever', 'love' etc.,

## Model 1: Full Feature Model

First model was created by including the data transformations that have been performed and new features that have been added to the model.

Performance Metrics:

- Testing Accuracy: 77%
- Specificity: 76%
- Sensitivity: 81%
- Area Under Curve(AUC): 0.88

The inclusion of all transformed and engineered variables slightly improved the model's performance, as evidenced by the AUC of 0.88.

## Model 2: Backward stepwise selection

The second model was developed by performing backward stepwise, where insignificant variables were removed using a top-down approach based on statistical significance and contribution to the model.

Performance Metrics:

- Test Accuracy: 77%
- Specificity: 76%
- Sensitivity: 81%
- Area Under Curve (AUC): 0.88

This model has the same performance as model 1 but it is simple because all the insignificant variables were removed.

## Model 3: Principal Component Analysis (PCA)

The third model was created by performing Principal Component Analysis (PCA) to reduce number of variables and address multicollinearity issues.
Performance Metrics:

- Test Accuracy: 86%
- Specificity: 20%
- Sensitivity: 98%
- Area Under Curve (AUC): 0.59

The model's accuracy is high, which means it is good at predicting songs that will make it to the Top 10 (high sensitivity), but it struggles significantly with correctly identifying songs that won't make it to the Top 10 (low specificity). This means even though the model accuracy is high, it won't be able to differentiate well between the two classes (Top10 and Not Top10) across different thresholds.

## Final Model:

In reviewing the performance metrics of above three models, both Model 1 and Model 2 exhibit better performance. However, I chose to proceed with Model 2 as my final model due to its simplicity. While both models deliver comparable results, the reduced complexity of Model 2 is more practical without sacrificing significant predictive power. This balance of simplicity and effectiveness aligns well with the objectives of the analysis.

## Interpreting coefficients:

The coefficient of logistic regression is the amount by which the independent variable increases the log-odds of dependent variable.
For Example,
In our model, we are trying to predict the probability of a song being in Top10, the log-odds will be given by the ratio of probability of song being in Top10 to probability of song not being in Top10.

p = probability of song being in Top10
1-p = probability of song being in Top10
then the log-odds will be given by below equation
Log-odds = Log (p/(1-p))
Odds-ratio = e^(p/(1-p))

Below is the interpretation of 5 variables from the final model:

**Popular artist:**

The coefficient 2.2641 is the amount by which the variable 'popular_artist' increases the log odds of probability of song being in Top10. This translates to an odds ratio of e^(2.2641), which is approximately 9.63.

In other words, the odds ratio of a song being in the Top 10 are about 9.63 times higher if it is by a popular artist, compared to a song that is not by a popular artist, holding all other variables constant.

**Timesignature:**

The coefficient 0.1995 is the amount by which the variable 'timesignature' increases the log odds of probability of song being in Top10. This translates to an odds ratio of e^(0.1995), which is approximately 1.22.

In other words, the odds ratio of a song being in the Top 10 increase by 1.22 times (or) 22% for one unit increase in time signature, holding all other variables constant.

**Timbre_0_max:**

The coefficient of this variable is -0.1883. The log odds of probability of song being in Top10 decrease by 0.1883. This translates to an odds ratio of e^(-0. 1883), which is approximately 0.83.

In other words, the odds ratio of a song being in the Top 10 decrease by 0.8 times (or) 20% for one unit increase in variable 'Timbre_0_max', when holding all other variables constant.

**Year:**

The coefficient 0.0018 is the amount by which the variable 'year' increases the log odds of probability of song being in Top10. This translates to an odds ratio of e^(0. 0018), which is approximately 1.001.

In other words, the odds ratio of a song being in the Top 10 increase by 1.001 times (or) 0.01% for each increasing year, holding all other variables constant.

**Log_pitch:**

The coefficient of this variable is -39.3069. The log odds of probability of song being in Top10 decrease by 39.3069. This translates to an odds ratio of e^(-39.3069), which is approximately 8.75*10^(-18).

In other words, the odds ratio of a song being in the Top 10 drastically decreases for one unit increase in variable 'log_pitch', holding all other variables constant.