Rodolfo Maranhao
9/04/2023

Predicting Song Popularity on Spotify

Problem Statement

The question we are trying to answer in this project is: “Can we predict song popularity on Spotify?”. This could be potentially insightful for new and upcoming artists trying to grow a business in the music industry.

This can also answer the question “What do people like listening to the most?”. This is paramount for companies that want to make their advertisements more effective or want to appeal more to certain audiences.

Background on the subject matter

This is a perfect example of applied Data Science, since Spotify has an API with several metrics that could be used for different types of analysis. With clean datasets and useful metrics, we can run different types of models and try to understand the data in a new way.

In the past, Spotify has used their data to create recommendation systems (what a user would like to listen to next), campaigns like Spotify Wrapped (when the usage status of a user over the previous year is displayed with several metrics) and even recently, creating an AI that caters to each specific user by creating customized playlists with a virtual DJ.

Spotify has always been at the forefront of Data Science. Knowing this, we can use a dataset imported from the Spotify API to try to run our own personal analysis.

Dataset

Our dataset was acquired from Kaggle.com. The creator is Yamac Eren Ay, and his data source, as mentioned before, is the Spotify API. The last update on the dataset was in April 2021.

The file contains two datasets, one with information about 586,000 songs and their features, such as, “Song Name”, “Artist”, “Popularity”, “Acousticness”, “Liveness”, “Speechiness”, “Tempo”, “Valence” and so on. The other dataset contains information about 1.16 million artists and their features like “Artist Name”, “Genre”, “Artist Popularity” and “Followers” on Spotify.

Cleaning and Preprocessing

We started our cleaning and preprocessing by importing our datasets. While importing, we filtered out data that could skew our analysis, such as songs with 0 “Popularity” (this could make our dataset unbalanced), songs with over 0.66 “Speechiness” (this represents talk shows, podcasts, and not actual music) and finally, we removed songs with 0 “Time Signature” (this could be a mistake or experimental music).

Then, we merged our datasets on “id_artists”. Our second dataset did not have the column “Genre” filled in, so we imported only “Artist Popularity” and “Followers”. After that, we filtered out artists with 0 artist popularity and artists with 0 followers, again, because this could skew our data, and make our models biased towards predicting unpopular songs.

The last few steps in data cleaning involved creating a column with release year in the right numeric format and transforming song duration in milliseconds to song duration in seconds and minutes.

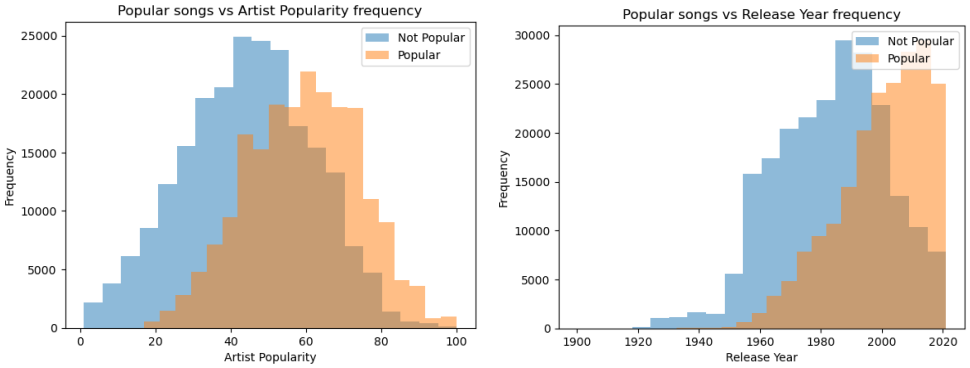
Now, for EDA (exploratory data analysis) all the features were plotted to analyze the distribution of our dataset. This was particularly important to decide on a cutoff point for our binary target variable (song popularity). Thus, according to the dataset the average song popularity (30) was a good cutoff point to determine whether a song is popular or not. In other words, if a song scored above 30 in popularity it is considered popular, if it scored 30 or less in popularity, it is considered not popular.

Rodolfo Maranhao
9/04/2023

Additionally, our EDA revealed some patterns inside our data. For example:

- Most of our songs were released after the 1980s. They were composed in the major scale, in 4/4 time signature, in the Keys of C, F#, G# and D, recorded in studio. They also are not explicit and have a duration between 2 and 6 minutes.
- As for Artists, the followers feature was highly skewed (with a minimum of 1 follower, maximum of 78 million and averaging 1.1 million). Artist Popularity was also skewed (with a minimum of 1, maximum of 100 and averaging 51). However, the most popular artists (Justin Bieber, Taylor Swift, Bad Bunny…) were not necessarily the artists with the most followers (Ed Sheeran, Ariana Grande, Drake…).

Our EDA also showed the most dissimilarity between popular and unpopular songs in two features, release year and artist popularity:



Lastly, we will feature engineer “NOS_artist”, which means number of songs per artist. This feature could be useful to predict whether or not a song is popular. It may reveal patterns where the number of songs an artist publishes could impact song popularity.

Modeling

Before modelling, we should start by checking for correlated features and doing feature selection. By plotting a heatmap with our correlation matrix we find that the most obvious correlated features are “duration in seconds” and “duration in minutes” since these two are the same variable with a different unit of measurement. We should drop “duration in minutes” and only carry on with “duration in seconds”.

Other features that were correlated are “energy”, “acousticness” and “loudness”. It makes sense that all 3 of these variables are correlated, since “energy” is the song’s overall perceived energy, “acousticness” is how likely the song is to be acoustic, and “loudness” is the overall volume the song has. A song with high “energy” will tend to be loud and not acoustic, an acoustic song will tend to be calm and not have a high volume. For these reasons, it makes sense to drop 2 of these variables and keep only one. I decided to keep “energy” since it is more comprehensive, and it may preserve characteristics of the other two features.

After running our feature selection, we should check the p values of our variables to identify if any of our features’ observations were due to chance. By setting a p value of 0.05 we should drop “Key”, “Time Signature” and “Valence” from our models.

Finally, after splitting our data into train/test sets and scaling our data, we are ready to start our models. Since our target variable is binary, it makes sense to start with Logistic Regression and Decision Tree classifier. Both models work well when predicting a binary outcome (0 or 1, or in this case, a popular song or not).

Our Logistic Regression performed decently. The model also showed that “Artist Popularity” was the most important feature when predicting song popularity, followed by “Release Year” and “NOS_artist”. The first two features have a positive correlation with the outcome, which means that the more popular the artist is and the

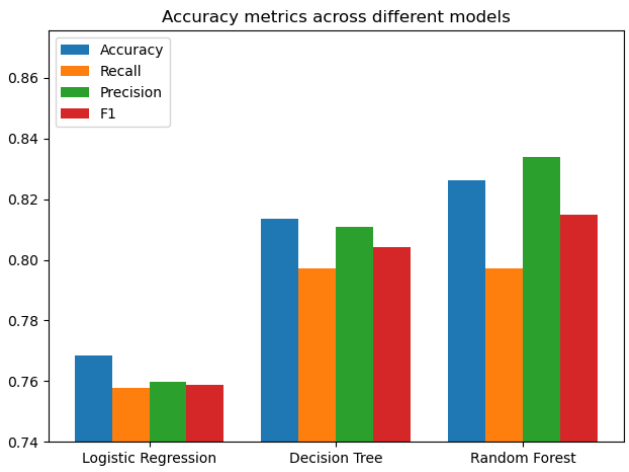
more recently the song was released, the more likely the song is to be popular. The “NOS_artist” has released had a negative correlation with song popularity, which means that it probably has a non-linear correlation with our outcome. Therefore, there is probably a point where releasing more songs is detrimental to the songs’ popularity.

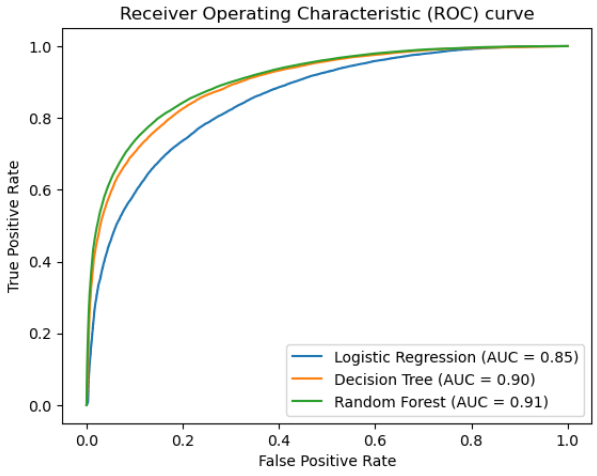
Then, we ran our Decision Tree with even better performance. Since this was the case, following up with a Random Forest model makes sense. This is a model that combines several different Decision Trees and compiles a final more accurate Decision Tree.

Results

Here we can look at the graphs to see our results. Logistic Regression was the lowest scoring model, its accuracy, precision, recall and F1 score were all about 76%. For Decision Tree, our performance metrics were revolving around 81% and for Random Forest, our metrics were about 83% (except for recall, sitting at about 80%).

Our Receiver Operating Characteristic curve and Area Under Curve (the balance between getting the popular song right and getting the unpopular songs right) also showed the same tendencies, Logistic Regression was the lowest performing model and Decision Tree performed just a little worse than Random Forest.





Key takeaways

- Popular songs tend to have higher energy, be less explicit and last between 2 to 6 minutes.
- Not necessarily the artist with most followers on Spotify scores higher on Artist Popularity.
- Songs released after the 2000s have a much higher tendency to become popular.
- Artist popularity matters, people view songs as popular if they were made by artists they already know.
- With an accuracy of about 83% our Random Forest was our best performing model.

Conclusion

To summarize, predicting the popularity of songs can be very challenging. We have unpredictable one hit wonders and songs by famous artists that failed in an industry that’s always changing. However, through our comprehensive analysis of this matter, we have identified several crucial insights and developed models that are reasonably effective in predicting a song’s level of popularity.

We also tested the effectiveness of our models, making sure that they are valid and could provide insights to future projects; furthermore, creating value to artists, companies, and the music industry in general.